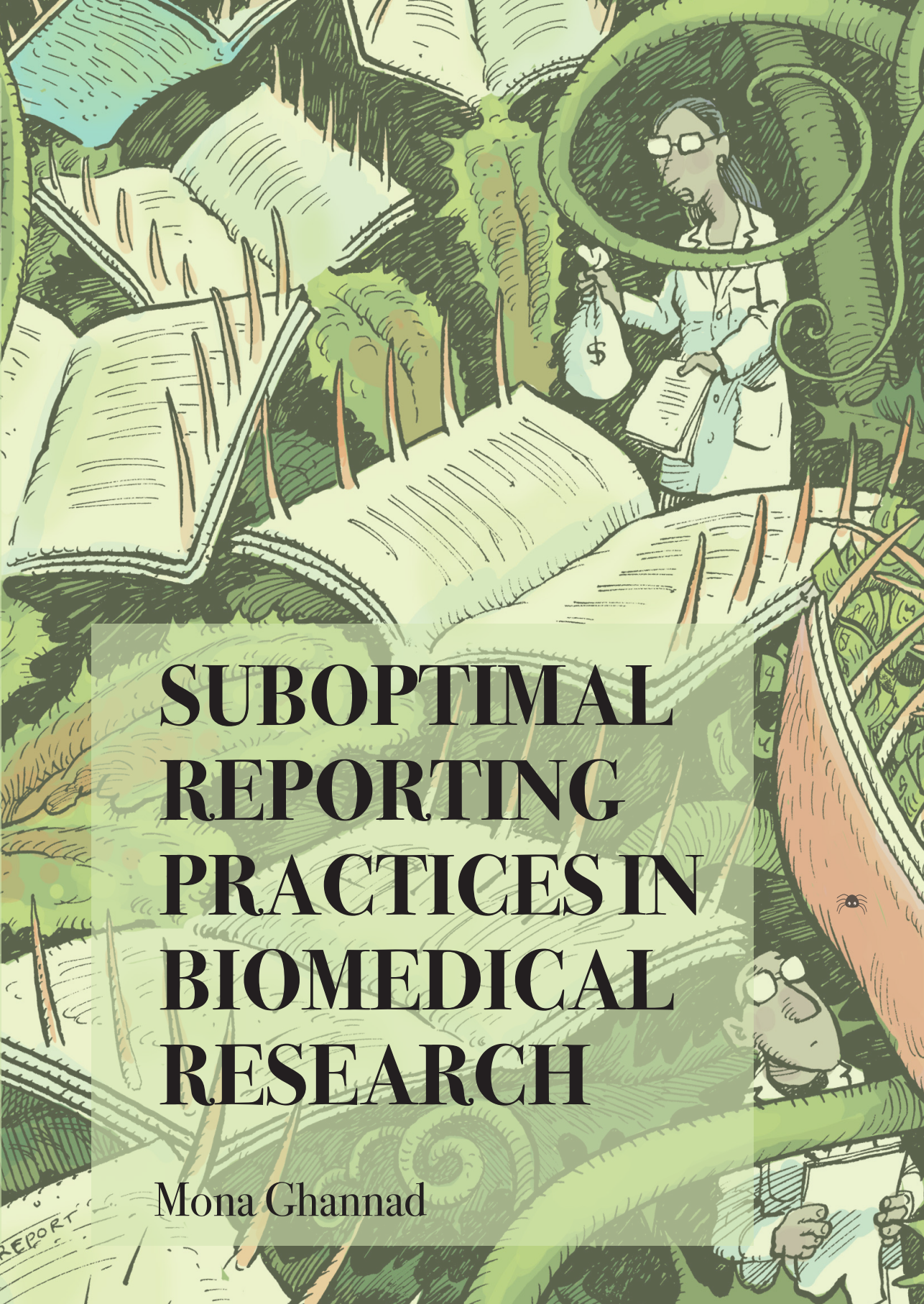
Suboptimal reporting practices in biomedical research

Ghannad, M.

Publication date
2021
Document Version
Final published version



SUBOPTIMAL REPORTING PRACTICES IN BIOMEDICAL RESEARCH

Mona Ghannad

Suboptimal reporting practices in biomedical research

Mona Ghannad

ISBN: 978-94-6332-763-3

Cover design: David Parkins
Printed by GVO drukkers & vormgevers B.V. Proefschrften.nl

Suboptimal reporting practices in biomedical research

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 9 juli 2021, te 14.00 uur
door

Mona Ghannad

geboren te Tehran

Promotiecommissie

Promotores:	prof. dr. P.M.M. Bossuyt	AMC-UvA
	prof. dr. I. Boutron	Université de Paris
Overige leden:	prof. dr. Ph. I. Spuls	AMC-UvA
	prof. dr. F. Miedema	Universiteit Utrecht
	dr. K. M. Smits	Universiteit Maastricht
	prof. dr. M. McInnes	University of Ottawa
	prof. dr. G. Collins	University of Oxford
	prof. dr. M. Chalumeau	Université de Paris

Faculteit der Geneeskunde

Dit proefschrift is tot stand gekomen in het kader van het Europese project 'MiRoR' GA No. 676207, met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid in de Faculteit der Geneeskunde van de Universiteit van Amsterdam en in het Centre de Recherche Épidémiologie et Statistiques van de Université de Paris.

This thesis has been written within the framework of the European project 'MiRoR' GA No. 676207, with the purpose of obtaining a joint doctorate degree. The thesis was prepared in the Faculty of Medicine at the University of Amsterdam and in the Centre de Recherche Épidémiologie et Statistiques at the University of Paris.

To Simon, Maman (in memoriam) & Baba

Acknowledgments

This thesis would not have been possible without the support of so many people, to whom I am infinitely grateful.

First, I like to thank my super supervisors, Patrick Bossuyt and Isabelle Boutron. I have grown both professionally and personally over the last four years – especially thanks to you.

Dear Patrick, looking back through-out this past four years of the PhD, I always felt encouraged and energized, despite various challenges. You allowed me the freedom to explore and take charge, even when I felt uncertain. Any set-backs encountered became a learning experience, which together with your clear directions and guidance, in a calm and caring manner, were paramount in me developing both professionally and personally. I always looked forward to our weekly discussions and all your advice, for insight and inspiration. I am sincerely and incredibly grateful for your guidance, your infinite patience, your encouragement, and your support.

Dear Isabelle, I thank you for your supervisorship of my PhD project as well as your truly inspirational leadership of this ambitious MiRoR project. I am whole-heartedly and incredibly grateful for your guidance, patience and unwavering support in the last four years, and especially during my secondment at Cochrane France, and in 2019, during a hard time that was marked with pain and loss.

I will be forever grateful to David Moher, for the opportunity, mentorship, encouragement and support that has not only propelled and shaped my career as an epidemiologist, but has also led me on a personally transformative journey towards becoming the person I always hoped I would be. Dear David, I would not be where I am today, if it weren't for you.

Dear Lotty Hooft – I sincerely thank you for all your support and patience while finalizing my thesis especially during this challenging (Covid) year. It is not just your support and mentorship that I value, but also the special synergy and openness of our connection. I have really enjoyed working together, and I hope in the future we find ways to continue to do so.

I am indebted to Bill Cameron for the opportunity that tremendously helped me in my transition to Amsterdam for the PhD.

I am immensely grateful to Sara Schroter and Adrian Aldcroft and BMJ Open to for hosting me during my second secondment and the opportunity to develop an editorial intervention to reduce 'spin', which became the highlight of my thesis.

