Clinical Chemistry 63:8
1353–1362 (2017)

Overinterpretation of Research Findings: Evidence of "Spin" in Systematic Reviews of Diagnostic Accuracy Studies

Trevor A. McGrath,[1] Matthew D.F. McInnes,[2*] Nick van Es,[3] Mariska M.G. Leeflang,[4] Daniël A. Korevaar,[4] and Patrick M.M. Bossuyt[4]

BACKGROUND: We wished to assess the frequency of overinterpretation in systematic reviews of diagnostic accuracy studies.

METHODS: MEDLINE was searched through PubMed from December 2015 to January 2016. Systematic reviews of diagnostic accuracy studies in English were included if they reported one or more metaanalyses of accuracy estimates. We built and piloted a list of 10 items that represent actual overinterpretation in the abstract and/or full-text conclusion, and a list of 9 items that represent potential overinterpretation. Two investigators independently used the items to score each included systematic review, with disagreements resolved by consensus.

RESULTS: We included 112 systematic reviews. The majority had a positive conclusion regarding the accuracy or clinical usefulness of the investigated test in the abstract (n = 83; 74%) and full-text (n = 83; 74%). Of the 112 reviews, 81 (72%) contained at least 1 actual form of overinterpretation in the abstract, and 77 (69%) in the full-text. This was most often a "positive conclusion, not reflecting the reported summary accuracy estimates," in 55 (49%) abstracts and 56 (50%) full-texts and a "positive conclusion, not taking high risk of bias and/or applicability concerns into account," in 47 abstracts (42%) and 26 full-texts (23%). Of these 112 reviews, 107 (96%) contained a form of potential overinterpretation, most frequently "nonrecommended statistical methods for metaanalysis performed" (n = 57; 51%).

CONCLUSIONS: Most recent systematic reviews of diagnostic accuracy studies present positive conclusions and a majority contain a form of overinterpretation. This may lead to unjustified optimism about test performance and erroneous clinical decisions and recommendations.

In their 2015 report, the US National Academy of Medicine identified better understanding of diagnostics as the next imperative for patient safety *(1, 2)*. Lack of understanding of the diagnostic accuracy of tests can negatively impact patients by leading to inaccurate estimates of the risk of false negatives (missed diagnoses) or false positives (potential overdiagnosis).

Diagnostic accuracy research impacts virtually all aspects of healthcare; for example, physical examination, imaging, pathology, microbiology, and laboratory medicine. Systematic reviews of diagnostic accuracy studies synthesize data from multiple studies to provide more precise estimates of the ability of medical tests to detect a target condition and to understand reasons for variability in test performance. The number of such systematic reviews has increased in recent years, with hundreds now conducted annually *(3)*. Clinicians and practice-guideline authors commonly regard systematic reviews as high-level evidence. Like any research, systematic reviews should be free from bias and present objective summaries of completed research.

Major concerns have been raised about overrepresentation in the biomedical literature of randomized trial results favoring interventions *(4–6)*. Recently, a phenomenon called "spin," or distortion of the study findings, has been identified. This refers to reporting practices that mislead readers by being more optimistic about the intervention than the results justify *(7)*. Several studies evaluated spin among reports of trials of medical interventions with statistically nonsignificant results for the

[1] University of Ottawa, Ottawa, Ontario, Canada; [2] University of Ottawa Department of Radiology. Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada; [3] Department of Vascular Medicine, Academic Medical Center, Amsterdam, the Netherlands; [4] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center Amsterdam, the Netherlands.

* Address correspondence to this author at: University of Ottawa Department of Radiology. Clinical Epidemiology Program, Ottawa Hospital Research Institute, Rm. c159 Ottawa Hospital Civic Campus, 1053 Carling Ave., Ottawa ON, K1Y 4E9. Fax +613761-4476; e-mail mmcinnes@toh.on.ca.
Received January 17, 2017; accepted March 15, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.271544

primary outcome, and found at least one spin strategy in 84% of these trial reports *(7–15)*.

Spin, or "overinterpretation," is also common in reports of diagnostic accuracy studies *(16, 17)*. Such studies are often accompanied by optimistic claims about the accuracy or clinical usefulness of the investigated test. Examples of spin are use of language that exaggerates the magnitude of the findings, stronger conclusions in the abstract than in the full-text, and selective presentation of favorable results in the abstract *(16)*.

Overinterpretation can bias readers' interpretation of scientific results *(18)*. For diagnostic accuracy studies, this may lead to inappropriate application of tests, which could harm patients and unnecessarily increase healthcare costs. The purpose of this study was to assess the frequency of overinterpretation in systematic reviews of diagnostic accuracy studies.

Methods

SEARCH AND SELECTION

MEDLINE was searched through PubMed using the following search strategy: systematic[sb] AND (sensitivity and specificity[mesh] OR sensitivit*[tw] OR specifit*[tw] OR accur*[tw] or ROC[tw] or AUC[tw] or likelihood[tw]). In line with previous evaluations of spin, we aimed for a sample size of approximately 100–150 reports *(16, 17)*. The search was therefore restricted to articles published in December 2015 and January 2016 and performed on March 12, 2016.

