STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies

Incomplete reporting has been identified as a major source of avoidable waste in biomedical research. Essential information is often not provided in study reports, impeding the identification, critical appraisal, and replication of studies. To improve the quality of reporting of diagnostic accuracy studies, the Standards for Reporting Diagnostic Accuracy (STARD) statement was developed. Here we present STARD 2015, an updated list of 30 essential items that should be included in every report of a diagnostic accuracy study. This update incorporates recent evidence about sources of bias and variability in diagnostic accuracy and is intended to facilitate the use of STARD. As such, STARD 2015 may help to improve completeness and transparency in reporting of diagnostic accuracy studies.

Patrick M Bossuyt¹, Johannes B Reitsma², David E Bruns³, Constantine A Gatsonis⁴, Paul P Glasziou⁵, Les Irwig⁶, Jeroen G Lijmer⁷, David Moher⁸⁹, Drummond Rennie¹⁰¹¹, Henrica C W de Vet¹², Herbert Y Kressel¹³¹⁴, Nader Rifai¹⁵¹⁶, Robert M Golub¹⁷¹⁸, Douglas G Altman¹⁹, Lotty Hooft²⁰, Daniël A Korevaar¹, Jérémie F Cohen¹²¹, for the STARD Group

¹Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands; ²Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, the Netherlands; ³Department of Pathology, University of Virginia School of Medicine, Charlottesville, VA, USA; ⁴Center for Statistical Sciences, Brown University School of Public Health, Providence, RI, USA; ⁵Centre for Research in Evidence-Based Practice, Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Queensland, Australia; ⁶Screening and Diagnostic Test Evaluation Program, School of Public Health, University of Sydney, Sydney, New South Wales, Australia; ⁷Department of Psychiatry, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands; ⁸Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada; ⁹School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada; ¹⁰Peer Review Congress, Chicago, IL, USA; ¹¹Philip R Lee Institute for Health Policy Studies, University of California, San Francisco, CA, USA; ¹²Department of Epidemiology and Biostatistics, Emory University School of Medicine, Emory University Medical Center, Atlanta, Georgia, USA; ¹³Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA; ¹⁴Radiology Editorial Office, Boston, MA, USA; ¹⁵Department of Laboratory Medicine, Boston Children’s Hospital, Harvard Medical School, Boston, MA, USA; ¹⁶Clinical Chemistry Editorial Office, Washington, DC, USA; ¹⁷Division of General Internal Medicine and Geriatrics and Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; ¹⁸JAMA Editorial Office, Chicago, IL, USA; ¹⁹Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK; ²⁰Dutch Cochrance Centre, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, the Netherlands; ²¹INSERM UMR 1153 and Department of Pediatrics, Necker Hospital, AP-HP, Paris Descartes University, Paris, France.

As researchers, we talk and write about our studies, not just because we are happy—or disappointed—with the findings, but also to allow others to appreciate the validity of our methods, to enable our colleagues to replicate what we did, and to disclose our findings to clinicians, other health care professionals, and decision makers, all of whom rely on the results of strong research to guide their actions. Unfortunately, deficiencies in the reporting of research have been highlighted in several areas of clinical medicine. Essential elements of study methods are often poorly described and sometimes completely omitted, making both critical appraisal and replication difficult, if not impossible. Sometimes study results are selectively reported, and other times researchers cannot resist unwarranted optimism in interpretation of their findings. These practices limit the value of the research and any downstream products or activities, such as systematic reviews and clinical practice guidelines.
Reports of studies of medical tests are no exception. A growing number of evaluations have identified deficiencies in the reporting of test accuracy studies. These are studies in which a test is evaluated against a clinical reference standard, or gold standard; the results are typically reported as estimates of the test’s sensitivity and specificity, which express how good the test is in correctly identifying patients as having the target condition. Other accuracy statistics can be used as well, such as the area under the receiver operating characteristics (ROC) curve or positive and negative predictive values.

Despite their apparent simplicity, such studies are at risk of bias. If not all patients undergoing testing are included in the final analysis, for example, or if only healthy controls are included, the estimates of test accuracy may not reflect the performance of the test in clinical applications. Yet such crucial information is often missing from study reports.

It is now well established that sensitivity and specificity are not fixed test properties. The relative number of false positive and false negative test results varies across settings, depending on how patients present and which tests they have already undergone. Unfortunately, many authors also fail to completely report the clinical context and when, where, and how they identified and recruited eligible study participants. In addition, sensitivity and specificity estimates can differ because of variable definitions of the reference standard against which the test is being compared. Thus this information should be available in the study report.

The 2003 STARD statement

To assist in the completeness and transparency of reporting diagnostic accuracy studies, a group of researchers, editors, and other stakeholders developed a minimum list of essential items that should be included in every study report. The guiding principle for developing the list was to select items that, if described, would help readers to judge the potential for bias in the study and appraise the applicability of the study findings and the validity of the authors’ conclusions and recommendations.

