Chapter 1

Overinterpretation and misreporting of diagnostic accuracy studies: Evidence of “Spin”.

Eleanor A. Ochodo, Margriet C. de Haan, Johannes B. Reitsma, Lotty Hooft, Patrick M. Bossuyt, Mariska M.G. Leeflang

Abstract

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.

Results: Of the 126 articles, 39 (31%; 95% confidence interval [CI]: 23, 39) contained a form of actual overinterpretation, including 29 (23%; 95% CI: 16, 30) with an overly optimistic abstract, 10 (8%; 96% CI: 3%, 13%) with a discrepancy between the study aim and conclusion, and eight with conclusions based on selected subgroups. In our analysis of potential overinterpretation, authors of 89% (95% CI: 83%, 94%) of the studies did not include a sample size calculation, 88% (95% CI: 82%, 94%) did not state a test hypothesis, and 57% (95% CI: 48%, 66%) did not report CIs of accuracy measurements. In 43% (95% CI: 34%, 52%) of studies, authors were unclear about the intended role of the test, and in 3% (95% CI: 0%, 6%) they used inappropriate statistical tests. A subgroup analysis of imaging studies showed 16 (30%; 95% CI: 17%, 43%) and 53 (100%; 95% CI: 92%, 100%) contained forms of actual and potential overinterpretation, respectively.

Conclusion: Overinterpretation and misreporting of results in diagnostic accuracy studies is frequent in journals with high impact factors.
1.1 Introduction
Reporting that distorts or misrepresents study results in order to make the interventions look favourable may lead to overinterpretation of study results. Such reporting is also referred to as ‘spin’(1). Authors may overinterpret scientific reports by using exaggerated language or by presenting an overoptimistic abstract compared to the main text, or by drawing favourable conclusions from results of selected subgroups (2,3). Overinterpretation may also be introduced by methodological shortcomings such as failure to specify a study hypothesis, not making a sample size calculation, or using statistical tests that produce desirable results (1,3–6). These forms of misleading representation of the results of scientific research may compromise decision making in health care and thus the well being of patients.

Overinterpretation has been shown to be common in randomized controlled trials. Boutron and colleagues described how they had identified overinterpretation in reports of randomized clinical trials with a clearly identified primary outcome showing statistically non-significant results. More than 40% of the reports had distorted interpretation in at least two sections of the main text (7).

Overinterpretation may also play a role in diagnostic accuracy studies. Such studies evaluate the ability of a test or marker to correctly identify those with the target condition. The clinical use of tests based on inflated conclusions may trigger physicians to make incorrect clinical decisions, thereby compromising patient safety. Exaggerated conclusions could also lead to unnecessary testing and avoidable health care costs (8). The purpose of this study was to estimate the frequency of distorted presentation and overinterpretation of results “spin” in diagnostic accuracy studies.

1.2 Materials and Methods
This study was based on a systematic search of the literature for diagnostic accuracy studies and the evaluation of included study reports.
1.2.1. Literature search

Two authors (E.O-PhD student; 2-years experience and M.L-Assistant professor, clinical epidemiology; 8-years experience) independently searched MEDLINE in January 2011 for diagnostic accuracy studies published between January and June 2010 in journals with an impact factor of 4 or higher. We focused on these journals because a prior study found that overinterpretation of the clinical applicability of molecular diagnostic tests was more likely in journals with higher impact factors (9). This impact factor cut-off was based on a previously published analysis of accuracy studies (10). We limited our search to retrieve the most recent studies indexed in MEDLINE.

The search combined a previously validated search strategy for diagnostic accuracy studies (“sensitivity AND specificity.sh” OR “specificit*.tw” OR “false negative.tw” OR “accuracy.tw” (where ”.sh” indicates subject heading and “.tw” indicates text word) (11) and a list of 622 international standard serial numbers (ISSN) of journals with an impact factor of 4 or higher obtained from the medical library of the University of Amsterdam (See Appendix-1).

Eligible for inclusion were primary studies that evaluated the diagnostic accuracy of one or more tests against a clinical reference standard. Excluded were non-English studies, animal studies and studies that did not report any accuracy measure.

