Reporting standards:
STARD & TRIPOD

Patrick MM Bossuyt
Disclosure

• I have no actual or potential conflict of interest in relation to this presentation, or the tests and biomarkers mentioned in it.

• I have helped in the development of the STARD guidelines.
Waste, Leaks, and Failures in the Biomarker Pipeline

John P.A. Ioannidis1* and Patrick M.M. Bossuyt2

BACKGROUND: The large, expanding literature on biomarkers is characterized by almost ubiquitous significant results, with claims about the potential importance, but few of the discovered biomarkers are used in routine clinical care.

CONTENT: The pipeline of biomarker development includes several specific stages: discovery, validation, clinical translation, evaluation, implementation (and, in the case of nonutility, deimplementation). Each of these stages can be plagued by problems that cause failures of the overall pipeline. Some problems are nonspecific challenges for all biomedical investigation, while others are specific to the peculiarities of biomarker research. Discovery suffers from poor methods and incomplete and selective reporting. External independent validation is limited. Selection for clinical translation is often shaped by irrational choices. Evaluation is sparse and the clinical utility of many biomarkers remains unknown. The regulatory environment for biomarkers remains weak and guidelines can reach biased or divergent recommendations. Removing inefficient or even harmful biomarkers that have been entrenched in clinical care can meet with major resistance.

Biomarkers in many diseases. There is hope that biomarkers will improve our ability to identify, manage, or prevent a wide range of conditions that jeopardize health.

Research in this field has expanded over the years to include measurements of increasing numbers of proteins (the more typical type of biomarker) and other types of molecules (metabolites, DNA genetic variants, different types of RNA molecules) that may serve as biomarkers. For proteins, mass spectrometry allows measurement of multiple analytes with possibly high sensitivity and selectivity, at impressive speed. Multiple peptides, proteins, and their isoforms can be analyzed simultaneously, seemingly allowing even more refined classifications of patients into different classes, or the monitoring of patients during the course of their disease or the management thereof (1). Similarly, recent technical advances in metabolome, genome, and transcriptome measurements have been impressive and raise new methodological and clinical challenges for harnessing this information (2, 3).

Biomarkers are typically measured in biospecimens, but an expanded definition includes also other bioinformation, e.g., procured by imaging (4), sensors, or other measurement tools.
Research
Increasing value, reducing waste

It has been estimated that 85% of research is wasted, usually because it asks the wrong questions, is badly designed, not published or poorly reported. This diminishes the value of research and also represents a significant financial loss. However, many causes of this waste are simple problems that could easily be fixed, such as appropriate randomisation or blinding of a clinical trial. A first step towards increasing the value of research and reducing waste is to monitor the problems and develop solutions that aim to fix them.

researchwaste.net is a place to share and exchange documentation, information, and resources on how to increase the value of both basic and applied research and reduce or avoid wasting research. It is based on a series of articles that were published in the medical journal The Lancet in 2014.

Access articles
Waste at four stages of research

1. Questions relevant to clinicians & patients?
   - Low priority questions addressed
   - Important outcomes not assessed
   - Clinicians and patients not involved in setting research agendas
   - 50%

2. Appropriate design and methods?
   - Over 50% studies designed without reference to systematic reviews of existing evidence
   - Over 50% of studies fail to take adequate steps to reduce biases, e.g. uncontrolled treatment allocation

3. Accessible full publication?
   - Over 50% of studies never published in full
   - Biased under-reporting of studies with disappointing results
   - 50%

4. Unbiased and usable report?
   - Over 30% of trial interventions not sufficiently described
   - Over 50% of planned study outcomes not reported
   - Most new research not interpreted in context of systematic assessment of other relevant evidence
   - 50%

85% Research waste = over $100 Billion / year
Diagnostic Accuracy

• How good is the test in correctly classifying patients as having the target condition?
Diagnostic Accuracy

Medical Test

Gold Standard
Diagnostic Accuracy

Medical Test

Reference standard
Diagnostic Accuracy Study

1. Series of Patients
2. Medical Test
3. Reference standard
4. Cross-classification
Rapid quantitative D-dimer to exclude pulmonary embolism: a prospective cohort management study

S. M. BATES,∗† S. TAKACH LAPNER,∗ J. D. DOUKETIS,∗† C. KEARON,∗† J. JULIAN,‡ S. PARPIA,‡ S. SCHULMAN,∗† J. I. WEITZ,∗† L. A. LINKINS,†‡ M. CROWThER,∗† W. LIM,†‡ F. A. SPENCER,∗† A. Y. Y. LEE,§ P. L. GROSS,∗§ and J. GINSBERG∗§

∗Department of Medicine, McMaster University; †Thrombosis and Atherosclerosis Research Institute, McMaster University; ‡Ontario Clinical Oncology Group, Juravinski Hospital, Hamilton, ON; and §Department of Medicine, University of British Columbia, Vancouver, Canada


Essentials

- It is not known if D-dimer testing alone can safely exclude pulmonary embolism (PE).
- We studied the safety of using a quantitative latex agglutination D-dimer to exclude PE in 808 patients.
- 52% of patients with suspected PE had a negative D-dimer test and were followed for 3 months.
- The negative predictive value of D-dimer testing alone was 99.8%, suggesting it may safely exclude PE.