I would also like to thank Carl Heneghan, Jeffrey Aronson, Liz Wager, Ruth Davis, and Joanna Lach for making my secondment at the Centre for Evidence-Based Medicine at the University of Oxford possible. A warm thank you to Kamal Mahtani, Annette Pluddemann, Georgia Richards, Claire Friedemann Smith, David Nunan, Iain Chalmers, and all others at CEBM for making my stay such a wonderful and memorable experience.

Reza Ramezan: I am so incredibly grateful for your time, and your help with the modeling, statistical support, and collaboration on the project with CEBM. I don't know what I would have done without your collaboration, and I am so grateful that I didn't have to find out. You are truly awesome, brilliant and one of a kind in your generosity, Professor Reza!

Dear Robin Haring and Matthew Page, thank you for the beautiful opportunity of collaborating together.

Dear Montse Rué, Pauline Grimm and Erica Ison, I will never forget your friendship during my first months in Amsterdam and during the course in Florence and my stay at Oxford. Thank you for modeling the extraordinary kindness, attention, and strength that make me want to be a better and more generous person. The world is a better place because it is graced by caring souls such as yours.

I would like to thank all the MiRoR PhD students, thank you for being amazing colleagues, travel companions, friends, and sources of cheer and strength.

Members of the BiTE group – I found our weekly discussions very enjoyable and educational, and will miss them. A special thanks to Maria Olsen, Bada Yang and Mariska Leeflang for the collaboration, and cherished colleagues, Yasaman, Jenny, Amber, Miranda, Hadi, Linda P, Marileen and Attila.

To my PhD sisters, Linda N, Eli, Noémie, Camila, Ceci, Anna, Mel, Van, Keti, Evi, Alice, Yasaman, and Amber, I am super thankful for all the laughter, tears, and conversations we have had, and your friendship through the good times and the bad times.

Dearest Linda N, Camila, Lorenzo, and Christopher, and Eli, your genuine friendship, the hearty conversations, and the times we have had exploring Amsterdam and Paris, has made this PhD journey all the more meaningful and fun!

Amir Ghannad: Dear Amir, with all my heart I thank you for your time and energy, all your advice, and for believing in me and coaching me with a perfect balance of positivity and wisdom. You truly are the embodiment of a 'transformative leader'.

Martine Laurens, Bert (and Levi) van Wijk: Thank you for your never-ending patience with me during the trying times – especially during 2019 (and beginning of the pandemic year 2020!), and for holding me under the big collective warmth of your family. I needed it, and I am grateful. Dear Martine, if it weren't for you, I would not have yet discovered my love of painting and drawing, which has truly been one of the best presents I have ever received.

Arjen Boerstra: Thank you for being there for me during a pivotal time, whilst keeping so calm, with optimism, encouragement and enthusiasm. I have grown to love your spontaneous visits, and appreciate you always being there when called on.

To my sisters Maryam and Marjan, I have shared with you my childhood and several hard seasons of pain and loss, that has span over decades and journeyed across oceans from East to West. One thing I know for sure, despite our 12- and 13-year age difference, my life's experience has been uniquely interconnected with yours, in its joys and woes. I am grateful to you both for the times you have been able to see things from my perspective, and offer compassion and support.

I am grateful to my nieces, the ever-insightful Noureen, as well as Yasmin and Aliya who I hear are both ever-talented– for keeping the energy of our family so bright and happy and full of hope!

Thank you to my father, Akbar Ghannad, for all your sacrifice, love and support, and for teaching me about hard work, perseverance, and dedication to a cause that is of societal value. I am forever grateful to my beloved late mother, Zary Barzegar, for her unconditional love, support and sacrifices. Dear Maman and Baba, I am everything I am because of your love.

Lastly but certainly not least, to my partner and my best friend, Simon Boerstra, I thank you for your endless support and love, and your presence by my side. You are my heart … *Always*!

Table of Contents

General introduction

Interpretation of data is subjective and can lead to bias

The last twelve months of this PhD project coincided with the Covid-19 global pandemic. Prior to the pandemic, despite spending the last 3.5 years of my PhD project evaluating interpretation bias and strategies aimed at attenuating such practices, I only now realize how much I fell short of grasping the urgency and implications of currently existing subjective interpretation of data culminating to misleading conclusions.

Before the pandemic, I had drafted the first introductory paragraph of my thesis, with an excerpt from Kaptchuk's article on the natural existence of the effect of interpretive bias on research evidence, quoting his argument that "*good science is embodied in the tension between the empiricism of concrete data and the rationalism of deeply held convictions*"[1]. Similarly, like any researcher, I empathized with the notion presented by Ioannidis and colleagues that a major challenge for scientists is balancing the ability to see novel and unexpected patterns in data, while simultaneously avoiding apophenia: the tendency to see structure or patterns in random data.[2]

The combination of apophenia and interpretative biases can easily lead us to false conclusions.[2] Indeed, the human element in the interpretative process in science is subjective and prone to bias.[1] After all, it is in line with one of my personally most frequently used phrases when accounting responsibility for faltering or error: 'it is only human'. It is thus understandable that scientific interpretation may also be based on good judgment or error, and the distinction can only be observed retrospectively. However, what if the system is off course, and errors and biases are not mitigated to the extent that is possible – through robust methodology and good reporting practices – thus leading to excessive waste in research efforts and mistrust in science?

Most recently, an article in the journal STAT, published in July 2020, discussed such issues in the 1200 registered clinical trials since the start of January 2020 that were aimed at testing treatment and prevention strategies against Covid-19.[3] One analysis found that one in every six trials was designed to evaluate the malaria drugs hydroxychloroquine or chloroquine, despite evidence of lack of benefit in hospitalized patients.[3]

Resch and colleagues documented an example of confirmation bias in a randomized controlled study, in which 398 researchers were unknowingly randomized to evaluate fictitious reports of treatment for obesity for a respected journal. The reports only differed in their description of treatment intervention: an unproved but credible treatment or an unconventional treatment. Reviewers showed a significant bias in favour of the credible treatment, disfavouring a technically good but unconventional report.[4]

Experimental results are commonly judged by expectations, and evidence that is inconsistent with well confirmed principles may be discounted by selectively finding faults in the study design or conduct.[1] When early randomized controlled trials of hormone replacement therapy (HRT) did not show a reduction in risk of coronary heart disease[5], advocates argued that the disease was far too advanced in the study population to benefit from the treatment, deeming it still valuable for primary prevention[1]. The early negative evidence supporting hormone replacement therapy

may have been more readily accepted if the pathophysiological mechanism had not created a strong expectation that the cardiovascular system is benefited by oestrogens.[6]

Potential biases can also occur before data are collected. Being convinced of the hypothesis may affect data collection, thus leading to orientation bias. Psychology graduate students found that rats specially bred for maze brightness performed superior to those bred for maze dullness, despite both groups being standard laboratory rats assigned at random.[7]

Articles published in *The Lancet* illustrated the problem of research waste during various stages of research encompassing design, conduct and reporting. [8, 9] Given that much of this waste is avoidable, there is a need to develop and implement remedies. [8] Of these, accurate interpretation and presentation of results in published data is essential in order to avoid producing misleading studies and waste valuable resources.