We included systematic reviews of diagnostic accuracy studies in humans that were in English and reported ≥1 metaanalysis of accuracy estimates. Reviews of a test's predictive or prognostic accuracy were excluded.

Two independent investigators reviewed the titles and abstracts of search results (T.A. McGrath, 100%, M.D.F. McInnes and D.A. Korevaar, 50% each). Full-texts of reports identified as potentially eligible by ≥1 of the investigators were independently reviewed for inclusion; discrepancies were resolved through discussion.

ESTABLISHING ITEMS

A published list of items developed for evaluating the occurrence of overinterpretation in primary diagnostic accuracy studies *(16)* was combined, with a published list of items deemed to establish spin in systematic reviews of medical interventions *(19)*. Items on this longlist were assessed for relevance by 3 investigators (M.D.F. McInnes, D.A. Korevaar, P.M.M. Bossuyt). Overinterpretation was defined as reporting of diagnostic accuracy systematic reviews that makes test performance look more favorable than the results justify. The longlist, with reasons for exclusion, is provided in Tables 1 and 2 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol63/issue8.

Based on this longlist, a shortlist of items was established. The same 3 investigators categorized items as representing "actual overinterpretation" or "potential overinterpretation." The latter refers to practices that facilitate overinterpretation but make a formal assessment impossible, such as not reporting CIs around summary estimates. Because most of the items that represent actual overinterpretation rely on the presence of a positive conclusion, a classification scheme for assessing the overall degree of positivity of each conclusion was established.

The 3 investigators developed criteria for assessing overinterpretation for each item on the shortlist and for assessing the overall degree of positivity of each conclusion. The prefinal list of criteria was piloted by all 6 investigators who were involved in data extraction. They applied the criteria on 3 systematic reviews, and the criteria for assessing the degree of positivity of conclusions on 15 abstracts, all not part of the final data set. The results were discussed until consensus was reached, after which the criteria were refined and finalized.

FINAL LIST OF CRITERIA

The final list consists of 10 items representing actual overinterpretation, and 9 items representing potential overinterpretation (Table 1). Seven items representing actual overinterpretation refer to "positive conclusions." When assessing the overall degree of positivity of a conclusion, we classified them as "summary statements" about the accuracy of the test or "implications for practice" about the clinical utility. Each summary statement and implication for practice reported in a conclusion could be "positive," "negative," "neutral," or "not applicable." A conclusion was considered positive if it contained both a positive summary statement and a positive implication for practice, or if one of these was positive and the other neutral or not applicable. Detailed criteria for assessing the degree of positivity of the conclusion, and for the classification of summary statements and implications for practice, are in online Supplemental Tables 3, 4, and 5.

DATA EXTRACTION

One author (T.A. McGrath) extracted the following data from included systematic reviews: first author, journal, and the journal impact factor in 2014 from the Web of Knowledge. Two investigators (T.A. McGrath, 100%; M.D.F. McInnes, N. van Es, M.M.G. Leeflang, D.A. Korevaar, or P.M.M. Bossuyt, 20% each) independently extracted information about the type of test under evaluation (imaging vs laboratory vs other) and design [comparative (multiple index tests) vs noncomparative (1 index test)].

Table 1a. Actual overinterpretation practice criteria in abstracts or full-texts of systematic reviews of diagnostic accuracy studies.

