BMJ Open

# STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration



| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2016-012799 |
| Article Type: | Research |
| Date Submitted by the Author: | 26-May-2016 |
| Complete List of Authors: | Cohen, Jérémie; Academic Medical Centre, University of Amsterdam, Department of Clinical Epidemiology, Biostatistics and Bioinformatics<br>Korevaar, Daniël; University of Amsterdam, Academic Medical Centre<br>Altman, Doug; Centre for Statistics in Medicine<br>Bruns, David; University of Virginia School of Medicine, Department of Pathology<br>Gatsonis, Constantine; Brown School of Public Health<br>Hooft, Lotty; University Medical Center Utrecht, University of Utrecht, Cochrane Netherlands<br>Irwig, Les; University of Sydney, Sydney Medical School<br>Levine, Deborah; Beth Israel Deaconess Medical Center, Department of Radiology<br>Reitsma, Johannes; University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care<br>de Vet, Riekie; VU University Medical Center<br>Bossuyt, Patrick; Academic Medical Center; University of Amsterdam, Dept. Clinical Epidemiology and Biostatistics |
| Primary Subject Heading: | Medical publishing and peer review |
| Secondary Subject Heading: | Diagnostics, Epidemiology, Evidence based practice, Research methods |
| Keywords: | reporting quality, sensitivity and specificity, diagnostic accuracy, research waste, peer review, medical publishing |
| | |

# STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration

Jérémie F. Cohen*, Daniël A. Korevaar*, Douglas G. Altman, David E. Bruns, Constantine A. Gatsonis,

Lotty Hooft, Les Irwig, Deborah Levine, Johannes B. Reitsma, Henrica C.W. de Vet, Patrick M.M. Bossuyt

***Both authors contributed equally to this manuscript and share first authorship.**

**Authors' names, academic degrees, positions, affiliations, and email addresses:**

**Jérémie F. Cohen***, MD PhD, *Postdoctoral research fellow*

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands; INSERM UMR 1153 and Department of Pediatrics, Necker Hospital, AP-HP, Paris Descartes University, Paris, France

jeremie.cohen@inserm.fr

**Daniël A. Korevaar***, MD, *PhD candidate*

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands

d.a.korevaar@amc.uva.nl

**Douglas G. Altman**, DSc, *Professor of statistics in medicine*

Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

doug.altman@csm.ox.ac.uk

**David E. Bruns**, MD, *Professor of pathology*

Department of Pathology, University of Virginia School of Medicine, Charlottesville, Virginia, USA

dbruns@virginia.edu

**Constantine A. Gatsonis**, PhD, Professor of biostatistics

Department of Biostatistics, Brown University School of Public Health, Providence, Rhode Island, USA

gatsonis@stat.brown.edu

**Lotty Hooft**, PhD, *Associate professor / Co-director*

Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, the Netherlands

l.hooft@umcutrecht.nl

**Les Irwig**, MBBS, PhD, *Professor of epidemiology*

Screening and Diagnostic Test Evaluation Program, School of Public Health, University of Sydney, Sydney, New South Wales, Australia

les.irwig@sydney.edu.au

**Deborah Levine**, MD, *Professor of radiology*

Department of Radiology, Beth Israel Deaconess Medical Center, Boston, MA, USA; Radiology Editorial Office, Boston, MA, USA.

dlevine@bidmc.harvard.edu

**Johannes B. Reitsma**, MD PhD, *Associate professor of clinical epidemiology*

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, the Netherlands

j.b.reitsma-2@umcutrecht.nl

**Henrica C.W. de Vet**, PhD, *Professor of clinimetrics*

Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, the Netherlands

hcw.devet@vumc.nl

**Patrick M.M. Bossuyt**, PhD, *Professor of clinical epidemiology*

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands

p.m.bossuyt@amc.uva.nl

**Corresponding author:** Prof. Patrick M.M. Bossuyt

Department of Clinical Epidemiology, Biostatistics and Bioinformatics

Academic Medical Center - University of Amsterdam

PO Box 22700, 1100 DE Amsterdam, The Netherlands

Email: p.m.bossuyt@amc.uva.nl Phone: +31(20)566 3240 Fax: +31(20)691 2683

**Word count (text only):** 9,316.

**Keywords**: reporting quality; sensitivity and specificity; diagnostic accuracy; research waste; peer review; medical publishing

## ABSTRACT

Diagnostic accuracy studies are, like other clinical studies, at risk of bias due to shortcomings in design and conduct, and the results of a diagnostic accuracy study may not apply to other patient groups and settings. Readers of study reports need to be informed about study design and conduct, in sufficient detail to judge the trustworthiness and applicability of the study findings.

The STARD statement (Standards for Reporting of Diagnostic Accuracy Studies) was developed to improve the completeness and transparency of reports of diagnostic accuracy studies. STARD contains a list of essential items that can be used as a checklist, by authors, reviewers and other readers, to ensure that a report of a diagnostic accuracy study contains the necessary information.

STARD was recently updated. Here we present the STARD 2015 explanation and elaboration document. Through commented examples of appropriate reporting, we clarify the rationale for each of the 30 items on the STARD 2015 checklist, and describe what is expected from authors in developing sufficiently informative study reports.


## STRENGTHS AND LIMITATIONS OF THIS STUDY

Not applicable to this explanation and elaboration document.

## INTRODUCTION

Diagnostic accuracy studies are at risk of bias, not unlike other clinical studies. Major sources of bias originate in methodological deficiencies, in participant recruitment, data collection, executing or interpreting the test, or in data analysis.[1 2] As a result, the estimates of sensitivity and specificity of the test that is compared against the reference standard can be flawed, deviating systematically from what would be obtained in ideal circumstances (see key terminology in Table 1). Biased results can lead to improper recommendations about testing, negatively affecting patient outcomes or health care policy.

Diagnostic accuracy is not a fixed property of a test. A test's accuracy in identifying patients with the target condition typically varies between settings, patient groups, and depending on prior testing.[2] These sources of variation in diagnostic accuracy are relevant for those who want to apply the findings of a diagnostic accuracy study to answer a specific question about adopting the test in his or her environment. Risk of bias and concerns about the applicability are the two key components of QUADAS-2, a quality assessment tool for diagnostic accuracy studies.[3]

Readers can only judge the risk of bias and applicability of a diagnostic accuracy study if they find the necessary information to do so in the study report. The published study report has to contain all the essential information to judge the trustworthiness and relevance of the study findings, in addition to a complete and informative disclose about the study results.

Unfortunately, several surveys have shown that diagnostic accuracy study reports often fail to transparently describe core elements.[4-6] Essential information about included patients, study design and the actual results is frequently missing, and recommendations about the test under evaluation are often generous and too optimistic.

To facilitate more complete and transparent reporting of diagnostic accuracy studies the STARD statement was developed: Standards for Reporting of Diagnostic Accuracy Studies.[7] Inspired by the Consolidated Standards for the Reporting of Trials or CONSORT statement for reporting randomized

controlled trials,[8 9] STARD contains a checklist of items that should be reported in any diagnostic accuracy study.

The STARD statement was initially released in 2003 and updated in 2015.[10] The objectives of this update were to include recent evidence about sources of bias and variability and other issues in complete reporting, and make the STARD list easier to use. The updated STARD 2015 list now has 30 essential items (Table 2).

Below we present an explanation and elaboration of STARD 2015. This is an extensive revision and update of a similar document that was prepared for the STARD 2003 version.[11] Through commented examples of appropriate reporting, we clarify the rationale for each item and describe what is expected from authors.

We are confident that these descriptions can further assist scientists in writing fully informative study reports, and help peer reviewers, editors and other readers in verifying that submitted and published manuscripts of diagnostic accuracy studies are sufficiently detailed.

## STARD 2015 ITEMS: EXPLANATION AND ELABORATION

### Title or abstract

*Item 1. Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)*

**Example.** "Main outcome measures: Sensitivity and specificity of CT colonography in detecting individuals with advanced neoplasia (i.e., advanced adenoma or colorectal cancer) 6 mm or larger."[12]

**Explanation.** When searching for relevant biomedical studies on a certain topic, electronic databases such as Medline and Embase are indispensable. To facilitate retrieval of their article, authors can explicitly identify it as a report of a diagnostic accuracy study. This can be done by using terms in the title and/or abstract that refer to measures of diagnostic accuracy, such as "sensitivity", "specificity",

"positive predictive value", "negative predictive value", "area under the ROC curve (AUC)", or "likelihood ratio".

In 1991, Medline introduced a specific keyword (MeSH heading) for indexing diagnostic studies: "Sensitivity and Specificity." Unfortunately, the sensitivity of using this particular MeSH heading to identify diagnostic accuracy studies can be as low as 51%.[13] As of May 2015, Embase's thesaurus (Emtree) has 38 check tags for study types; "diagnostic test accuracy study" is one of them, but was only introduced in 2011.

*In the example*, the authors mentioned the terms "sensitivity" and "specificity" in the abstract. The article will now be retrieved when using one of these terms in a search strategy, and will be easily identifiable as one describing a diagnostic accuracy study.

## Abstract

*Item 2. Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)*

**Example.** See STARD for Abstracts (*manuscript in preparation; checklist will be available at http://www.equator-network.org/reporting-guidelines/stard/*).

**Explanation.** Readers use abstracts to decide whether they should retrieve the full study report and invest time in reading it. In cases where access to the full study report cannot be obtained or where time is limited, it is conceivable that clinical decisions are based on the information provided in abstracts only.

In two recent literature surveys, abstracts of diagnostic accuracy studies published in high-impact journals or presented at an international scientific conference were found insufficiently informative, because key information about the research question, study methods, study results, and the implications of findings were frequently missing.[14 15]

Informative abstracts help readers to quickly appraise critical elements of study validity (risk of bias) and applicability of study findings to their clinical setting (generalisability). Structured abstracts, with separate headings for objectives, methods, results and interpretation, allow readers to find essential information more easily.[16]

Building on STARD 2015, the newly developed STARD for Abstracts provides a list of essential items that should be included in journal and conference abstracts of diagnostic accuracy studies *(list finalized; manuscript under development)*.

## Introduction

*Item 3. Scientific and clinical background, including the intended use and clinical role of the index test*

**Example.** "The need for improved efficiency in the use of emergency department radiography has long been documented. This need for selectivity has been identified clearly for patients with acute ankle injury, who generally are all referred for radiography, despite a yield for fracture of less than 15%. The referral patterns and yield of radiography for patients with knee injuries have been less well described but may be more inefficient than for patients with ankle injuries. […] The sheer volume of low-cost tests such as plain radiography may contribute as much to rising health care costs as do high-technology, low-volume procedures. […] If validated in subsequent studies, a decision rule for knee-injury patients could lead to a large reduction in the use of knee radiography and significant health care savings without compromising patient care."[17]

**Explanation.** In the introduction of scientific study reports, authors should describe the rationale for their study. In doing so they can refer to previous work on the subject, remaining uncertainty, and the clinical implications of this knowledge gap. To help readers in evaluating the implications of the study, authors can clarify the intended use and the clinical role of the test under evaluation, which is referred to as the index test.

The intended use of a test can be diagnosis, screening, staging, monitoring, surveillance, prognosis, treatment selection, or other purposes.[18] The clinical role of the test under evaluation refers to its anticipated position relative to other tests in the clinical pathway.[19] A triage test, for example, will be used before an existing test because it is less costly or burdensome, but often less accurate as well. An add-on test will be used after existing tests, to improve the accuracy of the total test strategy by identifying false positives or false negatives of the initial test. In other cases, a new test may be used to replace an existing test.

Defining the intended use and clinical role of the test will guide the design of the study and the targeted level of sensitivity and specificity; from these definitions follow the eligibility criteria, how and where to identify eligible participants, how to perform tests, and how to interpret test results.[19]

Specifying the clinical role is helpful in assessing the relative importance of potential errors (false positives and false negatives) made by the index test. A triage test to rule out disease, for example, will need very high sensitivity, whereas one that mainly aims to rule in disease will need very high specificity.

*In the example*, the intended use is diagnosis of knee fractures in patients with acute knee injuries, and the potential clinical role is triage test; radiography, the existing test, would only be performed in those with a positive outcome of the newly developed decision rule. The authors outline the current scientific and clinical background of the health problem studied, and their reason for aiming to develop a triage test: this would reduce the number of radiographs and, consequently, healthcare costs.

*Item 4. Study objectives and hypotheses*

**Example (1).** "The objective of this study was to evaluate the sensitivity and specificity of 3 different diagnostic strategies: a single rapid antigen test, a rapid antigen test with a follow-up rapid antigen test if negative (rapid-rapid diagnostic strategy), and a rapid antigen test with follow-up culture if negative (rapid-culture) — the AAP diagnostic strategy—all compared with a 2-plate culture gold standard. In

addition, […] we also compared the ability of these strategies to achieve an absolute diagnostic test sensitivity of >95%."[20]

**Example (2).** "Our 2 main hypotheses were that rapid antigen detection tests performed in physician office laboratories are more sensitive than blood agar plate cultures performed and interpreted in physician office laboratories, when each test is compared with a simultaneous blood agar plate culture processed and interpreted in a hospital laboratory, and rapid antigen detection test sensitivity is subject to spectrum bias".[21]

**Explanation.** Clinical studies may have a general aim (a long term goal, such as "to improve the staging of oesophageal cancer"), specific objectives (well defined goals for this particular study), and testable hypotheses (statements than can be falsified by the study results).

In diagnostic accuracy studies, statistical hypotheses are typically defined in terms of acceptability criteria for single tests (minimum levels of sensitivity, specificity, or other measures). In those cases, hypotheses generally include a quantitative expression of the expected value of the diagnostic parameter. In other cases, statistical hypotheses are defined in terms of equality or non-inferiority in accuracy when comparing two or more index tests.

A priori specification of the study hypotheses limits the chances of post-hoc data-dredging with spurious findings, premature conclusions about the performance of tests, or subjective judgment about the accuracy of the test. Objectives and hypotheses also guide sample size calculations. An evaluation of 126 reports of diagnostic test accuracy studies published in high-impact journals in 2010 revealed that 88% did not state a clear hypothesis.[22]

*In the first example*, the authors' objective was to evaluate the accuracy of three diagnostic strategies; their specific hypothesis was that the sensitivity of any of these would exceed the pre-specified value of 95%. *In the second example*, the authors explicitly describe the hypotheses they want to explore in their study. The first hypothesis is about the comparative sensitivity of two index tests (rapid antigen

detection test vs. culture performed in physician office laboratories); the second is about variability of rapid test performance according to patient characteristics (spectrum bias).

## Methods

*Item 5. Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)*

**Example.** "We reviewed our database of patients who underwent needle localization and surgical excision with digital breast tomosynthesis guidance from April 2011 through January 2013. […] The patients' medical records and images of the 36 identified lesions were then reviewed retrospectively by an author with more than 5 years of breast imaging experience after a breast imaging fellowship."[23]

**Explanation.** If authors define the study question before index test and reference standards are performed, they can take appropriate actions for optimizing procedures according to the study protocol and for dedicated data collection.[24]

Sometimes the idea for a study originates when patients have already undergone the index test and the reference standard. If so, data collection relies on reviewing patient charts or extracting data from registries. Though such retrospective studies can sometimes reflect routine clinical practice better than prospective studies, they may fail to identify all eligible patients, and often result in data of lower quality, with more missing data points.[24] A reason for this could be, for example, that in daily clinical practice, not all patients undergoing the index test may proceed to have the reference standard.

*In the example*, the data were clearly collected retrospectively: participants were identified through database screening, clinical data were abstracted from patients' medical records, though images were re-interpreted.

*Item 6. Eligibility criteria*

**Example (1).** "Patients eligible for inclusion were consecutive adults (≥18 years) with suspected pulmonary embolism, based on the presence of at least one of the following symptoms: unexplained (sudden) dyspnoea, deterioration of existing dyspnoea, pain on inspiration, or unexplained cough. We excluded patients if they received anticoagulant treatment (vitamin K antagonists or heparin) at presentation, they were pregnant, follow-up was not possible, or they were unwilling or unable to provide written informed consent."[25]

**Example (2).** "The cross-sectional cohort included 529 patients with Alzheimer dementia and 304 cognitively healthy controls, and the longitudinal cohort 750 patients with Mild Cognitive Impairment."[26].

**Explanation.** Since a diagnostic accuracy study describes the behavior of a test under particular circumstances, a report of the study must include a complete description of the criteria that were used to identify eligible participants. Eligibility criteria are usually related to the nature and stage of the target condition and the intended future use of the index test; they often include the signs, symptoms, or previous test results that generate the suspicion about the target condition. Additional criteria can be used to exclude participants for reasons of safety, feasibility, and ethical arguments.

Excluding patients with a specific condition or receiving a specific treatment known to adversely affect the way the test works can lead to inflated diagnostic accuracy estimates.[27] An example is the exclusion of patients using beta-blockers in studies evaluating the diagnostic accuracy of exercise electrocardiography.

Some studies have one set of eligibility criteria for all study participants; these are sometimes referred to as single-gate or cohort studies. Other studies have one set of eligibility criteria for participants with the target condition, and (an)other set(s) of eligibility criteria for those without the target condition; these are called multiple-gate or case-control studies.[28]

*In the first example*, the eligibility criteria list presenting signs and symptoms, an age limit, and exclusion based on specific conditions and treatments. Because the same set of eligibility criteria applies to all study participants, this is an example of a single-gate study.

*In the second example*, the authors used different eligibility criteria for participants with and without the target condition: one group consisted of patients with a confirmed diagnosis of Alzheimer disease, one group consisted of patients with confirmed Mild Cognitive Impairment, and one group of participants consisted of healthy controls; this is an example of a multiple-gate study. Extreme contrasts between severe cases and healthy controls can lead to inflated estimates of accuracy.[6 29]

*Item 7. On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)*

**Example.** "We reviewed our database of patients who underwent needle localization and surgical excision with digital breast tomosynthesis guidance from April 2011 through January 2013."[23]

**Explanation.** The eligibility criteria specify who can participate in the study, but they do not describe how the study authors identified eligible participants. This can be done in various ways.[30] A general practitioner may evaluate every patient for eligibility that he sees during office hours. Researchers can go through registries in an emergency department, to identify potentially eligible patients. In other studies, patients are only identified after having been subjected to the index test. Still other studies start with patients in whom the reference standard was performed. Many retrospective studies include participants based on searching hospital databases for patients that underwent both the index test and the reference standard.[31]

Differences in methods for identifying eligible patients can affect the spectrum and prevalence of the target condition in the study group, as well as the range and relative frequency of alternative conditions in patients without the target condition.[32] These differences can influence the estimates of diagnostic accuracy.

*In the example*, participants were identified through searching a patient database and were included if they underwent both the index test and the reference standard.