One author (E.O) independently identified potentially eligible articles by reading titles and abstracts. To ensure that no articles were missed, a second author (M.L) independently screened a random sample of titles and abstracts of 1,000 articles resulting from the total articles that the search strategy yielded. We aimed to score a random sample of the potentially eligible articles as outlined in the analysis and results sections. A summary of the search process is outlined in Figure 1.
Overinterpretation of diagnostic accuracy studies

1.2.2. Definition of overinterpretation in diagnostic accuracy studies

Diagnostic accuracy studies vary in study question, design, type of test evaluated and number of tests evaluated (12,13). We aimed to use a definition of overinterpretation based on common features that could apply to a wide range of tests.

We defined overinterpretation as any reporting of diagnostic accuracy studies that makes tests look more favorable than the results justify. We further distinguished between actual and potential forms of overinterpretation. We defined actual overinterpretation as explicit overinterpretation of study results.
and potential overinterpretation as practices that facilitate overinterpretation such as incompletely reporting applied study methods and assumptions or using inappropriate methods. Incomplete reporting of data may hinder objective appraisal of a paper and mislead readers into thinking that tests are favorable (3,4,14).

This definition of overinterpretation was based on items extracted from published literature on spin (1–3,5,6), the Standards for Reporting of Diagnostic Accuracy (STARD) (8) and experience of content experts in the team. We first listed the potential items that could introduce overinterpretation in diagnostic accuracy studies, based on the experience of content experts in the team. We then searched MEDLINE for key literature that had been published on poor reporting and interpretation of scientific reports up to January 2011. We identified key articles on misrepresentation of study findings in randomized control trials (7), molecular diagnostic tests (9), and in scientific research generally (1–3,5,6,15,16). From these articles, we extracted a list of potential items that could identify overinterpretation in diagnostic accuracy studies.

We then designed and pre-tested a data collection form with the potential list of items that may introduce overinterpretation on 10 diagnostic accuracy studies published in 2011, which were independently evaluated by five authors. These studies were not included in the final analysis. This process of identifying overinterpretation is outlined in figure 2.
## Potential list of spin items
(From content experts[^7], STARD[^4, 8] & published literature[^1])

<table>
<thead>
<tr>
<th>Item</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study hypothesis not stated[^4, 8]</td>
<td></td>
</tr>
<tr>
<td>Sample size calculation not stated[^3, 4, 6, 8]</td>
<td></td>
</tr>
<tr>
<td>Study design: case control study, non-consecutive or non random sampling[^4, 8, 14]</td>
<td></td>
</tr>
<tr>
<td>Subgroups not pre-specified[^4, 18]</td>
<td></td>
</tr>
<tr>
<td>Selective reporting of subgroups[^4]</td>
<td></td>
</tr>
<tr>
<td>Thresholds of continuous test not pre-specified[^4, 8]</td>
<td></td>
</tr>
<tr>
<td>Confidence intervals not reported[^4, 8]</td>
<td></td>
</tr>
<tr>
<td>Favorable recommendations not reflecting reported accuracy measures[^4, 14]</td>
<td></td>
</tr>
<tr>
<td>Recommendations based on other test characteristics (e.g. cost, shelf life, storage, ease of use, user acceptability, adverse events[^9])[^6]</td>
<td></td>
</tr>
<tr>
<td>Overoptimistic title[^6]</td>
<td></td>
</tr>
<tr>
<td>Over interpretation of p-values[^1, 7]</td>
<td></td>
</tr>
<tr>
<td>Over-interpretation of clinical applicability of tests[^14]</td>
<td></td>
</tr>
</tbody>
</table>

## Extra items included after pretesting data collection form
- Role of test under evaluation not stated
- Discrepancy between aim and conclusion

## Items excluded (reasons)
- *Over interpretation of p-values (difficult to score as many diagnostic studies don’t report p-values)
- Over-interpretation of clinical applicability of tests (this is contextual and can’t be objectively scored when evaluating a wide range of tests)

## Items excluded after pretesting of data collection form (reasons)
- Overoptimistic title (diagnostic accuracy articles had neutral ways of writing titles)
- Study design: case control study, non-consecutive or non random sampling
- Favorable recommendations not reflecting accuracy measures (difficult to standardize due to wide range of tests)
- Recommendations based on other test characteristics (difficult to score due to varied reporting)

