Journal of Clinical Epidemiology 116 (2019) 9−17

Journal of Clinical Epidemiology

REVIEW

A systematic review finds that spin or interpretation bias is abundant in evaluations of ovarian cancer biomarkers

Mona Ghannad[a,b,*], Maria Olsen[a,b], Isabelle Boutron[b], Patrick M. Bossuyt[a]

[a]*Amsterdam UMC, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, Amsterdam Public Health Research Institute, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands*

[b]*Université de Paris, CRESS, INSERM, INRA, F-75004 Paris, France*

Accepted 15 July 2019; Published online 19 July 2019

Abstract

Background: In the scientific literature, ‘‘spin’’ refers to reporting practices that make the study findings appear more favorable than results justify. The practice of ‘‘spin’’ or misrepresentation and overinterpretation may lead to an imbalanced and unjustified optimism in the interpretation of study results about performance of putative biomarkers. We aimed to classify spin (i.e., misrepresentation and overinterpretation of study findings) in recent clinical studies evaluating the performance of biomarkers in ovarian cancer.

Methods: We searched PubMed systematically for all evaluations of ovarian cancer biomarkers published in 2015. Studies eligible for inclusion reported the clinical performance of prognostic, predictive, or diagnostic biomarkers.

Results: Our search identified 1,026 studies; 326 studies met all eligibility criteria, of which we evaluated the first 200 studies. Of these, 140 (70%) contained one or more form of spin in the title, abstract, or main-text conclusion, exaggerating the performance of the biomarker. The most frequent forms of spin identified were (1) other purposes of biomarker claimed not investigated (65; 32.5%); (2) mismatch between intended aim and conclusion (57; 28.5%); and (3) incorrect presentation of results (40; 20%).

Conclusion: Our study provides evidence of misrepresentation and overinterpretation of finding in recent clinical evaluations of ovarian cancer biomarkers.

Keywords: Misrepresentation; Misinterpretation; Bias; Spin; Biomarker; Ovarian cancer

1. Introduction

Research in cancer biomarkers has expanded in recent years leading to growing and large literature. However, despite major investments and advances in technology, the current biomarker pipeline is found to be too prone to failures [1,2]. Similarly, much research has been dedicated to the discovery of ovarian cancer biomarkers. However, despite many biomarkers being evaluated, very few have been successfully introduced in clinical care [3]. Likely reasons for failure have been documented at each of the stages of biomarker evaluation [1−3].

It has been argued that biomarker discovery studies sometimes suffer from weak study designs, limited sample size, and incomplete or biased reporting, which can render them vulnerable to exaggerated interpretation of biomarker performance [1,4]. Authors may claim favorable performance and clinical effectiveness of biomarkers based on selective reporting of significant findings, or present study results with an overly positive conclusion in the abstract compared with the main text [5]. Specific study features could facilitate distorted study results, such as not prespecifying a biomarker threshold or lacking a specific study objective.

Spin, or misrepresentation and misinterpretation of study findings, not necessarily intentional, is any reporting practice that makes the study findings appear more favorable than the results justify [6,7]. Several studies

Conflict of interest: None.

This work was supported by the European Union’s Horizon 2020, European Union research and innovation program under the Marie Sklodowska-Curie grant agreement no. 676207.

The funder/sponsor had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Previous presentation of the manuscript: This study was presented at the International Congress on Peer Review and Scientific Publication 2017, Evidence Live 2018, and the Cochrane Colloquium 2018 meeting.

* Corresponding author. Tel.: +31625583959; fax: +31206912683.

E-mail address: m.ghannad@amc.uva.nl (M. Ghannad).

https://doi.org/10.1016/j.jclinepi.2019.07.011

What is new?

Key findings

- Much research has been dedicated to the discovery of ovarian cancer biomarkers, but few are successfully introduced in clinical care. A number of factors, such as poor study design and bias, have been attributed to the lack of success in identifying clinically relevant biomarkers.
- In this study, we investigate biased reporting and interpretation in published articles as a potential contributing factor, which has not previously been characterized in ovarian cancer biomarkers. The practice of frequent misrepresentation or overinterpretation of study findings may lead to an imbalanced and unjustified optimism in the interpretation of study results about performance of putative biomarkers.