Summary. Background: Strategies are needed to exclude pulmonary embolism (PE) efficiently without the need for imaging tests. Although validated rules for clinical probability assessment can be combined with D-dimer testing to safely exclude PE, the rules can be complicated or partially subjective, which limits their use. Objectives: To determine if PE can be safely excluded in patients with a negative D-dimer without incorporating clinical probability assessment. Patients/Methods: We imaging tests for PE. All patients in whom PE was excluded had anticoagulant therapy withheld and were followed for 3 months for venous thromboembolism (VTE). Suspected events during follow-up were adjudicated centrally. Results: Eight hundred and eight patients were enrolled, of whom 99 (12%) were diagnosed with VTE at presentation. Four hundred and twenty (52%) patients had a negative D-dimer level at presentation and were not treated with anticoagulants; of these, one had VTE during follow-up. The negative predictive value of D-dimer testing for PE was 99.8% (95% confidence interval, 98.7–99.9%). Conclusions: A negative latex agglutination D-dimer assay is seen in about one-half of patients with suspected PE and reliably excludes PE as a stand-alone test.

Keywords: D-dimer; humans; probability; pulmonary embolism; sensitivity and specificity; venous thromboembolism.
# The results

<table>
<thead>
<tr>
<th>D-dimer</th>
<th>Reference Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PE</td>
</tr>
<tr>
<td>Positive</td>
<td>103</td>
</tr>
<tr>
<td>Negative</td>
<td>1</td>
</tr>
</tbody>
</table>
D-dimer in non PE patients
outpatients

Dotted line, ELISA; dashed line, microlatex; solid line, sensitivity = (1 – specificity); ●, cutoff values (0.5 mg/L FEU).

inpatients
<table>
<thead>
<tr>
<th>Variable risk factors</th>
<th>No. of points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical signs and symptoms of DVT</td>
<td>3</td>
</tr>
<tr>
<td>An alternative diagnosis deemed less likely than PE</td>
<td>3</td>
</tr>
<tr>
<td>Heart rate &gt;100 beats/min</td>
<td>1.5</td>
</tr>
<tr>
<td>Immobilization or surgery in the previous 4 wk</td>
<td>1.5</td>
</tr>
<tr>
<td>Previous DVT or PE</td>
<td>1.5</td>
</tr>
<tr>
<td>Hemoptysis</td>
<td>1</td>
</tr>
<tr>
<td>Cancer (receiving treatment, treated in past 6 mo, or palliative care)</td>
<td>1</td>
</tr>
</tbody>
</table>

### Clinical probability

- **Low**: <2
- **Intermediate**: 2-6
- **High**: >6

**Key:** DVT, deep vein thrombosis; PE, pulmonary embolism.

Systematic review of D-dimer in suspected PE

Di Nisio M.
J Thromb Haemost.
Quality of Reporting of Diagnostic Accuracy Studies

PURPOSE: To evaluate quality of reporting in diagnostic accuracy articles published in 2000 in journals with impact factor of at least 4 by using items of Standards for Reporting of Diagnostic Accuracy (STARD) statement published later in 2003.

MATERIALS AND METHODS: English-language articles on primary diagnostic accuracy studies in 2000 were identified with validated search strategy in MEDLINE. Articles published in journals with impact factor of 4 or higher that regularly publish articles on diagnostic accuracy were selected. Two independent reviewers evaluated quality of reporting by using STARD statement, which consists of 25 items and encourages use of a flow diagram. Total STARD score for each article was calculated by summing number of reported items. Subgroup analyses were performed for study design (case-control or cohort study) by using Student t tests for continuous outcomes and χ² tests for dichotomous outcomes.

RESULTS: Included were 124 articles published in 2000 in 12 journals: 33 case-control and 91 cohort studies. Only 41% of articles (51 of 124) reported on more than 50% of STARD items, while no articles reported on more than 80%. A flow chart was presented in two articles. Assessment of reporting on individual items of STARD statement revealed wide variation, with some items described in 11% of articles and others in 92%. Mean STARD score (0–25 points available) was 11.9 (range, 3.5–19.5). Mean difference in STARD score between cohort studies and case-control studies was 1.53 (95% confidence interval: 0.24, 2.82).

CONCLUSION: Quality of reporting in diagnostic accuracy articles published in 2000 is less than optimal, even in journals with high impact factor. Authors, editors, and reviewers should pay more attention to reporting by checking STARD statement items and including a flow diagram to represent study design and patient flow. © RSNA, 2005
<table>
<thead>
<tr>
<th>Category and Item No.</th>
<th>All Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>((n = 124)^*)</td>
</tr>
</tbody>
</table>

### Methods

3. Study population: Inclusion and exclusion criteria, setting, and locations where data were collected.

4. Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had undergone the index tests or the reference standard?

5. Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>103</td>
<td>83</td>
</tr>
<tr>
<td>70</td>
<td>56</td>
</tr>
</tbody>
</table>
Diagnostic Accuracy Study

- Patients with suspected PE
  - D-dimer
  - CTPA
  - Cross-classification
Patients with suspected PE → D-dimer → CTPA → Cross-classification
D-dimer

Cross-classification

PE Patients

Healthy controls
Patients with suspected PE → D-dimer → CTPA, Follow-up → Cross-classification
Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests

Jeroen G. Lijmer, MD
Ben Willem Mol, MD, PhD
Siem Heisterkamp, PhD
Gouke J. Bonsel, MD, PhD
Martin H. Prins, MD, PhD
Jan H. P. van der Meulen, MD, PhD
Patrick M. M. Bossuyt, PhD

context The literature contains a large number of potential biases in the evaluation of diagnostic tests. Strict application of appropriate methodological criteria would invalidate the clinical application of most study results.

objective To empirically determine the quantitative effect of study design shortcomings on estimates of diagnostic accuracy.

design and setting Observational study of the methodological features of 184 original studies evaluating 218 diagnostic tests. Meta-analyses on diagnostic tests were identified through systematic search of the literature using MEDLINE, EMBASE, and DARE databases and the Cochrane Library (1996-1997). Associations between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model.