The resulting Standards for Reporting Diagnostic Accuracy (STARD) statement appeared in 2003 in two dozen journals. It was accompanied by editorials and commentaries in several other publications and endorsed by many more. Since the publication of STARD, several evaluations have pointed to small but statistically significant improvements in reporting accuracy studies (mean gain 1.4 items (95% confidence interval 0.7 to 2.2)). Gradually, more of the essential items are being reported, but the situation remains far from optimal.

Methods for developing STARD 2015

The STARD steering committee periodically reviews the literature for potentially relevant studies to inform a possible update. In 2013, the steering committee decided that the time was right to update the checklist.

Updating had two major goals: first, to incorporate recent evidence about sources of bias, applicability concerns, and factors facilitating generous interpretation in test accuracy research, and, second, to make the list easier to use. In making modifications, we also considered harmonization with other reporting guidelines, such as Consolidated Standards of Reporting Trials (CONSORT) 2010.

A complete description of the updating process and the justification for the changes are available on the Enhancing the Quality and Transparency of Health Research (EQUATOR) website at www.equator-network.org/reporting-guidelines/stard.

In short, we invited the 2003 STARD group members to participate in the updating process, nominate new members, and comment on the general scope of the update. Suggested new members were contacted. As a result, the STARD group has now grown to 85 members that include researchers, editors, journalists, evidence synthesis professionals, funders, and other stakeholders.

STARD group members were then asked to suggest, and later to endorse, proposed changes in a two round, web based survey. This served to prepare a draft list of essential items, which was discussed in the steering committee in a two day meeting in Amsterdam in September 2014. The list was then piloted in different groups: starting and advanced researchers, peer reviewers, and editors.

The general structure of STARD 2015 is similar to that of STARD 2003. A one page document presents 30 items, grouped under sections that follow the introduction, methods, results, and discussion (IMRAD) structure of a scientific article (see table 1). Several of the STARD 2015 items are identical to the ones in the 2003 version. Others have been reworded, combined, or (if complex) split. A few have been added (see table 2) for a summary of new items and table 3 (for key terms). A diagram to describe the flow of participants through the study is now expected in all reports (figure).
same, and STARD can help in reporting the study in an
informative way. Other reporting guidelines target more specific
forms of tests, such as Transparent Reporting of a Multivariable
Prediction Model for Individual Prognosis or Diagnosis
(TRIPOD) for multivariable prediction models.14
Although STARD focuses on full study reports of test accuracy
studies, the items can also be helpful when writing conference
abstracts, including information in trial registries, and
developing protocols for such studies. Additional initiatives are
underway to provide more specific guidance for each of these
applications.

STARD extensions and applications
The STARD statement was designed to apply to all types of
medical tests. The STARD group believed that a single checklist,
for all diagnostic accuracy studies, would be more widely
disseminated and more easily accepted by authors, peer
reviewers, and journal editors than separate lists for different
types of tests such as imaging, biochemistry, or histopathology.
Having a general list may necessitate additional instructions for
informative reporting, with more information for specific types
of tests, specific applications, or specific forms of analysis. Such
guidance could describe the preferred methods for studying and
reporting measurement uncertainty, for example, without
changing any of the other STARD items. The STARD group
welcomes the development of such STARD extensions and
invites interested groups to contact the STARD executive
committee before developing them.
Other groups may want to develop additional guidance to
facilitate the use of STARD for specific applications. An example
of such a STARD application was prepared for history
taking and physical examination.15 Another type of application
is the use of STARD for specific target conditions such as
deficits.16

Availability
The new STARD 2015 list and related documents can be
found on the STARD pages of the EQUATOR website.
EQUATOR is an international initiative that seeks to improve
the value of published health research literature by promoting
transparent and accurate reporting and wider use of robust
reporting guidelines.14 The STARD group believes that
working more closely with EQUATOR and other reporting
guideline developers will help us to better reach shared
objectives. We have updated the 2003 explanation and
elaboration document, which can also be found at the
EQUATOR website. This document explains the rationale for
each item and gives examples.
The STARD list is released under a Creative Commons license.
This allows everyone to use and distribute the work if they
acknowledge the source. The STARD statement was originally
reported in English, but several groups have worked on
translations in other languages. We welcome such translations,
which are preferably developed by groups of researchers, by
use of a cyclical development process, with back-translation to
the original language and user testing.19 We have also applied
for a trademark for STARD to ensure that the steering committee
has the exclusive right to use the word “STARD” to identify
goods or services.