## 2nd list
- Role of test under evaluation not stated
- Study hypothesis not stated
- Sample size calculation not stated
- Subgroups not pre-specified
- Study conclusions based on selective subgroups
- Thresholds of continuous test not pre-specified
- Confidence intervals not reported
- Discrepancy between aim and conclusion studying main text
- Discrepancy between abstract and main text (overoptimistic results in abstract)

## Final list
### Active spin
- Discrepancy between abstract and main text (overoptimistic results in abstract)
- Study conclusions based on selective subgroups
- Discrepancy between aim and conclusion studying main text

### Passive spin
- Role of test under evaluation not stated
- Study hypothesis not stated
- Sample size calculation not stated
- Subgroups not pre-specified
- Thresholds of continuous tests not pre-specified
- Confidence intervals not reported
- Inappropriate statistical tests

---

**Figure 2: Flow chart of the development of the definition of overinterpretation “spin” in diagnostic accuracy studies**

*In diagnostic tests, p-values can be used to compare the accuracy measures of an index test with a pre-specified value or to compare the accuracy of multiple tests. In this case overinterpretation can occur when tests with non-significant differences are reported as significant, or when setting a low pre-specified value to compare against or a poor comparator test, and then show nicely statistically significant results. However, scoring this was difficult as this depends on the pre-specified value (which was rarely stated in our sample of tests) and the comparator.*
1.2.3 Data Extraction

From the included articles we extracted data on study characteristics and items that can introduce overinterpretation, using the pre-tested data-extraction form (See Appendix-2). We looked for items that can introduce overinterpretation in the abstract (with special focus on the results and conclusion section), and main text (introduction, methods, results and conclusion sections).

The actual forms of overinterpretation that we extracted included:

- **An overoptimistic abstract.** This was considered to be present when the abstract only reported the best results while the main text had an array of results; or when stronger test recommendations or conclusions were reported in the abstract compared to the main text. For the latter, we evaluated the language that was used to make the recommendations. If the authors used affirmative language in the abstract while they used conditional language in the main text to make recommendations or conclusions, we scored this as a stronger abstract. Affirmative language included words such as ‘is definitely’, ‘should be’, ‘excellent for use’, ‘strongly recommended’; conditional language included words such as ‘appears to be’, ‘may be’, ‘could be’, and ‘probably should’.

- **Favorable study conclusions or test recommendations based on selected subgroups.** We scored this as overinterpretation when multiple subgroups were reported in the methods or results section, while the recommendations were only based on one of these subgroups. Caution must be employed when analyzing results of subgroup analysis as they may have a high false positive rate due to the effect of multiple testing (17,18).

- **Discrepancy between study aim and conclusion.** A conclusion that does not reflect the aim of the report may be indicative of flawed results (3,5). We evaluated the main text of the article to note if the conclusion was in line with the specified aim of the study.

The potential forms of overinterpretation were characterized as follows:
• **Not stating a test hypothesis.** We evaluated whether a specific statistical test hypothesis was stated like one test being superior over another, or whether a specific measure of diagnostic accuracy surpasses a pre-specified value. The minimally acceptable accuracy values (null hypothesis) and anticipated performance values (alternative hypothesis) of a test under evaluation depend on the clinical context. These performance values can be obtained from pilot data, prior studies or in cases of novel studies, from experts who may give estimates of clinically relevant performance values. A priori specifying a hypothesis limits the chances of post hoc or subjective judgment about the test's accuracy and intended role (4,19). The anticipated or desirable performance measures also guide sample size calculations.

• **Not reporting a sample size calculation.** We assessed if the sample size required to give a study the power to estimate test accuracy with sufficient precision and the method used to calculate the sample size were reported. By failing to report the sample size calculation used at the outset of the study, readers cannot know if the sample size used in the study was sufficient to estimate the accuracy measures with enough precision (19).

• **Not stating or unclearly stating the intended role of the test under evaluation.** We assessed whether the role of the test was clearly defined in the main text. Before evaluating and recommending a test its intended role ought to be clearly defined. A new test may be used to replace the existing test, or may be used before (triage) or after (add-on) an existing test (20,21). The preferred accuracy value of a test depends on its role.

• **Not pre-specifying groups for subgroup analysis a priori in the methods section.** We assessed whether the subgroups presented in the results section were pre-specified at the start of the study. Failure to pre-specify subgroups can lead to post-hoc analyses motivated by initial inspection of data and may give room for manipulation of results to look favourable (17,18,22).