main outcome measures Relative diagnostic odds ratio (RDOR), which compared the diagnostic odds ratios of studies of a given test that lacked a particular methodological feature with those without the corresponding shortcomings in design.

results Fifteen (6.8%) of 218 evaluations met all 8 criteria; 64 (30%) met 6 or more. Studies evaluating tests in a diseased population and a separate control group overestimated the diagnostic performance compared with studies that used a clinical population (RDOR, 3.0; 95% confidence interval [CI], 2.0-4.5). Studies in which different reference tests were used for positive and negative results of the test underestimated the diagnostic performance compared with studies using a single reference test for all patients (RDOR, 2.2; 95% CI, 1.5-3.3). Diagnostic performance was also overestimated when the reference test was interpreted with knowledge of the test result (RDOR, 1.3; 95% CI, 1.0-1.9), when no criteria for the test were described (RDOR, 1.7; 95% CI, 1.1-2.5), and when no description of the population under study was provided (RDOR, 1.4; 95% CI, 1.1-1.7).

conclusion These data provide empirical evidence that diagnostic studies with methodological shortcomings may overestimate the accuracy of a diagnostic test, particularly those including nonrepresentative patients or applying different reference standards.

JAMA. 1999;282:1061-1066
STARD Project Plan

• Formation of steering committee (Dec 1999)
• Meeting preparations
• Consensus meeting (16-17 Sep 2000)
KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN
STARD Project Plan

• Formation of steering committee (Dec 1999)
• Meeting preparations
• Consensus meeting (16-17 Sep 2000)
• Piloting and Testing (2001-2002)
• Publication in 24 journals (January 2003)
Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative

Patrick M. Bossuyt,1∗ Johannes B. Reitsma,1 David E. Bruns,2,3 Constantine A. Gatsonis,4 Paul P. Glasziou,5 Les M. Irwig,6 Jeroen G. Lijmer,1 David Moher,7 Drummond Rennie,5,9 and Henrica C.W. de Vet,10 for the STARD Group
<table>
<thead>
<tr>
<th></th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participants</strong></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Eligibility criteria</td>
</tr>
<tr>
<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>8</td>
<td>Where and when potentially eligible participants were identified (setting, location and dates)</td>
</tr>
<tr>
<td>9</td>
<td>Whether participants formed a consecutive, random or convenience series</td>
</tr>
</tbody>
</table>
Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative

Patrick M Bosuuy, Johannes B Reitsma, David E Bruns, Constantine A Gatsonis, Paul P Glasziou, Les M Irwig, Jerem G Limper, David Mohr, Drummond Renne, Henrica C W de Vet, for the STARD group

**Summary**

Background: The world of diagnostic tests is highly dynamic. Now tests are developed at a fast rate and the technology of existing tests is constantly evolving.

**Introduction**

The accuracy of diagnostic tests is highly dynamic. New tests are developed at a fast rate and the technology of existing tests is constantly evolving.

**Methods**

This STARD (Standards for Reporting of Diagnostic Accuracy) initiative aims to improve the transparency and completeness of reporting of diagnostic studies.

**Results**

The search for published guidelines on diagnostic research yielded 38 previously published checklists, from which we extracted a list of 75 potential items. The consensus meeting shortened the list to 25 items, using evidence on bias whenever available. A prototype flow diagram provides information about the method of patient recruitment, the order of test execution, and the numbers of patients undergoing the test under evaluation, the reference standard, or both.

**Conclusion**

Evaluation of research depends on complete and accurate reporting. If medical journals adopt the checklist and flow diagram, the quality of reporting of studies of diagnostic accuracy should improve to the advantage of clinicians, researchers, reviewers, editors, and the public.

The world of diagnostic tests is highly dynamic. Now tests are developed at a fast rate and the technology of existing tests is constantly evolving.
Estudios de precisión diagnóstica (STARD) y pronóstica (REMARK)

迈向完整、准确的诊断准确性研究报告：STARD计划

L’iniziativa STARD per la produzione di studi di accuratezza diagnostica: spiegazione e commenti*

診断精度的研究の完全で正確な報告に向けて：STARDイニシアチブ：日本語翻訳文書

استانداردی برای گزارش صحت آزمون‌های تشخیصی

Vollständiges und präzises Berichten von Studien zur diagnostischen Genauigkeit: Die STARD-Initiative
Reporting Diagnostic Accuracy Studies: Some Improvements after 10 Years of STARD

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

Materials and Methods: PubMed was searched for studies published in 12 high-impact-factor journals in 2012 that evaluated the accuracy of one or more diagnostic tests against a clinical reference standard. Two independent reviewers scored reporting completeness of each article with the 25-item STARD checklist. Mixed-effects modeling was used to analyze differences in reporting with previous evaluations from articles published in 2000 and 2004.
<table>
<thead>
<tr>
<th>Category and Item No.</th>
<th>No. of Articles Published in 2000, before STARD ($n = 124$)*</th>
<th>No. of Articles Published in 2004, after STARD ($n = 141$)*</th>
<th>No. of Articles Published in 2012, after STARD ($n = 112$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Study population: Inclusion and exclusion criteria, setting and location of data collection</td>
<td>35 (28)</td>
<td>30 (21)</td>
<td>73 (65)</td>
</tr>
<tr>
<td>4. Participant recruitment: Recruitment on the basis of symptoms, results from previous tests, or the fact participants had received the index test or the reference standard</td>
<td>103 (83)</td>
<td>130 (92)</td>
<td>106 (95)</td>
</tr>
<tr>
<td>5. Participant sampling: Was the study population a consecutive series of participants? If not, specify how participants were further selected</td>
<td>70 (57)</td>
<td>108 (77)</td>
<td>62 (55)</td>
</tr>
</tbody>
</table>
Figure 2  Forest plot for studies included in meta-analysis comparing adherence post-Standards for Reporting of Diagnostic Accuracy Studies (STARD) and pre-STARD. *Wilczynski \(^{10}\) evaluated only 13 STARD items; the other studies evaluated 25 STARD items. **Results of the studies on obstetrics. ***Results of the studies on gynaecology.
Adherence to STARD