What this adds to what was known?

- We performed a systematic review of 200 recent evaluations of ovarian cancer biomarkers and observed that 70% had at least one form of spin (i.e., misrepresentation or overinterpretation of study findings) in the title, abstract, or main-text conclusion, exaggerating the performance of the biomarker.

What is the implication and what should change now?

- This review indicates that biased reporting and interpretation is prevalent in recent clinical evaluations of biomarkers in ovarian cancer. These results indicate a need for strategies to minimize biased reporting and interpretation.

have shown that authors of clinical studies may commonly present and interpret their research findings with a form of spin [5,7–10]. A consequence of biased representation of results in scientific reports is that the published literature may suggest stronger evidence than is justified [11]. Misrepresentation of study findings may also lead to serious implications for patients, health care providers, and policy makers [12].

The primary aim of our study was to evaluate the presence of spin, further categorized as misrepresentation and overinterpretation of study findings, in recent clinical studies evaluating the performance of biomarkers in ovarian cancer. In addition, we also evaluated facilitators of spin (i.e., practices that would facilitate overinterpretation of results), as well as a number of potential determinants of spin.

2. Methods

We performed a systematic review to document the prevalence of spin in recent evaluations of the clinical performance of biomarkers in ovarian cancer.

2.1. Literature search

MEDLINE was searched through PubMed on December 22, 2016, for all studies evaluating the performance of biomarkers in ovarian cancer published in 2015. The search terms and strategy were developed in collaboration with a medical information specialist (R.S.), using a combination of terms that express the clinical performance of biomarkers in ovarian cancer (Appendix A). We included all markers of ovarian cancer risk, screening, prognosis, or treatment response in body fluid, tissue, or imaging measurements. Reviews, animal studies, and cell line studies were excluded.

Two authors (M.G. and M.O.) independently reviewed the titles and abstracts to identify potentially eligible articles. Thereafter, full-texts of reports identified as potentially eligible were independently reviewed by the same two authors for inclusion. All disagreements were resolved through discussion or by third party arbitration (P.M.B.). We analyzed the first 200 consecutive studies, ranked according to publication date, to have a sample size comparable with previous systematic reviews of spin [6,13].

2.2. Establishing criteria and data extraction

Biomarker studies in ovarian cancer vary by study design, biomarker clinical application, type and number of tests evaluated [14,15]. Within the evaluation process, several components can be assessed, such as analytical performance, clinical performance, clinical effectiveness, cost-effectiveness, and all other consequences beyond clinical effectiveness and cost-effectiveness. We developed a definition of spin that encompassed common features applicable to all the various biomarker types and study designs. We defined spin as reporting practices that make the clinical performance of markers look more favorable than results justify. This definition of spin was based on criteria extracted from key articles on misrepresentation and misinterpretation of study findings [5–7,9,10,13,16,17].

To evaluate the frequency of spin, we established a preliminary list incorporating previously established items that represent spin as well [5–7,16,17]. We then established a preliminary list of criteria to evaluate the frequency of spin and optimized our criteria through a gradual data extraction process. A set of 20 articles were fully verified by a second reviewer (M.O.), and points of disagreements were discussed with a third investigator (I.B. and P.M.B.) to fine-tune the scoring criteria and clarify the coding scheme. Through this process and discussions that ensued, a final list of items was established with content experts (P.M.B. and I.B.), categorizing items as representing "spin" or "facilitator of spin." Each of the categories encompassed several forms of spin.

We further classified spin into two categories: "misrepresentation" and "misinterpretation", to distinguish between distorted presentation and incorrect interpretation of findings with special focus on the abstract and main-text conclusions. As the presence of a positive conclusion is interdependent with the items that represent spin, we assessed the overall positivity of the main-text conclusion by using a previously established classification scheme [13]. The overall positivity was classified according to the summary statement in the main-text conclusion about the biomarker's analytical performance or clinical utility. We used the same criteria defined by McGrath et al. [13] to assess the main-text conclusion as "positive", "positive with qualifier", "neutral", or "negative". A qualifier attenuates the summary statement or its implication for practice [13]. Examples include but are not limited to the use of conjunctions such as "may" in the summary statement or statements such as "limited evidence is available" in the same paragraph as the summary statement.