*Item 8. Where and when potentially eligible participants were identified (setting, location, and dates)*

**Example.** "The study was conducted at the Emergency Department of a university-affiliated children's hospital between January 21, 1996, and April 30, 1996."[33]

**Explanation.** The results of a diagnostic accuracy study reflect the performance of a test in a particular clinical context and setting. A medical test may perform differently in a primary, secondary or tertiary care setting, for example. Authors should therefore report the actual setting in which the study was performed, as well as the exact locations: names of the participating centers, city and country. The spectrum of the target condition as well as the range of other conditions that occur in patients suspected of the target condition can vary across settings, depending on which referral mechanisms are in play.[34-36]

Since test procedures, referral mechanisms, and the prevalence and severity of diseases can evolve over time, authors should also report the start and end dates of participant recruitment.

This information is essential for readers who want to evaluate the generalisability of the study findings, and their applicability to specific questions, for those who would like to use the evidence generated by the study to make informed health care decisions.

*In the example*, study setting and study dates were clearly defined.

*Item 9. Whether participants formed a consecutive, random or convenience series*

**Example.** "All subjects were evaluated and screened for study eligibility by the first author (E.N.E.) prior to study entry. This was a convenience sample of children with pharyngitis; the subjects were enrolled when the first author was present in the emergency department."[37]

**Explanation.** The included study participants may be either a consecutive series of all patients evaluated for eligibility at the study location and satisfying the inclusion criteria, or a subselection of these. A subselection can be purely random, produced by using a random numbers table, or less random, if patients are only enrolled on specific days or during specific office hours. In that case, included participants may not be considered a representative sample of the targeted population, and the generalisability of the study results may be jeopardized.[2 29]

*In the example*, the authors explicitly described a convenience series where subjects were enrolled based on their accessibility to the clinical investigator.

*Item 10a. Index test, in sufficient detail to allow replication*

*Item 10b. Reference standard, in sufficient detail to allow replication*

**Example.** "An intravenous line was inserted in an antecubital vein and blood samples were collected into serum tubes before (baseline), immediately after, and 1.5 and 4.5 h after stress testing. Blood samples were put on ice, processed within 1 h of collection, and later stored at -80 °C before analysis. The samples had been through 1 thaw–freeze cycle before cardiac troponin I (cTnI) analysis. We measured cTnI by a prototype hs assay (ARCHITECT *STAT* high-sensitivity troponin, Abbott Diagnostics) with the capture antibody detecting epitopes 24–40 and the detection antibody epitopes 41–49 of cTnI. The limit of detection (LoD) for the high sensitivity (hs) cTnI assay was recently reported by other groups to be 1.2 ng/L, the 99th percentile 16 ng/L, and the assay 10% coefficient of variation (CV) 3.0 ng/L. […] Samples with concentrations below the range of the assays were assigned values of 1.2 […] for cTnI. […] ."[38]

**Explanation.** Differences in the execution of the index test or reference standard are a potential source of variation in diagnostic accuracy.[39 40] Authors should therefore describe the methods for executing the index test and reference standard, in sufficient detail to allow other researchers to replicate the study,

and to allow readers to assess (1) the feasibility of using the index test in their own setting, (2) the adequacy of the reference standard, and (3) the applicability of the results to their clinical question. The description should cover key elements of the test protocol, including details of:

a. the pre-analytical phase, for example, patient preparation such as fasting/feeding status prior to blood sampling, the handling of the sample prior to testing and its limitations (such as sample instability), or the anatomic site of measurement;
b. the analytical phase, including materials and instruments and analytical procedures;
c. the post-analytical phase, such as calculations of risk scores using analytical results and other variables.

Between-study variability in measures of test accuracy due to differences in test protocol has been documented for a number of tests, including the use of hyperventilation prior to exercise electrocardiography and the use of tomography for exercise thallium scintigraphy.[27 40]

The number, training and expertise of the persons executing and reading the index test and the reference standard may also be critical. Many studies have shown between-reader variability, especially in the field of imaging.[41 42] The quality of reading has also been shown to be affected in cytology and microbiology by professional background, expertise, and prior training to improve interpretation and to reduce inter-observer variation.[43-45] Information about the amount of training of the persons in the study who read the index test can help readers to judge whether similar results are achievable in their own settings.

In some cases, a study depends on multiple reference standards. Patients with lesions on an imaging test under evaluation may, for example, undergo biopsy with a final diagnosis based on histology, whereas patients without lesions on the index test undergo clinical follow-up as reference standard. This could be a potential source of bias, so authors should specify which patient groups received which reference standard.[2 3]

More specific guidance for specialized fields of testing, or certain types of tests, will be developed in future STARD extensions. Whenever available, these extensions will be made available on the STARD pages at the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) website (http://www.equator-network.org/).

*In the example*, the authors described how blood samples were collected and processed in the laboratory. They also report analytical performance characteristics of the index test device, as obtained in previous studies.

*Item 11. Rationale for choosing the reference standard (if alternatives exist)*

**Example.** "The MINI [Mini International Neuropsychiatric Inventory] was developed as a short and efficient diagnostic interview to be used in both research and clinical settings *(reference supporting this statement provided by the authors)*. It has good reliability and validity rates compared with other gold standard diagnostic interviews, such as the Structured Clinical Interview for DSM [Diagnostic and Statistical Manual of Mental Disorders] Disorders (SCID) and the Composite International Diagnostic Interview *(references supporting this statement provided by the authors)*."[46]

**Explanation.** In diagnostic accuracy studies, the reference standard is used for establishing the presence or absence of the target condition in study participants. Several reference standards may be available to define the same target condition. In such cases authors are invited to provide their rationale for selecting the specific reference standard from the available alternatives. This may depend on the intended use of the index test, the clinical relevance, or practical and/or ethical reasons.

Alternative reference standards are not always in perfect agreement. Some reference standards are less accurate than others. In other cases, different reference standards reflect related but different manifestations or stages of the disease, as in confirmation by imaging (first reference standard) versus clinical events (second reference standard).

*In the example*, the authors selected the MINI, a structured diagnostic interview commonly used for psychiatric evaluations, as the reference standard for identifying depression and suicide risk in adults with epilepsy. As a rationale for their choice, they claimed that the MINI test was short to administer, efficient both for clinical and research purposes, reliable, and valid as compared to alternative diagnostic interviews.

*Item 12a. Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory*

*Item 12b. Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory*

**Example.** "We also compared the sensitivity of the risk-model at the specificity that would correspond to using a fixed FIT [fecal immunochemical test] positivity threshold of 50 ng/ml. We used a threshold of 50 ng/ml because this was the anticipated cut-off for the Dutch screening programme at the time of the study."[47]

**Explanation.** Test results in their original form can be dichotomous (positive versus negative), have multiple categories (as in high, intermediate, or low risk), or be continuous (interval or ratio scale). For tests with multiple categories, or continuous results, the outcomes from testing are often reclassified into positive (disease confirmed) and negative (disease excluded). This is done by defining a threshold: the test positivity cut-off. Results that exceed the threshold would then be called positive index test results. In other studies, an ROC curve is derived, by calculating the sensitivity-specificity pairs for all possible cutoffs.

To evaluate the validity and applicability of these classifications, readers would like to know these positivity cut-offs or result categories, how they were determined, and whether they were defined prior to the study or after collecting the data. Pre-specified thresholds can be based on (1) previous studies, (2) cutoffs used in clinical practice, (3) thresholds recommended by clinical practice guidelines, or (4)

thresholds recommended by the manufacturer. If no such thresholds exist, the authors may be tempted to explore the accuracy for various thresholds after the data have been collected.

If the authors selected the positivity cut-off after performing the test, choosing the one that maximized test performance, there is an increased risk that the resulting accuracy estimates are overly optimistic, especially in small studies.[48 49] Subsequent studies may fail to replicate the findings.[50 51]

*In the example*, the authors stated the rationale for their selection of cut-offs.

*Item 13a. Whether clinical information and reference standard results were available to the performers or readers of the index test*

*Item 13b. Whether clinical information and index test results were available to the assessors of the reference standard*

**Example.** "Images for each patient were reviewed by two fellowship-trained genitourinary radiologists with 12 and 8 years of experience, respectively, who were blinded to all patient information, including the final histopathologic diagnosis."[52]

**Explanation.** Some medical tests, such as most forms of imaging, require human handling, interpretation and judgment. These actions may be influenced by the information that is available to the reader.[1 53 54] This can lead to artificially high agreement between tests, or between the index test and the reference standard.

If the reader of a test has access to information about signs, symptoms and previous test results, the reading may be influenced by this additional information, but this may still represent how the test is used in clinical practice.[2] The reverse may also apply, if the reader does not have enough information for a proper interpretation of the index test outcome. In that case, test performance may be affected downwards, and the study findings may have limited applicability. Either way, readers of the study report should know to which extent such additional information was available to test readers and may have influenced their final judgment.

In other situations the assessors of the reference standard may have had access to the index test results. In those cases, the final classification may be guided by the index test result, and the reported accuracy estimates for the index test will be too high.[1 2 27] Tests that require subjective interpretation are particularly susceptible to this bias.

Withholding information from the readers of the test is commonly referred to as "blinding" or "masking". The point of this reporting item is not that blinding is desirable or undesirable, but, rather, that readers of the study report need information about blinding for both the index test and the reference standard to be able to interpret the study findings.

*In the example*, the readers of unenhanced CT for differentiating between renal angiomyolipoma and renal cell carcinoma did not have access to clinical information, nor to the results of histopathology, the reference standard in this study.

*Item 14. Methods for estimating or comparing measures of diagnostic accuracy*

**Example.** "Statistical tests of sensitivity and specificity were conducted by using the McNemar test for correlated proportions. All tests were two sided, testing the hypothesis that stereoscopic Digital Mammography performance differed from that of Digital Mammography. A p-value of .05 was considered as the threshold for significance."[55]

**Explanation.** Multiple measures of diagnostic accuracy exist to describe the performance of a medical test, and their calculation from the collected data is not always straightforward.[56] Authors should report the methods used for calculating the measures that they considered appropriate for their study objectives.

Statistical techniques can be used to test specific hypotheses, following from the study's objectives. In single test evaluations, authors may want to evaluate if the diagnostic accuracy of the tests exceeds a pre-specified level (e.g. sensitivity of at least 95%, see Item 4).

Diagnostic accuracy studies can also compare two or more index tests. In such comparisons, statistical hypothesis testing usually involves assessing the superiority of one test over another, or the non-inferiority.[57] For such comparisons, authors should indicate what measure they specified to make the comparison; these should match their study objectives, and the purpose and role of the index test relative to the clinical pathway. Examples are the relative sensitivity, the absolute gain in sensitivity, and the relative diagnostic odds ratio.[58]

*In the example*, the authors used McNemar's test statistic to evaluate whether the sensitivity and specificity of stereoscopic Digital Mammography differed from that of Digital Mammography in patients with elevated risk for breast cancer. In itself, the resulting p-value is not a quantitative expression of the relative accuracy of the two investigated tests. Like any p-value it is influenced by both the magnitude of the difference in effect and the sample size. In the example, the authors could have calculated the relative or absolute difference in sensitivity and specificity, including a 95% confidence interval that takes into account the paired nature of the data.

*Item 15. How indeterminate index test or reference standard results were handled*

**Example.** "Indeterminate results were considered false-positive or false-negative and incorporated into the final analysis. For example, an indeterminate result in a patient found to have appendicitis was considered to have had a negative test result."[59]

**Explanation.** Indeterminate results refer to those that are neither positive or negative.[60] Such results can occur both on the index test and the reference standard, and are a challenge when evaluating the performance of a diagnostic test.[60-63] The occurrence of indeterminate test results varies from test to test, but frequencies up to 40% have been reported.[62]

There are many underlying causes for indeterminate test results.[62 63] A test may fail because of technical reasons or an insufficient sample, for example, in the absence of cells in a needle biopsy from a tumor.[43 64 65] Sometimes test results are not reported as just positive or negative, as in the case of ventilation-

perfusion scanning in suspected pulmonary embolism, where the findings are classified in three categories: normal, high probability, or inconclusive.[66]

In itself, the frequency of indeterminate test results is an important indicator of the feasibility of the test, and typically limits the overall clinical usefulness; therefore, authors are encouraged to always report the respective frequencies with reasons, as well as failures to complete the testing procedure. This applies both to the index test and the reference standard.

Ignoring indeterminate test results can produce biased estimates of accuracy if these results do not occur at random. Clinical practice may guide the decision on how to handle indeterminate results. There are multiple ways for handling indeterminate test results in the analysis when estimating accuracy and expressing test performance.[63] They can be ignored altogether, be reported but not accounted for, or handled as a separate test result category. Handling these results as a separate category may be useful when indeterminate results occur more often, for example, in those without the target condition than in those with the target condition. It is also possible to reclassify all such results: as false positives or false negatives, depending on the reference standard result ("worst-case scenario"), or as true positives and true negatives ("best-case scenario").

*In the example*, the authors explicitly chose a conservative approach by considering all indeterminate results from the index test as being false-negative (in those with the target condition) or false-positive (in all others), a strategy sometimes referred to as the "worst-case scenario".

*Item 16. How missing data on the index test and reference standard were handled*

**Example.** "One vessel had missing $FFR_{CT}$ and 2 had missing CT data. Missing data were handled by exclusion of these vessels as well as by the worst-case imputation."[67]

**Explanation.** Missing data are common in any type of biomedical research. In diagnostic accuracy studies, they can occur for both the index test and reference standard. There are several ways to deal with them when analyzing the data.[68] Many researchers exclude participants without an observed test

result. This is known as “complete case” or “available case” analysis. This may lead to a loss in precision and can introduce bias, especially if having a missing index test or reference standard result is related to having the target condition.

Participants with missing test results can be included in the analysis if missing results are imputed.[68-70] Another option is to assess the impact of missing test results on estimates of accuracy by considering different scenarios. For the index test, for example, in the “worst-case scenario”, all missing index test results are considered false-positive or false-negative depending on the reference standard result; in the “best-case scenario”, all missing index test results are considered true-positive or true-negative.

*In the example*, the authors explicitly reported how many cases with missing index test data they encountered and how they handled these data: they excluded them, but also applied a “worst-case scenario”.

*Item 17. Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory*

**Example.** “To assess the performance of urinary indices or their changes over the first 24 hours in distinguishing transient AKI [acute kidney injury] from persistent AKI, we plotted the receiver-operating characteristic curves for the proportion of true positives against the proportion of false positives, depending on the prediction rule used to classify patients as having persistent AKI. The same strategy was used to assess the performance of indices and their changes over time in two predefined patient subgroups; namely, patients who did not receive diuretic therapy and patients without sepsis.”[71]

**Explanation.** The relative proportion of false positive or false-negative results of a diagnostic test may vary depending on patient characteristics, experience of readers, the setting, and previous test results.[2 3] Researchers may therefore want to explore possible sources of variability in test accuracy within their study. In such analyses, investigators typically assess differences in accuracy across subgroups of participants, readers or centers.

Post hoc analyses, done after looking at the data, carry a high risk for spurious findings. The results are especially likely not to be confirmed by subsequent studies. Analyses that were pre-specified in the protocol, before data were collected, have greater credibility.[72]

*In the example*, the authors reported that the accuracy of the urinary indices was evaluated in two subgroups that were explicitly pre-specified.

*Item 18. Intended sample size and how it was determined*

**Example.** "Study recruitment was guided by an expected 12% prevalence of adenomas 6 mm or larger in a screening cohort and a point estimate of 80% sensitivity for these target lesions. We planned to recruit approximately 600 participants to achieve margins of sampling error of approximately 8 percentage points for sensitivity. This sample would also allow 90% power to detect differences in sensitivity between computed tomographic colonography and optical colonoscopy of 18 percentage points or more."[73]

**Explanation.** Performing sample size calculations when developing a diagnostic accuracy study may ensure that a sufficient amount of precision is reached. Sample size calculations also take into account the specific objectives and hypotheses of the study.

Readers may want to know how the sample size was determined, and whether the assumptions made in this calculation are in line with the scientific and clinical background, and the study objectives. Readers will also want to learn whether the study authors were successful in recruiting the targeted number of participants. Methods for performing sample size calculations in diagnostic research are widely available,[74-76] but such calculations are not always performed or provided in reports of diagnostic accuracy studies.[77 78]

Many diagnostic accuracy studies are small; a systematic survey of studies published in eight leading journals in 2002 found a median sample size of 118 participants (interquartile range 71-350).[77] Estimates

of diagnostic accuracy from small studies tend to be imprecise, with wide confidence intervals around them.

*In the example*, the authors reported in detail to achieve a desired level of precision for an expected sensitivity of 80%.

## Results

*Item 19. Flow of participants, using a diagram*

**Example.** "Between 1 June 2008 and 30 June 2011, 360 patients were assessed for initial eligibility and invited to participate. The figure shows the flow of patients through the study, along with the primary outcome of advanced colorectal neoplasia. Patients who were excluded (and reasons for this) or who withdrew from the study are noted. In total, 229 patients completed the study, a completion rate of 64%."[79] (See Figure 1)

**Explanation.** Estimates of diagnostic accuracy may be biased if not all eligible participants undergo both the index test and the desired reference standard.[80-86] This includes studies in which not all study participants undergo the reference standard, as well as studies where some of the participants receive a different reference standard.[70] Incomplete verification by the reference standard occurs in up to 26% of diagnostic studies; it is especially common when the reference standard is an invasive procedure.[84]

To allow the readers to appreciate the potential for bias, authors are invited to build a diagram to illustrate the flow of participants through the study. Such a diagram also illustrates the basic structure of the study. An example of a prototypical STARD flow diagram is presented in Figure 2.

By providing the exact number of participants at each stage of the study, including the number of true positive, false positive, true negative, and false negative index test results, the diagram also helps identifying the correct denominator for calculating proportions such as sensitivity and specificity. The diagram should also specify the number of participants that were assessed for eligibility, the number of

subjects who did not receive either the index test and/or the reference standard, and the reasons for that. This helps readers to judge the risk of bias, but also the feasibility of the evaluated testing strategy, and the applicability of the study findings.

*In the example*, the authors very briefly described the flow of participants, and referred to a flow diagram in which the number of participants and corresponding test results at each stage of the study were provided, as well as detailed reasons for excluding participants (Figure 1).

*Item 20. Baseline demographic and clinical characteristics of participants*

**Example.** "The median age of participants was 60 years (range 18–91), and 209 participants (54.7%) were female. The predominant presenting symptom was abdominal pain, followed by rectal bleeding and diarrhea, whereas fever and weight loss were less frequent. At physical examination, palpation elicited abdominal pain in almost half the patients, but palpable abdominal or rectal mass was found in only 13 individuals (Table X)."[87] (See Table 3)

**Explanation.** The diagnostic accuracy of a test can depend on the demographic and clinical characteristics of the population in which it is applied.[2 3 88-92] These differences may reflect variability in the extent or severity of disease, which affects sensitivity, or in the alternative conditions that are able to generate false positive findings, affecting specificity.[85]

An adequate description of the demographic and clinical characteristics of study participants allows the reader to judge whether the study can adequately address the study question, and whether the study findings apply to the reader's clinical question.