Increasing value, reducing waste
The STARD steering committee is aware that building a list of
essential items is not sufficient to achieve substantial
improvements in reporting completeness, as the modest
improvement after introduction of the 2003 list has shown. We
see this list not as the final product, but as the starting point for
building more specific instruments to stimulate complete and
transparent reporting, such as a checklist and a writing aid for
authors, tools for reviewers and editors, instruction videos, and
teaching materials, all based on this STARD list of essential
items.
Incomplete reporting has been identified as one of the sources
of avoidable waste in biomedical research.9 Since STARD was
initiated, several other initiatives have been undertaken to
enhance the reproducibility of research and promote greater
transparency.20 Multiple factors are at stake, but incomplete
reporting is one of them. We hope that this update of STARD,
together with additional implementation initiatives, will help
authors, editors, reviewers, readers, and decision makers to
collect, appraise, and apply the evidence needed to strengthen
decisions and recommendations about medical tests. In the end,
we are all to benefit from more informative and transparent
reporting: as researchers, as healthcare professionals, as payers,
and as patients.

This article is being simultaneously published in October 2015 by The
BMJ, Radiology, and Clinical Chemistry. This article is published under the
Creative Commons CC BY-NC license http://creativecommons.org/
licenses/by-nc/4.0/

STARD Group collaborators: Todd Alonso, Douglas G Altman, Augusto
Azuara-Blanco, Lucas Bachmann, Jeffrey Blume, Patrick M Bosuayt,
Isabelle Boutron, David Bruns, Harry Büller, Frank Buntinx, Sarah Byron,
Stephanie Chang, Jérémie F Cohen, Richelle Cooper, Joris de Groot,
Henrica C W de Vet, Jon Deeks, Nandini Dendukuri, Jac Dennes,
Kenneth Fleming, Constantine A Gatsonis, Paul P Glasziou, Robert M
Golub, Gordon Guyatt, Carl Heneghan, Jørgen Hilden, Lotty Hooft, Rita
Horvath, Myriam Hunink, Chris Hyde, John Ioannidis, Les Irwig, Holly
Janes, Jos Kleijnen, André Knottnerus, Daniël A Korevaar, Herbert Y
Kressel, Stefan Lange, Mariska Leeflang, Jeroen G Lijmer, Sally Lord,
Blanca Lumbraeras, Petra Macaskill, Erik Magid, Susan Mallett, Matthew
Mclnnes, Barbara McNeil, Matthew McQueen, David Moher, Karel
Moons, Katie Reem, Reem Mustafa, Nancy Obuchowski, Eleanor
Ochodo, Andrew Onderdonk, John Overbeke, Nitika Pai, Rosanna
Peeling, Margaret Pepe, Steffen Petersen, Christopher Price, Philippe
Ravard, Johannes B. Reitsma, Drummond Rennie, Nader Rifai, Anne
Rutjes, Holger Schunemann, David Simel, Iveta Simera, Nynke Smidt,
Ewout Steyerberg, Sharon Straus, William Summerskill, Yemisi
Takwoingi, Matthew Thompson, Ann van de Bruiel, Hans van Maanen,
Andrew Vickers, Gianni Virgili, Stephen Walter, Wim Weber, Marie
Westwood, Penny Whiting, Nancy Wilczynski, Andreas Ziegler.

Contributors: All authors confirm they have contributed to the intellectual
content of this paper and have met the following 3 requirements: (a)
significant contributions to the conception and design, acquisition of
data, or analysis and interpretation of data; (b) drafting or revising
the article for intellectual content; and (c) final approval of the published
article.
Funding: There was no explicit funding for the development of STARD
2015. The Academic Medical Center of the University of Amsterdam,
the Netherlands, partly funded the meeting of the STARD steering
group but had no influence on the development or dissemination of the list
of essential items. STARD steering group members and STARD group
members covered additional personal costs individually.
Competing interests: All authors have completed the Clinical Chemistry author disclosure form: N Rifai works for Clinical Chemistry, AACC; C A Gatsonisis a member of RSNA Research Development Committee.

2 Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA 2010;303:2058-64.