Background and objectives

"Spin" is a standard concept in public relations and politics, achieved through providing a biased interpretation of an event in order to slew public opinion (https://en.wikipedia.org/wiki/Spin_(propaganda)). For instance, the way in which news is reported may contain bias and distortion, and so, modify the perception of an event, through tactics such as selectively presenting specific facts (i.e., "cherry picking"), or understating potentially negative information.

The concept of "spin" has also been investigated in scientific communications. Authors have a wide latitude in interpreting and reporting their findings.[10] "Spin" has been defined as a way of reporting, not necessarily intentional, "that fails to faithfully reflect the nature and range of findings and that could affect the impression that the results produce in readers", i.e., a way to distort science reporting without actually lying.[11] Several studies have shown that authors of clinical studies may commonly present and interpret their research findings with a form of spin.[10, 12-17] "Spin", biased representation or interpretation of results in scientific reports, can harm patients and constitutes as a source of avoidable waste in research. [2, 8]

The overarching aim of this PhD project was to identify and document suboptimal reporting practices in published reports and to suggest preferred strategies to overcome these. We focused on three key topics: (1) investigating suboptimal reporting practices, such as mis-representation and over-interpretation of study findings (also known as spin) and inadequate study design or methods, in diagnostic/prognostic biomarker studies and randomized trials (Chapters 1-3); (2) developing an intervention to reduce spin and evaluated the feasibility of the proposed strategy, by conducting a collaborative field trial at The BMJ publishing group (London, UK) (Chapter 4); and (3) looking at other aspects of suboptimal reporting practices leading to bias and waste in scientific publications (Chapters 5 and 6).

Documenting suboptimal reporting and design features

As mentioned above, previous studies have characterized a high level of spin in published reports of randomized controlled trials, nonrandomized studies, diagnostic test accuracy studies, and systematic reviews.[10, 13, 17-23] Additionally, findings from previous methodological research suggest that inconsistencies in treatment effect estimates may be driven by methodological differences related to study design, sample size or participant characteristics [24, 25]. We investigated the presence of spin, further categorized as misrepresentation and overinterpretation of study findings in ovarian cancer biomarkers (**Chapter 1**), and analyzed practices that facilitate spin, such as suboptimal design features and inadequate reporting of methods (**Chapter 2**). We then evaluated the association between reported trial characteristics (e.g., related to study design, sample size, sequence generation, blinding, funding and conflict of interest) and treatment effect estimates in randomized trials of testosterone therapy in men (**Chapter 3**).

Developing an intervention to reduce spin

To date, there has been no documented additional intervention shown to clearly mitigate or reduce the prevalence of spin in biomedical literature. Having documented the level of spin in previous study, it was also relevant to develop and evaluate the effectiveness of an intervention that guides authors to reduce spin in their published articles. To estimate the effect of the intervention compared to the usual peer-review process on reducing spin in the abstract of biomedical study reports, we conducted a two-arm, parallel-group RCT in a sample of primary research manuscripts submitted to *BMJ Open* (**Chapter 4**). In the intervention group, authors received short instructions as part of the decision letter alongside the peer reviewers' comments to check for and remove spin in the abstract of their revised manuscript. In the control group, the authors received recommended editorial revisions and reviewers' comments in their usual manner.

Assessing other aspects of publication practices

Where the previous projects focused on issues in the reporting and methodological deficiencies in published articles, we also focused on the publication culture. Challenges that threaten the validity and credibility of published reports span beyond attenuating spin in published articles. For example, entities that have become known as 'predatory' journals and publishers are permeating the world of scholarly publishing, yet little is known about the articles they publish. We examined nearly 2000 biomedical studies from more than 200 journals thought likely to be predatory, recording their study designs and their epidemiological and reporting characteristics (**Chapter 5**).

Publication of articles in scientific journals is not exclusively for the scientific community and academic progress; it also serves the purpose of disseminating scientific findings to the public. Alternative metrics, such as Altmetric scores, have been developed to measure the attention publications receive from social news media and blogs, in an attempt to measure how often journal articles and other scholarly outputs are discussed and used around the world. Lifestyle factors and

their association with health and longevity have always been of great public interest, and generate significant attention from social and news media.[26, 27] We wondered whether the high level of interest in dietary interventions and differences is a persisting phenomenon, and performed an analysis of the Altmetric scores of nutritional studies, relative to other interventions by evaluating more than 300 articles published in medical journals in 2019 with an Altmetric score of more than 50. This project is reported in **Chapter 6**. The final chapter, (**Chapter 7),** provides a summary of the findings and highlights potential strategies to avoid these problems and deficiencies in the publishing process, with the ultimate goal of increasing confidence and value in published reports of clinical research.

Chapter 1

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

Mona Ghannad

Maria Olsen

Isabelle Boutron

Patrick M. Bossuyt

Journal of Clinical Epidemiology 2019;116:9-17

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

ABSTRACT

Background In the scientific literature, 'spin' refers to reporting practices that make the study findings appear more favourable 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 1026 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.

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.[28, 29] 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.[30] Likely reasons for failure have been documented at each of the stages of biomarker evaluation.[28-30]

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.[28, 31] Authors may claim favourable 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 to the main text.[18] Specific study features could facilitate distorted study results, such as not pre-specifying 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 favourable than the results justify.[10, 17] Several studies have shown that authors of clinical studies may commonly present and interpret their research findings with a form of spin.[10, 18, 20, 21, 32] A consequence of biased representation of results in scientific reports is that the published literature may suggest stronger evidence than is justified.[2] Misrepresentation of study findings may also lead to serious implications for patients, healthcare providers, and policy makers.[33]

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. We documented the prevalence of actual forms of spin misrepresentation and misinterpretation. 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.

METHODS

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

Literature search

MEDLINE was searched through PubMed on December 22nd 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 (RS), 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 (MG, MO) 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 (PB). We analyzed the first 200 consecutive studies, ranked according to publication date, to have a sample size comparable to previous systematic reviews of spin.[17, 22]

Establishing criteria and data extraction

Biomarker studies in ovarian cancer vary by study design, biomarker clinical application, type and number of tests evaluated.[34, 35] 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 favourable than results justify. This definition of spin was based on criteria extracted from key articles on misrepresentation and misinterpretation of study findings.[10, 13, 17, 18, 20, 22, 23, 32]

To evaluate the frequency of spin, we established a preliminary list incorporating previously established items that represent spin as well.[10, 13, 17, 18, 23] 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 (MO), and points of disagreements were discussed with a third investigator (IB, PB) 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 (PB, IB), categorizing items as representing ‘spin’ or ‘facilitator of spin’. Each of the categories encompassed several forms of spin. The list of items and the criteria are shown in Table 2.

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.[22] 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 and colleagues[22] 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.[22] 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 pre-specified and/or investigated, (5) mismatch between intended aim and abstract or main text conclusion, (6) other benefits of biomarkers claimed not pre-specified 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 pre-specifying 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.[17]

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, reported in Table 4.

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 (MG), and all uncertainties were resolved in discussions with a second reviewer (PB, MO).