*In the example*, the authors presented the demographic and clinical characteristics of the study participants in a separate table, a commonly used, informative way of presenting key participant characteristics (Table 3).

*Item 21a. Distribution of severity of disease in those with the target condition*

*Item 21b. Distribution of alternative diagnoses in those without the target condition*

**Example.** "Of the 170 patients with coronary disease, one had left main disease, 53 had three vessel disease, 64 two vessel disease, and 52 single vessel disease. The mean ejection fraction of the patients with coronary disease was 64% (range 37-83). The other 52 men with symptoms had normal coronary arteries or no significant lesions at angiography."[93]

**Explanation.** Most target conditions are not fixed states, either present or absent; many diseases cover a continuum, ranging from minute pathological changes to advanced clinical disease. Test sensitivity is often higher in studies in which more patients have advanced stages of the target condition, as these cases are often easier to identify by the index test.[28 85] The type, spectrum and frequency of alternative diagnoses in those without the target condition may also influence test accuracy; typically, the healthier the patients without the target condition, the less frequently one would find false-positive results of the index test.[28]

An adequate description of the severity of disease in those with the target condition and of the alternative conditions in those without it allows the reader to judge both the validity of the study, relative to the study question, and the applicability of the study findings to the reader's clinical question.

*In the example*, the authors investigated the accuracy of exercise tests for diagnosing coronary artery disease. They reported the distribution of severity of disease in terms of the number of vessels involved; the more vessels, the more severe the coronary artery disease would be. Sensitivity of test exercises was higher in those with more diseased vessels (39% for single vessel disease, 58% for two and 77% for three vessels).[91]

*Item 22. Time interval and any clinical interventions between index test and reference standard*

**Example.** "The mean time between arthrometric examination and MR imaging was 38.2 days (range, 0–107 days)."[94]

**Explanation.** Studies of diagnostic accuracy are essentially cross-sectional investigations. In most cases, one wants to know how well the index test classified patients in the same way as the reference standard, when both tests are performed in the same patients, at the same time.[30] When a delay occurs between the index test and the reference standard, the target condition and alternative conditions can change; conditions may worsen, or improve in the meanwhile, due to the natural course of the disease, or due to clinical interventions applied between the two tests. Such changes influence the agreement between the index test and the reference standard, which could lead to biased estimates of test performance.

The bias can be more severe if the delay differs systematically between test positives and test negatives, or between those with a high prior suspicion of having the target condition and those with a low suspicion.[1 2]

When follow-up is used as the reference standard, readers will want to know how long the follow-up period was.

*In the example*, the authors reported the mean number of days, and a range, between the index test and the reference standard.

*Item 23. Cross tabulation of the index test results (or their distribution) by the results of the reference standard*

**Example.** "Table X shows pain over speed bumps in relation to diagnosis of appendicitis."[95] (see Table 4)

**Explanation.** Research findings should be reproducible and verifiable by other scientists; this applies both to the testing procedures, to the conduct of the study, and to the statistical analyses.

A cross tabulation of index test results against reference standard results facilitates recalculating measures of diagnostic accuracy. It also facilitates recalculating the proportion of study group participants with the target condition, which is useful as the sensitivity and specificity of a test may vary

with disease prevalence.[32 96] It also allows for performing alternative or additional analyses, such as meta-analysis.

Preferably, such tables should include actual numbers, not just percentages, because mistakes made by study authors in calculating estimates for sensitivity and specificity are not rare.

*In the example*, the authors provided a contingency table from which the number of true positives, false positives, false negatives, and true negatives can be easily identified (Table 4).

*Item 24. Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)*

**Example.** "Forty-six patients had pulmonary fibrosis at CT, and sensitivity and specificity of MR imaging in the identification of pulmonary fibrosis were 89% (95% CI: 77%, 96%) and 91% (95% CI: 76%, 98%), respectively, with positive and negative predictive values of 93% (95% CI: 82%, 99%) and 86% (95% CI: 70%, 95%), respectively."[97]

**Explanation.** Diagnostic accuracy studies never determine a test's 'true' sensitivity and specificity; at best the data collected in the study can be used to calculate valid estimates of sensitivity and specificity. These estimates can be more or less precise, i.e. closer or farther away from the true value. The smaller the number of study participants, the less precise these estimates will be.[98] Although one never knows how far a single estimate is from the true, some more general statistical principles allow researchers to express how likely the estimates are to approximate the true value.

The most frequently used expression of precision is to report not just the estimates – sometimes referred to as point estimates - but also 95% confidence intervals around the estimates. If a series of studies each reports a 95% confidence interval, then 95% of these intervals include the true value. From this fact one could derive a statement about a single 95% confidence interval: it has a 95% chance to include the true value. Results from studies with imprecise estimates of accuracy should be interpreted with caution, as over-optimism lurks.[22]

*In the example*, where MRI is the index test and CT the reference standard, the authors reported both point estimates and 95% confidence intervals around them, for sensitivity, specificity, and positive and negative predictive value.

*Item 25. Any adverse events from performing the index test or the reference standard*

**Example.** "No significant adverse events occurred as a result of colonoscopy. Four (2%) patients had minor bleeding in association with polypectomy that was controlled endoscopically. Other minor adverse events are noted in the appendix."[79]

**Explanation.** Not all medical tests are equally safe, and in this they do not differ from many other medical interventions.[99 100] The testing procedure can lead to complications, such as perforations with endoscopy, contrast allergic reactions in CT imaging, or claustrophobia with MRI scanning. Measuring and reporting of adverse events in studies of diagnostic accuracy will provide additional information to clinicians, who may be reluctant to use them if they produce severe or frequent adverse events. Actual application of a test in clinical practice will not just be guided by the test's accuracy, but by several other dimensions as well, including feasibility and safety. This also applies to the reference standard.

*In the example*, the authors distinguished between "significant" and "minor" adverse events, and explicitly reported how often these were observed.

## Discussion

*Item 26. Study limitations, including sources of potential bias, statistical uncertainty, and generalisability*

**Example.** "This study had limitations. First, not all patients who underwent CT colonography (CTC) were assessed by the reference standard methods. […] However, considering that the 41 patients who were eligible but did not undergo the reference standard procedures had negative or only mildly positive CTC

findings, excluding them from the analysis of CTC diagnostic performance may have slightly overestimated the sensitivity of CTC (i.e., partial verification bias). Second, there was a long time interval between CTC and the reference methods in some patients, predominately those with negative CTC findings. […] If anything, the prolonged interval would presumably slightly underestimate the sensitivity and NPV of CTC for non-cancerous lesions, since some 'missed' lesions could have conceivably developed or increased in size since the time of CTC."[101]

**Explanation.** Like other clinical trials and studies, diagnostic accuracy studies are at risk of bias; they can generate estimates of the test's accuracy that do not reflect the true performance of the test, due to flaws or deficiencies in study design and analysis.[1 2] In addition, imprecise accuracy estimates, with wide confidence intervals, should be interpreted with caution. Because of differences in design, participants and procedures, the findings generated by one particular diagnostic accuracy study may not be obtained in other conditions; their generalisability may be limited.[102]

In the discussion section, authors should critically reflect on the validity of their findings, and address potential limitations. As bias can come down to over- or underestimation of the accuracy of the index test under investigation, authors should discuss the direction of potential bias, along with its likely magnitude. Readers are then informed of the likelihood that the limitations jeopardize the study's results and conclusions (see also Item 27).[103]

Some journals explicitly encourage authors to report on study limitations, but many are not specific about which elements should be addressed.[104] For diagnostic accuracy studies, we highly recommend that at least potential sources of bias are discussed, as well as imprecision, and concerns related to the selection of patients and the setting in which the study was performed.

*In the example*, the authors identified two potential sources of bias that are common in diagnostic accuracy studies: not all test results were verified by the reference standard, and there was a time interval between index test and reference standard, allowing the target condition to change. They also

discussed the magnitude of this potential bias, and the direction: whether this may have led to over- or underestimations of test accuracy.

*Item 27. Implications for practice, including the intended use and clinical role of the index test*

**Example.** "A Wells score of ≤4 combined with a negative point of care D-dimer test result ruled out pulmonary embolism in 4-5 of 10 patients, with a failure rate of less than 2%, which is considered safe by most published consensus statements. Such a rule-out strategy makes it possible for primary care doctors to safely exclude pulmonary embolism in a large proportion of patients suspected of having the condition, thereby reducing the costs and burden to the patient (for example, reducing the risk of contrast nephropathy associated with spiral computed tomography) associated with an unnecessary referral to secondary care."[25]

**Explanation.** To make the study findings relevant for practice, authors of diagnostic accuracy studies should elaborate on the consequences of their findings, taking into account the intended use (the purpose of testing) and clinical role of the test (how will the test be positioned in the existing clinical pathway).

A test can be proposed for diagnostic purposes, for susceptibility, screening, risk stratification, staging, prediction, prognosis, treatment selection, monitoring, surveillance, or other purposes. The clinical role of the test reflects its positioning relative to existing tests for the same purpose, within the same clinical setting: triage, add-on, or replacement.[19 105] Both the intended use and the clinical role of the index test should have been described in the introduction of the paper (Item 3).

The intended use and the proposed role will guide the desired magnitude of the measures of diagnostic accuracy. For ruling-out disease with an inexpensive triage test, for example, high sensitivity is required, and less-than-perfect specificity may be acceptable. If the test is supposed to rule-in disease, specificity may become much more important.[106]

In the Discussion section, authors should elaborate on whether or not the accuracy estimates are sufficient for considering the test to be 'fit for purpose'.

*In the example*, the authors concluded that the combination of a Wells score ≤4 and a negative point-of-care D-dimer result could reliably rule-out pulmonary embolism in a large proportion of patients seen in primary care.

## Other information

*Item 28. Registration number and name of registry*

**Example.** "The study was registered at http://www.clinicaltrials.org (NCT00916864)." [107]

**Explanation.** Registering study protocols before their initiation in a clinical trial registry, such as ClinicalTrials.gov or one of the WHO Primary Registries, ensures that existence of the studies can be identified.[108-112] This has many advantages, including avoiding overlapping or redundant studies, and allowing colleagues and potential participants to contact the study coordinators.

Additional benefits of study registration are the prospective definition of study objectives, outcome measures, eligibility criteria and data to be collected, allowing editors, reviewers and readers to identify deviations in the final study report. Trial registration also allows reviewers to identify studies that have been completed but were not yet reported.

Many journals require registration of clinical trials. A low but increasing number of diagnostic accuracy studies are also being registered. In a recent evaluation of 351 test accuracy studies published in high-impact journals in 2012, 15% had been registered.[113]

Including a registration number in the study report facilitates identification of the trial in the corresponding registry. It can also be regarded as a sign of quality, if the trial was registered before its initiation.

*In the example*, the authors reported that the study was registered at ClinicalTrials.gov. The registration number was also provided, so that the registered record could be easily retrieved.

*Item 29. Where the full study protocol can be accessed*

**Example.** "The design and rationale of the OPTIMAP study have been previously published in more detail [with reference to study protocol]."[114]

**Explanation.** Full study protocols typically contain additional methodological information that is not provided in the final study report, because of word limits, or because it has been reported elsewhere. This additional information can be helpful for those who want to thoroughly appraise the validity of the study, for researchers who want to replicate the study, and for practitioners who want to implement the testing procedures.

An increasing number of researchers share their original study protocol, often before enrollment of the first participant in the study. They may do so by publishing the protocol in a scientific journal, at an institutional or sponsor website, or as supplementary material on the journal website, to accompany the study report.

If the protocol has been published or posted online, authors should provide a reference or a link. If the study protocol has not been published authors should state from whom it can be obtained.[115]

*In the example*, the authors provided a reference to the full protocol, which had been published previously.

*Item 30. Sources of funding and other support; role of funders*

**Example.** "Funding, in the form of the extra diagnostic reagents and equipment needed for the study, was provided by Gen-Probe. The funders had no role in the initiation or design of the study, collection of samples, analysis, interpretation of data, writing of the paper, or the submission for publication. The study and researchers are independent of the funders, Gen-Probe."[116]

**Explanation.** Sponsorship of a study by a pharmaceutical company has been shown to be associated with results favoring the interests of that sponsor.[117] Unfortunately, sponsorship is often not disclosed in scientific articles, making it difficult to assess this potential bias. Sponsorship can consist of direct funding of the study, or of the provision of essential study materials, such as test devices.

The role of the sponsor, including the degree to which that sponsor was involved in the study varies. A sponsor could, for example, be involved in the design of the study, but also in the conduct, analysis, reporting, and decision to publish. Authors are encouraged to be explicit about sources of funding as well as the sponsors role(s) in the study, as this transparency helps readers to appreciate the level of independency of the researchers.

*In the example*, the authors were explicit about the contribution from the sponsor, and their independence in each phase of the study.

## ACKNOWLEDGEMENTS

We thank the STARD Group for helping us in identifying essential items for reporting diagnostic accuracy studies.

## COMPETING INTERESTS

All authors have completed the ICMJE Unified Competing Interest form atwww.icmje.org/coi_disclosure.pdf (available upon request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## AUTHOR CONTRIBUTIONS

JFC, DAK, PMMB: drafting of manuscript. DGA, DEB, CAG, LH, LI, DL, JBR, HCWdV: critical revision of manuscript.

## DATA SHARING STATEMENT

No additional data available.

## TABLES

**Table 1. Key STARD terminology.**

| Term | Explanation |
|---|---|
| Medical test | Any method for collecting additional information about the current or future health status of a patient. |
| Index test | The test under evaluation. |
| Target condition | The disease or condition that the index test is expected to detect. |
| Clinical reference standard | The best available method for establishing the presence or absence of the target condition. A gold standard would be an error-free reference standard. |
| Sensitivity | Proportion of those with the target condition who test positive with the index test. |
| Specificity | Proportion of those without the target condition who test negative with the index test. |
| Intended use of the test | Whether the index test is used for diagnosis, screening, staging, monitoring, surveillance, prediction, prognosis, or other reasons. |
| Role of the test | The position of the index test relative to other tests for the same condition (e.g. triage, replacement, add-on, new test). |
| Indeterminate results | Results that are neither positive or negative |

**Table 2. The STARD 2015 list.**[10]

| Section and topic | No | 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 3. Example of baseline demographic and clinical characteristics of participants in a study evaluating the accuracy of point-of-care fecal tests for diagnosis of organic bowel disease (adapted from Kok et al.[87], with permission).**

| Patient characteristics | n (%) |
|---|---|
| **Geographic region of residency in the Netherlands** | |
| Central (Gelderse Vallei) | 257 (66.6) |
| South (Oostelijke Mijnstreek) | 129 (33.4) |
| **Median age, years (range)** | 60 (18–91) |
| **Women** | 211 (54.7) |
| **Presenting symptoms** | |
| Rectal blood loss | 141 (37.7) |
| Abdominal pain | 267 (70.6) |
| Median duration of abdominal pain (range) | 150 days (1 day to 30 years) |
| Persistent diarrhea | 40 (16.9) |
| Diarrhea | 131 (37.2) |
| Fever | 40 (11.0) |
| Weight loss | 62 (17.1) |
| Bloating | 195 (53.6) |
| Constipation | 169 (46.6) |
| **Physical examination** | |
| Pain at palpation | 117 (46.8) |
| Palpable abdominal mass | 12 (3.0) |
| Palpable rectal mass | 1 (0.3) |

**Table 4. Example of contingency table from a study evaluating the accuracy of pain over speed bumps for diagnosis of appendicitis (adapted from Ashdown et al.[95], with permission).**

| | Appendicitis | | |
|---|---|---|---|
| **Pain over speed bumps** | Positive | Negative | Total |
| Positive | 33 | 21 | 54 |
| Negative | 1 | 9 | 10 |
| Total | 34 | 30 | 64 |

# FIGURES

**Figure 1. Example of flow diagram from a study evaluating the accuracy of faecal immunochemical testing for diagnosis of advanced colorectal neoplasia (from Collins et al.[79], with permission).**

**Figure 2. STARD 2015 flow diagram.**

## REFERENCES


1. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Annals of internal medicine 2004;**140**(3):189-202.
2. Whiting PF, Rutjes AW, Westwood ME, et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. Journal of clinical epidemiology 2013;**66**(10):1093-104.
3. 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):529-36.
4. Korevaar DA, van Enst WA, Spijker R, et al. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med 2014;**19**(2):47-54.
5. Korevaar DA, Wang J, van Enst WA, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. Radiology 2015;**274**(3):781-9.
6. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;**282**(11):1061-66.
7. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003;**49**(1):1-6.
8. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;**276**(8):637-39.
9. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;**340**:c332.
10. 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.
11. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Annals of internal medicine 2003;**138**(1):W1-12.
12. Regge D, Laudi C, Galatola G, et al. Diagnostic accuracy of computed tomographic colonography for the detection of advanced neoplasia in individuals at increased risk of colorectal cancer. JAMA 2009;**301**(23):2453-61.
13. 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):65-9.

14. Korevaar DA, Cohen JF, Hooft L, et al. Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies. Journal of clinical epidemiology 2015;**68**(6):708-15.

15. Korevaar DA, Cohen JC, de Ronde MW, et al. Reporting Weaknessess in Conference Abstracts of Diagnostic Accuracy Studies in Ophthalmology. Jama Ophthalmology 2015;**133**(12):1464-67.

16. A proposal for more informative abstracts of clinical articles. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. Annals of internal medicine 1987;**106**(4):598-604.

17. Stiell IG, Greenberg GH, Wells GA, et al. Derivation of a decision rule for the use of radiography in acute knee injuries. Ann Emerg Med 1995;**26**(4):405-13.

18. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clinica chimica acta; international journal of clinical chemistry 2014;**427**:49-57.

19. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. Bmj 2006;**332**(7549):1089-92.

20. Gieseker KE, Roe MH, MacKenzie T, et al. Evaluating the American Academy of Pediatrics diagnostic standard for Streptococcus pyogenes pharyngitis: backup culture versus repeat rapid antigen testing. Pediatrics 2003;**111**(6 Pt 1):e666-70.

21. Tanz RR, Gerber MA, Kabat W, et al. Performance of a rapid antigen-detection test and throat culture in community pediatric offices: implications for management of pharyngitis. Pediatrics 2009;**123**(2):437-44.

22. Ochodo EA, de Haan MC, Reitsma JB, et al. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". Radiology 2013;**267**(2):581-8.

23. Freer PE, Niell B, Rafferty EA. Preoperative Tomosynthesis-guided Needle Localization of Mammographically and Sonographically Occult Breast Lesions. Radiology 2015;**275**(2):377-83.

24. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. International journal of epidemiology 1996;**25**(2):435-42.

25. Geersing GJ, Erkens PM, Lucassen WA, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. BMJ (Clinical research ed 2012;**345**:e6564.