No.	Item	Criteria
a.1	Positive conclusion, not reflecting the reported summary accuracy estimates	Actual overinterpretation if all the following: 1. Positive conclusion 2. Lower confidence limits of the accuracy estimates were lower than the minimally acceptable criteria for test performance as defined by the authors of the review, OR no minimally acceptable criteria were defined but test performance was suboptimal (at least one of the accuracy estimates on which the conclusions were based had a point estimate below 0.90 and/or a lower confidence limit below 0.80) *Note (1): In case a conclusion is classified as "positive," but the positivity only applies to one accuracy estimate in a pair of estimates (e.g. "sensitivity was high" without a statement on specificity), we will only consider the conclusion as representing "actual overinterpretation" if the criteria above apply to that accuracy estimate about which the conclusion was positive.* *Note (2): In case a conclusion is classified as "positive," but the positivity only applies to one test while multiple were investigated (e.g. "test A was highly accurate, but test B was not"), we will only consider the conclusion as representing "actual overinterpretation" if the criteria above apply to that accuracy estimate about which the conclusion was positive.* *Note (3): In case a conclusion is classified as "positive," but refers to the superiority of one test over another in a systematic review that compares the accuracy of multiple index tests (e.g. "test A was more accurate than test B"), this item will be scored as "no actual overinterpretation."* *Note (4): For scoring actual overinterpretation in the abstract: if the abstract does not contain accuracy estimates that are proportions (e.g. DOR only), the focus will be on the corresponding accuracy estimates that are proportions in the full-text (e.g. sensitivity and specificity).*
a.2	Positive conclusion, not taking high risk of bias and/or applicability concerns into account	Actual overinterpretation if all the following: 1. Positive conclusion 2. Recommended tools were used to evaluate risk of bias and/or applicability concerns (i.e. QUADAS or QUADAS-2) 3. At least one question/item/domain (depends on how authors present it) showed high or unclear risk of bias and/or applicability concerns for at least half of the studies included in the final analysis 4. This was not taken into account in the conclusion, or the discussion thereof (e.g. risk of bias or applicability concerns acknowledged in discussion)
a.3	Positive conclusion, not taking heterogeneity into account	Actual overinterpretation if all the following: positive conclusion AND *for abstract*: (Potential) heterogeneity was identified or not taken into account in the full-text, and not taken into account in the abstract results, abstract conclusion or the abstract discussion (reporting on a comparison of subgroups in the abstract is sufficient); *for full-text:* (Potential) heterogeneity was identified or not assessed, but was not taken into account in the conclusion or the discussion thereof (e.g., heterogeneity acknowledged in discussion, or statement such as "accuracy was highly variable") *Note (1): For abstract: In case it was concluded in the full-text that heterogeneity was non-existing, but nothing was reported on heterogeneity in the abstract, this was scored as "no actual overinterpretation" in the abstract.*
a.4	Positive conclusion, focusing on the results of primary studies favoring the diagnostic accuracy of the test instead of the metaanalysis results	Actual overinterpretation if all the following: 1. Positive conclusion 2. Conclusion was only based on primary studies included in the review instead of the metaanalysis results, which showed lower accuracy
a.5	Positive conclusion, selectively focusing on a selection of subgroups, tests or accuracy estimates, while others were evaluated as well	Actual overinterpretation if all the following: 1. Positive conclusion 2. Multiple subgroup analyses were performed, multiple tests were evaluated, and/or one or more pairs of accuracy estimates (i.e. sensitivity and specificity, positive and negative predictive value, or positive and negative likelihood ratio, corresponding to one 2×2 table) were reported 3. Conclusion was only based on the subgroups in which the test performed best, on the test that performed best overall, and/or on the accuracy estimate that was highest

Continued on page 1356

Table 1a. Actual overinterpretation practice criteria in abstracts or full-texts of systematic reviews of diagnostic accuracy studies. *(Continued from page 1355)*

No.	Item	Criteria
a.6	Positive conclusion, inappropriately extrapolated to a wider population or setting	Actual overinterpretation if all the following: 1. Positive conclusion 2. Conclusion was extrapolated to a wider population or setting than focused on by the review (as determined by the study aim)
a.7	Positive conclusion, inappropriately extrapolated as surrogates for improvement in patient important outcomes	Actual overinterpretation if all the following: positive conclusion AND conclusion or discussion contain statements in which it is claimed that the use of the test under investigation in clinical practice will lead to improved patient important outcomes (e.g. mortality), while the review only focused on diagnostic accuracy
a.8	Stronger conclusion in abstract than full text	Actual overinterpretation if one of the following: 1. The overall degree of positivity in the conclusion (as defined in Appendix 2) was "positive" for the abstract, but "positive with qualifier" or "negative" for the full-text 2. The overall degree of positivity in the conclusion was "positive with qualifier" for the abstract, but "negative" for the full-text
a.9	Conclusion claiming test equivalence or superiority based on indirect comparisons *(only applicable for reviews that compare the accuracy of multiple index tests)*	Actual overinterpretation if the following: it was claimed in the conclusion that a test was equally accurate as, or superior to, another test, based on indirect comparisons only
a.10	Conclusion claiming test equivalence or superiority without performing statistical comparisons; or claiming test equivalence for non-statistically significant results *(only applicable for reviews that compare the accuracy of multiple index tests)*	Actual overinterpretation if one of the following: 1. It was claimed in the conclusion that a test was equally accurate as, or superior to, another test, without making statistical comparisons 2. It was claimed in the conclusion that a test was equally accurate as another test, based on a non-significant difference in diagnostic accuracy estimates *Note (1): If 95% CIs were provided for both diagnostic tests and they were mutually exclusive (no overlap), then this item was scored as "no actual overinterpretation."*

The same investigators independently applied the criteria for assessing overinterpretation to each included systematic review. Items representing actual overinterpretation were separately applied to the abstract and full-text, whereas items representing potential overinterpretation were applied to the whole article (abstract and full-text); disagreements resolved through discussion.

DATA ANALYSIS

κ Values were calculated for interassessor agreement for the overall degree of positivity of the conclusions, and for items of actual and potential overinterpretation. For each actual and potential overinterpretation practice, the number and percentage of reviews applying them was calculated. We also calculated the overall number of review abstracts or full-texts applying at least 1 actual and at least 1 potential overinterpretation practice.

Comparisons of the overall number of reviews applying at least 1 actual or at least 1 potential overinterpretation practice between subgroups based on journal impact factor (below vs above the median), type of test (imaging vs laboratory vs other), and design (comparative vs noncomparative) were performed with χ^2 test statistics.