Proportion of articles reporting the items

Cumulation of number of STARD items (0-25) reported

Pre-STARD (2000)
Post-STARD (2004)
Post-STARD (2012)
In the clinical assessment of a medical test, the evaluation of its diagnostic accuracy is an essential step. In an evaluation of diagnostic accuracy, the results of the test are compared with the results of the reference standard in the same patients. Yet, one cannot unconditionally take the results from any particular study at face value. Many authors have pointed out the multiple risks of bias in diagnostic accuracy studies (1,2). Critical appraisal of published studies is therefore essential.

Unfortunately, researchers in many published studies fail to report essential elements of study design and the checklist to evaluate the completeness of reporting and the flow diagram to simplify reporting. Editors and reviewers could use the checklist in a similar way to appraise submitted manuscripts.

After pilot testing, the STARD statement was published in January 2003 in Radiology and several other major scientific journals, including BMJ, Clinical Chemistry, Lancet, and Annals of Internal Medicine (6–8). A large number of journals have included the STARD statement in their instructions to authors (details on the Web site at http://www.stard-statement.org/).
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 of 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 Bossuyt, Johannes B Reitsma, David E Bruns, Constantine A Gatsonis, Paul P Glasziou, Les Irwig, Jeroen G Lijmer, David Moher, Drummond Rennie, and Henrica C W de Vet, for the STARD Group
Overinterpretation and Misreporting of Diagnostic Accuracy Studies: Evidence of “Spin”¹

Purpose:
To estimate the frequency of distorted presentation and overinterpretation of results in diagnostic accuracy studies.

Materials and Methods:
MEDLINE was searched for diagnostic accuracy studies published between January and June 2010 in journals with an impact factor of 4 or higher. Articles included were primary studies of the accuracy of one or more tests in which the results were compared with a clinical reference standard. Two authors scored each article independently by using a pretested data-extraction form to identify actual overinterpretation and practices that facilitate overinterpretation, such as incomplete reporting of study methods or the use of inappropriate methods (potential overinterpretation). The frequency of overinterpretation was estimated in all studies and in a subgroup of imaging studies.
ClinicalTrials.gov

418 evaluations of tests & markers registered

01-2006 – 12-2010

Excluding 94 registered after completion

N=324

(Daniel Korevaar et al. 2014)
<table>
<thead>
<tr>
<th>Section and Topic</th>
<th>No.</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE OR ABSTRACT</td>
<td>1</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>ABSTRACT</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>INTRODUCTION</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>METHODS</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>Study design</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>10a</td>
<td>Index test, in sufficient detail to allow replication</td>
</tr>
<tr>
<td></td>
<td>10b</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>12a</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>12b</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>13a</td>
<td>Whether clinical information and reference standard results were available to the performers/readers of the index test</td>
</tr>
<tr>
<td></td>
<td>13b</td>
<td>Whether clinical information and index test results were available to the assessors of the reference standard</td>
</tr>
<tr>
<td>Analysis</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>RESULTS</td>
<td>19</td>
<td>Flow of participants, using a diagram</td>
</tr>
<tr>
<td>Participants</td>
<td>20</td>
<td>Baseline demographic and clinical characteristics of participants</td>
</tr>
<tr>
<td></td>
<td>21a</td>
<td>Distribution of severity of disease in those with the target condition</td>
</tr>
<tr>
<td></td>
<td>21b</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>DISCUSSION</td>
<td>26</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>OTHER INFORMATION</td>
<td>28</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>
BMJ Open
STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration

Jérémie F Cohen,1,2 Daniel A Koveevar,1 Douglas G Altman,1 David E Bruns,4 Constance A Gatsonis,3 Lofti Hoelt,4 Les Inig,4 Deborah Levitt,4 Johannes B Betsnma,4 Henrica C W de Vet,1 Patrick M Mossi Bossuyt1

ABSTRACT
Diagnostic accuracy studies are, on the other clinical studies, at risk of bias due to shortcomings in design and flaws in the reporting. The STARD (Standards for Reporting of Diagnostic Accuracy Studies) guideline was developed to improve standards for such studies. STARD contains a list of essential items that can be used as a checklist, by authors, reviewers and editors, 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 http://www.equator-network.org/reporting-guidelines/stard/
Here, we present the STARD 2015 explanation and elaboration document. Through comments 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.

INTRODUCTION
Diagnostic accuracy studies are at risk of bias, not unlike other clinical studies. Major sources of bias originate in methodological deficiencies, in the selection of study participants and in the accuracy of the outcome measure. This is particularly true for diagnostic accuracy studies, which often suffer from an excessive use of vague terms. To improve study quality, STARD was developed. The STARD checklist contains 30 specific items, of which 2 is about bias. In order to improve these items, we have prepared the STARD checklist.

The STARD checklist was initially released in 2003 and updated in 2015. The objectives of this update were to include recent evidence about bias and validity, and other issues in complete reporting, and to make the STARD checklist easier to use. The updated STARD 2015 checklist has 30 items, of which 25 are specific to diagnostic accuracy studies.