26. Mattsson N, Rosen E, Hansson O, et al. Age and diagnostic performance of Alzheimer disease CSF biomarkers. Neurology 2012;**78**(7):468-76.

27. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. The American journal of cardiology 1980;**46**(5):807-12.

28. Rutjes AW, Reitsma JB, Vandenbroucke JP, et al. Case-control and two-gate designs in diagnostic accuracy studies. Clinical chemistry 2005;**51**(8):1335-41.

29. Rutjes AW, Reitsma JB, Di Nisio M, et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;**174**(4):469-76.

30. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. Journal of clinical epidemiology 2003;**56**(11):1118-28.

31. van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. Journal of clinical epidemiology 1995;**48**(3):417-22.

32. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. Journal of clinical epidemiology 2009;**62**(1):5-12.

33. Attia M, Zaoutis T, Eppes S, et al. Multivariate predictive models for group A beta-hemolytic streptococcal pharyngitis in children. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine 1999;**6**(1):8-13.

34. Knottnerus JA, Knipschild PG, Sturmans F. Symptoms and selection bias: the influence of selection towards specialist care on the relationship between symptoms and diagnoses. Theor Med 1989;**10**:67-81.

35. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. Journal of clinical epidemiology 1992;**45**(10):1143-54.

36. Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. Scandinavian journal of primary health care 1993;**11**(4):241-6.

37. Ezike EN, Rongkavilit C, Fairfax MR, et al. Effect of using 2 throat swabs vs 1 throat swab on detection of group A streptococcus by a rapid antigen detection test. Archives of pediatrics & adolescent medicine 2005;**159**(5):486-90.

38. Rosjo H, Kravdal G, Hoiseth AD, et al. Troponin I measured by a high-sensitivity assay in patients with suspected reversible myocardial ischemia: data from the Akershus Cardiac Examination (ACE) 1 study. Clinical chemistry 2012;**58**(11):1565-73.

39. Irwig L, Bossuyt P, Glasziou P, et al. Designing studies to ensure that estimates of test accuracy are transferable. BMJ (Clinical research ed 2002;**324**(7338):669-71.

40. Detrano R, Gianrossi R, Froelicher V. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. Progress in cardiovascular diseases 1989;**32**(3):173-206.

41. Brealey S, Scally AJ. Bias in plain film reading performance studies. The British journal of radiology 2001;**74**(880):307-16.

42. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretations of mammograms. The New England journal of medicine 1994;**331**(22):1493-9.

43. Ronco G, Montanari G, Aimone V, et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. Cytopathology : official journal of the British Society for Clinical Cytology 1996;**7**(3):151-8.

44. Cohen MB, Rodgers RP, Hales MS, et al. Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver operating characteristics curve analysis. Archives of pathology & laboratory medicine 1987;**111**(6):518-20.

45. Fox JW, Cohen DM, Marcon MJ, et al. Performance of rapid streptococcal antigen testing varies by personnel. Journal of clinical microbiology 2006;**44**(11):3918-22.

46. Gandy M, Sharpe L, Perry KN, et al. Assessing the efficacy of 2 screening measures for depression in people with epilepsy. Neurology 2012;**79**(4):371-5.

47. Stegeman I, de Wijkerslooth TR, Stoop EM, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut 2014;**63**(3):466-71.

48. Leeflang MM, Moons KG, Reitsma JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem 2008;**54**(4):729-37.

49. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. Journal of clinical epidemiology 2006;**59**(8):798-801.

50. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine 1999;**130**(6):515-24.

51. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;**15**(4):361-87.

52. Hodgdon T, McInnes MD, Schieda N, et al. Can Quantitative CT Texture Analysis be Used to Differentiate Fat-poor Renal Angiomyolipoma from Renal Cell Carcinoma on Unenhanced CT Images? Radiology 2015:142215.

53. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;**6**(4):411-23.

54. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. AJR American journal of roentgenology 1981;**137**(5):1055-8.

55. D'Orsi CJ, Getty DJ, Pickett RM, et al. Stereoscopic digital mammography: improved specificity and reduced rate of recall in a prospective clinical trial. Radiology 2013;**266**(1):81-8.

56. Knottnerus JA, Buntinx F. *The evidence base of clinical diagnosis: Theory and methods of diagnostic research*. 2nd ed: BMJ Books, 2008.

57. Pepe M. Study design and hypothesis testing. The statistical evaluation of medical tests for classification and prediction. Oxford, UK: Oxford University Press, 2003:214-51.

58. Hayen A, Macaskill P, Irwig L, et al. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. Journal of clinical epidemiology 2010;**63**(8):883-91.

59. Garcia Pena BM, Mandl KD, Kraus SJ, et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. JAMA 1999;**282**(11):1041-6.

60. Simel DL, Feussner JR, DeLong ER, et al. Intermediate, indeterminate, and uninterpretable diagnostic test results. Medical decision making : an international journal of the Society for Medical Decision Making 1987;**7**(2):107-14.

61. Philbrick JT, Horwitz RI, Feinstein AR, et al. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. Jama 1982;**248**(19):2467-70.

62. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. Journal of chronic diseases 1986;**39**(8):575-84.

63. Shinkins B, Thompson M, Mallett S, et al. Diagnostic accuracy studies: how to report and analyse inconclusive test results. Bmj 2013;**346**:f2778.

64. Pisano ED, Fajardo LL, Tsimikas J, et al. Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial: The Radiologic Diagnostic Oncology Group 5 Study. The RDOG5 investigators. Cancer 1998;**82**(4):679-88.

65. Giard RW, Hermans J. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer 1992;**69**(8):2104-10.

66. Investigators P. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). JAMA 1990;**263**(20):2753-9.

67. Min JK, Leipsic J, Pencina MJ, et al. Diagnostic accuracy of fractional flow reserve from anatomic CT angiography. Jama 2012;**308**(12):1237-45.

68. Naaktgeboren CA, de Groot JA, Rutjes AW, et al. Anticipating missing reference standard data when planning diagnostic accuracy studies. Bmj 2016;**352**:i402.

69. van der Heijden GJ, Donders AR, Stijnen T, et al. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. Journal of clinical epidemiology 2006;**59**(10):1102-9.

70. de Groot JA, Bossuyt PM, Reitsma JB, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. Bmj 2011;**343**:d4770.

71. Pons B, Lautrette A, Oziel J, et al. Diagnostic accuracy of early urinary index changes in differentiating transient from persistent acute kidney injury in critically ill patients: multicenter cohort study. Crit Care 2013;**17**(2):R56.

72. Sun X, Ioannidis JP, Agoritsas T, et al. How to use a subgroup analysis: users' guide to the medical literature. JAMA 2014;**311**(4):405-11.

73. Zalis ME, Blake MA, Cai W, et al. Diagnostic accuracy of laxative-free computed tomographic colonography for detection of adenomatous polyps in asymptomatic adults: a prospective evaluation. Annals of internal medicine 2012;**156**(10):692-702.

74. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. Journal of clinical epidemiology 2005;**58**(8):859-62.

75. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press Inc., New York, 2003.

76. Vach W, Gerke O, Hoilund-Carlsen PF. Three principles to define the success of a diagnostic study could be identified. Journal of clinical epidemiology 2012;**65**(3):293-300.

77. Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. Bmj 2006;**332**(7550):1127-9.

78. Bochmann F, Johnson Z, Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. The British journal of ophthalmology 2007;**91**(7):898-900.

79. Collins MG, Teo E, Cole SR, et al. Screening for colorectal cancer and advanced colorectal neoplasia in kidney transplant recipients: cross sectional prevalence and diagnostic accuracy study of faecal immunochemical testing for haemoglobin and colonoscopy. Bmj 2012;**345**:e4657.

80. Cecil MP, Kosinski AS, Jones MT, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. Journal of clinical epidemiology 1996;**49**(7):735-42.

81. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. Journal of clinical epidemiology 1992;**45**(6):581-6.

82. Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. Medical decision making : an international journal of the Society for Medical Decision Making 1992;**12**(1):22-31.

83. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. Journal of chronic diseases 1986;**39**(5):343-55.

84. Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. Investigative radiology 1985;**20**(7):751-6.

85. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. The New England journal of medicine 1978;**299**(17):926-30.

86. Zhou XH. Effect of verification bias on positive and negative predictive values. Statistics in medicine 1994;**13**(17):1737-45.

87. Kok L, Elias SG, Witteman BJ, et al. Diagnostic accuracy of point-of-care fecal calprotectin and immunochemical occult blood tests for diagnosis of organic bowel disease in primary care: the Cost-Effectiveness of a Decision Rule for Abdominal Complaints in Primary Care (CEDAR) study. Clinical chemistry 2012;**58**(6):989-98.

88. Harris JM, Jr. The hazards of bedside Bayes. Jama 1981;**246**(22):2602-5.

89. Hlatky MA, Pryor DB, Harrell FE, Jr., et al. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. The American journal of medicine 1984;**77**(1):64-71.

90. Lachs MS, Nachamkin I, Edelstein PH, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Annals of internal medicine 1992;**117**(2):135-40.

91. Moons KG, van Es GA, Deckers JW, et al. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;**8**(1):12-7.

92. O'Connor PW, Tansay CM, Detsky AS, et al. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. Neurology 1996;**47**(1):140-4.

93. Deckers JW, Rensing BJ, Tijssen JG, et al. A comparison of methods of analysing exercise tests for diagnosis of coronary artery disease. British heart journal 1989;**62**(6):438-44.

94. Naraghi AM, Gupta S, Jacks LM, et al. Anterior cruciate ligament reconstruction: MR imaging signs of anterior knee laxity in the presence of an intact graft. Radiology 2012;**263**(3):802-10.

95. Ashdown HF, D'Souza N, Karim D, et al. Pain over speed bumps in diagnosis of acute appendicitis: diagnostic accuracy study. Bmj 2012;**345**:e8012.

96. Leeflang MM, Rutjes AW, Reitsma JB, et al. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013;**185**(11):E537-44.

97. Rajaram S, Swift AJ, Capener D, et al. Lung morphology assessment with balanced steady-state free precession MR imaging compared with CT. Radiology 2012;**263**(2):569-77.

98. Lang TAS, M. *Generalizing from a sample to a population: Reporting estimates and confidence intervals.* Philadelphia: American College of Physicians, 1997.

99. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Annals of internal medicine 2004;**141**(10):781-8.

100. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. Jama 2001;**285**(4):437-43.

101. Park SH, Lee JH, Lee SS, et al. CT colonography for detection and characterisation of synchronous proximal colonic lesions in patients with stenosing colorectal cancer. Gut 2012;**61**(12):1716-22.

102. J.G. ILMBPMGPPGCL. Designing studies to ensure that estimates of test accuracy will travel. In: J.A. K, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:95-116.

103. Ter Riet G, Chesley P, Gross AG, et al. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. PLoS One 2013;**8**(11):e73623.

104. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. Journal of clinical epidemiology 2007;**60**(4):324-9.

105. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Annals of internal medicine 2006;**144**(11):850-5.

106. Pewsner D, Battaglia M, Minder C, et al. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. Bmj 2004;**329**(7459):209-13.

107. Foerch C, Niessner M, Back T, et al. Diagnostic accuracy of plasma glial fibrillary acidic protein for differentiating intracerebral hemorrhage and cerebral ischemia in patients with symptoms of acute stroke. Clinical chemistry 2012;**58**(1):237-45.

108. Altman DG. The time has come to register diagnostic and prognostic research. Clinical chemistry 2014;**60**(4):580-2.

109. Hooft L, Bossuyt PM. Prospective registration of marker evaluation studies: time to act. Clinical chemistry 2011;**57**(12):1684-6.

110. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. Clin Chem 2008;**54**(7):1101-3.

111. Korevaar DA, Ochodo EA, Bossuyt PM, et al. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov. Clin Chem 2014;**60**(4):651-9.

112. 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):1146-52.

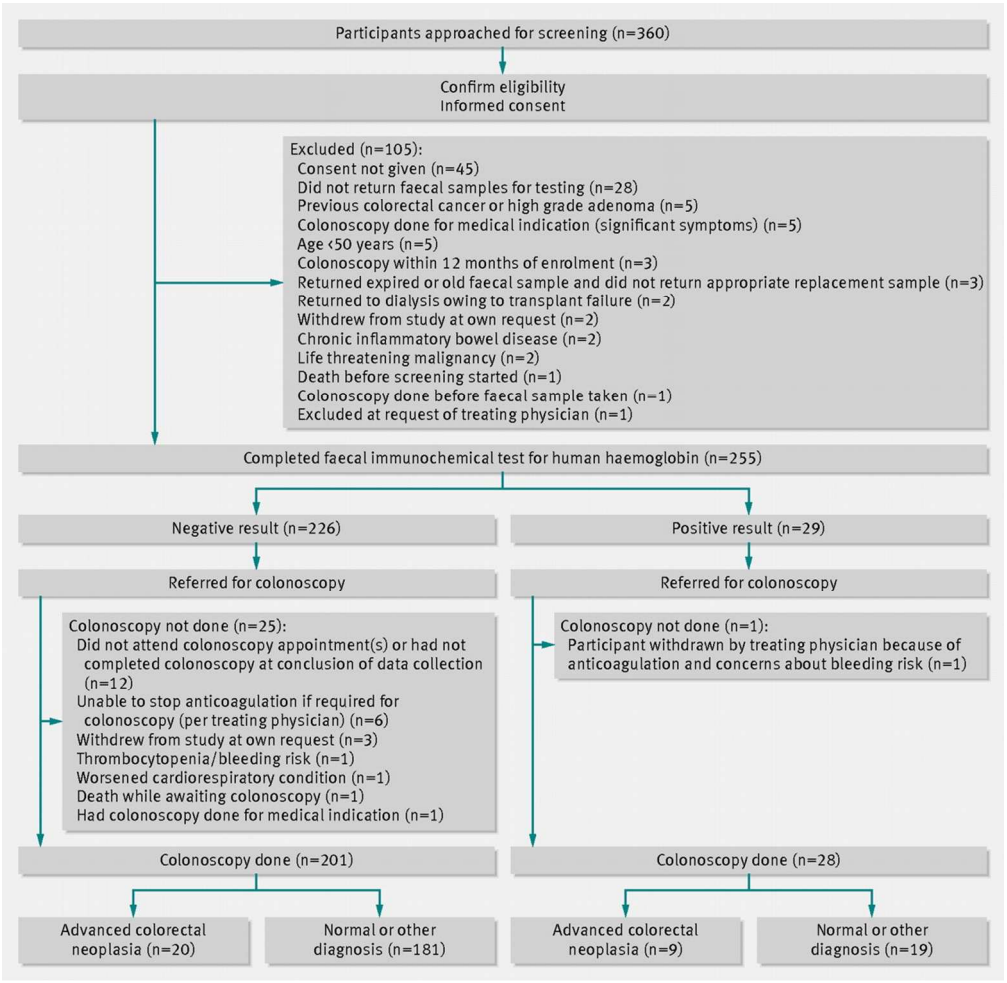
113. Korevaar DA, Bossuyt PM, Hooft L. Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports. BMJ open 2014;**4**(1):e004596.

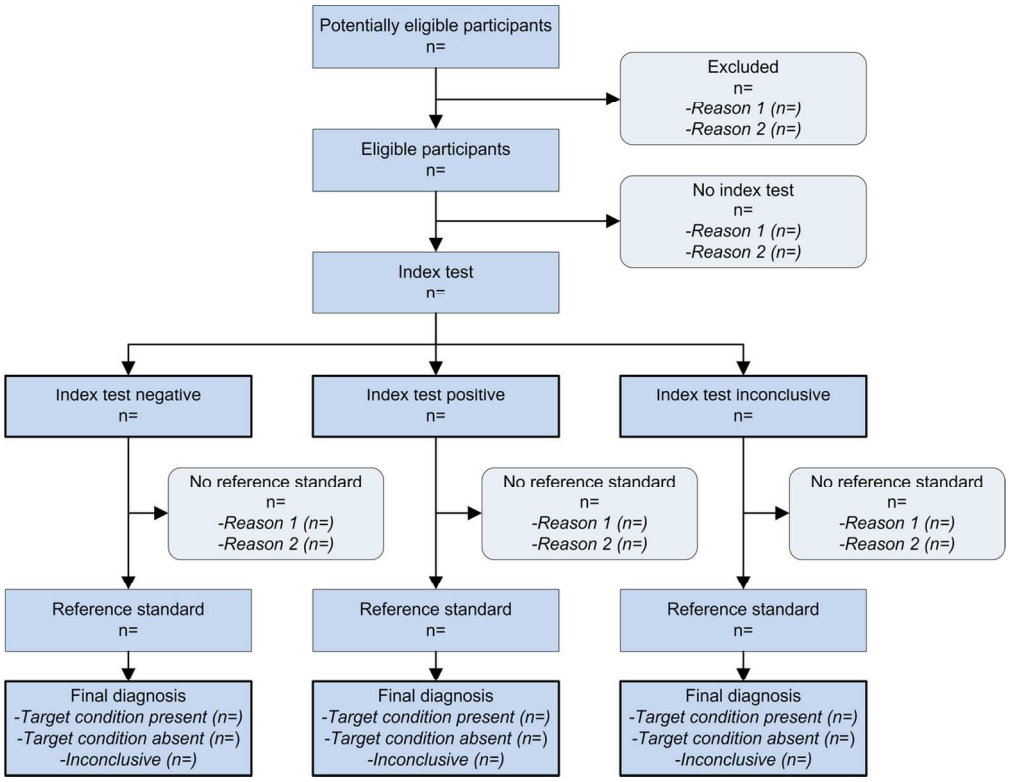
114. Leeuwenburgh MM, Wiarda BM, Wiezer MJ, et al. Comparison of imaging strategies with conditional contrast-enhanced CT and unenhanced MR imaging in patients suspected of having appendicitis: a multicenter diagnostic performance study. Radiology 2013;**268**(1):135-43.

115. Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. Lancet 2014;**383**(9913):257-66.

116. Stewart CM, Schoeman SA, Booth RA, et al. Assessment of self taken swabs versus clinician taken swab cultures for diagnosing gonorrhoea in women: single centre, diagnostic accuracy study. Bmj 2012;**345**:e8107.

117. Sismondo S. Pharmaceutical company funding and its consequences: a qualitative systematic review. Contemporary clinical trials 2008;**29**(2):109-13.