We defined misrepresentation as misreporting and/or distorted presentation of the study results in the title, abstract, or the main text, in a way that could mislead the reader. This category of spin encompassed (1) incorrect presentation of results in the abstract or main-text conclusion, (2) mismatch between results reported in abstract and main text, and (3) mismatch between results reported and the title.

We defined misinterpretation as an interpretation of the study results in the abstract or main-text conclusion that is not consistent and/or is an extrapolation of the actual study results. This category of spin encompassed (4) other purposes of biomarker claimed not prespecified and/or investigated, (5) mismatch between intended aim and abstract or main-text conclusion, (6) other benefits of biomarkers claimed not prespecified and/or investigated, and (7) extrapolation from study participants to a larger or a different population.

We defined "facilitators of spin" as practices that facilitate spin that, but due to various elements, do not allow for a formal assessment and classification as actual spin. For example, in our study, we considered not prespecifying a positivity threshold for continuous biomarker as a facilitator of spin. Stating a threshold value after data collection and analysis may leave room in the representation and interpretation of the data to maximize performance characteristics [6].

In addition to spin and facilitators of spin, we extracted the following information on study characteristics: country, biomarker intended use, author affiliations, conflict disclosures declared, and source of funding. To evaluate which of the factors we identified may be associated with the manifestation of spin, we counted the occurrence of spin corresponding to each of the determinants.

Actual forms of spin, facilitators of spin, and potential determinants of spin were recorded in all studies reporting the performance of the discovered biomarker. Items were scored independently by the first reviewer (M.G.), and all uncertainties were resolved in discussions with a second reviewer (P.M.B. and M.O.).

2.3. Analysis

For each of the items on spin, facilitators of spin, and potential determinants of spin, we report the frequency in our sample of biomarker evaluations, with 95% confidence intervals.

3. Results

3.1. Search results

Our search identified 1,026 citations in PubMed. After title and abstract screening, 516 citations were selected for full-text evaluation. Of these, 326 studies met all eligibility criteria, and the first 200 studies, ranked according to publication date, were included in our analysis (Fig. 1).

3.2. Characteristics of included studies

A description of included studies is presented in Table 1. The studies originated from a total of 32 countries, with the majority of the studies coming from China ($n = 69$, 34.5%) and USA ($n = 41$, 20.5%). The remaining 30 countries had a distribution range of 1 to 14 articles per country. The studies were published in 94 journals in total (Appendix B).

Of all the studies evaluated in the included articles, prognostic ($n = 89$, 44.5%) and diagnostic ($n = 40$, 20%) markers comprised the largest group. Authors of almost all included studies had an affiliation with a clinical department ($n = 194$, 97%), but only 34 of these (17.5%) had one or more authors affiliated with a statistical or bioinformatics department.

Nearly all the included studies ($n = 193$, 96.5%) reported a positive conclusion in the main text, with only seven studies (3.5%) reporting a negative or neutral

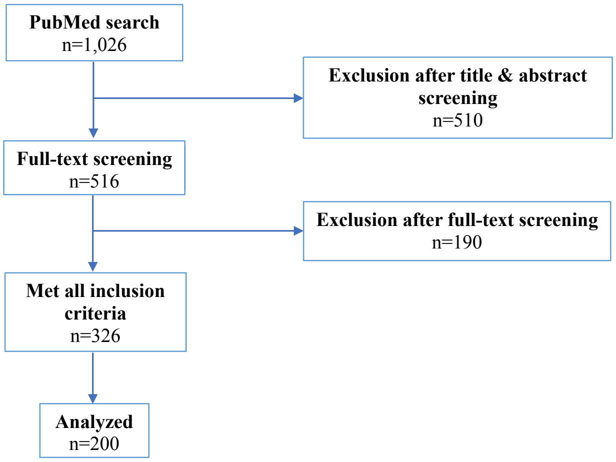
PubMed search
n=1,026
Exclusion after title & abstract screening
n=510
Full-text screening
n=516
Exclusion after full-text screening
n=190
Met all inclusion criteria
n=326
Analyzed
n=200

Fig. 1. Flow chart of search results.