• **Not pre-specifying positivity thresholds of tests.** For continuous tests, we assessed if the optimal threshold or cut-off value at which a test result is
either considered positive or negative was pre-specified before the start
of the study. Stating a threshold value after data collection and analysis
may give room for manipulation to maximize a test characteristic
(4,23,24).

- Not stating confidence intervals of accuracy measures. We assessed if the
  confidence intervals of the accuracy estimates were reported. Confidence
  intervals enable readers to appreciate the precision of the accuracy
  measures. Without these data, it is difficult to assess the clinical
  applicability of the tests (25–27).

- Using inappropriate statistical tests. Here we evaluated the tests of
  significance used to compare the accuracy measures of the index test and
  reference standard or those that compared the accuracy measures of
  multiple tests. In diagnostic accuracy studies, the test of significance to be
  used depends on the role of test under evaluation and whether the tests
  are performed in different groups of patients (unpaired design) or the
  same patients (paired design) (28).

Two authors scored each article independently. The abstract and main text
sections (Introduction, Methods, Results, discussion and conclusion) were read.
One author (E.O) scored all the selected articles. The other five authors (M.H-
Resident in Radiology; 4-years experience), (JBR-Associate Professor, clinical
epidemiology; 18-years experience), (L.H-Assistant professor, clinical
epidemiology; 13-years experience), (P.B-Professor of clinical epidemiology; 25-
years experience) and (M.L) scored the same articles in predetermined
proportions. Disagreement was resolved by consensus or by a third party when
needed.

1.2.4. Analysis
We calculated the proportion of papers with actual and potential
overinterpretation, and the level of inter-rater agreement in scoring both.
Anticipating 40% of papers to have a distorted presentation or
overinterpretation (7), we calculated that evaluating 100 papers would produce
a two-sided 95% confidence interval that extends from 30% to 50% around the sample proportion, using the exact (Clopper-Pearson) method (29,30).

We analyzed all the included studies and the subset of imaging studies to estimate the frequencies of actual and potential overinterpretation using SAS version 9.2 (SAS Institute Inc, Cary, North Carolina).

1.3 Results
1.3.1 Search results
Our initial search yielded 6,978 articles. After reading the titles and abstracts, 6,558 articles were deemed ineligible. Of the remaining 420 potentially eligible articles, we randomly selected 140 articles for evaluation using STATA version 10.0 with the code “sample 140, count”. We sampled more articles than indicated by our sample size calculation to make up for any false positive articles that would be present in the random selection. After assessing the full texts of these 140 articles, 14 studies were excluded (See Figure 1). In total, 126 studies were included in the final analysis.