125x96mm (300 x 300 DPI)

BMJ Open

# STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration



| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2016-012799.R1 |
| Article Type: | Research |
| Date Submitted by the Author: | 03-Aug-2016 |
| Complete List of Authors: | Cohen, Jérémie; Academic Medical Centre, University of Amsterdam, Department of Clinical Epidemiology, Biostatistics and Bioinformatics<br>Korevaar, Daniël; University of Amsterdam, Academic Medical Centre<br>Altman, Doug; Centre for Statistics in Medicine<br>Bruns, David; University of Virginia School of Medicine, Department of Pathology<br>Gatsonis, Constantine; Brown School of Public Health<br>Hooft, Lotty; University Medical Center Utrecht, University of Utrecht, Cochrane Netherlands<br>Irwig, Les; University of Sydney, Sydney Medical School<br>Levine, Deborah; Beth Israel Deaconess Medical Center, Department of Radiology<br>Reitsma, Johannes; University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care<br>de Vet, Riekie; VU University Medical Center<br>Bossuyt, Patrick; Academic Medical Center; University of Amsterdam, Dept. Clinical Epidemiology and Biostatistics |
| Primary Subject Heading: | Medical publishing and peer review |
| Secondary Subject Heading: | Diagnostics, Epidemiology, Evidence based practice, Research methods |
| Keywords: | Reporting quality, Sensitivity and specificity, Diagnostic accuracy, Research waste, Peer review, Medical publishing |
| | |

# STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration

Jérémie F. Cohen*, Daniël A. Korevaar*, Douglas G. Altman, David E. Bruns, Constantine A. Gatsonis,

Lotty Hooft, Les Irwig, Deborah Levine, Johannes B. Reitsma, Henrica C.W. de Vet, Patrick M.M. Bossuyt

***Both authors contributed equally to this manuscript and share first authorship.**

**Authors' names, academic degrees, positions, affiliations, and email addresses:**

**Jérémie F. Cohen***, MD PhD, *Postdoctoral research fellow*

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands; INSERM UMR 1153 and Department of Pediatrics, Necker Hospital, AP-HP, Paris Descartes University, Paris, France

jeremie.cohen@inserm.fr

**Daniël A. Korevaar***, MD, *PhD candidate*

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands

d.a.korevaar@amc.uva.nl

**Douglas G. Altman**, DSc, *Professor of statistics in medicine*

Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

doug.altman@csm.ox.ac.uk

**David E. Bruns**, MD, *Professor of pathology*

Department of Pathology, University of Virginia School of Medicine, Charlottesville, Virginia, USA

dbruns@virginia.edu

**Constantine A. Gatsonis**, PhD, Professor of biostatistics

Department of Biostatistics, Brown University School of Public Health, Providence, Rhode Island, USA

gatsonis@stat.brown.edu

**Lotty Hooft**, PhD, *Associate professor / Co-director*

Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, the Netherlands

l.hooft@umcutrecht.nl

**Les Irwig**, MBBS, PhD, *Professor of epidemiology*

Screening and Diagnostic Test Evaluation Program, School of Public Health, University of Sydney, Sydney, New South Wales, Australia

les.irwig@sydney.edu.au

**Deborah Levine**, MD, *Professor of radiology*

Department of Radiology, Beth Israel Deaconess Medical Center, Boston, MA, USA; Radiology Editorial Office, Boston, MA, USA.

dlevine@bidmc.harvard.edu

**Johannes B. Reitsma**, MD PhD, *Associate professor of clinical epidemiology*

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, the Netherlands

j.b.reitsma-2@umcutrecht.nl

**Henrica C.W. de Vet**, PhD, *Professor of clinimetrics*

Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, the Netherlands

hcw.devet@vumc.nl

**Patrick M.M. Bossuyt**, PhD, *Professor of clinical epidemiology*

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands

p.m.bossuyt@amc.uva.nl

**Corresponding author:** Prof. Patrick M.M. Bossuyt

Department of Clinical Epidemiology, Biostatistics and Bioinformatics

Academic Medical Center - University of Amsterdam

PO Box 22700, 1100 DE Amsterdam, The Netherlands

Email: p.m.bossuyt@amc.uva.nl Phone: +31(20)566 3240 Fax: +31(20)691 2683

**Word count (text only):** 9,316.

**Keywords**: reporting quality; sensitivity and specificity; diagnostic accuracy; research waste; peer review; medical publishing

## ABSTRACT

Diagnostic accuracy studies are, like other clinical studies, at risk of bias due to shortcomings in design and conduct, and the results of a diagnostic accuracy study may not apply to other patient groups and settings. Readers of study reports need to be informed about study design and conduct, in sufficient detail to judge the trustworthiness and applicability of the study findings.

The STARD statement (Standards for Reporting of Diagnostic Accuracy Studies) was developed to improve the completeness and transparency of reports of diagnostic accuracy studies. STARD contains a list of essential items that can be used as a checklist, by authors, reviewers and other readers, to ensure that a report of a diagnostic accuracy study contains the necessary information.

STARD was recently updated. All updated STARD materials, including the checklist, are available at www.equator-network.org/reporting-guidelines/stard. Here we present the STARD 2015 explanation and elaboration document. Through commented examples of appropriate reporting, we clarify the rationale for each of the 30 items on the STARD 2015 checklist, and describe what is expected from authors in developing sufficiently informative study reports.


## STRENGTHS AND LIMITATIONS OF THIS STUDY

Not applicable to this explanation and elaboration document.

## INTRODUCTION

Diagnostic accuracy studies are at risk of bias, not unlike other clinical studies. Major sources of bias originate in methodological deficiencies, in participant recruitment, data collection, executing or interpreting the test, or in data analysis.[1,2] As a result, the estimates of sensitivity and specificity of the test that is compared against the reference standard can be flawed, deviating systematically from what would be obtained in ideal circumstances (see key terminology in Table 1). Biased results can lead to improper recommendations about testing, negatively affecting patient outcomes or health care policy.

Diagnostic accuracy is not a fixed property of a test. A test's accuracy in identifying patients with the target condition typically varies between settings, patient groups, and depending on prior testing.[2] These sources of variation in diagnostic accuracy are relevant for those who want to apply the findings of a diagnostic accuracy study to answer a specific question about adopting the test in his or her environment. Risk of bias and concerns about the applicability are the two key components of QUADAS-2, a quality assessment tool for diagnostic accuracy studies.[3]

Readers can only judge the risk of bias and applicability of a diagnostic accuracy study if they find the necessary information to do so in the study report. The published study report has to contain all the essential information to judge the trustworthiness and relevance of the study findings, in addition to a complete and informative disclose about the study results.

Unfortunately, several surveys have shown that diagnostic accuracy study reports often fail to transparently describe core elements.[4-6] Essential information about included patients, study design and the actual results is frequently missing, and recommendations about the test under evaluation are often generous and too optimistic.

To facilitate more complete and transparent reporting of diagnostic accuracy studies the STARD statement was developed: Standards for Reporting of Diagnostic Accuracy Studies.[7] Inspired by the Consolidated Standards for the Reporting of Trials or CONSORT statement for reporting randomized

controlled trials,[8,9] STARD contains a checklist of items that should be reported in any diagnostic accuracy study.

The STARD statement was initially released in 2003 and updated in 2015.[10] The objectives of this update were to include recent evidence about sources of bias and variability and other issues in complete reporting, and make the STARD list easier to use. The updated STARD 2015 list now has 30 essential items (Table 2).

Below we present an explanation and elaboration of STARD 2015. This is an extensive revision and update of a similar document that was prepared for the STARD 2003 version.[11] Through commented examples of appropriate reporting, we clarify the rationale for each item and describe what is expected from authors.

We are confident that these descriptions can further assist scientists in writing fully informative study reports, and help peer reviewers, editors and other readers in verifying that submitted and published manuscripts of diagnostic accuracy studies are sufficiently detailed.

## STARD 2015 ITEMS: EXPLANATION AND ELABORATION

### Title or abstract

*Item 1. Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)*

**Example.** "Main outcome measures: Sensitivity and specificity of CT colonography in detecting individuals with advanced neoplasia (i.e., advanced adenoma or colorectal cancer) 6 mm or larger."[12]

**Explanation.** When searching for relevant biomedical studies on a certain topic, electronic databases such as Medline and Embase are indispensable. To facilitate retrieval of their article, authors can explicitly identify it as a report of a diagnostic accuracy study. This can be done by using terms in the title and/or abstract that refer to measures of diagnostic accuracy, such as "sensitivity", "specificity",

"positive predictive value", "negative predictive value", "area under the ROC curve (AUC)", or "likelihood ratio".

In 1991, Medline introduced a specific keyword (MeSH heading) for indexing diagnostic studies: "Sensitivity and Specificity." Unfortunately, the sensitivity of using this particular MeSH heading to identify diagnostic accuracy studies can be as low as 51%.[13] As of May 2015, Embase's thesaurus (Emtree) has 38 check tags for study types; "diagnostic test accuracy study" is one of them, but was only introduced in 2011.

*In the example*, the authors mentioned the terms "sensitivity" and "specificity" in the abstract. The article will now be retrieved when using one of these terms in a search strategy, and will be easily identifiable as one describing a diagnostic accuracy study.

## Abstract

*Item 2. Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)*

**Example.** See STARD for Abstracts (*manuscript in preparation; checklist will be available at http://www.equator-network.org/reporting-guidelines/stard/*).

**Explanation.** Readers use abstracts to decide whether they should retrieve the full study report and invest time in reading it. In cases where access to the full study report cannot be obtained or where time is limited, it is conceivable that clinical decisions are based on the information provided in abstracts only.

In two recent literature surveys, abstracts of diagnostic accuracy studies published in high-impact journals or presented at an international scientific conference were found insufficiently informative, because key information about the research question, study methods, study results, and the implications of findings were frequently missing.[14,15]

Informative abstracts help readers to quickly appraise critical elements of study validity (risk of bias) and applicability of study findings to their clinical setting (generalisability). Structured abstracts, with separate headings for objectives, methods, results and interpretation, allow readers to find essential information more easily.[16]

Building on STARD 2015, the newly developed STARD for Abstracts provides a list of essential items that should be included in journal and conference abstracts of diagnostic accuracy studies *(list finalized; manuscript under development)*.

## Introduction

*Item 3. Scientific and clinical background, including the intended use and clinical role of the index test*

**Example.** "The need for improved efficiency in the use of emergency department radiography has long been documented. This need for selectivity has been identified clearly for patients with acute ankle injury, who generally are all referred for radiography, despite a yield for fracture of less than 15%. The referral patterns and yield of radiography for patients with knee injuries have been less well described but may be more inefficient than for patients with ankle injuries. […] The sheer volume of low-cost tests such as plain radiography may contribute as much to rising health care costs as do high-technology, low-volume procedures. […] If validated in subsequent studies, a decision rule for knee-injury patients could lead to a large reduction in the use of knee radiography and significant health care savings without compromising patient care."[17]

**Explanation.** In the introduction of scientific study reports, authors should describe the rationale for their study. In doing so they can refer to previous work on the subject, remaining uncertainty, and the clinical implications of this knowledge gap. To help readers in evaluating the implications of the study, authors can clarify the intended use and the clinical role of the test under evaluation, which is referred to as the index test.

The intended use of a test can be diagnosis, screening, staging, monitoring, surveillance, prognosis, treatment selection, or other purposes.[18] The clinical role of the test under evaluation refers to its anticipated position relative to other tests in the clinical pathway.[19] A triage test, for example, will be used before an existing test because it is less costly or burdensome, but often less accurate as well. An add-on test will be used after existing tests, to improve the accuracy of the total test strategy by identifying false positives or false negatives of the initial test. In other cases, a new test may be used to replace an existing test.

Defining the intended use and clinical role of the test will guide the design of the study and the targeted level of sensitivity and specificity; from these definitions follow the eligibility criteria, how and where to identify eligible participants, how to perform tests, and how to interpret test results.[19]

Specifying the clinical role is helpful in assessing the relative importance of potential errors (false positives and false negatives) made by the index test. A triage test to rule out disease, for example, will need very high sensitivity, whereas one that mainly aims to rule in disease will need very high specificity.

*In the example*, the intended use is diagnosis of knee fractures in patients with acute knee injuries, and the potential clinical role is triage test; radiography, the existing test, would only be performed in those with a positive outcome of the newly developed decision rule. The authors outline the current scientific and clinical background of the health problem studied, and their reason for aiming to develop a triage test: this would reduce the number of radiographs and, consequently, healthcare costs.

*Item 4. Study objectives and hypotheses*

**Example (1).** "The objective of this study was to evaluate the sensitivity and specificity of 3 different diagnostic strategies: a single rapid antigen test, a rapid antigen test with a follow-up rapid antigen test if negative (rapid-rapid diagnostic strategy), and a rapid antigen test with follow-up culture if negative (rapid-culture) — the AAP diagnostic strategy—all compared with a 2-plate culture gold standard. In

addition, […] we also compared the ability of these strategies to achieve an absolute diagnostic test sensitivity of >95%."[20]

**Example (2).** "Our 2 main hypotheses were that rapid antigen detection tests performed in physician office laboratories are more sensitive than blood agar plate cultures performed and interpreted in physician office laboratories, when each test is compared with a simultaneous blood agar plate culture processed and interpreted in a hospital laboratory, and rapid antigen detection test sensitivity is subject to spectrum bias".[21]

**Explanation.** Clinical studies may have a general aim (a long term goal, such as "to improve the staging of oesophageal cancer"), specific objectives (well defined goals for this particular study), and testable hypotheses (statements than can be falsified by the study results).

In diagnostic accuracy studies, statistical hypotheses are typically defined in terms of acceptability criteria for single tests (minimum levels of sensitivity, specificity, or other measures). In those cases, hypotheses generally include a quantitative expression of the expected value of the diagnostic parameter. In other cases, statistical hypotheses are defined in terms of equality or non-inferiority in accuracy when comparing two or more index tests.

A priori specification of the study hypotheses limits the chances of post-hoc data-dredging with spurious findings, premature conclusions about the performance of tests, or subjective judgment about the accuracy of the test. Objectives and hypotheses also guide sample size calculations. An evaluation of 126 reports of diagnostic test accuracy studies published in high-impact journals in 2010 revealed that 88% did not state a clear hypothesis.[22]

*In the first example*, the authors' objective was to evaluate the accuracy of three diagnostic strategies; their specific hypothesis was that the sensitivity of any of these would exceed the pre-specified value of 95%. *In the second example*, the authors explicitly describe the hypotheses they want to explore in their study. The first hypothesis is about the comparative sensitivity of two index tests (rapid antigen

detection test vs. culture performed in physician office laboratories); the second is about variability of rapid test performance according to patient characteristics (spectrum bias).

## Methods

*Item 5. Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)*

**Example.** "We reviewed our database of patients who underwent needle localization and surgical excision with digital breast tomosynthesis guidance from April 2011 through January 2013. […] The patients' medical records and images of the 36 identified lesions were then reviewed retrospectively by an author with more than 5 years of breast imaging experience after a breast imaging fellowship."[23]

**Explanation.** There is great variability in the way the terms 'prospective' and 'retrospective' are defined and used in the literature. We believe it is therefore necessary to describe clearly whether data collection was planned before the index test and reference standard were performed, or afterwards. If authors define the study question before index test and reference standards are performed, they can take appropriate actions for optimizing procedures according to the study protocol and for dedicated data collection.[24] Sometimes the idea for a study originates when patients have already undergone the index test and the reference standard. If so, data collection relies on reviewing patient charts or extracting data from registries. Though such retrospective studies can sometimes reflect routine clinical practice better than prospective studies, they may fail to identify all eligible patients, and often result in data of lower quality, with more missing data points.[24] A reason for this could be, for example, that in daily clinical practice, not all patients undergoing the index test may proceed to have the reference standard.

*In the example*, the data were clearly collected retrospectively: participants were identified through database screening, clinical data were abstracted from patients' medical records, though images were re-interpreted.

*Item 6. Eligibility criteria*

**Example (1).** "Patients eligible for inclusion were consecutive adults (≥18 years) with suspected pulmonary embolism, based on the presence of at least one of the following symptoms: unexplained (sudden) dyspnoea, deterioration of existing dyspnoea, pain on inspiration, or unexplained cough. We excluded patients if they received anticoagulant treatment (vitamin K antagonists or heparin) at presentation, they were pregnant, follow-up was not possible, or they were unwilling or unable to provide written informed consent."[25]

**Example (2).** "Eligible cases had symptoms of diarrhoea and both a positive result for toxin by enzyme immunoassay and a toxigenic *C difficile* strain detected by culture (in a sample taken less than seven days before the detection round). We defined diarrhoea as three or more loose or watery stool passages a day. We excluded children and adults on intensive care units or haematology wards. Patients with a first relapse after completing treatment for a previous *C difficile* infection were eligible but not those with subsequent relapses. […] For each case we approached nine control patients. These patients were on the same ward as and in close proximity to the index patient. Control patients did not have diarrhoea, or had diarrhoea but a negative result for *C difficile* toxin by enzyme immunoassay and culture (in a sample taken less than seven days previously)."[26]

**Explanation.** Since a diagnostic accuracy study describes the behavior of a test under particular circumstances, a report of the study must include a complete description of the criteria that were used to identify eligible participants. Eligibility criteria are usually related to the nature and stage of the target condition and the intended future use of the index test; they often include the signs, symptoms, or previous test results that generate the suspicion about the target condition. Additional criteria can be used to exclude participants for reasons of safety, feasibility, and ethical arguments.

Excluding patients with a specific condition or receiving a specific treatment known to adversely affect the way the test works can lead to inflated diagnostic accuracy estimates.[27] An example is the exclusion

of patients using beta-blockers in studies evaluating the diagnostic accuracy of exercise electrocardiography.

Some studies have one set of eligibility criteria for all study participants; these are sometimes referred to as single-gate or cohort studies. Other studies have one set of eligibility criteria for participants with the target condition, and (an)other set(s) of eligibility criteria for those without the target condition; these are called multiple-gate or case-control studies.[28]

*In the first example*, the eligibility criteria list presenting signs and symptoms, an age limit, and exclusion based on specific conditions and treatments. Because the same set of eligibility criteria applies to all study participants, this is an example of a single-gate study.

*In the second example*, the authors used different eligibility criteria for participants with and without the target condition: one group consisted of patients with a confirmed diagnosis of *Clostridium difficile*, and one group consisted of healthy controls. This is an example of a multiple-gate study. Extreme contrasts between severe cases and healthy controls can lead to inflated estimates of accuracy.[6,29]

*Item 7. On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)*

**Example.** "We reviewed our database of patients who underwent needle localization and surgical excision with digital breast tomosynthesis guidance from April 2011 through January 2013."[23]

**Explanation.** The eligibility criteria specify who can participate in the study, but they do not describe how the study authors identified eligible participants. This can be done in various ways.[30] A general practitioner may evaluate every patient for eligibility that he sees during office hours. Researchers can go through registries in an emergency department, to identify potentially eligible patients. In other studies, patients are only identified after having been subjected to the index test. Still other studies start with patients in whom the reference standard was performed. Many retrospective studies include

participants based on searching hospital databases for patients that underwent both the index test and the reference standard.[31]

Differences in methods for identifying eligible patients can affect the spectrum and prevalence of the target condition in the study group, as well as the range and relative frequency of alternative conditions in patients without the target condition.[32] These differences can influence the estimates of diagnostic accuracy.

*In the example*, participants were identified through searching a patient database and were included if they underwent both the index test and the reference standard.