Table 1. Study Characteristics

Characteristic	No. (%) (all studies *n* = 200)
Number of journals	94
Origin	
Asia	101 (51%)
North America	51 (26%)
Europe	39 (20%)
Other (Australia, Brazil, Chile)	9 (5%)
Biomarker clinical application	
Prognosis	89 (45%)
Diagnosis	40 (20%)
Prediction of therapeutic response	26 (13%)
Risk susceptibility, monitoring, screening	17 (9%)
Multiple	28 (14%)
Author affiliations	
Clinical department only	194 (97%)
Clinical **and either** statistical department **or** bioinformatics/ computational biology (*affiliation with statistical department or bioinformatics/computational biology are **not** mutually exclusive)	34 (17%)
Positivity of conclusions	
Positive	113 (57%)
Positive with qualifier	80 (40%)
Negative	5 (3%)
Neutral	2 (1%)
Conflict disclosure	
No	151 (76%)
Not reported	38 (19%)
Yes	11 (6%)
Funding source	
Nonprofit	135 (68%)
Not reported	53 (27%)
No funding	6 (3%)
For profit	4 (2%)
Mix (for profit and nonprofit)	2 (1%)

conclusion. Of the 193 studies with a positive conclusion, 80 studies had a qualifier, stating a positive summary statement with a qualifier, for example, with a conjunction such as ''may'', and thereby attenuating the statement. Eleven studies (5.5%) declared a conflict of interest, 38 (19%) did not report if they had a conflict of interest. The funding source was mainly nonprofit (*n* = 135, 67.5%). However, 53 of the included studies (27%) did not report source of funding.

3.3. Actual forms of spin

In our 200 analyzed studies, 140 (70%) contained one or more forms of spin; 75 had two or more forms of spin. Sixty studies (30%) had no form of spin in the article, based on our criteria. Table 2 lists the prevalence for each form of spin (i.e., misrepresentation or misinterpretation) from the articles in our set, with examples presented in Appendix D.

We identified incorrect presentation of results in abstract or main-text conclusion in 40 study reports (20%). We observed this more frequently in the main-text conclusion (*n* = 37, 18.5%) than in the abstract conclusion (*n* = 14, 7%). These were reports in which a positive conclusion was made about the biomarker that was not supported by the study results, or not accompanied by a test for statistical significance or an appropriate expression of precision, such as 95% confidence intervals. Examples were a study that claimed a multivariable algorithm had been validated, despite poor results (the study presents positive results on biomarkers, but these were not included in the algorithm), and a study that claimed a ''high specificity'', while the corresponding estimate was only 58% [18,19].

Several studies claimed superiority in performance in the absence of tests for statistical significance [20,21]. In 33 study reports (16.5%), there was a mismatch in results reported in the abstract and the main text. The most frequent examples were studies that selectively reported findings in the abstract, including only the most positive or statistically significant results in the study abstract. In few studies, we observed a mismatch between results reported in abstract and results reported in the main text. In 11 articles (5.5%), we observed a mismatch in the title.

Apart from these forms of misrepresentation of study findings, we also looked at forms of misinterpretation. In 65 study reports (32.5%), biomarker purposes were suggested, which had not been investigated in the actual study. We also observed this more frequently in the main-text conclusion (*n* = 60, 30%) than in the abstract conclusion (*n* = 36, 20.5%). An example was a study that claimed in the conclusion of the abstract that a biomarker ''showed strong promise as a diagnostic tool for large-scale screening'', although the marker had only been evaluated in a diagnostic setting, with symptomatic patients [22].

In addition, we identified a mismatch between the intended aim of the biomarker and one of the conclusions of the study report in 57 cases (28.5%). This form of misinterpretation was also more frequently observed in the abstract section (*n* = 41, 20.5%) than the main text section (*n* = 31, 15.5%). A typical example was a claim about clinical usefulness in a study where the report only included an expression of performance in a nonclinical setting, discriminating between cases and noncases, based on the biomarker [22]. In 10 studies (5%), biomarker benefits were claimed that had not been evaluated, such as a reduction in health care costs. In 10 articles (5%), there was an unsupported extrapolation from the study group to a different population. An example was a study that concluded that a spectroscopy technique was useful for the early detection of biomarker, although the study had only evaluated patients undergoing surgery [23].