1.3.2 Characteristics of included articles
Details of the study characteristics are outlined in Table 1. In summary, the median impact factor of all the included articles was 5.5 (range 4.0 to 16.2) and the median sample size was 151 (range 12 to 20,765). Of all the tests evaluated in the included articles, imaging tests formed the largest group (n=53/126, 42%).
Table 1. Summary of study characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Number (% [95% CI])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N=126</strong></td>
<td></td>
</tr>
<tr>
<td>Journal Impact factor, median [range]</td>
<td>5.491 [4.014-16.225]</td>
</tr>
<tr>
<td>Sample size, median [range]</td>
<td>150.5 [12 -20765]</td>
</tr>
<tr>
<td>Type of test evaluated</td>
<td></td>
</tr>
<tr>
<td>Clinical diagnosis (history/physical examination)</td>
<td>13 (10) [5-16]</td>
</tr>
<tr>
<td>Imaging test</td>
<td>53 (42) [33-51]</td>
</tr>
<tr>
<td>Biochemical test</td>
<td>13 (10) [5-16]</td>
</tr>
<tr>
<td>Molecular test</td>
<td>14 (11) [6-17]</td>
</tr>
<tr>
<td>Immunodiagnostic test</td>
<td>14 (11) [6-17]</td>
</tr>
<tr>
<td>Other</td>
<td>19 (15) [9-21]</td>
</tr>
<tr>
<td>Study design</td>
<td></td>
</tr>
<tr>
<td>Single test</td>
<td>64 (51) [42-60]</td>
</tr>
<tr>
<td>Comparator test</td>
<td>62 (49) [40-58]</td>
</tr>
<tr>
<td>Study design</td>
<td></td>
</tr>
<tr>
<td>Cross sectional</td>
<td>102 (81) [74-88]</td>
</tr>
<tr>
<td>Longitudinal (with follow up)</td>
<td>17 (13) [7-19]</td>
</tr>
<tr>
<td>Diagnostic randomized design</td>
<td>1 (1) [0-2]</td>
</tr>
<tr>
<td>Case control</td>
<td>2 (2) [0-4]</td>
</tr>
<tr>
<td>Unclear</td>
<td>5 (3) [0-6]</td>
</tr>
<tr>
<td>Sampling method</td>
<td></td>
</tr>
<tr>
<td>Consecutive series</td>
<td>60 (48) [39-56]</td>
</tr>
<tr>
<td>Random series</td>
<td>6 (5) [1-8]</td>
</tr>
<tr>
<td>Convenience sampling</td>
<td>23 (18) [11-25]</td>
</tr>
<tr>
<td>Multistage stratified sampling</td>
<td>1 (1) [0-2]</td>
</tr>
<tr>
<td>Unclear</td>
<td>36 (28) [21-37]</td>
</tr>
<tr>
<td>Number of groups patients are sampled from</td>
<td></td>
</tr>
<tr>
<td>One group</td>
<td>87 (69) [61-77]</td>
</tr>
<tr>
<td>Different groups</td>
<td>30 (24) [16-31]</td>
</tr>
<tr>
<td>Unclear</td>
<td>9 (7) [3-12]</td>
</tr>
</tbody>
</table>

1.3.3 Agreement

The inter-rater agreements for scoring both actual and potential overinterpretation are outlined in Appendix 3.

1.3.4 Actual overinterpretation

Of 126 included articles, 39(31%[95% CI 23-39]) contained a form of actual overinterpretation (Table 2). The most frequent form of overinterpretation was an overoptimistic abstract (n=29/126, 23% [95% CI 16-30). Of these, 22 articles (17%[95% CI 11-24) had stronger test recommendations or conclusions in the abstract than in the main text while 7 papers (6%[95% CI 2-10) selectively reported favorable results in the abstract.
Of the included imaging studies (n=53), 16 articles (30% [95% CI 17-43]) contained a form of actual overinterpretation (Table 2). Similar to the overall studies, the most frequent form of actual overinterpretation was an overoptimistic abstract (n=13/53, 25% [95% CI 13-37]).

Table 2. Actual overinterpretation in diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Forms of actual overinterpretationa</th>
<th>All studies</th>
<th>Imaging studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. (% [95% CI])</td>
<td>No. (%[95% CI])</td>
</tr>
<tr>
<td>Overoptimistic abstract</td>
<td>29 (23) [16-30]</td>
<td>13 (25) [13-37]</td>
</tr>
<tr>
<td>Stronger recommendations or conclusions in the abstract</td>
<td>22 (17) [11-24]</td>
<td>9 (17) [9-30]</td>
</tr>
<tr>
<td>Selective reporting of results in the abstract</td>
<td>7 (6) [2-10] [3-20]</td>
<td></td>
</tr>
<tr>
<td>Study conclusions based on selected subgroups</td>
<td>8/80b (10) [5-19]</td>
<td>3/41c (7) [2-21]</td>
</tr>
<tr>
<td>Presence of discrepancy between aim and conclusion</td>
<td>10 (8) [3-13]</td>
<td>2 (4) [0-9]</td>
</tr>
<tr>
<td>Overall proportion of articles with one or more forms of actual overinterpretation</td>
<td>39 (31) [23-39]</td>
<td>16 (30) [17-43]</td>
</tr>
</tbody>
</table>

aOne study may fall under multiple categories
bEighty articles out of the overall 126 articles analyzed subgroups
cForty one articles out of the 53 imaging articles analyzed subgroups

1.3.5 Potential overinterpretation

Details of the potential forms of overinterpretation are outlined in Table 3. Of the 126 included articles, only 14 (11%) reported a sample size calculation. Only 15 articles (12%) reported an explicit study hypothesis. All imaging studies (n=53) contained a form of potential overinterpretation. Of these, only 5 included
articles (9%) reported a sample size calculation. Only 6 articles (11%) reported an explicit study hypothesis. Examples of overinterpretation are provided in Table 4.