Data extraction was done using Distiller SR (Evidence Partners). Data analysis was done using R (The R Foundation for Statistical Computing).

Results

The literature search yielded 795 unique reports, of which 112 articles were included in the study. The study selection process is detailed in online Supplemental Fig. 1, with references of included reports in the References provided in the online Supplemental Data file.

The median impact factor of publishing journals was 2.8 [interquartile ratio (IQR) 1.7–5.2]. Forty reviews (36%) assessed the accuracy of imaging tests, 50 reviews (45%) assessed laboratory tests, and 22 reviews (20%) assessed other types of tests, such as cognitive screening tests or physical exam maneuvers *(20, 21)*. Of the reviews, 36 (32%) were comparative.

κ Values for assessing the overall positivity of conclusions for abstracts and full-texts were 0.69 and 0.68, respectively. The median κ value for assessing the 10 actual overinterpretation items was 0.64 (IQR 0.58–0.66). The median κ value for assessing the 9 potential overinterpretation items was 0.69 (IQR 0.46–0.77). A

Table 1b. Potential overinterpretation practice criteria in abstracts or full-texts of systematic reviews of diagnostic accuracy studies.

No.	Item	Criteria
p.1	Intended role of index test in clinical pathway unclear	Potential overinterpretation if the following: Unclear where the test will be located among other tests in the existing clinical pathway (e.g. triage test, add-on test or replacement test; *at least one test in the existing clinical pathway and the (potential) role of the test under investigation in relation to it is sufficient*)
p.2	No or inadequate assessment of risk of bias and applicability concerns	Potential overinterpretation if one of the following: 1. No risk of bias and applicability concerns assessment was performed 2. Risk of bias and applicability concerns assessment was performed using tools that were not specifically designed for diagnostic accuracy studies (i.e. any other tool than QUADAS or QUADAS-2; e.g., The Cochrane Collaboration's tool for assessing risk of bias in randomized controlled trials) Risk of bias and applicability concerns assessment was performed using tools specifically designed for diagnostic accuracy studies (i.e. QUADAS or QUADAS-2), but results were not provided per QUADAS(-2)-question/-item/-domain, so that item a.2 could not be adequately assessed
p.3	Recommended statistical methods for metaanalysis performed	Potential overinterpretation if one of the following: 1. Traditional methods for metaanalysis were used (i.e., univariate fixed or random effects models or SROC),[a] instead of recommended ones (i.e. bivariate model or HSROC) 2. Traditional and recommended methods for metaanalysis were used, but the summary accuracy estimates on which the conclusion was based were produced by traditional methods But no potential overinterpretation if the following: Traditional methods for metaanalysis were used, but recommended ones could not be applied (e.g. because the specificity of the investigated test is generally accepted to be 100%, and only the sensitivity was analyzed); Traditional methods and recommended methods for metaanalysis were used, and the summary accuracy estimates on which the conclusion was based were produced by recommended methods
p.4	Failure to report the number of studies and patients contributing to the meta-analyses in abstract	Potential overinterpretation if one of the following: the number of studies and/or patients was lower in the metaanalysis than in the overall systematic review, but only the overall number was reported; the number of studies and patients was not reported *(number of lesions or samples instead of patients is insufficient)*
p.5	No CIs around summary accuracy estimates in abstract	Potential overinterpretation if the following: No CIs were reported in the abstract around the summary accuracy estimates on which the conclusions were based
p.6	No CIs around summary accuracy estimates in full-text	Potential overinterpretation if the following: No CIs were reported in the full-text around the summary accuracy estimates on which the conclusions were based
p.7	No statistical assessment of heterogeneity performed	Potential overinterpretation if the following: No results of a statistical assessment of heterogeneity across studies was performed (e.g. tau, prediction intervals in ROC space, I^2)
p.8	No review limitations discussed	Potential overinterpretation if the following: No review limitations were reported in the discussion section *(at least one statement that can be considered as a limitation is sufficient)*
p.9	Unclear conflict of interests	Potential overinterpretation if the following: No conflict of interest statement was provided in the study report *(a statement of funding of the study only is insufficient)*

[a] SROC, summary ROC; HSROC, hierarchical summary ROC.

Table 2. Summary of overall degree of positivity of conclusions.

Degree of conclusion positivity	Reviews, n (%)	≥1 actual overinterpretation practice n (%) [95% CI]
Abstract	n = 112	n = 112
Positive	56 (50)	55 (98) [94–100]
Positive with qualifier	27 (24)	23 (85) [72–98]
Neutral/not applicable	16 (14)	3 (19) [0–38]
Negative	13 (12)	2 (15) [0–34]
Full-text	n = 112	n = 112
Positive	28 (25)	27 (96) [89–100]
Positive with qualifier	55 (49)	46 (84) [74–94]
Neutral/not applicable	15 (14)	2 (13) [0–30]
Negative	14 (12)	2 (14) [0–32]

summary of the assessment of the overall degree of positivity of abstract and full-text conclusions is presented in Table 2.