Table 2. Actual forms of spin in clinical studies evaluating performance of biomarkers in ovarian cancer

Category of spin	Form of spin	Criteria	Spin frequency, *n* = 200; *n* (% [95% CI])
Misrepresentation	1. Incorrect presentation of results in the abstract or main-text conclusion	Abstract conclusion OR main-text conclusion for BM's clinical performance is not in accordance with or is stronger than results justify. Actual spin if the following: a. Exaggerating the performance of the BM in the conclusion despite low-performance measures reported in the results b. Claiming effect of the BM despite statistically nonsignificant results c. Claiming effect despite not providing imprecision or statistical test (confidence interval or *P*-values) between different biomarker models tested or patient groups (subgroups)	**Total: 40 (20% [15–26%])** Frequency in abstract conclusion: 14 (7% [4–12%]) Frequency in main-text conclusion: 37 (18.5% [14–25%])
	2. Mismatch between results reported in abstract and main text	Results reported in the abstract are not in accordance with results reported in main text. Actual spin if the following: a. Results reported in the abstract contains statement in which statistical significance is claimed, despite not providing imprecision or test of significant (CI or *P*-values) in results reported in the main text b. Selective reporting of statistically significant outcomes in the abstract compared to the results reported in the main text c. Results reported in the abstract that do not match results provided in the main text	**33 (16.5% [12–23%])**
	3. Mismatch between results reported and the title	The title contains wording misrepresenting BM's clinical performance compared with results in the main text.	**11 (5.5% [3–10%])**
Misinterpretation	4. Other purposes of biomarker claimed not prespecified and/or investigated	Abstract conclusion OR main text conclusion contains statement suggesting BM purposes not prespecified and/or investigated.	**Total: 65 (32.5% [26–40%])** Frequency in abstract conclusion: 36 (20.5% [13–24%]) Frequency in main text conclusion: 60 (30% [24–37%])
	5. Mismatch between intended aim and abstract or main-text conclusion	Abstract conclusion OR main-text conclusion for BM's clinical performance is stronger than study design. Actual spin if the following: a. The main-text conclusion contains statement in which BM utility is claimed despite not evaluating clinical effectiveness (i.e., useful) b. The main-text conclusion contains statement in which BM performance improvement is claimed despite not evaluating incremental measures (i.e., improve)	**Total: 57 (28.5% [23–35%])** Frequency in abstract conclusion: 41 (20.5% [15–27%]) Frequency in main-text conclusion: 31 (15.5% [11–21%])

(*Continued*)

Table 2. Continued

Category of spin	Form of spin	Criteria	Spin frequency, $n = 200$; n (% [95% CI])
		c. The main-text conclusion contains statement that uses causal language for BM(s) being assessed despite the use of a nonrandomized design	
	6. Other benefits of BM claimed not prespecified and/or investigated	The main-text conclusion contains statement claiming BM benefits not prespecified and/or investigated.	**10 (5% [3–9%])**
	7. Extrapolation from study participants to a larger or a different population	The main-text conclusion contains statement that extrapolates BM's clinical performance to a larger or a different population, not supported by recruited subjects.	**10 (5% [3–9%])**

Abbreviations: BM, biomarker; HR, hazard ratio; OS, overall survival; PFS, progression-free survival.
Note: We evaluated all results reported in the abstract and main text of the article, excluding supplementary material.

3.4. Facilitators of spin

Details of our analysis of potential facilitators of spin are presented in Table 3. Of the 200 analyzed studies, none reported a sample size justification or any potential harms. Only half of the studies prespecified a positivity threshold for the continuous biomarker evaluated.

3.5. Potential determinants of spin

We investigated potential determinants of spin in the 200 articles in our data set (Table 4). Articles from China (75%) and Japan (86%) were more frequently observed to have spin (Appendix C). Diagnostic accuracy studies (80%) and articles that reported multiple clinical utility of the biomarker (79%) were more often associated with spin. Studies that reported affiliations with a statistical or bioinformatics department (59%) were less likely to have spin in the report than studies that did not report an affiliation with a statistical or bioinformatics department (73%). Studies that failed to report whether there was a conflict of interest (82%) more often had spin than studies that declared no conflict of interest (67%).