Analysis

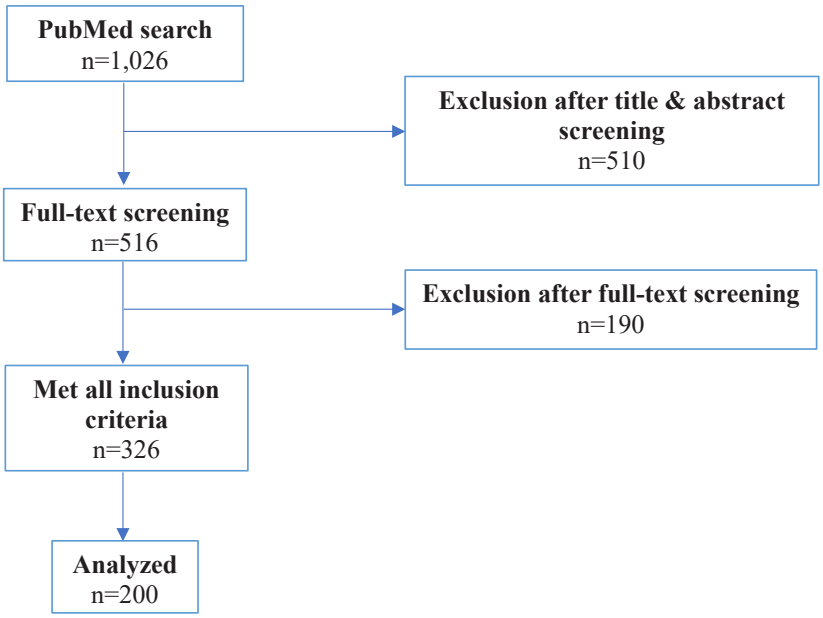
For each of the items on spin, facilitators of spin, and potential determinants of spin, we report the frequency in our sample of biomarker evaluations, with 95% confidence intervals.

RESULTS

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 (Figure 1).

Figure 1. Flow chart of search results



Characteristics of Included Studies

A description of included studies is 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 7 studies (3.5%) reporting a negative or neutral 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 non-profit (n=135, 67.5%). However, 53 of the included studies (27%) did not report source of funding.

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 of interest	
No	151 (76%)
Not reported	38 (19%)
Yes	11 (6%)
Funding source	
Non-profit	135 (68%)
Not reported	53 (27%)
No funding	6 (3%)
For-profit	4 (2%)
Mix (for-profit and non-profit)	2 (1%)

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%.[36, 37]

Several studies claimed superiority in performance in the absence of tests for statistical significance.[38, 39] In 33 study reports (16.5%) there was a mismatch in results reported in the abstract and the main text. Most frequent example 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 that 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", while the marker had only been evaluated in a diagnostic setting, with symptomatic patients.[40]

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%) compared to 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 non-clinical setting, discriminating between cases and non-cases, based on the biomarker.[40] 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 study that concluded that a spectroscopy technique was useful for the early detection of disease, while the study had only evaluated patients undergoing surgery.[41]

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 non-significant 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);	**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 is 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 to results in the main text.	**11 (5.5% [3% - 10%])**

Table 2 continued.

Category of spin	Form of spin	Criteria	Spin frequency, n = 200 n (%) [95% CI]
Misinterpretation 4	Other purposes of biomarker claimed not pre-specified and/or investigated	Abstract conclusion OR main text conclusion contains statement suggesting BM purposes not pre-specified 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); c. The main text conclusion contains statement that uses causal language for BM(s) being assessed despite the use of a nonrandomized design;	**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%])
6	Other benefits of BM claimed not pre-specified and/or investigated	The main text conclusion contains statement claiming BM benefits not pre-specified 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%])**

* All results presented in abstract and main text, excluding supplementary material.

**Abbreviations: BM, biomarker; HR, hazard ratio; OS, overall survival; PFS, progression-free survival.

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 pre-specified a positivity threshold for the continuous biomarker evaluated.

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 pre-specifying a positivity threshold for continuous biomarker	84/164* (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%])

* 164 articles included evaluation of continuous biomarkers.

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 compared to 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, compared to studies that declared no conflict of interest (67%).

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					
Non-profit	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 (non-profit and for-profit)	2	1	1	0	2 (100% [20% - 100%])

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 non-clinical setting. These two forms of misinterpretation were more prevalent in the abstract conclusion compared to 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 pre-specify a positivity threshold for continuous biomarkers.

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, whilst 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.

While some of the forms of spin in or 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.[10, 13, 17-23] The reasons behind biased and incomplete reporting are probably multifaceted and complex. Yavchitz and colleagues 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.[21] The reward system currently used in biomedical science can also be held responsible, as it focuses greatly on quantity of publications rather than quality.[10]

It has previously been shown that spin in articles may indeed hinder the ability of readers to confidently appraise results. Boutron and colleagues[20] 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.[42]

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[43], journal editors do not explicitly recommend the use of reporting guidelines in the review process[44]. 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 pre-specified hypothesis and other relevant studies” in their discussion[45], and item 4 in the STARD guideline for diagnostic accuracy studies recommending authors to “specify the objective and

hypothesis" in their introduction[46]. 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 evidenced based change in practice for the review process.

The development of biomarkers holds great promise for early detection, diagnosis and treatment of cancer patients. 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.

ACKNOWLEDGEMENTS

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

Chapter 2

Shortcomings in the evaluation of biomarkers in ovarian cancer: a systematic review

Maria Olsen

Mona Ghannad

Christianne Lok

Patrick M. Bossuyt

Clinical Chemistry and Laboratory Medicine 2019;58(1):3-10

Supplementary data to this article can be found online at:
https://doi.org/10.1515/cclm-2019-0038

ABSTRACT

Background Shortcomings in study design have been hinted at as one of the possible causes of failures in translation of discovered biomarkers into the care of ovarian cancer patients, but systematic assessments of biomarker studies are scarce. We aimed to document study design features of recently reported evaluations of biomarkers in ovarian cancer.

Methods We performed a systematic search in PubMed (MEDLINE) for reports of studies evaluating the clinical performance of putative biomarkers in ovarian cancer. We extracted data on study designs and characteristics.

Results Our search resulted in 1,026 studies; 329 (32%) were found eligible after screening, of which we evaluated the first 200. Of these, 93 (47%) were single center studies. Few studies reported eligibility criteria (17%), sampling methods (10%) or a sample size justification or power calculation (3%). Studies often used disjoint groups of patients, sometimes with extreme phenotypic contrasts; 46 studies included healthy controls (23%), but only 5 (3%) had exclusively included advanced stage cases.

Conclusions Our findings confirm the presence of suboptimal features in clinical evaluations of ovarian cancer biomarkers. This may lead to premature claims about the clinical value of these markers or, alternatively, the risk of discarding potential biomarkers that are urgently needed.

Key message: This review shows that design shortcomings in the clinical evaluations of ovarian cancer biomarkers are frequent. These include limited sample size and the recruitment of multiple, disjoint groups. Such shortcomings may hinder successful translation of ovarian cancer biomarkers.

INTRODUCTION

Epithelial ovarian cancer (EOC) is the gynecologic malignancy with the highest mortality rate. With an overall 5-year survival of 95% for early stages and only 30% for advanced disease, efforts to change survival rate in ovarian cancer has let to minor improvements over the past 25 years. Of the different histological EOC subtypes, high grade serous adenocarcinoma is the most frequent. Ovarian cancer is often asymptomatic or has specific symptoms in early-stage disease. As 70-80% of patients are diagnosed with advanced disease, prognosis is typically poor[47]. Using biomarkers for detection at an early curative stage is therefore a pressing unmet clinical need[48]. Biomarkers can also be used to evaluate treatment and to detect recurrence of EOC.

Considerable investments in ovarian cancer biomarker research have been made in the last decades. Despite claims from numerous studies, few markers have been successfully implemented in practice since the discovery of CA-125[49].