Below, we present an explanation and elaboration of STARD 2015. This is an extensive revision and update of a similar STARD appendix that was prepared for the STARD 2005 version.10 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 explanations can further assist scientists in writing full 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 specificity, sensitivity, positive predictive value or AUC)

Example. "This case/control study evaluated the accuracy of a test for autoimmune disease, and measured the results in terms of sensitivity and specificity."

Explanation. When searching for relevant biomedical studies on a certain topic, electronic databases such as MEDLINE and Embase are indispensable. To facilitate retrieval of such articles, authors can explicitly identify it as a report of a diagnostic accuracy study. This can be performed by using terms in the title and/or abstract that refer to measures of diagnostic accuracy, such as sensitivity, specificity, positive and negative predictive value. AUC is a measure of accuracy that includes these measures.

In 1991, MEDLINE introduced a specific keyword (MeSH heading) for indexing diagnostic studies: "Sensitivity and Specificity." Unfortunately, the sensitivity of this particular MeSH heading to identify diagnostic accuracy studies can be as low as 51%. As of May 2013, Embase’s thesaurus (Emtree) has 38 check tags for study types, and only one of them, "Evaluation of diagnostic accuracy study," is one of them, but was only introduced in 2011.

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

Abstract
Item 2: Structured summary of study design, methods, results and conclusions (for specific guidelines, 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 the abstract.

In two recent literature surveys, abstracts of diagnostic accuracy studies published in high-impact journals or presented at an international scientific conference were found to be insufficiently informative, because key information about the research question, study methods, study results and the implications of findings were frequently missing.11,12

Instructive abstracts help readers to quickly appraise critical elements of study validity (risk of bias) and applicability of study findings to their clinical setting (generalizability). Structured abstracts, with separate headings for objectives, methods, results and interpretation, allow readers to find essential information more easily.13

Building on STARD 2015, the newly developed STARD for Abstracts provides a list of core elements that should be included in journal and conference abstracts of diagnostic accuracy studies (see final checklist, manuscript under development).

Introduction
Item 2: Scientific and clinical background, including the intended use and clinical role of the index test
STARD for Abstracts: essential items for reporting diagnostic accuracy studies in journal or conference abstracts

Jérémie F Cohen,1,2 Daniël A Korevaar,1 Constantine A Gatsonis,3 Paul P Glasziou,4 Lotty Hooft,5 David Moher,5,7 Johannes B Reitsma,8 Henrica CW de Vet,9 Patrick M Bossuyt1

Many abstracts of diagnostic accuracy studies are currently insufficiently informative. We extended the STARD (Standards for Reporting Diagnostic Accuracy) statement by developing a list of essential items that authors should consider when reporting diagnostic accuracy studies in journal or conference abstracts. After a literature review of published guidance for reporting biomedical studies, we identified 39 items potentially relevant to report in an abstract. We then selected essential items through a two-round Delphi process. Ineligibility criteria, study setting, patient sampling procedures, and confidence intervals around accuracy estimates were reported in less than half of the abstracts. In some cases, study abstracts may be the only information available to clinicians, researchers, reviewers, guideline developers, or policy makers. In evaluations, the proportion of diagnostic accuracy studies presented as conference abstracts that are eventually reported in articles was found to be as low as 39%.

We recently evaluated the quality of reporting of abstracts of diagnostic accuracy studies published in several high impact journals and abstracts presented at a major ophthalmology conference. In line with previous authors, we found that many of these abstracts were insufficiently informative. Key items, such as eligibility criteria, study setting, patient sampling procedures, and confidence intervals around accuracy estimates were reported in less than half of the abstracts. This makes it difficult for readers to assess the validity and applicability of the study findings.
Facilitating Prospective Registration of Diagnostic Accuracy Studies: A STARD Initiative


Although the introduction of prospective trial registration policies has been successful in reducing waste in research, diagnostic accuracy studies are rarely registered. We describe why diagnostic accuracy studies should be registered, and what can be done. Advantages of registration include the identification of unpub- lished studies, prevention of selective outcome reporting, prevention of unnecessary duplication of research, collaboration between researchers, and linkage of study mate- rials. In a survey among representatives of 16 major trial registries, such as ClinicalTrials.gov, ANZCTR (Australian New Zealand Clinical Trials Registry), and the UK-based ISRCTN registry (International Standard Randomised Controlled Trial Number), 13 responded, of which 8 (62%) indicated they always accept registration of diag- nostic accuracy studies and 5 (38%) do so in some cases. How- ever, all but one of them (92%) indicated that their registry currently does not provide specific guidance for registering diagnostic accuracy studies. A second survey among the 85 members of the STARD Group (Standards for Reporting of Diagnostic Accuracy) resulted in the definition of 14 essential protocol items and was used for developing a guide on how these items can be registered in existing major trial registries. We propose that investigation responsible for dia- gnostic accuracy studies should register their study, before recruiting patients, in 1 of the existing major trial registries that are willing to host such studies. We also propose that governmental, research, and academic institutions that pro- vide funding for and journals that publish diagnostic accu- racy studies require such registration.