4. Discussion

Our review systematically documented spin in recent clinical studies evaluating performance of biomarkers in ovarian cancer. We identified spin in the title, abstract, result, and conclusion of the main text. Of the 200 studies we evaluated, all but seven reported a positive conclusion about the performance of the biomarker. We found that only one-third of these 200 reports were free of spin, one-third contained one form of spin, and another third contained two or more forms of spin.

The most frequent form of spin was claiming other purposes for the biomarker, outside of the study aim and not investigated, adding that the biomarker could be used for other clinical purposes that were not investigated. The second most frequent form of spin we identified was a mismatch between intended aim and study conclusions, concluding on the biomarker's clinical usefulness, for example, despite the fact that the study had only evaluated classification in a nonclinical setting. These two forms of misinterpretation were more prevalent in the abstract conclusion than the main-text conclusion. The third most frequent form of spin was incorrect presentation of results in the conclusion, with some authors reporting an unjustified positive conclusion about the biomarker's performance, using terms such as ''significantly associated'' or ''highly specific'' without providing the test of significance or lacking support by the study results. This form of misrepresentation was more prevalent in the main-text than in the abstract conclusion.

In terms of facilitators of spin, we observed that none of the studies reported a justification for the sample size or discussed any potential harms, and most of the articles did not prespecify a positivity threshold for continuous biomarkers.

Table 3. Facilitators of spin in clinical studies evaluating performance of biomarkers in ovarian cancer

Potential facilitators of spin	Spin frequency, $n = 200$; n (% [95% CI])
Not stating sample size calculations	200 (100% [98–100%])
Not mentioning potential harms	200 (100% [98–100%])
Not prespecifying a positivity threshold for continuous biomarker	84/164[a] (51.2% [43–59%])
Incomplete or not reporting imprecision or statistical test for data shown	26 (13% [9–19%])
Study objective not reported or unclear	24 (12% [8–18%])

[a] 164 articles included evaluation of continuous biomarkers.

Table 4. Potential determinants of spin

Determinant	No. of articles with determinant	Number of articles with determinant and occurrence of spin			Number of articles with determinant and overall occurrence of spin; *n* (% [95% CI])
		1 occurrence of spin	2 occurrences of spin	>2 occurrences of spin	
Origin					
Asia (including Turkey and Israel)	101	39	20	19	78 (77% [68–85%])
North America	51	13	13	6	32 (63% [48–76%])
Europe	39	9	9	3	21 (54% [37–70%])
Other (Australia, Brazil, Chile)	9	4	3	2	9 (100% [63–100%])
Biomarker clinical application					
Prognosis	89	36	14	9	59 (66% [56–76%])
Diagnosis	40	4	14	14	32 (80% [64–90%])
Prediction of therapeutic response	26	11	4	3	18 (69% [48–85%])
Risk susceptibility, monitoring, screening	17	4	5	0	9 (53% [29–76%])
Multiple	28	10	8	4	22 (79% [59–91%])
Affiliation between clinical department and statistical or bioinformatics department					
No	160	53	37	27	117 (73% [65–80%])
Yes	34	12	6	2	20 (59% [41–75%])
Conflict of interest					
No	151	51	29	21	101 (67% [59–74%])
Not reported	38	12	12	7	31 (82% [65–92%])
Yes	11	2	4	2	8 (73% [39–93%])
Funding source					
Nonprofit	135	46	25	21	92 (68% [60–76%])
Not reported	53	18	13	7	38 (72% [57–83%])
No funding	6	0	3	1	4 (67% [24–94%])
For profit	4	0	3	1	4 (100% [40–100%])
Mix (nonprofit and for profit)	2	1	1	0	2 (100% [20–100%])

Our study had several strengths. A particular feature of our work was that we comprehensively included all markers of ovarian cancer risk, screening, prognosis, or treatment response in body fluid, tissue, or imaging measurements. To evaluate spin in a wide variety of biomarkers and study designs, we optimized our definition of spin in terms of common features that apply to most biomarker studies. We also used a definition of spin that is very broad and encompasses all forms of spin ranging from misreporting, misrepresentation, to linguistic spin, while developing a classification scheme that aims to limits subjectivity.