The bench-to-bedside process of biomarker development is a complex and multistep process. It has several distinct phases, ranging from the discovery and analytical validation, to clinical marker evaluation, and final implementation. Each phase holds different primary objectives, methods, and study designs[50-54]. Discovery studies usually show an association between marker values and clinical entities. In contrast, evaluations of clinical performance will be used to inform clinical decision making, as in recommendations for using the biomarker to guide further testing, start treatment, choice of treatment.

To properly inform decision-making, a clinical evaluation of a biomarker would include a single group of consecutive participants, recruited in a clinical setting, identified by pre-defined and clear eligibility criteria, preferably from multiple centers, to facilitate generalizability and with a sufficiently large sample size for precise estimates, justified by a power calculation[55-57].

Shortcomings in the design of clinical evaluation studies have been hinted at as one of the possible causes of failures in translation of discovered biomarkers into the care of ovarian cancer patients. This has been described mostly in commentaries, based on anecdotal evidence, but more systematic assessments of biomarker studies are scarce. The use of sub-optimal designs features may introduce bias in the estimated performance of a marker or limit the applicability of study findings, subsequently leading to unjustified optimism or premature rejection, contributing to translational failure[58-62].

We here report a systematic review of study design features used in recent evaluations of the clinical performance of ovarian cancer biomarkers.

METHODS

Literature search

We performed a search on 22.12.2016 for reports of studies evaluating biomarkers in ovarian cancer in PubMed (MEDLINE). The search was limited to 2015 to obtain recent studies already indexed in MEDLINE.

The search strategy was developed in collaboration with a medical information specialist (RS) (Supplementary 1). Based on sample sizes from similar systematic reviews, we aimed to include 200 studies[63].

Study selection

Articles were eligible if they reported a primary clinical study, evaluating one or more biomarkers, and included adult women diagnosed, screened, treated, or monitored for any type of ovarian cancer. To distinguish clinical evaluation studies from studies of other phases (primarily discovery studies) we defined a clinical study as a study that included the assessment of a previously discovered biomarker and reported a clinical performance measure that could be used to inform clinical decision-making.

We relied on the 1998 National Institutes of Health definition of a biomarker[64], including not only markers from body fluids but also imaging markers, such as ultrasound, CT, MRI and other modalities. Screening of titles and abstracts and full text evaluations was done in duplicate by two independent reviewers (MG and MO). Disagreements were solved through discussion; a third reviewer (PB) was consulted if consensus was not reached.

Data extraction

The study features were identified from previous commentaries, studies, checklists, and quality assessment tools[49, 55, 61, 65-68] (Table 1). Data extraction was performed with a dedicated form by one reviewer (MO); unclear items were discussed with two other reviewers (MG and PB). Extraction guidance, as used in data-extraction, is provided in Supplementary Table 2.

Statistics

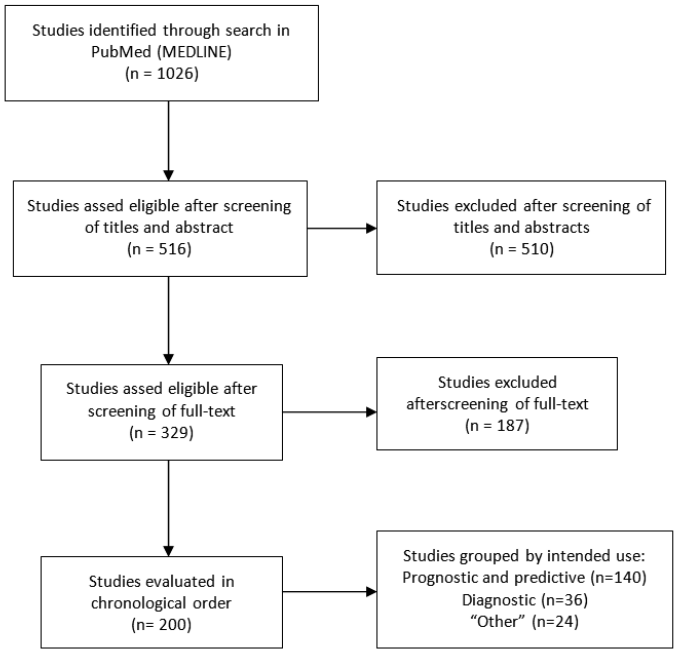
We calculated the proportion of studies with each respective feature, presented as estimates and 95% confidence intervals. We used Fisher's Exact test to evaluate differences and a Kruskal-Wallis test for differences in sample size between subgroups. Two-sided P-values below 0.05 were considered as pointing to statistically significant differences. Calculations were performed in R (version i386 3.4.3).

RESULTS

Search and study selection

Our search resulted in 1,026 articles, of which 516 (49%) reports were considered potentially eligible after screening titles and abstract, and 329 eligible (32%) after reading the full text (Figure 1). Of these, we evaluated the first 200, in chronological order of publication, starting January 1st 2015 towards most recent. The evaluated studies had been published in 95 journals from January 2015 until January 2016 and with a distribution ranging from 1 to 13 articles per journal (Supplementary Table 3) within both pre-clinical/translational and clinical journals.

Figure 1. Flow diagram of studies.



Shows search results and study flow, including the distribution of intended use among the evaluated studies.

The largest group of studies reported on prognostic and predictive biomarkers (70%). The second largest group consisted of studies describing markers for diagnostic purposes (18%) (Table 1). Across applications, we found a variety of different types of biomarkers and biomarker profiles including but not limited to clinical (risk) factors, as BMI and menopausal status, genetic profiles/mutations, as BRCA1/2, protein biomarkers, as CA-125 and HE4, clinical risk scores, as ROMA and RMI. The most frequently evaluated biomarkers were CA-125, HE4, and risk scores,

evaluated either alone or in combinations. E-cadherin and clinical prognostic factors were also among the most frequently evaluated (Supplementary Table 4).

The most frequently reported performance measures expressed the strength of associations, for example as hazard ratios or odds ratios, often accompanied by Kaplan-Meier survival analysis (54%). Other studies reported classification statistics, such as the area under the receiver operating characteristic (ROC) curve and ROC-statistics (24%).

Study design features

1. Recruitment of study participants

To evaluate the validity and applicability of the performance measures, study reports should include clear eligibility criteria and the methods for recruiting study participants. Of the 200 included study reports, 34 (17%) explicitly reported eligibility criteria and 19 (10%) sampling methods. Only 12 articles (6%) referred to an existing protocol (Table 1). As illustrated in Supplementary Table 5, the information provided on the identification and selection of study participants was often limited (Supplementary Table 5, Ex. 1) and even less detailed in analyses based on registries (Supplementary Table 5, Ex. 2).

Whenever the study group was described in study reports (n=59, 30%), this was often done in rather broad and general terms, such as "sampled from the general population" (n=1) or "in women/patients with ovarian cancer/tumor" (n=10). In other cases, this was described by nationality (n=8), subtype (n=18), or symptom(s) (n=3). In contrast, a few studies had a description very specific to treatment or outcome (n=6).