*Item 8. Where and when potentially eligible participants were identified (setting, location, and dates)*

**Example.** "The study was conducted at the Emergency Department of a university-affiliated children's hospital between January 21, 1996, and April 30, 1996."[33]

**Explanation.** The results of a diagnostic accuracy study reflect the performance of a test in a particular clinical context and setting. A medical test may perform differently in a primary, secondary or tertiary care setting, for example. Authors should therefore report the actual setting in which the study was performed, as well as the exact locations: names of the participating centers, city and country. The spectrum of the target condition as well as the range of other conditions that occur in patients suspected of the target condition can vary across settings, depending on which referral mechanisms are in play.[34-36]

Since test procedures, referral mechanisms, and the prevalence and severity of diseases can evolve over time, authors should also report the start and end dates of participant recruitment.

This information is essential for readers who want to evaluate the generalisability of the study findings, and their applicability to specific questions, for those who would like to use the evidence generated by the study to make informed health care decisions.

*In the example*, study setting and study dates were clearly defined.

*Item 9. Whether participants formed a consecutive, random or convenience series*

**Example.** "All subjects were evaluated and screened for study eligibility by the first author (E.N.E.) prior to study entry. This was a convenience sample of children with pharyngitis; the subjects were enrolled when the first author was present in the emergency department."[37]

**Explanation.** The included study participants may be either a consecutive series of all patients evaluated for eligibility at the study location and satisfying the inclusion criteria, or a subselection of these. A subselection can be purely random, produced by using a random numbers table, or less random, if patients are only enrolled on specific days or during specific office hours. In that case, included participants may not be considered a representative sample of the targeted population, and the generalisability of the study results may be jeopardized.[2,29]

*In the example*, the authors explicitly described a convenience series where subjects were enrolled based on their accessibility to the clinical investigator.

*Item 10a. Index test, in sufficient detail to allow replication*

*Item 10b. Reference standard, in sufficient detail to allow replication*

**Example.** "An intravenous line was inserted in an antecubital vein and blood samples were collected into serum tubes before (baseline), immediately after, and 1.5 and 4.5 h after stress testing. Blood samples were put on ice, processed within 1 h of collection, and later stored at -80 °C before analysis. The samples had been through 1 thaw–freeze cycle before cardiac troponin I (cTnI) analysis. We measured cTnI by a prototype hs assay (ARCHITECT *STAT* high-sensitivity troponin, Abbott Diagnostics) with the capture antibody detecting epitopes 24–40 and the detection antibody epitopes 41–49 of cTnI. The limit of detection (LoD) for the high sensitivity (hs) cTnI assay was recently reported by other groups to be 1.2 ng/L, the 99th percentile 16 ng/L, and the assay 10% coefficient of variation (CV) 3.0 ng/L. […]

Samples with concentrations below the range of the assays were assigned values of 1.2 […] for cTnI. […]."[38]

**Explanation.** Differences in the execution of the index test or reference standard are a potential source of variation in diagnostic accuracy.[39,40] Authors should therefore describe the methods for executing the index test and reference standard, in sufficient detail to allow other researchers to replicate the study, and to allow readers to assess (1) the feasibility of using the index test in their own setting, (2) the adequacy of the reference standard, and (3) the applicability of the results to their clinical question. The description should cover key elements of the test protocol, including details of:

a. the pre-analytical phase, for example, patient preparation such as fasting/feeding status prior to blood sampling, the handling of the sample prior to testing and its limitations (such as sample instability), or the anatomic site of measurement;
b. the analytical phase, including materials and instruments and analytical procedures;
c. the post-analytical phase, such as calculations of risk scores using analytical results and other variables.

Between-study variability in measures of test accuracy due to differences in test protocol has been documented for a number of tests, including the use of hyperventilation prior to exercise electrocardiography and the use of tomography for exercise thallium scintigraphy.[27,40]
The number, training and expertise of the persons executing and reading the index test and the reference standard may also be critical. Many studies have shown between-reader variability, especially in the field of imaging.[41,42] The quality of reading has also been shown to be affected in cytology and microbiology by professional background, expertise, and prior training to improve interpretation and to reduce inter-observer variation.[43-45] Information about the amount of training of the persons in the study who read the index test can help readers to judge whether similar results are achievable in their own settings.

In some cases, a study depends on multiple reference standards. Patients with lesions on an imaging test under evaluation may, for example, undergo biopsy with a final diagnosis based on histology, whereas patients without lesions on the index test undergo clinical follow-up as reference standard. This could be a potential source of bias, so authors should specify which patient groups received which reference standard.[2,3]

More specific guidance for specialized fields of testing, or certain types of tests, will be developed in future STARD extensions. Whenever available, these extensions will be made available on the STARD pages at the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) website (http://www.equator-network.org/).

*In the example*, the authors described how blood samples were collected and processed in the laboratory. They also report analytical performance characteristics of the index test device, as obtained in previous studies.

*Item 11. Rationale for choosing the reference standard (if alternatives exist)*

**Example.** "The MINI [Mini International Neuropsychiatric Inventory] was developed as a short and efficient diagnostic interview to be used in both research and clinical settings *(reference supporting this statement provided by the authors)*. It has good reliability and validity rates compared with other gold standard diagnostic interviews, such as the Structured Clinical Interview for DSM [Diagnostic and Statistical Manual of Mental Disorders] Disorders (SCID) and the Composite International Diagnostic Interview *(references supporting this statement provided by the authors)*."[46]

**Explanation.** In diagnostic accuracy studies, the reference standard is used for establishing the presence or absence of the target condition in study participants. Several reference standards may be available to define the same target condition. In such cases authors are invited to provide their rationale for selecting the specific reference standard from the available alternatives. This may depend on the intended use of the index test, the clinical relevance, or practical and/or ethical reasons.

Alternative reference standards are not always in perfect agreement. Some reference standards are less accurate than others. In other cases, different reference standards reflect related but different manifestations or stages of the disease, as in confirmation by imaging (first reference standard) versus clinical events (second reference standard).

*In the example*, the authors selected the MINI, a structured diagnostic interview commonly used for psychiatric evaluations, as the reference standard for identifying depression and suicide risk in adults with epilepsy. As a rationale for their choice, they claimed that the MINI test was short to administer, efficient both for clinical and research purposes, reliable, and valid as compared to alternative diagnostic interviews.

*Item 12a. Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory*

*Item 12b. Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory*

**Example.** "We also compared the sensitivity of the risk-model at the specificity that would correspond to using a fixed FIT [fecal immunochemical test] positivity threshold of 50 ng/ml. We used a threshold of 50 ng/ml because this was the anticipated cut-off for the Dutch screening programme at the time of the study."[47]

**Explanation.** Test results in their original form can be dichotomous (positive versus negative), have multiple categories (as in high, intermediate, or low risk), or be continuous (interval or ratio scale). For tests with multiple categories, or continuous results, the outcomes from testing are often reclassified into positive (disease confirmed) and negative (disease excluded). This is done by defining a threshold: the test positivity cut-off. Results that exceed the threshold would then be called positive index test results. In other studies, an ROC curve is derived, by calculating the sensitivity-specificity pairs for all possible cutoffs.

To evaluate the validity and applicability of these classifications, readers would like to know these positivity cut-offs or result categories, how they were determined, and whether they were defined prior to the study or after collecting the data. Pre-specified thresholds can be based on (1) previous studies, (2) cutoffs used in clinical practice, (3) thresholds recommended by clinical practice guidelines, or (4) thresholds recommended by the manufacturer. If no such thresholds exist, the authors may be tempted to explore the accuracy for various thresholds after the data have been collected.

If the authors selected the positivity cut-off after performing the test, choosing the one that maximized test performance, there is an increased risk that the resulting accuracy estimates are overly optimistic, especially in small studies.[48,49] Subsequent studies may fail to replicate the findings.[50,51]

*In the example*, the authors stated the rationale for their selection of cut-offs.

*Item 13a. Whether clinical information and reference standard results were available to the performers or readers of the index test*

*Item 13b. Whether clinical information and index test results were available to the assessors of the reference standard*

**Example.** "Images for each patient were reviewed by two fellowship-trained genitourinary radiologists with 12 and 8 years of experience, respectively, who were blinded to all patient information, including the final histopathologic diagnosis."[52]

**Explanation.** Some medical tests, such as most forms of imaging, require human handling, interpretation and judgment. These actions may be influenced by the information that is available to the reader.[1,53,54] This can lead to artificially high agreement between tests, or between the index test and the reference standard.

If the reader of a test has access to information about signs, symptoms and previous test results, the reading may be influenced by this additional information, but this may still represent how the test is used in clinical practice.[2] The reverse may also apply, if the reader does not have enough information for

a proper interpretation of the index test outcome. In that case, test performance may be affected downwards, and the study findings may have limited applicability. Either way, readers of the study report should know to which extent such additional information was available to test readers and may have influenced their final judgment.

In other situations the assessors of the reference standard may have had access to the index test results. In those cases, the final classification may be guided by the index test result, and the reported accuracy estimates for the index test will be too high.[1,2,27] Tests that require subjective interpretation are particularly susceptible to this bias.

Withholding information from the readers of the test is commonly referred to as "blinding" or "masking". The point of this reporting item is not that blinding is desirable or undesirable, but, rather, that readers of the study report need information about blinding for both the index test and the reference standard to be able to interpret the study findings.

*In the example*, the readers of unenhanced CT for differentiating between renal angiomyolipoma and renal cell carcinoma did not have access to clinical information, nor to the results of histopathology, the reference standard in this study.

*Item 14. Methods for estimating or comparing measures of diagnostic accuracy*

**Example.** "Statistical tests of sensitivity and specificity were conducted by using the McNemar test for correlated proportions. All tests were two sided, testing the hypothesis that stereoscopic Digital Mammography performance differed from that of Digital Mammography. A p-value of .05 was considered as the threshold for significance."[55]

**Explanation.** Multiple measures of diagnostic accuracy exist to describe the performance of a medical test, and their calculation from the collected data is not always straightforward.[56] Authors should report the methods used for calculating the measures that they considered appropriate for their study objectives.

Statistical techniques can be used to test specific hypotheses, following from the study's objectives. In single test evaluations, authors may want to evaluate if the diagnostic accuracy of the tests exceeds a pre-specified level (e.g. sensitivity of at least 95%, see Item 4).

Diagnostic accuracy studies can also compare two or more index tests. In such comparisons, statistical hypothesis testing usually involves assessing the superiority of one test over another, or the non-inferiority.[57] For such comparisons, authors should indicate what measure they specified to make the comparison; these should match their study objectives, and the purpose and role of the index test relative to the clinical pathway. Examples are the relative sensitivity, the absolute gain in sensitivity, and the relative diagnostic odds ratio.[58]

*In the example*, the authors used McNemar's test statistic to evaluate whether the sensitivity and specificity of stereoscopic Digital Mammography differed from that of Digital Mammography in patients with elevated risk for breast cancer. In itself, the resulting p-value is not a quantitative expression of the relative accuracy of the two investigated tests. Like any p-value it is influenced by both the magnitude of the difference in effect and the sample size. In the example, the authors could have calculated the relative or absolute difference in sensitivity and specificity, including a 95% confidence interval that takes into account the paired nature of the data.

*Item 15. How indeterminate index test or reference standard results were handled*

**Example.** "Indeterminate results were considered false-positive or false-negative and incorporated into the final analysis. For example, an indeterminate result in a patient found to have appendicitis was considered to have had a negative test result."[59]

**Explanation.** Indeterminate results refer to those that are neither positive or negative.[60] Such results can occur both on the index test and the reference standard, and are a challenge when evaluating the performance of a diagnostic test.[60-63] The occurrence of indeterminate test results varies from test to test, but frequencies up to 40% have been reported.[62]

There are many underlying causes for indeterminate test results.[62,63] A test may fail because of technical reasons or an insufficient sample, for example, in the absence of cells in a needle biopsy from a tumor.[43,64,65] Sometimes test results are not reported as just positive or negative, as in the case of ventilation-perfusion scanning in suspected pulmonary embolism, where the findings are classified in three categories: normal, high probability, or inconclusive.[66]

In itself, the frequency of indeterminate test results is an important indicator of the feasibility of the test, and typically limits the overall clinical usefulness; therefore, authors are encouraged to always report the respective frequencies with reasons, as well as failures to complete the testing procedure. This applies both to the index test and the reference standard.

Ignoring indeterminate test results can produce biased estimates of accuracy if these results do not occur at random. Clinical practice may guide the decision on how to handle indeterminate results. There are multiple ways for handling indeterminate test results in the analysis when estimating accuracy and expressing test performance.[63] They can be ignored altogether, be reported but not accounted for, or handled as a separate test result category. Handling these results as a separate category may be useful when indeterminate results occur more often, for example, in those without the target condition than in those with the target condition. It is also possible to reclassify all such results: as false positives or false negatives, depending on the reference standard result ("worst-case scenario"), or as true positives and true negatives ("best-case scenario").

*In the example*, the authors explicitly chose a conservative approach by considering all indeterminate results from the index test as being false-negative (in those with the target condition) or false-positive (in all others), a strategy sometimes referred to as the "worst-case scenario".

*Item 16. How missing data on the index test and reference standard were handled*

**Example.** "One vessel had missing $FFR_{CT}$ and 2 had missing CT data. Missing data were handled by exclusion of these vessels as well as by the worst-case imputation."[67]

**Explanation.** Missing data are common in any type of biomedical research. In diagnostic accuracy studies, they can occur for both the index test and reference standard. There are several ways to deal with them when analyzing the data.[68] Many researchers exclude participants without an observed test result. This is known as "complete case" or "available case" analysis. This may lead to a loss in precision and can introduce bias, especially if having a missing index test or reference standard result is related to having the target condition.

Participants with missing test results can be included in the analysis if missing results are imputed.[68-70] Another option is to assess the impact of missing test results on estimates of accuracy by considering different scenarios. For the index test, for example, in the "worst-case scenario", all missing index test results are considered false-positive or false-negative depending on the reference standard result; in the "best-case scenario", all missing index test results are considered true-positive or true-negative.

*In the example*, the authors explicitly reported how many cases with missing index test data they encountered and how they handled these data: they excluded them, but also applied a "worst-case scenario".

*Item 17. Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory*

**Example.** "To assess the performance of urinary indices or their changes over the first 24 hours in distinguishing transient AKI [acute kidney injury] from persistent AKI, we plotted the receiver-operating characteristic curves for the proportion of true positives against the proportion of false positives, depending on the prediction rule used to classify patients as having persistent AKI. The same strategy was used to assess the performance of indices and their changes over time in two predefined patient subgroups; namely, patients who did not receive diuretic therapy and patients without sepsis."[71]

**Explanation.** The relative proportion of false positive or false-negative results of a diagnostic test may vary depending on patient characteristics, experience of readers, the setting, and previous test results.[2,3] Researchers may therefore want to explore possible sources of variability in test accuracy within their

study. In such analyses, investigators typically assess differences in accuracy across subgroups of participants, readers or centers.

Post hoc analyses, done after looking at the data, carry a high risk for spurious findings. The results are especially likely not to be confirmed by subsequent studies. Analyses that were pre-specified in the protocol, before data were collected, have greater credibility.[72]

*In the example*, the authors reported that the accuracy of the urinary indices was evaluated in two subgroups that were explicitly pre-specified.

*Item 18. Intended sample size and how it was determined*

**Example.** "Study recruitment was guided by an expected 12% prevalence of adenomas 6 mm or larger in a screening cohort and a point estimate of 80% sensitivity for these target lesions. We planned to recruit approximately 600 participants to achieve margins of sampling error of approximately 8 percentage points for sensitivity. This sample would also allow 90% power to detect differences in sensitivity between computed tomographic colonography and optical colonoscopy of 18 percentage points or more."[73]

**Explanation.** Performing sample size calculations when developing a diagnostic accuracy study may ensure that a sufficient amount of precision is reached. Sample size calculations also take into account the specific objectives and hypotheses of the study.

Readers may want to know how the sample size was determined, and whether the assumptions made in this calculation are in line with the scientific and clinical background, and the study objectives. Readers will also want to learn whether the study authors were successful in recruiting the targeted number of participants. Methods for performing sample size calculations in diagnostic research are widely available,[74-76] but such calculations are not always performed or provided in reports of diagnostic accuracy studies.[77,78]

Many diagnostic accuracy studies are small; a systematic survey of studies published in eight leading journals in 2002 found a median sample size of 118 participants (interquartile range 71-350).[77] Estimates of diagnostic accuracy from small studies tend to be imprecise, with wide confidence intervals around them.

*In the example*, the authors reported in detail to achieve a desired level of precision for an expected sensitivity of 80%.

## Results

*Item 19. Flow of participants, using a diagram*

**Example.** "Between 1 June 2008 and 30 June 2011, 360 patients were assessed for initial eligibility and invited to participate. The figure shows the flow of patients through the study, along with the primary outcome of advanced colorectal neoplasia. Patients who were excluded (and reasons for this) or who withdrew from the study are noted. In total, 229 patients completed the study, a completion rate of 64%."[79] (See Figure 1)

**Explanation.** Estimates of diagnostic accuracy may be biased if not all eligible participants undergo both the index test and the desired reference standard.[80-86] This includes studies in which not all study participants undergo the reference standard, as well as studies where some of the participants receive a different reference standard.[70] Incomplete verification by the reference standard occurs in up to 26% of diagnostic studies; it is especially common when the reference standard is an invasive procedure.[84]

To allow the readers to appreciate the potential for bias, authors are invited to build a diagram to illustrate the flow of participants through the study. Such a diagram also illustrates the basic structure of the study. An example of a prototypical STARD flow diagram is presented in Figure 2.

By providing the exact number of participants at each stage of the study, including the number of true positive, false positive, true negative, and false negative index test results, the diagram also helps

identifying the correct denominator for calculating proportions such as sensitivity and specificity. The diagram should also specify the number of participants that were assessed for eligibility, the number of subjects who did not receive either the index test and/or the reference standard, and the reasons for that. This helps readers to judge the risk of bias, but also the feasibility of the evaluated testing strategy, and the applicability of the study findings.

*In the example*, the authors very briefly described the flow of participants, and referred to a flow diagram in which the number of participants and corresponding test results at each stage of the study were provided, as well as detailed reasons for excluding participants (Figure 1).

*Item 20. Baseline demographic and clinical characteristics of participants*

**Example.** "The median age of participants was 60 years (range 18–91), and 209 participants (54.7%) were female. The predominant presenting symptom was abdominal pain, followed by rectal bleeding and diarrhea, whereas fever and weight loss were less frequent. At physical examination, palpation elicited abdominal pain in almost half the patients, but palpable abdominal or rectal mass was found in only 13 individuals (Table X)."[87] (See Table 3)

**Explanation.** The diagnostic accuracy of a test can depend on the demographic and clinical characteristics of the population in which it is applied.[2,3,88-92] These differences may reflect variability in the extent or severity of disease, which affects sensitivity, or in the alternative conditions that are able to generate false positive findings, affecting specificity.[85]

An adequate description of the demographic and clinical characteristics of study participants allows the reader to judge whether the study can adequately address the study question, and whether the study findings apply to the reader's clinical question.