We acknowledge potential limitations of this study. In our analysis, we focused on mismatches between results presented in the main text and conclusions made in the study abstract or the main text. This definition does not include other forms of generous presentation or interpretation. We did not include specific deficiencies in study design and conduct, data collection, statistical analysis and phrasing of statistical results, or the total body of knowledge about the biomarker to check validity of conclusions made. There may have been other limitations in the study design or conduct that would warrant caution in the conclusions but were not identified by us. Several of the studies had multiple elements, also encompassing a preclinical phase of evaluations. We did not evaluate statements related to the preclinical elements. Similarly, the actual clinical application was not included in our evaluation. For example, a study may claim predictive use of an evaluated biomarker, but the strength of the association may be so limited that the biomarker will not be of value in clinical practice.

Although some of the forms of spin in our analysis could be objectively demonstrated, like a mismatch between results in the main body of the article and results in the abstract, others relied more on interpretation. As in other evaluations of spin, we have tried to minimize the subjectivity of these classifications by having a stepwise development process of the criteria, including multiple reviewers and explicit discussions of scoring results.

Previous studies have documented a high prevalence of spin in published reports of randomized controlled trials, nonrandomized studies, diagnostic test accuracy studies, and systematic reviews [5−8,10,13,16,17,24]. The reasons behind biased and incomplete reporting are probably multifaceted and complex. Yavchitz et al. discussed that (1) lack of awareness of scientific standards, (2) naïveté and impressionability of junior researchers, (3) unconscious bias, or (4) in some instances willful intent to positively influence readers may all be factors giving rise to spin in published literature [8]. The reward system currently used in biomedical science can also be held responsible, as it focuses greatly on quantity of publications rather than quality [7].

It has previously been shown that spin in articles may indeed hinder the ability of readers to confidently appraise results. Boutron et al [10] evaluated the impact of spin in the abstract section of articles reporting results in the field of cancer. The studies selected were randomized control trials in cancer with statistically nonsignificant primary outcomes. Boutron observed that clinicians rated the experimental treatment as being more beneficial for abstracts with spin in the conclusion. Scientific articles with spin were also more frequently misrepresented in press releases and news [25].

To detect and limit spin and thus minimize biased and exaggerated reporting of clinical studies, we need to better understand drivers and strategies of spin. Efforts to prevent or reduce biased and incomplete reporting in biomedical research should be undertaken with vigor and in unison, given the intricate complexities that involve multiple players. Researchers and authors, peer reviewers, and journal editors unboundedly share responsibility. The role of institutions and senior researchers is integral in disseminating research integrity and best research practices. Existing educational programs for early career researchers can be enriched by implementing mentoring and training initiatives, making authors aware of forms and facilitators of spin and its impact. Another strategy to consider may be assembling diverse and multidisciplinary teams, including statisticians, to help ensure the rigorous and robust conduct of research methodology. In our review, studies that reported affiliations with statistical departments for at least one author less often had spin.

Despite emerging evidence that use of reporting guidelines is associated with more complete reporting [26], journal editors do not explicitly recommend the use of reporting guidelines in the review process [27]. In synergy with improving completeness of reporting, guidelines may also help reduce spin, although they are unlikely to fully eliminate it. Example of items in currently existing reporting guidelines that may help reduce spin include item 19 in the REMARK guideline for prognostic studies recommending authors to “interpret the results in the context of the prespecified hypothesis and other relevant studies” in their discussion [28], and item 4 in the STARD guideline for diagnostic accuracy studies recommending authors to “specify the objective and hypothesis” in their introduction [29]. Expanding currently existing reporting guidelines with items that prompt reviewers to check for manifestation of spin and evaluating the feasibility of the guidelines to limit spin may provide incentives for editors to prompt evidence-based change in practice for the review process.

The development of biomarkers holds great promise for early detection, diagnosis, and treatment of patients with cancer. Yet that promise can only be fulfilled with strong evaluations of the performance of putative markers, complete reporting of the study design and conduct, and a fair and balanced interpretation of study findings. This review of spin in recent evaluations of biomarker performance shows that there is room for improvement.

CRediT authorship contribution statement

Mona Ghannad: Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Maria Olsen:** Methodology, Investigation, Writing - review & editing. **Isabelle Boutron:** Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision. **Patrick M. Bossuyt:** Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Supervision.