2. Single versus multiple groups

In evaluations of the clinical performance of biomarkers, study participants should represent the intended use population. Of the 200 studies in our sample, 113 (57%) had indeed included a single group of study participants (i.e., groups of comparison originated from one single study group). In contrast, 66 (33%) studies had recruited patients in multiple, disjoint groups (i.e., groups of comparison originated from separate study groups). Forty-six studies (23%) reported on healthy controls, although the definition of a healthy control varied between studies (Supplementary Table 5, Ex. 3,4). The groups that were included, other than ovarian cancer patients, ranged from patients with benign conditions to participants with other diseases and conditions, also referred to as "controls" (Supplementary Table 5, Ex. 5, 6). In one study, patients served as their own control (Supplementary Table 5, Ex. 7). At the other end of the spectrum, 5 (3%) studies had exclusively included patients with advanced stages (III-IV), which was not entirely consistent with the stated target population and study objective (Supplementary Table 5, Ex. 8).

Table 1. Frequencies of study design features

Design features and collection characteristics	Total	95% CI	Prognostic and Predictive	Diagnostic	Other	p-value
Intended use	n = 200		n = 140 (70%)	n = 36 (18%)	n = 24 (12%)	
Reporting of eligibility	34 (17%)	[12% to 23%]	30/140 (22%)	3/36 (8%)	1/24 (4%)	p = 0.04
Reporting of sampling method	19 (10%)	[6% to 14%]	12/140 (9%)	5/36 (14%)	2/24 (8%)	p = 0.55
Protocol	12 (6%)	[3% to 10%]	6/140 (4%)	4/36 (11%)	2/24 (8%)	p = 0.16
Power calculation	5 (3%)	[1% to 6%]	2/140 (1%)	1/36 (3%)	2/24 (8%)	p = 0.10
Multi-group	66 (33%)	[27% to 40%]	37/140 (26%)	12/36 (33%)	16/24 (67%)	p < 0.01
Single-group	113 (57%)	[49% to 64%]	91/140 (65%)	15/36 (42%)	7/24 (29%)	p < 0.01
Unclear	21 (11%)	[7% to 16%]	12/140 (9%)	9/36 (25%)	1/24 (4%)	p = 0.02
Healthy controls	46 (23%)	[17% to 30%]	17/140 (12%)	10/36 (28%)	19/24 (79%)	p < 0.01
Exclusively advanced stages as cases	5 (3%)	[1% to 6%]	5/140 (4%)	0/36	0/24	p = 0.78
Multi-center	93 (47%)	[39% to 54%]	60/140 (43%)	16/36 (44%)	17/24 (71%)	p = 0.04
Single-center	93 (47%)	[39% to 54%]	72/140 (51%)	16/36 (44%)	5/24 (21%)	p = 0.02
Unclear	14 (7%)	[4% to 12%]	8/140 (6%)	4/36 (11%)	2/24 (8%)	p = 0.46
Primary data	14 (7%)	[4% to 12%]	7/140 (5%)	4/36 (11%)	3/24 (13%)	p = 0.19
Secondary data	182 (91%)	[86% to 95%]	133/140 (95%)	29/36 (81%)	20/24 (83%)	p < 0.01
Routinely collected	130/182 (71%)	[64% to 78%]	95/133 (71%)	22/29 (76%)	13/20 (65%)	p = 0.74
Including retrospective data	176 (88%)	[83% to 92%]	129/140 (92%)	27/36 (75%)	20/24 (88%)	p = 0.01
Including prospective data	21 (11%)	[7% to 16%]	11/140 (8%)	7/36 (19%)	3/24 (4%)	p = 0.10
Unclear	3 (2%)	[0% to 4%]	0/140 (0%)	2/36 (8%)	1/24 (4%)	p = 0.03
Median sample size	156		132	145	657	p < 0.01
Min to max	13-50,078		13-6,556	26-2,665	31-50,078	
IQR	97-357		89-214	100-309	227-2366	
Smallest sample size used in analysis (median)	28 (of 102)		34 (of 70)	20 (of 20)	12 (of 12)	

The table shows results for the total studies (n=200) and in subgroups of intended use. Testing for differences between subgroups where performed by Fisher's Exact test (two-sided), Kruskal Wallis test for medians, and binomial test for 95% CI. "Others" include risk stratification (11), screening (4), monitoring (1), and studies with multiple use (8).

3. Single-center versus multi-center

If data for clinical evaluation are collected in a single center, there may be a concern about a lack of generalizability; multi-center studies with prospective data collection are therefore preferred. We found that samples and data had often been acquired from a single center (93 studies; 47%). The majority of studies (182; 91%) relied on previously collected samples (Table 1). Of these, 130 studies (71%) used samples collected during routine clinical care (Supplementary Table 5, Ex. 1) while 31 (17%) used data from external registries of molecular data, of which 21 (68%) had used The Cancer Genome Atlas (TCGA) registry (Supplementary Table 5, Ex. 2). Most studies analysed retrospectively collected data (176 studies; 88%) only 21 (11%) had collected data prospectively (Table 1).

Sample size

The number of patients in biomarker studies should be high enough to arrive at sufficiently precise estimates or to have enough power to test statistical hypotheses. In this review, the median sample size was 156 patients, ranging from 13 to 50,078, with an interquartile range from 97 to 357. Only 5 (3%) studies justified sample size, for example by reporting a power calculation such as "*A preliminary power analysis was performed to determine the number of patients needed to generate solid, meaningful data using Cochran's formulas [35]. Based on this model under a 90% confident level, 0.5 Standard deviation and ±10% confidence interval, 68 EOC patients are needed to obtain confident results.*"[69] or justified by a sample sizes used in previous, similar studies such as "*The number of sequenced individuals is within an acceptable range used previously to obtain significant results*."[70]

Subgroup analysis

To assess whether frequencies of the design features differed between groups of biomarker studies defined by intended use, we classified the studies into seven groups (Supplementary Table 6). We found significant differences depending on the intended use of the biomarker in reporting of eligibility criteria, multi-group and single-groups, use of healthy controls, multi-center and single-center, use of secondary and retrospective collected data, and median sample size. Studies of biomarkers used for purposes other than prognostic, predictive or diagnostic more often included multiple groups, healthy controls, were often larger and designed as multi-center trials. In contrast, prognostic and predictive studies more frequently reported eligibility criteria and used a single group in their design.
In the 200 studies, we found one (0.5%) multi-center study that had recruited a single group of ovarian cancer patients (no separate controls) and clearly reported eligibility criteria, sampling method, and sample size justification.

DISCUSSION

In general, the field of biomarker research and medical tests is less well developed than the evaluation of pharmaceuticals and other interventions[52]. Despite the relatively large volume of published studies in ovarian cancer biomarker research, many putative markers have not been translated into clinical use[49, 61]. Shortcomings and deficiencies in study design have been suggested as a partial explanation for this translational failure. Our analysis of recently published evaluations of putative biomarkers provides systematic evidence for this hypothesis. Most studies in our sample were limited in size, performed in a single-center, and had often recruited multiple, disjoint groups of ovarian cancer cases and non-cancer controls.

As defined by Ransohoff and Gourlay, 2010, bias is " *a systematic difference between the compared groups*", which can give rise to differences caused by other factors than the one in question[71]. To this end, several authors, for example, have stressed the importance of identifying and selecting appropriate study participants and samples: those that represent the target population for the intended use. Failure to do so can lead to selection bias[51, 56, 58, 60, 71, 72].