ASSESSMENT OF OVERINTERPRETATION

Of the 112 reviews, 81 (72%) contained at least 1 actual overinterpretation practice in the abstract and 77 (69%) in the full-text. The frequency of actual overinterpretation practices with a breakdown by overall degree of positivity of conclusions can be found in Table 2 and a full per item breakdown of actual overinterpretation practices in Table 3.

The most frequent actual overinterpretation practice was a "positive conclusion, not reflecting the reported summary accuracy estimates" in 55 (49%) abstracts and 56 (50%) full-texts. A "positive conclusion, not taking high risk of bias and/or applicability concerns into account" was observed in 47 abstracts (42%) and 26 full-texts (23%), and a "positive conclusion, not taking heterogeneity into account" in 44 abstracts (39%) and 14 full-texts (12%). Thirty-two reviews (29%) had a "stronger conclusion in abstract than in full-text," such as conclusions that were judged as "positive with qualifier" or "neutral" in the full-text, but "positive" in the abstract. Examples of actual overinterpretation practices are provided in Table 4.

Of the reviews assessed, 107 (96%) contained at least one potential form of potential overinterpretation. A full per item breakdown of potential overinterpretation practices can be found in Table 3. Most common were "nonrecommended statistical methods for metaanalysis performed" (n = 57; 51%), "failure to report the number of studies and patients contributing to the metaanalyses in the abstract" (n = 54; 48%), and "intended role of index test in clinical pathway unclear" (51; 46%).

SENSITIVITY AND SUBGROUP ANALYSES

With the original criterion (a.1 in Table 1) of a threshold of 90% for summary estimates of sensitivity and specificity, and 80% for lower CIs unless stated otherwise by the author, 81 (72%) abstracts contained ≥1 actual overinterpretation practice; with a relaxed criteria of a threshold of 80% for summary estimates and 70% for lower CIs unless stated otherwise by the author, the number was 79 (71%). When we completely omitted this criterion, the number was 78 (70%). With the original criterion, 77 (69%) full-texts contained ≥1 actual overinterpretation practice; with the relaxed criterion—62 (55%), with the criterion omitted completely—53 (47%).

Overinterpretation practices by review subgroups are provided in eTable 6. There were no differences in actual overinterpretation based on impact factor median split ($P = 0.82$), type of test ($P = 0.44$), and study design ($P = 0.58$). There were no differences in potential overinterpretation based on impact factor median split ($P = 1.0$), type of test ($P = 0.10$), and study design ($P = 0.28$).

Discussion

Most recently published diagnostic accuracy systematic reviews have a positive conclusion and contain one or more forms of actual or potential overinterpretation. No significant associations with journal impact factor, type of test, or study design were identified.

We evaluated contemporary systematic reviews published in a broad spectrum of journals. In our inclusion criteria, we did not apply any restrictions on journal impact factor. Consequently, quite a number of systematic reviews published in lower impact factor journals were included. Interestingly, no differences in overinterpretation rates were observed between higher and lower im-

Table 3. Overinterpretation practices in systematic reviews of diagnostic accuracy studies.

Overinterpretation item	Abstracts scored as "yes" n (%) [95% CI]	Full-texts scored as "yes" n (%) [95% CI]	Reviews scored as "yes" n (%) [95% CI]
Total number of systematic reviews	112 (100)	112 (100)	112 (100)
a.1 Positive conclusion, not reflecting the reported summary accuracy estimates[a]	55 (49) [40–58]	56 (50) [41–59]	
a.2 Positive conclusion, not taking high risk of bias and/or applicability concerns into account	47 (42) [33–51]	26 (23) [15–31]	
a.3 Positive conclusion, not taking heterogeneity into account	44 (39) [30–48]	14 (12) [6–18]	
a.4 Positive conclusion, focusing on the results of primary studies favoring the diagnostic accuracy of the test instead of the metaanalysis results	0 (0)	0 (0)	
a.5 Positive conclusion, selectively focusing on a selection of subgroups, tests or accuracy estimates, while others were evaluated as well	12 (11) [5–17]	12 (11) [5–17]	
a.6 Positive conclusion, inappropriately extrapolated to a wider population or setting	3 (3) [0–6]	2 (2) [0–5]	
a.7 Positive conclusion, inappropriately extrapolated as surrogates for improvement in patient important outcomes	1 (1) [0–4]	2 (2) [0–5]	
a.8 Stronger conclusion in abstract than full text	32 (29) [21–37]		
a.9 Conclusion claiming test equivalence or superiority based on indirect comparisons	6 (5) [1–9]	6 (5) [1–9]	
a.10 Conclusion claiming test equivalence or superiority without performing statistical comparisons; or claiming test equivalence for non-statistically significant results	15 (13) [7–19]	16 (14) [8–20]	
p.1 Intended role of test in clinical pathway unclear			51 (46) [37–55]
p.2 No or inadequate assessment of risk of bias and applicability concerns			23 (21) [13–29]
p.3 Traditional statistical methods for metaanalysis performed			57 (51) [42–60]
p.4 Failure to report the number of studies and patients actually contributing to the meta-analyses in abstract			54 (48) [39–57]
p.5 No CIs around summary accuracy estimates in abstract			24 (21) [13–29]
p.6 No CIs around summary accuracy estimates in full text			3 (3) [0–6]
p.7 No statistical assessment of heterogeneity performed			16 (14) [8–20]
p.8 No review limitations discussed			9 (8) [3–13]
p.9 Unclear conflict of interests			14 (12) [6–18]