Acknowledgments

The authors thank Rene Spijker for developing the search strategy and Simon Boerstra for providing help with data analysis.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.07.011.

References

[1] Ioannidis JPA, Bossuyt PMM. Waste, leaks, and failures in the biomarker pipeline. Clin Chem 2017;63:963−72.

[2] Ioannidis JP. Biomarker failures. Clin Chem 2013;59:202−4.

[3] Diamandis EP. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? BMC Med 2012;10:87.

[4] Pepe MS, Feng ZD. Improving biomarker identification with better designs and reporting. Clin Chem 2011;57:1093−5.

[5] 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(20):2058−64.

[6] 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.

[7] Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A 2018;115:2613−9.

[8] Yavchitz A, Ravaud P, Altman DG, Moher D, Hrobjartsson A, Lasserson T, et al. 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.

[9] 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.

[10] 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 SPIIN randomized controlled trial. J Clin Oncol 2014;32:4120–6.

[11] Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav 2017;1:0021.

[12] Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383:101–4.

[13] McGrath TA, McInnes MDF, van Es N, Leeflang MMG, Korevaar DA, Bossuyt PMM. Overinterpretation of research findings: evidence of ‘‘spin’’ in systematic reviews of diagnostic accuracy studies. Clin Chem 2017;63:1353–62.

[14] Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta 2014;427:49–57.

[15] Pavlou MP, Diamandis EP, Blasutig IM. The long journey of cancer biomarkers from the bench to the clinic. Clin Chem 2013;59:147–57.

[16] Chiu K, Grundy Q, Bero L. ‘Spin’ in published biomedical literature: a methodological systematic review. PLoS Biol 2017;15(9): e2002173.

[17] 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.

[18] Dorn J, Bronger H, Kates R, Slotta-Huspenina J, Schmalfeldt B, Kiechle M, et al. OVSCORE - a validated score to identify ovarian cancer patients not suitable for primary surgery. Oncol Lett 2015; 9(1):418–24.

[19] Masoumi-Moghaddam S, Amini A, Wei AQ, Robertson G, Morris DL. Sprouty 2 protein, but not Sprouty 4, is an independent prognostic biomarker for human epithelial ovarian cancer. Int J Cancer 2015;137:560–70.

[20] Wilailak S, Chan KK, Chen CA, Nam JH, Ochiai K, Aw TC, et al. Distinguishing benign from malignant pelvic mass utilizing an algorithm with HE4, menopausal status, and ultrasound findings. J Gynecol Oncol 2015;26(1):46–53.

[21] Fujiwara H, Suzuki M, Takeshima N, Takizawa K, Kimura E, Nakanishi T, et al. Evaluation of human epididymis protein 4 (HE4) and risk of ovarian malignancy algorithm (ROMA) as diagnostic tools of type I and type II epithelial ovarian cancer in Japanese women. Tumour Biol 2015;36(2):1045–53.

[22] Shadfan BH, Simmons AR, Simmons GW, Ho A, Wong J, Lu KH, et al. A multiplexable, microfluidic platform for the rapid quantitation of a biomarker panel for early ovarian cancer detection at the point-of-care. Cancer Prev Res (Phila) 2015;8(1):37–48.

[23] Lima KM, Gajjar KB, Martin-Hirsch PL, Martin FL. Segregation of ovarian cancer stage exploiting spectral biomarkers derived from blood plasma or serum analysis: ATR-FTIR spectroscopy coupled with variable selection methods. Biotechnol Prog 2015;31(3):832–9.

[24] Lockyer S, Hodgson R, Dumville JC, Cullum N. ‘‘Spin’’ in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. Trials 2013;14:371.

[25] Haneef R, Yavchitz A, Ravaud P, Baron G, Oransky I, Schwitzer G, et al. Interpretation of health news items reported with or without spin: protocol for a prospective meta-analysis of 16 randomised controlled trials. BMJ Open 2017;7(11):e017425.

[26] Cobo E, Cortes J, Ribera JM, Cardellach F, Selva-O’Callaghan A, Kostov B, et al. Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. BMJ 2011;343:d6783.

[27] Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. PLoS One 2012;7:e35621.

[28] McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer 2005;93:387–91.

[29] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.