© 2017 American Association for Clinical Chemistry

Table 2. WHO’s Trial Registration Data Set (TRDS), modified for diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Item on TRDS</th>
<th>Modifications for diagnostic accuracy studies</th>
<th>Examples from existing registered records of diagnostic accuracy studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Primary Registry and Trial Identification Numbers</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>2. Date of Registration in Primary Registry</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>3. Secondary Identifying Numbers</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>4. Source(s) of Monetary or Material Support</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>5. Primary Sponsor</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>6. Secondary Sponsor(s)</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>7. Contact for Public Queries</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>8. Contact for Scientific Queries</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>9. Public Title</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>10. Scientific Title</td>
<td>Include “diagnostic accuracy” or one or more accuracy measures in the title (e.g., sensitivity, specificity, predictive value, likelihood ratio, AUC).</td>
<td></td>
</tr>
<tr>
<td>11. Countries of Recruitment</td>
<td>As defined for trials in WHO’s TRDS.</td>
<td></td>
</tr>
<tr>
<td>12. Health Condition(s) or Problem(s) Studied</td>
<td>Item 2: List the target condition that the diagnostic tests should detect, or the target event to predict.</td>
<td>“The diagnostic accuracy of exhaled breath fingerprinting by eNoSe in diagnosing asthma and atopy.” (NTR1399)</td>
</tr>
<tr>
<td>13. Intervention(s)</td>
<td>Item 4: Reference standard: Provide an informative description of the reference standard. Include (if applicable) the generic name and/or common name.</td>
<td>“Device: VitalScan Magnetocardiograph. A passive, non-contact, mobile medical device that measures, displays, stores, and retrieves magnetic fluctuations caused by heart activity at a patient’s bedside.” (NCT02921438)</td>
</tr>
<tr>
<td></td>
<td>Item 5: Index test: Describe the information available to the performers or readers of the index test (e.g., clinical information, other test results, results of the reference standard).</td>
<td>“Reference standard: Automated blood culture technology, in place as standard NHS care in microbiology laboratories at participating sites, and performed prospectively as part of usual clinical care.” (ISRCTN79977960)</td>
</tr>
<tr>
<td></td>
<td>Item 6: Reference standard: Describe the information available to the assessors of the reference standard (e.g., clinical information, other test results, results of the index test).</td>
<td>“All patients will undergo both of these diagnostic tests, though the order will be randomized. Interpretation and analysis of each test will be blinded to the results of the other.” (ACTRN12615001064426)</td>
</tr>
<tr>
<td></td>
<td>Item 7: Index test: Describe the information available to the performers or readers of the index test (e.g., clinical information, other test results, results of the reference standard).</td>
<td>“A second sonographer, blinded to all clinical information and the primary sonographer’s ultrasound interpretation, will review the diagnostic test sonograms.” (NCT02190981)</td>
</tr>
</tbody>
</table>

Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada.
© 2017 American Association for Clinical Chemistry

Continued on page XX
Enhancing the QUAlity and Transparency Of health Research

Your one-stop-shop for writing and publishing high-impact health research

Library for health research reporting

Search for reporting guidelines
Not sure which reporting guideline to use?
Reporting guidelines under development
Visit the library for more resources

Reporting guidelines for main study types

Randomised trials: CONSORT, Extensions, Other
Observational studies: STROBE, Extensions, Other
Systematic reviews: PRISMA, Extensions, Other
Case reports: CARE, CARE-E, Extensions, Other
Qualitative research: COREQ, COREQ-II, Other
Diagnostic/prognostic studies: STARD, TRIPD, Other
Quality improvement studies: SQUIRE, Other
Economic evaluations: CHEERS, Other
Animal pre-clinical studies: ARRIVE, Other
Study protocols: REBCT, PRISMA-PA, Other
Clinical practice guidelines: AGREE, AGREE-II, Other
See all 395 reporting guidelines

EQUATOR Oncology

Resources supporting complete, accurate and transparent research

Visit EQUATOR Oncology

Toolkits

Find practical help and resources to support you in
Writing research
Using guidelines in your journal
Peer reviewing research
Teaching research skills
View all Toolkits

EQUATOR highlights

30/04/2018 - Faido Terwen, Celebrating Iveta Simera’s decade with the EQUATOR Network

After ten years with the EQUATOR Network, Iveta Simera has moved on from her post as founding Deputy Director of the UK EQUATOR Centre, joining the Global Health Network. In this blog post, UK EQUATOR Centre Director Doug Altman summarises. Read More

12/01/2018 - EQUATOR 2017 highlights

2017 was another busy year for the UK EQUATOR Centre team! We were honoured to send Iveta Simera to the Council of Science Editors (CSE) Annual meeting, to accept the CSE’s 2017 Award for Meritorious Achievement on behalf of the EQUATOR Network. Read More

Interesting videos

Reproducibility and replicability of science

This video records an event held on 26 May 2016 at the Royal Society in London, that focussed on reproducibility and replicability in psychology. The video features a panel discussion with members of the EQUATOR Network at the event. Read More

News

The nuts-and-bolts website: promoting quality in dietary assessment data in epidemiological and clinical studies
30/06/2016

EQUATOR Network publishes review of reporting guidelines for cancer research
21/04/2018

Apply for the 2018 Cochrane-BEWARD prize for reducing waste in research
30/06/2016

Introducing MERIDIAN, home of reporting guidelines for research involving animals
24/01/2016

EQUATOR 2017 highlights
12/01/2018

Sign up to our newsletter
Keep up-to-date with the latest developments by email
RESEARCH METHODS & REPORTING

CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials

Kenneth F Schulz,1 Douglas G Altman,2 David Moher,3 for the CONSORT Group

The CONSORT statement is used worldwide to improve the reporting of randomised controlled trials. Kenneth Schulz and colleagues describe the latest version, CONSORT 2010, which updates the reporting guideline based on new methodological evidence and accumulating experience.