[a] For items a.1 to a.7 a "positive conclusion" refers to any review scored as "positive" or "positive with qualifier."

pact factor journals. Although our study was not sufficiently powered to evaluate specific "top" journals vs others, it should be noted that we did include several systematic reviews published in journals that have a high impact factor (IF) (>10, such as *Annals of Internal Medicine* IF 16.5) or are the journal with the highest impact factor in their field (*Radiology* IF 6.8, J *Nuclear Medicine* IF 6.1, *Neurology* IF 8.1). It seems that peer reviewers and editors may not be well-trained/equipped to detect and prevent overinterpretation *(22)*. Future studies could look more systematically at overinterpretation rates in high impact factor journals.

Table 4. Examples of actual overinterpretation practices.

a.1 Positive conclusion, not reflecting the reported summary accuracy estimates.
This review reports: "sensitivity was 0.59 and specificity was 0.89." Based on this, the authors have a strongly positive conclusion: "In conclusion, miR-29a may be a novel potential biomarker for CRC diagnosis." However, a sensitivity of 0.59 indicates that of every 100 diseased patients, 41 will be missed; a specificity of 0.89 indicates that of 100 every non-diseased patients, 11 will be wrongly diagnosed. Taking this into account, a positive conclusion is very optimistic.
a.2 Positive conclusion, not taking high risk of bias and/or applicability concerns into account.
The abstract of a review states: "In conclusion, the existing evidence demonstrated that methylated hTERT is effective in cancer detection." However, all 10 included primary studies were rated as having a "high" risk of bias with respect to the index test using the QUADAS-2 tool.
a.3 Positive conclusion, not taking heterogeneity into account.
The abstract of a review states ". . . statistical analysis of quantitative studies reveal the potential value of specific miRNAs in the diagnosis of AD." This conclusion does not address heterogeneity, and each of the 7 included primary studies used a different miRNA profile as an index test.
a.5 Positive conclusion, selectively focusing on a selection of subgroups, tests or accuracy estimates, while others were evaluated as well.
The study performed a global metaanalysis and evaluated 6 subgroups and concluded the abstract as "Conclusion: FeNO is accurate for the diagnosis of asthma in steroid-naive or nonsmoking patients, particularly in chronic cough patients."
a.6 Positive conclusion, inappropriately extrapolated to a wider population or setting.
The full-text conclusion reads: "Moreover, PCR detection of EBV DNA could be useful for pediatric patients who had an organ transplant because of their decreased immunity and higher risk of EBV infection." This select patient group was not analyzed in this study.
a.7 Positive conclusion, inappropriately extrapolated as surrogates for improvement in patient important outcomes.
The last sentence of the full-text conclusion reads: "We believe that the information obtained from this study will aid the decision making of physicians who take care of patients with possible M. tuberculosis infection." The objective of the study, however, was to summarize the diagnostic accuracy. Clinical decision making or "clinical utility" of a test is an outcome that is influenced by more factors than diagnostic accuracy alone, rendering it beyond the scope of this review [Bossuyt et al. *(32)*].
a.8 Stronger conclusion in abstract.
Conclusion in abstract: "The current analyses indicated that CA199 is a valuable marker in the diagnosis of pancreatic cancer."
Conclusion in full-text: "In conclusion, the present metaanalysis suggests a potential role for CA199 in screening and confirming a diagnosis of pancreatic cancer. Further well designed studies with large sample numbers should be performed to confirm the predictive value."
a.9 Conclusion claiming test equivalence or superiority based on indirect comparisons.
The review abstract states: ". . .the difference between PCNB and PNAB regarding diagnostic accuracy of benign or malignant pulmonary lesions is not obvious." However the 15 studies in the PCNB group and the 6 studies in the PNAB group were mutually exclusive. These statement of equivalence is solely based on indirect comparisons of diagnostic accuracy.
a.10 Conclusion claiming test equivalence or superiority without performing statistical comparisons; or claiming test equivalence for non-statistically significant results.
The abstract conclusion reads: "FIB-4 index with a 1.30 cutoff has better diagnostic accuracy than the FIB-4 index with a 3.25 cutoff, NFS and BARD score, despite showing its limited value for predicting NAFLD-related advanced fibrosis."
The FIB-4 index with a 1.30 cutoff is significantly more sensitive but significantly less specific than the FIB-4 index with a 3.25 cutoff. The reported AUC[a] for the 1.30 cutoff is 0.8496 (±0.0680) and the reported AUC for the 3.25 cutoff is 0.8445 (±0.0981). These values of diagnostic accuracy do not differ significantly.