Accepted: 18 September 2015
## Tables

Table 1 | The STARD 2015 list*

<table>
<thead>
<tr>
<th>Section and topic</th>
<th>No</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title or abstract</strong></td>
<td></td>
<td>Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)</td>
</tr>
<tr>
<td><strong>Abstract</strong></td>
<td>2</td>
<td>Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)</td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td>3</td>
<td>Scientific and clinical background, including the intended use and clinical role of the index test</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Study objectives and hypotheses</td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study design</td>
<td>5</td>
<td>Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)</td>
</tr>
<tr>
<td>Participants</td>
<td>6</td>
<td>Eligibility criteria</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Where and when potentially eligible participants were identified (setting, location, and dates)</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Whether participants formed a consecutive, random, or convenience series</td>
</tr>
<tr>
<td>Test methods</td>
<td>10</td>
<td>Index test, in sufficient detail to allow replication</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Reference standard, in sufficient detail to allow replication</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Rationale for choosing the reference standard (if alternatives exist)</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Whether clinical information and reference standard results were available to the performers or readers of the index test</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Whether clinical information and index test results were available to the assessors of the reference standard</td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Methods for estimating or comparing measures of diagnostic accuracy</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>How indeterminate index test or reference standard results were handled</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>How missing data on the index test and reference standard were handled</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>Intended sample size and how it was determined</td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>19</td>
<td>Flow of participants, using a diagram</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Baseline demographic and clinical characteristics of participants</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>Distribution of severity of disease in those with the target condition</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>Distribution of alternative diagnoses in those without the target condition</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>Time interval and any clinical interventions between index test and reference standard</td>
</tr>
<tr>
<td>Test results</td>
<td>23</td>
<td>Cross tabulation of the index test results (or their distribution) by the results of the reference standard</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>Any adverse events from performing the index test or the reference standard</td>
</tr>
<tr>
<td><strong>Discussion</strong></td>
<td></td>
<td>Study limitations, including sources of potential bias, statistical uncertainty, and generalisability</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>Implications for practice, including the intended use and clinical role of the index test</td>
</tr>
<tr>
<td><strong>Other information</strong></td>
<td></td>
<td>Registration number and name of registry</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>Where the full study protocol can be accessed</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>Sources of funding and other support; role of funders</td>
</tr>
</tbody>
</table>

*STARD stands for STrengthening the Reporting of OBservational studies in Epidemiology.*
Table 1 (continued)

<table>
<thead>
<tr>
<th>Section and topic</th>
<th>No</th>
<th>Item</th>
</tr>
</thead>
</table>

*At the start of each item row, authors should specify the page number of the manuscript where the item can be found.*
<table>
<thead>
<tr>
<th>No</th>
<th>Item</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Structured abstract</td>
<td>Abstracts are increasingly used to identify key elements of study design and results.</td>
</tr>
<tr>
<td>3</td>
<td>Intended use and clinical role of the test</td>
<td>Describing the targeted application of the test helps readers to interpret the implications of reported accuracy estimates.</td>
</tr>
<tr>
<td>4</td>
<td>Study hypotheses</td>
<td>Not having a specific study hypothesis may invite generous interpretation of the study results and “spin” in the conclusions.</td>
</tr>
<tr>
<td>18</td>
<td>Sample size</td>
<td>Readers want to appreciate the anticipated precision and power of the study and whether authors were successful in recruiting the targeted number of participants.</td>
</tr>
<tr>
<td>26-27</td>
<td>Structured discussion</td>
<td>To prevent jumping to unwarranted conclusions, authors are invited to discuss study limitations and draw conclusions keeping in mind the targeted application of the evaluated tests (see item 3).</td>
</tr>
<tr>
<td>28</td>
<td>Registration</td>
<td>Prospective test accuracy studies are trials, and, as such, they can be registered in clinical trial registries, such as ClinicalTrials.gov, before their initiation, facilitating identification of their existence and preventing selective reporting.</td>
</tr>
<tr>
<td>29</td>
<td>Protocol</td>
<td>The full study protocol, with more information about the predefined study methods, may be available elsewhere, to allow more fine grained critical appraisal.</td>
</tr>
<tr>
<td>30</td>
<td>Sources of funding</td>
<td>Awareness of the potentially compromising effects of conflicts of interest between researchers’ obligations to abide by scientific and ethical principles and other goals, such as financial ones; test accuracy studies are no exception.</td>
</tr>
<tr>
<td>Term</td>
<td>Explanation</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Medical test</td>
<td>Any method for collecting additional information about the current or future health status of a patient</td>
<td></td>
</tr>
<tr>
<td>Index test</td>
<td>The test under evaluation</td>
<td></td>
</tr>
<tr>
<td>Target condition</td>
<td>The disease or condition that the index test is expected to detect</td>
<td></td>
</tr>
<tr>
<td>Clinical reference standard</td>
<td>The best available method for establishing the presence or absence of the target condition; a gold standard would be an error-free reference standard</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>Proportion of those with the target condition who test positive with the index test</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>Proportion of those without the target condition who test negative with the index test</td>
<td></td>
</tr>
<tr>
<td>Intended use of the test</td>
<td>Whether the index test is used for diagnosis, screening, staging, monitoring, surveillance, prediction, prognosis, or other reasons</td>
<td></td>
</tr>
<tr>
<td>Role of the test</td>
<td>The position of the index test relative to other tests for the same condition (for example, triage, replacement, add-on, new test)</td>
<td></td>
</tr>
</tbody>
</table>
Figure

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