Despite the many initiatives to improve reporting and transparency, such as the reporting guidelines "Reporting Recommendations for Tumor Marker Prognostic Studies" (REMARK), "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE), "Biospecimen reporting for improved study quality" (BRISQ), "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis" (TRIPOD), and "Standards for Reporting of Diagnostic Accuracy Studies" (STARD)[73-77], we found that eligibility criteria and sampling methods were rarely reported As a consequence, we were not able to analyse in detail if the group of study participants actually matched the intended use population. Such incomplete reporting not only hampers secondary research but also the direct usefulness of a study report in clinical practice. However, the issues surrounding incomplete and non-transparent reporting have been addressed and documented elsewhere, by several other authors[78-80].

The use of multiple groups rather than a single group of study participants - preferably a consecutive series of patients - has been identified as a major source of bias in marker evaluations. Meta-epidemiological research has shown that the additional inclusion of other groups, in particular the recruitment of healthy controls, is prone to lead to an overestimation of performance in diagnostic studies[55, 65, 72]. We found that one in three ovarian cancer marker studies relied on multiple, disjoint groups. Almost one in four included some form of healthy controls. This may be surprising, since screening was not the intended use of most biomarkers, and application of the biomarker would not involve the testing of asymptomatic persons. The inclusion of healthy controls may be justified in the marker discovery phase, or for providing proof-of-principle, but the correct classification of these healthy, asymptomatic participants is not informative about the performance of the marker in clinical applications.

A majority of studies had used secondary and routinely collected data and many relied on retrospectively collected data. For the initial discovery phases, such convenient and readily accessible data and bio-specimen may be used. For a clinical evaluation, however, the data collection setting and conditions may not correspond to the clinical question[81, 82]. Single-

center studies were also relatively frequent, potentially limiting the generalizability of procedures and findings.

With a median sample size of slightly more than a hundred patients, most studies were relatively small, and, in particular without sample size justification, the uncertainty around the estimated performance measures may still be considerable, hampering strong conclusions about the value of putative markers, or the lack thereof.

We investigated shortcomings in ovarian cancer, as this is a disease with a great clinical need and substantial potential for the use of biomarkers. However, as the selected design features in our study are generic for studies that evaluate biomarkers, we believe that similar shortcomings exist in biomarker evaluations in other cancers as well.

The included studies were published in a variety of different journals and we found only one study that were free of deficiencies. For these reasons, we believe that our results reflect the general practice in biomarker evaluations rather than being related to the journals in which the studies were published.

Proposals for diagnostic, prognostic and predictive biomarker studies have been made before[56]. An impressive number of authors, statisticians and others have written about the designs and analysis of biomarker evaluations. Many of the design limitations that we observed could therefore be explained by a lack of awareness in biomedical research. This could be addressed through more extensive training, promoting the use of study protocols, encouraging the assembly of multidisciplinary teams, involving experienced biostatisticians from the initial discovery phase, and fostering large international collaborations, such as The Ovarian Tumor Tissue Analysis (OTTA) consortium, and The Ovarian Cancer Association Consortium (OCAC)[83, 84]. Such consortia could also help to achieve the targeted sample size for rare subtypes of ovarian cancer. Moreover, journal editors could demand better compliance to the reporting guidelines for primary studies, also as this may inform authors of how to better design a study for the individual clinical question.

Future commentaries and editorials in scientific journals about specific markers could additionally help to improve the practice of biomarker research, if they not only highlight the great potential of the putative biomarker, but also discuss the limitations in the research performed so far. These commentaries could, more consistently, highlight the need for real-world studies of the actual performance of biomarkers and the design of trials to document incremental effectiveness in improving patient outcomes, keeping the clinical context at the focus throughout biomarker development[52, 85]. As in intervention trials, involved stakeholders, such as companies that develop markers and funders, also need to facilitate such studies and trials.

We acknowledge a number of potential limitations of our own analysis. The data extraction form used to identify study features had not been used before. It was developed in close collaboration between two authors who also piloted it extensively, and most features were relatively easy to identify from the study reports, if reported at all. Reporting was often limited, hampering identification of some of the critical study features. Our set of design features evaluated in this review does not cover all aspects of methodological quality of the included

studies; we focused primarily on recruitment and sampling, and selected features because they had been highlighted before in commentaries and methodological analyses of other areas of testing and biomarker research.

CONCLUSION

The search for new biomarkers, fuelled by the impressive advances in omics-research, continues to hold great promise for clinical medicine. Yet, to fulfil this promise we need to increase the number of well-executed studies, with properly selected participants recruited in sufficient numbers. Although almost half of the studies were multi-center and more than half were single-group studies, we found only one study that was free of the selected shortcomings. Working in cooperation, in multidisciplinary groups and in larger consortia, could therefore be the way forward, starting fewer but higher-quality studies that can produce results that are at low risk of bias and more readily interpretable. This may avoid premature claims of biomarker performance, prevent the unwarranted removal of promising markers, and eventually produce the new tools that ovarian cancer patients can benefit from.

Chapter 3

No evidence found for an association between trial characteristics and treatment effects in randomized trials of testosterone therapy in men: meta-epidemiological study

Robin Haring

Mona Ghannad

Lorenzo Bertizzolo

Matthew J. Page

Journal of Clinical Epidemiology 2020;122:12-19

Supplementary data to this article can be found online at:
https://doi.org/10.1016/j.clinepi.2020.02.004

ABSTRACT

Objective: To identify potential trial characteristics associated with reported treatment effect estimates in randomized trials of testosterone therapy in adult men.

Design: Meta-epidemiological study.

Data source: MEDLINE was searched for meta-analyses of randomized trials of testosterone therapy in men published between 2008 and 2018.

Data extraction: Data on trial characteristics were extracted independently by two reviewers. The impact of trial characteristics on reported treatment effects was investigated using a two-step meta-meta-analytic approach.

Results: We identified 132 randomized trials, included in 19 meta-analyses, comprising data from 10,725 participants. None of the investigated design characteristics, including year of publication, sample size, trial registration status, centre status, regionality, funding source, and conflict of interest were statistically significantly associated with reported treatment effects of testosterone therapy in men. Although trials rated at high risk of bias overall reported treatment effects that were 21% larger compared to trials rated at low risk of bias overall, the 95% confidence interval included the null (ratio of odds ratio (ROR): 0.79, 95% confidence interval: 0.60 to 1.03).

Conclusions: The present study found no clear evidence that trial characteristics are associated with treatment effects in randomized trials of testosterone therapy in men. To establish stronger evidence about the treatment effects of testosterone therapy in men, future randomized trials should not only be adequately designed but also transparently reported.

Study registration: osf.io/x9g6m

INTRODUCTION

The clinical effects of the sex hormone testosterone inspired research and development since it was isolated and synthesized in a Nobel Prize winning effort in 1935 [86]. But despite eight decades of clinical use, considerable controversy exists regarding the risks and benefits of testosterone therapy in men [87, 88]. From measurement and diagnosis of low testosterone, over treatment formulations and duration, to treatment monitoring and goals, the safety, efficacy and effects of testosterone therapy in men are still under debate.

Consequently, guidelines for testosterone therapy in men demand for high quality evidence to strengthen recommendations for clinical decision-making about potential treatment [89]. Findings from previous methodological research suggests that inconsistencies in treatment effect estimates may be driven by methodological differences related to study design, sample size or participant characteristics [24, 25]. For example, systematic reviews of meta-epidemiological studies suggest that larger treatment effects are observed in randomized trials with inadequate sequence generation and allocation concealment, and in trials with a smaller sample size. Despite several trial characteristics being consistently found to affect the magnitude of treatment effects, their impact on the results of testosterone research is unknown. In addition, several factors, such a funding and conflicts of interest of study authors, have been investigated in comparatively fewer meta-epidemiological studies than other factors (e.g., sequence generation, blinding) [25].