[a] AUC, area under the ROC curve.

The criteria were piloted extensively, with substantial interobserver agreement *(23)*. We are aware that we propose a new classification for assessment of overall positivity in conclusions. Despite substantial interobserver agreement, this system is subjective, relying on interpretation of language. Although the overinterpretation items were extracted from previously published analyses of spin, and interobserver agreement was substantial, several items may be considered subjective, too strict, or situation dependent.

The high frequency of overinterpretation observed in this study is concordant with results for similar studies

performed on overinterpretation in primary diagnostic accuracy studies *(16, 17)*. These findings are worrisome, since systematic reviews are considered to be high-level evidence and are often used as guidance for funding decisions and to inform clinical practice guidelines.

Most items of actual overinterpretation relied on a positive conclusion regarding the accuracy or clinical usefulness of the test. This means that a systematic review with a neutral or negative conclusion, by definition, could not have overinterpreted results. Surprisingly, three fourths of the systematic reviews in our analysis had a positive conclusion, while only a small minority were clearly negative. Authors seem to have a strong tendency to provide positive statements regarding the accuracy of the evaluated test.

The most common actual overinterpretation practice was the reporting of a positive conclusion not reflecting the summary accuracy estimates, defined as a summary estimate lower than 0.90, a lower confidence limit below 0.80, or lower than predefined criteria. Positive conclusions about tests with low accuracy estimates may be inappropriate. In cases where it is appropriate to make a positive conclusion despite low accuracy estimates (i.e., the test is superior to alternate strategies), authors should explain the rationale for their conclusion.

Many reviews in our analysis had a positive conclusion but did not account for a high risk of bias or applicability concerns identified among included primary studies, or for heterogeneity. High risk of bias may lead to overestimations of test accuracy, while applicability concerns or heterogeneity may impact the applicability of the review findings *(24)*.

Many reviews had a more positive conclusion in the abstract than the full-text. Similar results have been reported for primary diagnostic accuracy studies *(16)*. This may partially stem from abstract word count limits. Reporting all the information necessary while respecting word count limits may be challenging. However, it is critical that the abstract provides a balanced summary of the results, since they are typically the most read section of biomedical manuscripts; clinicians, researchers, and policy-makers may only have access to the abstract. Previous research has shown that the interpretation of trial results by clinicians is influenced by spin in abstracts, and this could affect patient care *(18)*.

Almost all reviews had at least one 'potential overinterpretation practice, usually reflecting incomplete reporting *(25)*. Nearly half of studies failed to clearly indicate the intended role of the index test in the clinical pathway. The performance of a test may be different depending on the location of the test in the clinical pathway. A single test may be intended for use as a screening, add-on, or replacement test *(26)*. Without clear reporting of the intended role, conclusions about the applicability of review findings for specific clinical scenarios may be challenging.

Failure to perform a proper assessment of risk of bias or applicability concerns using a tool designed specifically for systematic reviews of diagnostic accuracy studies and with adequate reporting was observed in 1 of 5 reviews *(24, 27)*. These results are congruent with those of a previous study, which revealed that nearly three fourths of systematic reviews of diagnostic accuracy studies used QUADAS or QUADAS-2, tools designed for this purpose *(28)*. As bias in primary diagnostic accuracy studies is common, guidelines for performing diagnostic accuracy systematic reviews recommend assessing primary study design deficiencies *(24, 29)*.

The most common potential overinterpretation practice was performing nonrecommended methods for metaanalysis. Traditional methods for metaanalysis have been shown to overestimate summary estimates of diagnostic accuracy *(30)*.

Items regarding CIs and review sample size were reported variably. A small minority of reviews did not report CIs in the full-text, with several more not reporting them in the abstract. Nearly half did not report the number of studies and patients included for metaanalysis in the abstract. Sample size and CIs affect the precision of the identified summary estimates, and thereby confidence in the findings of a review.

Many forms of overinterpretation were linked to improper systematic review methodology and reporting issues. We advise authors who want to limit generosity in their conclusions to adhere to existing guidelines for performing and reporting reviews. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), while providing helpful guidance for systematic review authors, does not specifically focus on systematic reviews of diagnostic accuracy studies. Many aspects of a systematic review of diagnostic accuracy studies may not apply to all reviews and are thus not included in the PRISMA statement *(31)*. An extension of the PRISMA statement for diagnostic tests could alleviate some of these overinterpretation practices tied to underreporting and improper methods; these guidelines are currently in development *(25)*.

Reporting guidelines will help, but alone may not be sufficient to improve practice. Systematic reviews are an increasingly frequent form of evidence synthesis. If relevant, they are typically highly cited. Given their relative novelty, journal editors and reviewers are not always capable of identifying suboptimal practices and to identify overinterpretation. We call on them to become more demanding whenever systematic reviews are submitted, applying even higher standards when deciding to accept a report of a systematic review, compared to original accuracy studies.