*In the example*, the authors presented the demographic and clinical characteristics of the study participants in a separate table, a commonly used, informative way of presenting key participant characteristics (Table 3).

*Item 21a. Distribution of severity of disease in those with the target condition*

*Item 21b. Distribution of alternative diagnoses in those without the target condition*

**Example.** "Of the 170 patients with coronary disease, one had left main disease, 53 had three vessel disease, 64 two vessel disease, and 52 single vessel disease. The mean ejection fraction of the patients with coronary disease was 64% (range 37-83). The other 52 men with symptoms had normal coronary arteries or no significant lesions at angiography."[93]

**Explanation.** Most target conditions are not fixed states, either present or absent; many diseases cover a continuum, ranging from minute pathological changes to advanced clinical disease. Test sensitivity is often higher in studies in which more patients have advanced stages of the target condition, as these cases are often easier to identify by the index test.[28,85] The type, spectrum and frequency of alternative diagnoses in those without the target condition may also influence test accuracy; typically, the healthier the patients without the target condition, the less frequently one would find false-positive results of the index test.[28]

An adequate description of the severity of disease in those with the target condition and of the alternative conditions in those without it allows the reader to judge both the validity of the study, relative to the study question, and the applicability of the study findings to the reader's clinical question.

*In the example*, the authors investigated the accuracy of exercise tests for diagnosing coronary artery disease. They reported the distribution of severity of disease in terms of the number of vessels involved; the more vessels, the more severe the coronary artery disease would be. Sensitivity of test exercises was higher in those with more diseased vessels (39% for single vessel disease, 58% for two and 77% for three vessels).[91]

*Item 22. Time interval and any clinical interventions between index test and reference standard*

**Example.** "The mean time between arthrometric examination and MR imaging was 38.2 days (range, 0–107 days)."[94]

**Explanation.** Studies of diagnostic accuracy are essentially cross-sectional investigations. In most cases, one wants to know how well the index test classified patients in the same way as the reference standard, when both tests are performed in the same patients, at the same time.[30] When a delay occurs between the index test and the reference standard, the target condition and alternative conditions can change; conditions may worsen, or improve in the meanwhile, due to the natural course of the disease, or due to clinical interventions applied between the two tests. Such changes influence the agreement between the index test and the reference standard, which could lead to biased estimates of test performance.

The bias can be more severe if the delay differs systematically between test positives and test negatives, or between those with a high prior suspicion of having the target condition and those with a low suspicion.[1,2]

When follow-up is used as the reference standard, readers will want to know how long the follow-up period was.

*In the example*, the authors reported the mean number of days, and a range, between the index test and the reference standard.

*Item 23. Cross tabulation of the index test results (or their distribution) by the results of the reference standard*

**Example.** "Table X shows pain over speed bumps in relation to diagnosis of appendicitis."[95] (see Table 4)

**Explanation.** Research findings should be reproducible and verifiable by other scientists; this applies both to the testing procedures, to the conduct of the study, and to the statistical analyses.

A cross tabulation of index test results against reference standard results facilitates recalculating measures of diagnostic accuracy. It also facilitates recalculating the proportion of study group

participants with the target condition, which is useful as the sensitivity and specificity of a test may vary with disease prevalence.[32,96] It also allows for performing alternative or additional analyses, such as meta-analysis.

Preferably, such tables should include actual numbers, not just percentages, because mistakes made by study authors in calculating estimates for sensitivity and specificity are not rare.

*In the example*, the authors provided a contingency table from which the number of true positives, false positives, false negatives, and true negatives can be easily identified (Table 4).

*Item 24. Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)*

**Example.** "Forty-six patients had pulmonary fibrosis at CT, and sensitivity and specificity of MR imaging in the identification of pulmonary fibrosis were 89% (95% CI: 77%, 96%) and 91% (95% CI: 76%, 98%), respectively, with positive and negative predictive values of 93% (95% CI: 82%, 99%) and 86% (95% CI: 70%, 95%), respectively."[97]

**Explanation.** Diagnostic accuracy studies never determine a test's 'true' sensitivity and specificity; at best the data collected in the study can be used to calculate valid estimates of sensitivity and specificity. The smaller the number of study participants, the less precise these estimates will be.[98]

The most frequently used expression of imprecision is to report not just the estimates – sometimes referred to as point estimates - but also 95% confidence intervals around the estimates. Results from studies with imprecise estimates of accuracy should be interpreted with caution, as over-optimism lurks.[22]

*In the example*, where MRI is the index test and CT the reference standard, the authors reported both point estimates and 95% confidence intervals around them, for sensitivity, specificity, and positive and negative predictive value.

*Item 25. Any adverse events from performing the index test or the reference standard*

**Example.** "No significant adverse events occurred as a result of colonoscopy. Four (2%) patients had minor bleeding in association with polypectomy that was controlled endoscopically. Other minor adverse events are noted in the appendix."[79]

**Explanation.** Not all medical tests are equally safe, and in this they do not differ from many other medical interventions.[99,100] The testing procedure can lead to complications, such as perforations with endoscopy, contrast allergic reactions in CT imaging, or claustrophobia with MRI scanning. Measuring and reporting of adverse events in studies of diagnostic accuracy will provide additional information to clinicians, who may be reluctant to use them if they produce severe or frequent adverse events. Actual application of a test in clinical practice will not just be guided by the test's accuracy, but by several other dimensions as well, including feasibility and safety. This also applies to the reference standard.

*In the example*, the authors distinguished between "significant" and "minor" adverse events, and explicitly reported how often these were observed.

## Discussion

*Item 26. Study limitations, including sources of potential bias, statistical uncertainty, and generalisability*

**Example.** "This study had limitations. First, not all patients who underwent CT colonography (CTC) were assessed by the reference standard methods. […] However, considering that the 41 patients who were eligible but did not undergo the reference standard procedures had negative or only mildly positive CTC findings, excluding them from the analysis of CTC diagnostic performance may have slightly overestimated the sensitivity of CTC (i.e., partial verification bias). Second, there was a long time interval between CTC and the reference methods in some patients, predominately those with negative CTC findings. […] If anything, the prolonged interval would presumably slightly underestimate the sensitivity

and NPV of CTC for non-cancerous lesions, since some 'missed' lesions could have conceivably developed or increased in size since the time of CTC."[101]

**Explanation.** Like other clinical trials and studies, diagnostic accuracy studies are at risk of bias; they can generate estimates of the test's accuracy that do not reflect the true performance of the test, due to flaws or deficiencies in study design and analysis.[1,2] In addition, imprecise accuracy estimates, with wide confidence intervals, should be interpreted with caution. Because of differences in design, participants and procedures, the findings generated by one particular diagnostic accuracy study may not be obtained in other conditions; their generalisability may be limited.[102]

In the discussion section, authors should critically reflect on the validity of their findings, address potential limitations, and elaborate on why study findings may or may not be generalizable. As bias can come down to over- or underestimation of the accuracy of the index test under investigation, authors should discuss the direction of potential bias, along with its likely magnitude. Readers are then informed of the likelihood that the limitations jeopardize the study's results and conclusions (see also Item 27).[103]

Some journals explicitly encourage authors to report on study limitations, but many are not specific about which elements should be addressed.[104] For diagnostic accuracy studies, we highly recommend that at least potential sources of bias are discussed, as well as imprecision, and concerns related to the selection of patients and the setting in which the study was performed.

*In the example*, the authors identified two potential sources of bias that are common in diagnostic accuracy studies: not all test results were verified by the reference standard, and there was a time interval between index test and reference standard, allowing the target condition to change. They also discussed the magnitude of this potential bias, and the direction: whether this may have led to over- or underestimations of test accuracy.

*Item 27. Implications for practice, including the intended use and clinical role of the index test*

**Example.** "A Wells score of ≤4 combined with a negative point of care D-dimer test result ruled out pulmonary embolism in 4-5 of 10 patients, with a failure rate of less than 2%, which is considered safe by most published consensus statements. Such a rule-out strategy makes it possible for primary care doctors to safely exclude pulmonary embolism in a large proportion of patients suspected of having the condition, thereby reducing the costs and burden to the patient (for example, reducing the risk of contrast nephropathy associated with spiral computed tomography) associated with an unnecessary referral to secondary care."[25]

**Explanation.** To make the study findings relevant for practice, authors of diagnostic accuracy studies should elaborate on the consequences of their findings, taking into account the intended use (the purpose of testing) and clinical role of the test (how will the test be positioned in the existing clinical pathway).

A test can be proposed for diagnostic purposes, for susceptibility, screening, risk stratification, staging, prediction, prognosis, treatment selection, monitoring, surveillance, or other purposes. The clinical role of the test reflects its positioning relative to existing tests for the same purpose, within the same clinical setting: triage, add-on, or replacement.[19,105] Both the intended use and the clinical role of the index test should have been described in the introduction of the paper (Item 3).

The intended use and the proposed role will guide the desired magnitude of the measures of diagnostic accuracy. For ruling-out disease with an inexpensive triage test, for example, high sensitivity is required, and less-than-perfect specificity may be acceptable. If the test is supposed to rule-in disease, specificity may become much more important.[106]

In the Discussion section, authors should elaborate on whether or not the accuracy estimates are sufficient for considering the test to be 'fit for purpose'.

*In the example*, the authors concluded that the combination of a Wells score ≤4 and a negative point-of-care D-dimer result could reliably rule-out pulmonary embolism in a large proportion of patients seen in primary care.

## Other information

*Item 28. Registration number and name of registry*

**Example.** "The study was registered at http://www.clinicaltrials.org (NCT00916864)." [107]

**Explanation.** Registering study protocols before their initiation in a clinical trial registry, such as ClinicalTrials.gov or one of the WHO Primary Registries, ensures that existence of the studies can be identified.[108-112] This has many advantages, including avoiding overlapping or redundant studies, and allowing colleagues and potential participants to contact the study coordinators.

Additional benefits of study registration are the prospective definition of study objectives, outcome measures, eligibility criteria and data to be collected, allowing editors, reviewers and readers to identify deviations in the final study report. Trial registration also allows reviewers to identify studies that have been completed but were not yet reported.

Many journals require registration of clinical trials. A low but increasing number of diagnostic accuracy studies are also being registered. In a recent evaluation of 351 test accuracy studies published in high-impact journals in 2012, 15% had been registered.[113]

Including a registration number in the study report facilitates identification of the trial in the corresponding registry. It can also be regarded as a sign of quality, if the trial was registered before its initiation.

*In the example*, the authors reported that the study was registered at ClinicalTrials.gov. The registration number was also provided, so that the registered record could be easily retrieved.

*Item 29. Where the full study protocol can be accessed*

**Example.** "The design and rationale of the OPTIMAP study have been previously published in more detail [with reference to study protocol]."[114]

**Explanation.** Full study protocols typically contain additional methodological information that is not provided in the final study report, because of word limits, or because it has been reported elsewhere. This additional information can be helpful for those who want to thoroughly appraise the validity of the study, for researchers who want to replicate the study, and for practitioners who want to implement the testing procedures.

An increasing number of researchers share their original study protocol, often before enrollment of the first participant in the study. They may do so by publishing the protocol in a scientific journal, at an institutional or sponsor website, or as supplementary material on the journal website, to accompany the study report.

If the protocol has been published or posted online, authors should provide a reference or a link. If the study protocol has not been published authors should state from whom it can be obtained.[115]

*In the example*, the authors provided a reference to the full protocol, which had been published previously.

*Item 30. Sources of funding and other support; role of funders*

**Example.** "Funding, in the form of the extra diagnostic reagents and equipment needed for the study, was provided by Gen-Probe. The funders had no role in the initiation or design of the study, collection of samples, analysis, interpretation of data, writing of the paper, or the submission for publication. The study and researchers are independent of the funders, Gen-Probe."[116]

**Explanation.** Sponsorship of a study by a pharmaceutical company has been shown to be associated with results favoring the interests of that sponsor.[117] Unfortunately, sponsorship is often not disclosed in scientific articles, making it difficult to assess this potential bias. Sponsorship can consist of direct funding of the study, or of the provision of essential study materials, such as test devices.

The role of the sponsor, including the degree to which that sponsor was involved in the study varies. A sponsor could, for example, be involved in the design of the study, but also in the conduct, analysis,

reporting, and decision to publish. Authors are encouraged to be explicit about sources of funding as well as the sponsors role(s) in the study, as this transparency helps readers to appreciate the level of independency of the researchers.

*In the example*, the authors were explicit about the contribution from the sponsor, and their independence in each phase of the study.

## ACKNOWLEDGEMENTS

We thank the STARD Group for helping us in identifying essential items for reporting diagnostic accuracy studies.

## COMPETING INTERESTS

All authors have completed the ICMJE Unified Competing Interest form atwww.icmje.org/coi_disclosure.pdf (available upon request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## AUTHOR CONTRIBUTIONS

JFC, DAK, PMMB: drafting of manuscript. DGA, DEB, CAG, LH, LI, DL, JBR, HCWdV: critical revision of manuscript.

## DATA SHARING STATEMENT

No additional data available.

## TABLES

**Table 1. Key STARD terminology.**

| Term | Explanation |
|---|---|
| Medical test | Any method for collecting additional information about the current or future health status of a patient. |
| Index test | The test under evaluation. |
| Target condition | The disease or condition that the index test is expected to detect. |
| Clinical reference standard | The best available method for establishing the presence or absence of the target condition. A gold standard would be an error-free reference standard. |
| Sensitivity | Proportion of those with the target condition who test positive with the index test. |
| Specificity | Proportion of those without the target condition who test negative with the index test. |
| Intended use of the test | Whether the index test is used for diagnosis, screening, staging, monitoring, surveillance, prediction, prognosis, or other reasons. |
| Role of the test | The position of the index test relative to other tests for the same condition (e.g. triage, replacement, add-on, new test). |
| Indeterminate results | Results that are neither positive or negative |

**Table 2. The STARD 2015 list.**[10]

| Section and topic | No | 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 3. Example of baseline demographic and clinical characteristics of participants in a study evaluating the accuracy of point-of-care fecal tests for diagnosis of organic bowel disease (adapted from Kok et al.[87], with permission).**

| Patient characteristics | n (%) |
|---|---|
| **Geographic region of residency in the Netherlands** | |
| Central (Gelderse Vallei) | 257 (66.6) |
| South (Oostelijke Mijnstreek) | 129 (33.4) |
| **Median age, years (range)** | 60 (18–91) |
| **Women** | 211 (54.7) |
| **Presenting symptoms** | |
| Rectal blood loss | 141 (37.7) |
| Abdominal pain | 267 (70.6) |
| Median duration of abdominal pain (range) | 150 days (1 day to 30 years) |
| Persistent diarrhea | 40 (16.9) |
| Diarrhea | 131 (37.2) |
| Fever | 40 (11.0) |
| Weight loss | 62 (17.1) |
| Bloating | 195 (53.6) |
| Constipation | 169 (46.6) |
| **Physical examination** | |
| Pain at palpation | 117 (46.8) |
| Palpable abdominal mass | 12 (3.0) |
| Palpable rectal mass | 1 (0.3) |

**Table 4. Example of contingency table from a study evaluating the accuracy of pain over speed bumps for diagnosis of appendicitis (adapted from Ashdown et al.[95], with permission).**

| | Appendicitis | | |
|---|---|---|---|
| **Pain over speed bumps** | Positive | Negative | Total |
| Positive | 33 | 21 | 54 |
| Negative | 1 | 9 | 10 |
| Total | 34 | 30 | 64 |

# FIGURES

**Figure 1. Example of flow diagram from a study evaluating the accuracy of faecal immunochemical testing for diagnosis of advanced colorectal neoplasia (adapted from Collins et al.[79], with permission).**

**Figure 2. STARD 2015 flow diagram.**

## REFERENCES


1. 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.* Feb 3 2004;140(3):189-202.
2. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Group Q-S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol.* Oct 2013;66(10):1093-1104.
3. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* Oct 18 2011;155(8):529-536.
4. 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. *Evid Based Med.* Apr 2014;19(2):47-54.
5. Korevaar DA, Wang J, van Enst WA, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology.* Mar 2015;274(3):781-789.
6. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 9/15/1999 1999;282(11):1061-1066.
7. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem.* Jan 2003;49(1):1-6.
8. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA.* 8/28/1996 1996;276(8):637-639.
9. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010 2010;340:c332.
10. 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.
11. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med.* Jan 7 2003;138(1):W1-12.
12. Regge D, Laudi C, Galatola G, et al. Diagnostic accuracy of computed tomographic colonography for the detection of advanced neoplasia in individuals at increased risk of colorectal cancer. *JAMA.* Jun 17 2009;301(23):2453-2461.

13. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *Journal of clinical epidemiology.* Jan 2000;53(1):65-69.

14. 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.* Jun 2015;68(6):708-715.

15. Korevaar DA, Cohen JC, de Ronde MW, Virgili G, Dickersin K, Bossuyt PM. Reporting Weaknessess in Conference Abstracts of Diagnostic Accuracy Studies in Ophthalmology. *Jama Ophthalmology.* 2015;133(12):1464-1467.

16. A proposal for more informative abstracts of clinical articles. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. *Annals of internal medicine.* Apr 1987;106(4):598-604.

17. Stiell IG, Greenberg GH, Wells GA, et al. Derivation of a decision rule for the use of radiography in acute knee injuries. *Ann Emerg Med.* Oct 1995;26(4):405-413.

18. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clinica chimica acta; international journal of clinical chemistry.* Jan 1 2014;427:49-57.

19. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *Bmj.* May 6 2006;332(7549):1089-1092.

20. Gieseker KE, Roe MH, MacKenzie T, Todd JK. Evaluating the American Academy of Pediatrics diagnostic standard for Streptococcus pyogenes pharyngitis: backup culture versus repeat rapid antigen testing. *Pediatrics.* Jun 2003;111(6 Pt 1):e666-670.

21. Tanz RR, Gerber MA, Kabat W, Rippe J, Seshadri R, Shulman ST. Performance of a rapid antigen-detection test and throat culture in community pediatric offices: implications for management of pharyngitis. *Pediatrics.* Feb 2009;123(2):437-444.

22. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology.* May 2013;267(2):581-588.

23. Freer PE, Niell B, Rafferty EA. Preoperative Tomosynthesis-guided Needle Localization of Mammographically and Sonographically Occult Breast Lesions. *Radiology.* May 2015;275(2):377-383.

24. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *International journal of epidemiology.* Apr 1996;25(2):435-442.