Randomised controlled trials, when appropriately designed, conducted, and reported, represent the gold standard in evaluating healthcare interventions. However, randomised trials can yield biased results if they lack methodological rigour.1 To assess a trial accurately, readers of a published report need complete, clear, and transparent information on its methodology and findings. Unfortunately, attempted assessments frequently fail because authors of many trial reports neglect to provide lucid and complete descriptions of that critical information.2 4

That lack of adequate reporting fuelled the development of the original CONSORT (Consolidated Standards of Reporting Trials) statement in 1996 and its revision five years later.5 6 While those statements improved the reporting quality for some randomised controlled trials,7 8 many trial reports still remain inadequate.7 Furthermore, new methodological evidence and additional experience has accumulated since the last revision in 2001. Consequently, we organised a CONSORT Group meeting to update the 2001 statement.9 10 We introduce here the result of that process, CONSORT 2010.

Intent of CONSORT 2010
The CONSORT 2010 Statement is this paper including the 25 item checklist in the table and the flow diagram. It provides guidance for reporting all randomised controlled trials, but focuses on the most common design type—individually randomised, two group, parallel trials. Other trial designs, such as cluster randomised trials and non-inferiority trials, require varying amounts of additional information. CONSORT extensions for these designs,11 12 and other CONSORT products, can be found through the CONSORT website (www.consort-statement.org). Along with the CONSORT statement, we have updated the explanation and elaboration article,13 which explains the inclusion of each checklist item, provides methodological background, and gives published examples of transparent reporting.

Diligent adherence by authors to the checklist items facilitates clarity, completeness, and transparency of reporting. Explicit descriptions, not ambiguity or omission, best serve the interests of all readers. Note that the CONSORT 2010 Statement does not include recommendations for designing, conducting, and analysing trials. It solely addresses the reporting of what was done and what was found.
TRIPOD stands for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis. Besides the guidelines for authors and reviewers, TRIPOD encompasses various initiatives developed by the TRIPOD Group to alleviate the problems arising from inadequate reporting of prediction modeling studies in biomedical sciences.

The TRIPOD Statement

The main product of TRIPOD is the TRIPOD Statement, which is an evidence-based, minimum set of recommendations for reporting prediction modeling studies in biomedical sciences. This includes both prognostic and diagnostic prediction models as well as prediction model development, validation, updating or extending studies. It offers a standard way for reporting the results of prediction modeling studies and thus aiding their critical appraisal, interpretation and uptake by potential users.

The TRIPOD Statement comprises a 22-item checklist. The checklist items focus on reporting how the study was designed, conducted, analyzed and interpreted. The TRIPOD “Explanation and Elaboration” document explains the principles and background underlying each item of the TRIPOD Statement, each item illustrated by various examples of prediction model studies obtained throughout the field of biomedical sciences.

Endorsement of the TRIPOD Statement

The TRIPOD Statement is endorsed by a large number of prominent general medical journals, specialty medical journals, and leading editorial organizations. TRIPOD is part of a broader effort, to improve the reporting of different types of health research, and indeed, to improve the quality of research used in decision-making in healthcare, combined in the EQUATOR network.
<table>
<thead>
<tr>
<th>Section/Topic</th>
<th>Item</th>
<th>Development or Validation?</th>
<th>Checklist Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title and abstract</td>
<td>1 D</td>
<td>DV</td>
<td>Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.</td>
</tr>
<tr>
<td>Abstract</td>
<td>2 D</td>
<td>DV</td>
<td>Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.</td>
</tr>
<tr>
<td>Introduction Background and objectives</td>
<td>3a D</td>
<td>DV</td>
<td>Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.</td>
</tr>
<tr>
<td></td>
<td>3b D</td>
<td>DV</td>
<td>Specify the objectives, including whether the study describes the development or validation of the model or both.</td>
</tr>
<tr>
<td>Methods</td>
<td>4a D</td>
<td>DV</td>
<td>Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.</td>
</tr>
<tr>
<td></td>
<td>4b D</td>
<td>DV</td>
<td>Specify the key study dates, including start of accrual and end of accrual, and, if applicable, end of follow-up.</td>
</tr>
<tr>
<td>Participants</td>
<td>5a D</td>
<td>DV</td>
<td>Specify key elements of the study setting (e.g., primary care, secondary care, general population), including number and location of centres.</td>
</tr>
<tr>
<td></td>
<td>5b D</td>
<td>DV</td>
<td>Describe eligibility criteria for participants.</td>
</tr>
<tr>
<td></td>
<td>5c D</td>
<td>DV</td>
<td>Give details of treatments received, if relevant.</td>
</tr>
<tr>
<td>Outcome</td>
<td>6a D</td>
<td>DV</td>
<td>Clearly define the outcome that is predicted by the prediction model, including how and when assessed.</td>
</tr>
<tr>
<td>Predictors</td>
<td>6b D</td>
<td>DV</td>
<td>Report any actions to blind assessment of the outcome to be predicted.</td>
</tr>
<tr>
<td></td>
<td>6c D</td>
<td>DV</td>
<td>Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured.</td>
</tr>
<tr>
<td></td>
<td>6d D</td>
<td>DV</td>
<td>Report any actions to blind assessment of predictors for the outcome and other predictors.</td>
</tr>
<tr>
<td>Sample size</td>
<td>7a D</td>
<td>DV</td>
<td>Describe how the study size was arrived at.</td>
</tr>
<tr>
<td>Missing data</td>
<td>7b D</td>
<td>DV</td>
<td>Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.</td>
</tr>
<tr>
<td>Statistical analysis methods</td>
<td>8 D</td>
<td>DV</td>
<td>Describe how predictors were handled in the analyses.</td>
</tr>
<tr>
<td></td>
<td>9 D</td>
<td>DV</td>
<td>Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.</td>
</tr>
<tr>
<td></td>
<td>10a D</td>
<td>DV</td>
<td>Specify the model, all model-building procedures (including any predictor selection), and method for internal validation.</td>
</tr>
<tr>
<td></td>
<td>10b D</td>
<td>DV</td>
<td>Specify the model, all model-building procedures (including any predictor selection), and method for internal validation.</td>
</tr>
<tr>
<td></td>
<td>10c V</td>
<td>DV</td>
<td>For validation, describe how the predictions were calculated.</td>
</tr>
<tr>
<td></td>
<td>10d D</td>
<td>DV</td>
<td>Specify all measures used to assess model performance and, if relevant, to compare multiple models.</td>
</tr>
<tr>
<td></td>
<td>10e V</td>
<td>DV</td>
<td>Describe any model updating (e.g., recalibration) arising from the validation, if done.</td>
</tr>
<tr>
<td>Risk groups Development vs. validation</td>
<td>11 D</td>
<td>DV</td>
<td>Provide details on how risk groups were created, if done.</td>
</tr>
<tr>
<td></td>
<td>12 V</td>
<td>DV</td>
<td>For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.</td>
</tr>
</tbody>
</table>