Systematic reviews should provide a thorough appraisal of all the valid available evidence on a certain topic. Arriving at an objective and carefully balanced conclusion is crucial in this process. Overinterpretation of review findings, as observed in these systematic reviews, could lead to unjustified optimism about the performance of tests in clinical practice, and should therefore be prevented.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: T.A. McGrath, University of Ottawa Department of Radiology Summer Student research program; M.D.F. McInnes, Faculty Research Stipend Program.
Expert Testimony: None declared.
Patents: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, and final approval of manuscript.

Acknowledgments: We thank René Spijker, MSc (Medical Library, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands) for his comments on the search strategy.

References

1. Singh H, Graber ML. Improving diagnosis in health care–the next imperative for patient safety. N Engl J Med 2015;373:2493–5.
2. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving us adult populations. BMJ Qual Saf 2014;23:727–31.
3. Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016;94:485–514.
4. Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev 2007;(2):MR000005.
5. Schmucker C, Schell LK, Portalupi S, Oeller P, Cabrera L, Bassler D, et al. Extent of non-publication in cohorts of studies approved by research ethics committees or included in trial registries. PLoS One 2014;9:e114023.
6. Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Med Res Methodol 2009;9:79.
7. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA 2010;303:2058–64.
8. Latronico N, Metelli M, Turin M, Piva S, Rasulo FA, Minelli C. Quality of reporting of randomized controlled trials published in Intensive Care Medicine from 2001 to 2010. Intensive Care Med 2013;39:1386–95.
9. Le Fourn E, Giraudeau B, Chosidow O, Doutre MS, Lorette G. Study design and quality of reporting of randomized controlled trials of chronic idiopathic or autoimmune urticaria: review. PLoS One 2013;8:e70717.
10. Patel SV, Chadi SA, Choi J, Colquhoun PH. The use of "spin" in laparoscopic lower GI surgical trials with nonsignificant results: an assessment of reporting and interpretation of the primary outcomes. Dis Colon Rectum 2013;56:1388–94.
11. Patel SV, Van Koughnett JA, Howe B, Wexner SD. Spin is common in studies assessing robotic colorectal surgery: an assessment of reporting and interpretation of study results. Dis Colon Rectum 2015;58:878–84.
12. Arunachalam L, Hunter IA, Killeen S. Reporting of randomized controlled trials with statistically nonsignificant primary outcomes published in high-impact surgical journals. Ann Surg 2017;265:1141–5.
13. Gewandter JS, McKeown A, McDermott MP, Dworkin JD, Smith SM, Gross RA, et al. Data interpretation in analgesic clinical trials with statistically nonsignificant primary analyses: an ACTTION systematic review. J Pain 2015;16:3–10.
14. Lockyer S, Hodgson R, Dumville JC, Cullum N. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically nonsignificant primary outcome results or unspecified primary outcomes. Trials 2013;14:371.
15. Lazarus C, Haneef R, Ravaud P, Boutron I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. BMC Med Res Methodol 2015;15:85.
16. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". Radiology 2013;267:581–8.
17. Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JP, Hernández-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem 2009;55:786–94.
18. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIN randomized controlled trial. J Clin Oncol 2014;32:4120–6.
19. Yavchitz A, Ravaud P, Altman DG, Moher D, Hrobjartsson A, Lasserson T, Boutron I. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. J Clin Epidemiol 2016;75:56–65.
20. McGovern A, Pendlebury ST, Mishra NK, Fan Y, Quinn TJ. Test accuracy of informant-based cognitive screening tests for diagnosis of dementia and multidomain cognitive impairment in stroke. Stroke 2016;47:329–35.
21. Huang W, Zhang Y, Yao Z, Ma L. Clinical examination of anterior cruciate ligament rupture: a systematic review and meta-analysis. Acta Orthop Traumatol Turc 2016;50:22–31.
22. Lazarus C, Haneef R, Ravaud P, Hopewell S, Altman DG, Boutron I. Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. J Clin Epidemiol 2016;77:44–51.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
24. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.
25. McInnes M, Moher D, Bossuyt P. Development and implementation of a reporting guideline for systematic reviews and meta-analyses of diagnostic accuracy studies: the PRISMA-DTA initiative. http://www.equator-network.org/wp-content/uploads/2009/02/PRISMA-DTA-Executive-Summary.pdf (Accessed March 2017).
26. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013;185:E537–44.
27. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.
28. Ochodo EA, van Enst WA, Naaktgeboren CA, de Groot JA, Hooft L, Moons KG, et al. Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy reviews: a cross-sectional study. BMC Med Res Methodol 2014;14:33.
29. Deeks J, Bossuyt P, Gatsonis C. Cochrane handbook for systematic reviews of diagnostic test accuracy. 1.0.0 ed: The Cochrane Collaboration, 2013.
30. McGrath TA, McInnes MD, Korevaar DA, Bossuyt PM. Meta-analyses of diagnostic accuracy in imaging journals: analysis of pooling techniques and their effect on summary estimates of diagnostic accuracy. Radiology 2016;281:78–85.
31. Group P. The PRISMA statement. http://www.prisma-statement.org/ (Accessed November 2014).
32. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem 2012;58:1636–43.