Table 3. Potential overinterpretation in diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Forms of potential overinterpretation *</th>
<th>All studies</th>
<th>Imaging studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total number with potential spin (% [95% CI]) N=126</td>
<td>Total number with potential spin (% [95% CI]) N=53</td>
</tr>
<tr>
<td>Sample size calculation not reported</td>
<td>112 (89) [83-94]</td>
<td>48 (91) [82-99]</td>
</tr>
<tr>
<td>Test hypothesis not stated</td>
<td>111 (88) [82-94]</td>
<td>47 (89) [80-98]</td>
</tr>
<tr>
<td>Confidence interval of accuracy measures not reported</td>
<td>72 (57) [48-66]</td>
<td>26 (49) [35-63]</td>
</tr>
<tr>
<td>Role of test not stated or unclear</td>
<td>54 (43) [34-52]</td>
<td>17 (32) [19-45]</td>
</tr>
<tr>
<td>Groups for sub-group analysis not pre-specified in the methods section of main text</td>
<td>25/80 b (31) [21-42]</td>
<td>8/41 c (20) [8-36]</td>
</tr>
<tr>
<td>Positivity thresholds of continuous tests not reported</td>
<td>22/63 d (35) [24-48]</td>
<td>6/25 e (24) [10-45]</td>
</tr>
<tr>
<td>Use of inappropriate statistical tests</td>
<td>4 (3) [0-6]</td>
<td>3 (6) [0-12]</td>
</tr>
<tr>
<td>Overall proportion of articles with one or more forms of potential overinterpretation</td>
<td>125 (99) [98-100]</td>
<td>53 (100) [92-100]</td>
</tr>
</tbody>
</table>

*aOne study may fall under multiple categories

bEighty articles out of the overall 126 articles analyzed subgroups
cForty one articles out of the 53 imaging articles analyzed subgroups
dSixty-three articles out of the overall 126 articles evaluated continuous tests
eTwenty-five articles out of the 53 imaging articles evaluated continuous tests
Table 4. Examples of actual overinterpretation

1. An abstract with a stronger conclusion (31):

<table>
<thead>
<tr>
<th>Conclusion in Main text</th>
<th>Conclusion in Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Detection of antigen in BAL using the Mvista antigen appears to be a useful method (...) Additional studies are needed in patients with pulmonary histoplasmosis.”</td>
<td>“Detection of antigen in BAL fluid complements antigen detection in serum and urine as an objective test for histoplasmosis”</td>
</tr>
</tbody>
</table>

2. Conclusions drawn from selected subgroups (32):

A study evaluates the aptness of F-desmethoxyfallypride (F-DMFP) PET for the differential diagnosis of Idiopathic Parkinsonian Syndrome (IPS) and non-IPS in a series of 81 patients with a clinical diagnosis of Parkinsonism. The authors compared several F-DMFP PET indices for the discrimination of IPS and non-IPS and reported only the best sensitivity and specificity estimates. They concluded that F-DMFP PET was an accurate method for differential diagnosis.

3. Disconnect between the aim and conclusion of the study (33):

The study design described in this paper aimed to evaluate the sensitivity and specificity of the IgM anti-EV71 assay. However the conclusion is not on accuracy rather it focuses on other measures of diagnostic performance.

<table>
<thead>
<tr>
<th>Aim of study</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The aim of this study was to assess the performance of detecting IgM anti-EV71 for early diagnosis of patients with HFMD”</td>
<td>“The data here presented show that the detection of IgM anti-EV71 by ELISA affords a reliable convenient and prompt diagnosis of EV71. The whole assay takes 90 mins using readily available ELISA equipment, is easy to perform with low cost which make it suitable in clinical diagnosis as well as in public health utility”</td>
</tr>
</tbody>
</table>

Abbreviations

BAL Bronchoalveolar lavage; PET Positron Emission Tomography; IgM Immunoglobulin M: EV71 Enterovirus 71; HFMD Hand foot and Mouth Disease; ELISA Enzyme Linked Immunosorbent Assay

1.4 Discussion

Our study shows that about three out of ten studies of the diagnostic accuracy of biomarkers or other medical tests published in journals with an impact factor of 4 or higher overinterpret their results and 99% of studies contain practices that
facilitate overinterpretation. The most common form of actual overinterpretation is an overoptimistic abstract (about one in four) specifically reporting stronger conclusions or test recommendations in the abstract compared to the main text (about one in five). In terms of practices that facilitate overinterpretation, “potential overinterpretation”, the majority of studies failed to report an a priori formulated test hypothesis, did not include a corresponding sample size calculation, and did not include confidence intervals for accuracy measures.