### Results

| Participants                  | 13a D| DV                          | Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. |
|                               | 13b D| DV                          | Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome. |
|                               | 13c V| DV                          | For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome). |
| Modal development             | 14a D| DV                          | Specify the number of participants and outcome events in each analysis. |
|                               | 14b D| DV                          | If done, report the unadjusted association between each candidate predictor and outcome. |
| Modal specification           | 15a D| DV                          | Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point). |
| Modal performance             | 15b D| DV                          | Explain how to use the prediction model. |
| Modal updating                | 15c D| DV                          | Report performance measures (with CI) for the prediction model. |
|                               | 16 V | DV                          | If done, report the results from any model updating (i.e., modal specification, model performance). |

---

* Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D/V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Kareem G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Moons, PhD; Evert W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study, regardless of the study methods used. This explanation and elaboration document describes the rationale, clarifies the meaning of each item, and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tri pod-statement.org.


For author affiliations, see end of text.

For members of the TRIPOD Group, see the Appendices.

In medicine, numerous decisions are made by care providers, often in shared decision making, on the basis of an estimated probability that a specific disease or condition is present (diagnostic setting) or a specific event will occur in the future (prognostic setting) in an individual. In the diagnostic setting, the probability that a particular disease is present can be used, for example, to inform the referral of patients for further testing, to initiate treatment directly, or to reassure patients that a serious cause for their symptoms is unlikely. In the prognostic context, predictions can be used for planning lifestyle or therapeutic decisions on the basis of the risk for developing a particular outcome or state of health within a specific period (1–3). Such estimates of risk can also be used to risk-stratify participants in therapeutic intervention trials (4–7).

In both the diagnostic and prognostic setting, probability estimates are commonly based on combining information from multiple predictors observed or measured from an individual (1, 2, 8, 10). Information from a single predictor is often insufficient to provide reliable estimates of diagnostic or prognostic probabilities or risks (8, 11). In virtually all medical domains, diagnostic and prognostic multivariable (risk) prediction models are being developed, validated, updated, and implemented with the aim to assist doctors and individuals in estimating probabilities and potentially influence their decision making.

A multivariable prediction model is a mathematical equation that relates multiple predictors for a particular individual to the probability of or risk for the presence (diagnosis) or future occurrence (prognosis) of a particular outcome (10, 12). Other names for a prediction model include risk prediction model, predictive model, prognostic (or prediction) index or rule, and risk score (9).

Predictors are also referred to as covariates, risk indicators, prognostic factors, determinants, test results, or–more statistically–independent variables. They may range from demographic characteristics (for example, age and sex), medical history-taking, and physical examination results to results from imaging, electrophysiology, blood and urine measurements, pathologic examinations, and disease stages or characteristics, and results from genetics, proteomics, transcriptomics, pharmacogenomics, metabolomics, and other new biologic measurement platforms that continuously emerge.

Diagnostic and Prognostic Prediction Models

Multivariable prediction models fall into 2 broad categories: diagnostic and prognostic prediction models (Box A).

In a diagnostic model, multiple—or 2 or more—predictors (often referred to as diagnostic test results) are combined to estimate the probability that a certain condition or disease is present (or absent) at the moment of prediction. These are developed from and used to develop the model. They are intended to be used to estimate the probability of a particular outcome or event for example, mortality, disease recurrence, complication, or therapy response occurring in a certain period in the future. This period may range from hours (for example, predicting postoperative
Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View

Abstract

Background: As more and more researchers are turning to big data for new opportunities of biomedical discoveries, machine learning models, as the backbone of big data analysis, are mentioned more often in biomedical journals. However, owing to the inherent complexity of machine learning methods, they are prone to misuse. Because of the flexibility in specifying machine learning models, the results are often insufficiently reported in research articles, hindering reliable assessment of model validity and consistent interpretation of model output.

Objective: To attain a set of guidelines on the use of machine learning predictive models within clinical settings to make sure the models are correctly applied and sufficiently reported so that true discoveries can be distinguished from random coincidence.

Methods: A multidisciplinary panel of machine learning experts, clinicians, and traditional statisticians were interviewed, using an iterative process in accordance with the Delphi method.

Results: The process produced a set of guidelines that consists of (1) a list of reporting items to be included in a research article and (2) a set of practical sequential steps for developing predictive models.

Conclusion: A set of guidelines was generated to enable correct application of machine learning models and consistent reporting of model specifications and results in biomedical research. We believe that such guidelines will accelerate the adoption of big data analysis, particularly with machine learning methods, in the biomedical research community.

Decision tree
Lasso Regression
Random Forest
Support vector machines
Gradient Boosting Machines
Reporting standards: STARD & TRIPOD

Patrick MM Bossuyt