25. Geersing GJ, Erkens PM, Lucassen WA, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. *BMJ (Clinical research ed.* 2012;345:e6564.
26. Bomers MK, van Agtmael MA, Luik H, van Veen MC, Vandenbroucke-Grauls CM, Smulders YM. Using a dog's superior olfactory sensitivity to identify Clostridium difficile in stools and patients: proof of principle study. *Bmj.* 2012;345:e7396.
27. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. *The American journal of cardiology.* Nov 1980;46(5):807-812.
28. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem.* Aug 2005;51(8):1335-1341.
29. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ.* Feb 14 2006;174(4):469-476.
30. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *Journal of clinical epidemiology.* Nov 2003;56(11):1118-1128.
31. van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *Journal of clinical epidemiology.* Mar 1995;48(3):417-422.
32. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of clinical epidemiology.* Jan 2009;62(1):5-12.
33. Attia M, Zaoutis T, Eppes S, Klein J, Meier F. Multivariate predictive models for group A beta-hemolytic streptococcal pharyngitis in children. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine.* Jan 1999;6(1):8-13.
34. Knottnerus JA, Knipschild PG, Sturmans F. Symptoms and selection bias: the influence of selection towards specialist care on the relationship between symptoms and diagnoses. *Theor Med.* 1989;10:67-81.
35. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *Journal of clinical epidemiology.* Oct 1992;45(10):1143-1154.
36. Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. *Scandinavian journal of primary health care.* Dec 1993;11(4):241-246.
37. Ezike EN, Rongkavilit C, Fairfax MR, Thomas RL, Asmar BI. Effect of using 2 throat swabs vs 1 throat swab on detection of group A streptococcus by a rapid antigen detection test. *Archives of pediatrics & adolescent medicine.* May 2005;159(5):486-490.

38. Rosjo H, Kravdal G, Hoiseth AD, et al. Troponin I measured by a high-sensitivity assay in patients with suspected reversible myocardial ischemia: data from the Akershus Cardiac Examination (ACE) 1 study. *Clinical chemistry.* Nov 2012;58(11):1565-1573.

39. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ (Clinical research ed.* Mar 16 2002;324(7338):669-671.

40. Detrano R, Gianrossi R, Froelicher V. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. *Progress in cardiovascular diseases.* Nov-Dec 1989;32(3):173-206.

41. Brealey S, Scally AJ. Bias in plain film reading performance studies. *The British journal of radiology.* Apr 2001;74(880):307-316.

42. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *The New England journal of medicine.* Dec 1 1994;331(22):1493-1499.

43. Ronco G, Montanari G, Aimone V, et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. *Cytopathology : official journal of the British Society for Clinical Cytology.* Jun 1996;7(3):151-158.

44. Cohen MB, Rodgers RP, Hales MS, et al. Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver operating characteristics curve analysis. *Archives of pathology & laboratory medicine.* Jun 1987;111(6):518-520.

45. Fox JW, Cohen DM, Marcon MJ, Cotton WH, Bonsu BK. Performance of rapid streptococcal antigen testing varies by personnel. *Journal of clinical microbiology.* Nov 2006;44(11):3918-3922.

46. Gandy M, Sharpe L, Perry KN, et al. Assessing the efficacy of 2 screening measures for depression in people with epilepsy. *Neurology.* Jul 24 2012;79(4):371-375.

47. Stegeman I, de Wijkerslooth TR, Stoop EM, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. *Gut.* Mar 2014;63(3):466-471.

48. Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem.* Apr 2008;54(4):729-737.

49. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *Journal of clinical epidemiology.* Aug 2006;59(8):798-801.

50. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of internal medicine.* Mar 16 1999;130(6):515-524.

51. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* Feb 28 1996;15(4):361-387.
52. Hodgdon T, McInnes MD, Schieda N, Flood TA, Lamb L, Thornhill RE. Can Quantitative CT Texture Analysis be Used to Differentiate Fat-poor Renal Angiomyolipoma from Renal Cell Carcinoma on Unenhanced CT Images? *Radiology.* Apr 23 2015:142215.
53. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med.* Jun 1987;6(4):411-423.
54. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. *AJR. American journal of roentgenology.* Nov 1981;137(5):1055-1058.
55. D'Orsi CJ, Getty DJ, Pickett RM, et al. Stereoscopic digital mammography: improved specificity and reduced rate of recall in a prospective clinical trial. *Radiology.* Jan 2013;266(1):81-88.
56. Knottnerus JA, Buntinx F. *The evidence base of clinical diagnosis: Theory and methods of diagnostic research.* 2nd ed: BMJ Books; 2008.
57. Pepe M. Study design and hypothesis testing. *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press; 2003:214-251.
58. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. *Journal of clinical epidemiology.* Aug 2010;63(8):883-891.
59. Garcia Pena BM, Mandl KD, Kraus SJ, et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA.* Sep 15 1999;282(11):1041-1046.
60. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Medical decision making : an international journal of the Society for Medical Decision Making.* Apr-Jun 1987;7(2):107-114.
61. Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. *Jama.* Nov 19 1982;248(19):2467-2470.
62. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. *Journal of chronic diseases.* 1986;39(8):575-584.
63. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. *Bmj.* 2013;346:f2778.

64. Pisano ED, Fajardo LL, Tsimikas J, et al. Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial: The Radiologic Diagnostic Oncology Group 5 Study. The RDOG5 investigators. *Cancer.* Feb 15 1998;82(4):679-688.

65. Giard RW, Hermans J. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. *Cancer.* Apr 15 1992;69(8):2104-2110.

66. Investigators P. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA.* May 23-30 1990;263(20):2753-2759.

67. Min JK, Leipsic J, Pencina MJ, et al. Diagnostic accuracy of fractional flow reserve from anatomic CT angiography. *Jama.* Sep 26 2012;308(12):1237-1245.

68. Naaktgeboren CA, de Groot JA, Rutjes AW, Bossuyt PM, Reitsma JB, Moons KG. Anticipating missing reference standard data when planning diagnostic accuracy studies. *Bmj.* 2016;352:i402.

69. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology.* Oct 2006;59(10):1102-1109.

70. de Groot JA, Bossuyt PM, Reitsma JB, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. *Bmj.* 2011;343:d4770.

71. Pons B, Lautrette A, Oziel J, et al. Diagnostic accuracy of early urinary index changes in differentiating transient from persistent acute kidney injury in critically ill patients: multicenter cohort study. *Crit Care.* 2013;17(2):R56.

72. Sun X, Ioannidis JP, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. *JAMA.* Jan 22-29 2014;311(4):405-411.

73. Zalis ME, Blake MA, Cai W, et al. Diagnostic accuracy of laxative-free computed tomographic colonography for detection of adenomatous polyps in asymptomatic adults: a prospective evaluation. *Annals of internal medicine.* May 15 2012;156(10):692-702.

74. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol.* Aug 2005;58(8):859-862.

75. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press Inc., New York; 2003.

76. Vach W, Gerke O, Hoilund-Carlsen PF. Three principles to define the success of a diagnostic study could be identified. *Journal of clinical epidemiology.* Mar 2012;65(3):293-300.

77. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ.* May 13 2006;332(7550):1127-1129.

78. Bochmann F, Johnson Z, Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. *Br J Ophthalmol.* Jul 2007;91(7):898-900.

79. Collins MG, Teo E, Cole SR, et al. Screening for colorectal cancer and advanced colorectal neoplasia in kidney transplant recipients: cross sectional prevalence and diagnostic accuracy study of faecal immunochemical testing for haemoglobin and colonoscopy. *Bmj.* 2012;345:e4657.

80. Cecil MP, Kosinski AS, Jones MT, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. *Journal of clinical epidemiology.* Jul 1996;49(7):735-742.

81. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *Journal of clinical epidemiology.* Jun 1992;45(6):581-586.

82. Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. *Medical decision making : an international journal of the Society for Medical Decision Making.* Jan-Mar 1992;12(1):22-31.

83. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. *Journal of chronic diseases.* 1986;39(5):343-355.

84. Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. *Investigative radiology.* Oct 1985;20(7):751-756.

85. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *The New England journal of medicine.* Oct 26 1978;299(17):926-930.

86. Zhou XH. Effect of verification bias on positive and negative predictive values. *Statistics in medicine.* Sep 15 1994;13(17):1737-1745.

87. Kok L, Elias SG, Witteman BJ, et al. Diagnostic accuracy of point-of-care fecal calprotectin and immunochemical occult blood tests for diagnosis of organic bowel disease in primary care: the Cost-Effectiveness of a Decision Rule for Abdominal Complaints in Primary Care (CEDAR) study. *Clinical chemistry.* Jun 2012;58(6):989-998.

88. Harris JM, Jr. The hazards of bedside Bayes. *Jama.* Dec 4 1981;246(22):2602-2605.

89. Hlatky MA, Pryor DB, Harrell FE, Jr., Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *The American journal of medicine.* Jul 1984;77(1):64-71.

90. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Annals of internal medicine.* Jul 15 1992;117(2):135-140.

91. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology.* Jan 1997;8(1):12-17.

92. O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. *Neurology.* Jul 1996;47(1):140-144.

93. Deckers JW, Rensing BJ, Tijssen JG, Vinke RV, Azar AJ, Simoons ML. A comparison of methods of analysing exercise tests for diagnosis of coronary artery disease. *British heart journal.* Dec 1989;62(6):438-444.

94. Naraghi AM, Gupta S, Jacks LM, Essue J, Marks P, White LM. Anterior cruciate ligament reconstruction: MR imaging signs of anterior knee laxity in the presence of an intact graft. *Radiology.* Jun 2012;263(3):802-810.

95. Ashdown HF, D'Souza N, Karim D, Stevens RJ, Huang A, Harnden A. Pain over speed bumps in diagnosis of acute appendicitis: diagnostic accuracy study. *Bmj.* 2012;345:e8012.

96. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ.* Aug 6 2013;185(11):E537-544.

97. Rajaram S, Swift AJ, Capener D, et al. Lung morphology assessment with balanced steady-state free precession MR imaging compared with CT. *Radiology.* May 2012;263(2):569-577.

98. Lang TAS, M. *Generalizing from a sample to a population: Reporting estimates and confidence intervals.* Philadelphia: American College of Physicians; 1997.

99. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Annals of internal medicine.* Nov 16 2004;141(10):781-788.

100. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *Jama.* Jan 24-31 2001;285(4):437-443.

101. Park SH, Lee JH, Lee SS, et al. CT colonography for detection and characterisation of synchronous proximal colonic lesions in patients with stenosing colorectal cancer. *Gut.* Dec 2012;61(12):1716-1722.

102. J.G. ILMBPMGPPGCL. Designing studies to ensure that estimates of test accuracy will travel. In: J.A. K, ed. *The evidence base of clinical diagnosis*. London: BMJ Publishing Group; 2002:95-116.

103. Ter Riet G, Chesley P, Gross AG, et al. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PLoS One.* 2013;8(11):e73623.

104. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. *J Clin Epidemiol.* Apr 2007;60(4):324-329.

105. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Annals of internal medicine.* Jun 6 2006;144(11):850-855.

106. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. *Bmj.* Jul 24 2004;329(7459):209-213.

107. Foerch C, Niessner M, Back T, et al. Diagnostic accuracy of plasma glial fibrillary acidic protein for differentiating intracerebral hemorrhage and cerebral ischemia in patients with symptoms of acute stroke. *Clinical chemistry.* Jan 2012;58(1):237-245.

108. Altman DG. The time has come to register diagnostic and prognostic research. *Clinical chemistry.* Apr 2014;60(4):580-582.

109. Hooft L, Bossuyt PM. Prospective registration of marker evaluation studies: time to act. *Clin Chem.* Dec 2011;57(12):1684-1686.

110. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem.* Jul 2008;54(7):1101-1103.

111. Korevaar DA, Ochodo EA, Bossuyt PM, Hooft L. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov. *Clin Chem.* Apr 2014;60(4):651-659.

112. Rifai N, Bossuyt PM, Ioannidis JP, et al. Registering diagnostic and prognostic trials of tests: is it the right thing to do? *Clin Chem.* Sep 2014;60(9):1146-1152.

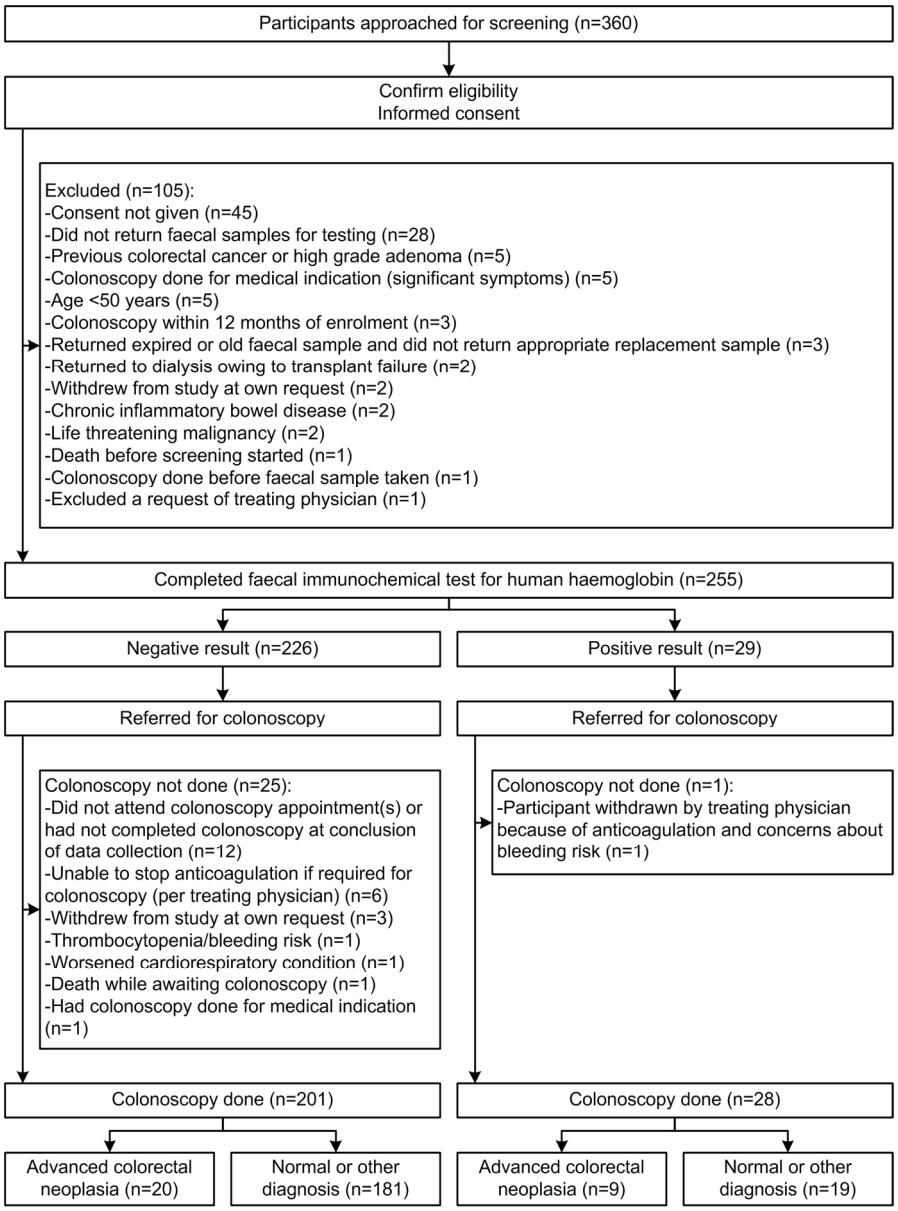
113. Korevaar DA, Bossuyt PM, Hooft L. Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports. *BMJ Open.* 2014;4(1):e004596.

114. Leeuwenburgh MM, Wiarda BM, Wiezer MJ, et al. Comparison of imaging strategies with conditional contrast-enhanced CT and unenhanced MR imaging in patients suspected of having appendicitis: a multicenter diagnostic performance study. *Radiology.* Jul 2013;268(1):135-143.

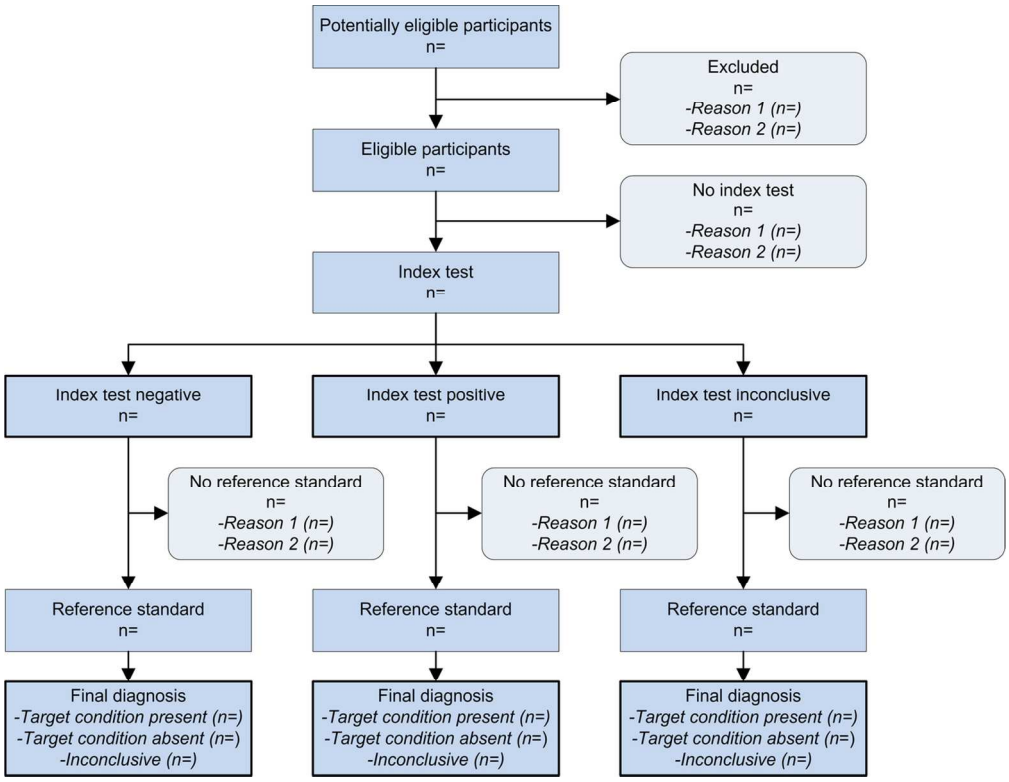
115. Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet.* Jan 18 2014;383(9913):257-266.

116. Stewart CM, Schoeman SA, Booth RA, Smith SD, Wilcox MH, Wilson JD. Assessment of self taken swabs versus clinician taken swab cultures for diagnosing gonorrhoea in women: single centre, diagnostic accuracy study. *Bmj.* 2012;345:e8107.

117. Sismondo S. Pharmaceutical company funding and its consequences: a qualitative systematic review. *Contemporary clinical trials.* Mar 2008;29(2):109-113.




172x233mm (300 x 300 DPI)




125x96mm (300 x 300 DPI)