A closely related study by Lumbreras and colleagues evaluated overinterpretation the clinical applicability of molecular diagnostic tests only. Of the 108 evaluated articles, 61 (56%) had overinterpreted the clinical applicability of the molecular test under study.

A defining strength of our study is that we analyzed a sample of diagnostic accuracy studies evaluating a wide range of tests and defined overinterpretation in terms of common features that apply to most tests. To limit subjectivity somewhat, we systematically searched for diagnostic studies with a validated search strategy. Scoring of the articles was done by two authors independently using a pretested data-extraction form.

The forms of overinterpretation that we found in our study may have several implications for diagnostic research and practice. One of the most important consequences might be that diagnostic accuracy studies with optimistic conclusions may be highly cited leading to a cascade of inflated and questionable evidence in the literature. Subsequently, this may translate to the premature adoption of tests into clinical practice. A recently published review by Ioannidis and Panagiotou reported that highly cited biomarker studies often had inflated results. Of the included highly cited studies in their review, 86% had larger effect sizes than in the largest study and 83% had larger effect sizes than in corresponding meta-analysis evaluating the same biomarker (34). Of note, the larger studies and meta-analysis received fewer citations (35).
Our study was largely limited to what was reported. For instance, there is no guarantee that because subgroups or thresholds were listed in the methods, they were indeed pre-specified. An alternative would be to look at study protocols, but unlike in randomized trials protocols of diagnostic accuracy studies are not always registered. Another limitation of our study was the considerable variation in the inter-rater agreement for scoring. The overall scoring of articles was difficult as many articles suffered from poor reporting. Many authors had not employed the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines in preparing their manuscripts. The suboptimal use of STARD has also been documented in previous reports (36–41).

Comprehensively evaluating overinterpretation in diagnostic studies depends on the context in which the test is used. For instance, overinterpretation can occur when positive recommendations are made for the clinical use of tests even if the accuracy measures do not justify this. Due to the wide range of tests evaluated in our study, it was difficult to come up with a standard cutoff measure to define low and high accuracy measures. Preferred accuracy measures differ and are dependent on the type of test used, the role of the test, the target condition and the setting in which the test is being evaluated and on the accuracy of other methods available. A sensitivity of 80% may be ‘definitely useful’ in one situation, while it may be useless in another situation.

Additionally, not reporting confidence intervals may be regarded as either actual or potential overinterpretation depending on the context. For instance, reporting very high point estimates such as 99% without confidence intervals based on a small sample size such as 10 cases may be regarded as actual overinterpretation. On the other hand, not reporting confidence intervals in cases of moderate or high estimates with large sample sizes or in comparative evaluations where a trial compares two tests and one is statistically superior, can be regarded as potential overinterpretation.

To curb the occurrence of overinterpretation and misreporting of results in diagnostic accuracy studies we recommend that journals continuously...
emphasize that accuracy manuscripts submitted to them must be reported according to the STARD reporting guidelines. This may also diminish the methodological conditions that may lead to overinterpretation. Readers largely depend on abstracts to draw conclusions about an article and sometimes when full texts are not available, decisions may be made on abstracts alone (7,42,43). Hence reviewers need to be more stringent when reading abstracts of submitted manuscripts to ensure that the abstracts are fair representations of the main texts. We hope that highlighting the forms of overinterpretation will enable peer reviewers correctly sieve overoptimistic reports of diagnostic accuracy studies and encourage investigators to be clearer in designing, more transparent in reporting, and more stringent in interpreting test accuracy studies.

References


34. Ioannidis JPA, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. JAMA 2011;305(21):2200–10.


42. Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. JAMA. ;281(12):1110-1.


Appendices

Appendices and supplemental material can be accessed at http://pubs.rsna.org/journal/radiology