Therefore, the aim of this study is to investigate the association between several trial characteristics (both commonly investigated and underexplored) and reported treatment effect estimates in randomized trials of testosterone therapy in men. Knowledge of these factors may help improve the design and conduct of future clinical trials to establish stronger evidence about treatment effects of testosterone therapy in men.

METHODS

We conducted this study in accordance with a study protocol we uploaded to the Open Science Framework in April 2018 (osf.io/x9g6m).

Search Strategy

Published systematic reviews with meta-analysis of randomized on testosterone therapy in men were identified from MEDLINE, via PubMed, using the following search strategy: ("testosterone"[All Fields] OR "TRT"[All Fields] OR "androgens"[All Fields] OR "sex hormone"[All Fields]) AND Meta-Analysis[ptyp] AND "2008/04/20"[PDAT] : "2018/04/17"[PDAT] AND "male"[MeSH Terms] AND English[lang].

Eligibility criteria and study selection

We included systematic reviews with meta-analyses of binary outcomes or meta-analyses of continuous outcomes, regardless of the specific dose/delivery of testosterone therapy or the comparator (placebo/control/standard of care) investigated in the included trials. We excluded systematic reviews with meta-analyses of individual participant data and meta-analyses computed using non-standard statistical methods (e.g., Bayesian). All records yielded from the

search were exported to Covidence software for screening. Screening of titles/abstracts and full text articles retrieved against the eligibility criteria was performed independently by two reviewers (RH and MJP). Once all eligible systematic reviews were identified, the list of randomized trials included in the largest meta-analysis was extracted from each review, and PubMed IDs (PMIDs) were assigned to each trial reference to identify duplicate trials. Sets of meta-analyses with overlapping trials were identified and all duplicate trials were removed, starting with the meta-analysis with the smallest number of trials and moving sequentially up to the meta-analysis with the largest number of trials. When this process led to only one or no unique trials being identified in the meta-analysis, the meta-analysis was excluded. This harmonisation process was used to generate a dataset without overlap between meta-analyses [90].

Data collection

Two reviewers (MG and LB) independently extracted data from all meta-analyses and randomized trials included in each meta-analysis by using a standardised data collection form. Disagreements were resolved via discussion with RH and MJP. For each meta-analysis, the following data were extracted: year of publication, medical specialty, number of randomized trials included in the meta-analysis, outcome domain (e.g., cardiovascular disease incidence), interventions compared (experimental and control group), meta-analysis model (fixed-effect or random-effects), and treatment effect estimates and 95% confidence intervals (CI) for each included randomized trial.

For each randomized trial included in the meta-analysis, information on trial characteristics (including year of publication, sample size, trial registration status, centre status, whether the corresponding study author was US-based or not, funding source and conflicts of interest of study authors) were extracted. All information was extracted from the trial report, and we did not contact the trials authors for clarification.

In addition, version 1 of the Cochrane risk of bias tool [91] was applied. We assessed the following domains of the tool: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, incomplete outcome data, and overall risk of bias. Each domain was judged as ‘low risk’, ‘unclear risk’ or ‘high risk of bias’. The overall risk of bias was judged as ‘low risk’ if all domains were considered ‘low risk’, ‘unclear risk’ if at least one domain was considered ‘unclear risk’ but none were considered ‘high risk’, and ‘high risk’ if at least one domain was judged to be ‘high risk’. We omitted the selective reporting bias domain given recent rethinking of the domain as a source of bias to be assessed at the level of meta-analyses rather than at the level of included studies [92].

Statistical analyses

We categorised all trail characteristics into the following pre-specified binary categories:

- Year of publication: published before 2000 versus published 2000 or later;
- Sample size: less than 50 participants versus 50 or more participants;
- Trial registration status: registered versus not registered (as declared in the trial report);
- Centre status: single-centre or not specified versus multi-centre;

- Region: US-based corresponding author versus non-US-based corresponding author;
- Funding: trial funded by industry versus trial not funded by industry;
- Conflicts of interest: trial with versus without an author with a conflict of interest declared in the trial report;
- Risk of bias: all domains classified as high/unclear versus low risk of bias.

Treatment effects for binary outcomes were estimated as odds ratios (ORs) and treatment effects for continuous outcomes as standardised mean differences (SMDs). The direction of effect was standardized so that an OR <1 or a SMD <0 indicated a beneficial effect of testosterone therapy. We analysed the association between each trial characteristic and the magnitude of a treatment effect using the two-step "meta-meta-analytic" approach for meta-epidemiological analyses described by Sterne et al. [93]. That is, we first estimated within each meta-analysis a ratio of odds ratio (ROR) for binary outcomes or a difference in standardized mean differences (dSMD) for continuous outcomes including standard errors between the two subgroup categories (e.g., industry versus no industry funding), by using a random-effects meta-regression model. Then, we estimated a combined ROR across meta-analyses and the 95% CI by using a random-effects meta-analysis model. To synthesise estimates for binary and continuous outcomes, we converted dSMDs to log RORs by multiplying by $\pi/\sqrt{3} = 1.814$ [94]. DerSimonian and Laird's method of moments estimator was used to estimate the between-meta-analysis variance. The inconsistency across RORs was quantified using the I^2 statistic and the between-meta-analysis variance estimated by τ^2. All analyses were stratified by "type of outcome" given previous studies [95, 96] showing a significant difference between the ROR for objectively measured vs. patient-reported subjective outcomes. We planned to control for potential confounding by other trial characteristics by adjusting the meta-regression models for all other characteristics investigated, but decided against doing so given the number of included trials was not considered sufficient to provide reliable estimates. All analyses were performed using the *metan* and *metareg* commands in the statistical software package Stata (version 15).

RESULTS

Our PubMed search yielded 123 records (Figure 1). After screening all titles/abstracts, we retrieved 28 full-text articles for review, of which 24 were initially considered eligible for inclusion. After removing duplicate trials across the meta-analyses, we were left with a total of 19 non-overlapping meta-analyses [97-115] including 132 trials (comprising data from 10,725 participants).

Of the 19 included meta-analyses, the median year of publication was 2014 (interquartile range (IQR) 2012-2015), and they included a median of 5 trials (IQR 4 to 8) (Table 1). The medical specialties investigated by the meta-analyses included cardiology (8 [42%]), urology (5 [26%]), endocrinology (4 [21%]), and psychiatry (2 [10%]). All meta-analyses compared testosterone replacement therapy with placebo. The majority of meta-analyses examined an objective outcome (12 [63%]), such as prostate-specific antigen (PSA) levels, HbA1c, body weight, and cardiovascular-related events. Patient-reported outcomes (investigated in 7 [37%] meta-analyses) included self-reported erectile dysfunction, mood, and subjective improvement

in cardiovascular or prostate-related symptoms. The type of outcome for most of the meta-analyses examined was continuous (13 [68%]).

Characteristics of the 132 trials are summarised in Table 2. Briefly, most of the trials were registered (79%), conducted in the year 2000 or later (77%), and comprised a sample size below N = 50 (48%). Across the six risk of bias domains, the percentage of trials rated at "unclear risk of bias" was: random sequence generation (42%), allocation concealment (62%), blinding of participants and personnel (41%), blinding of outcome assessors (20%), incomplete outcome data (17%), and overall risk of bias (56%).

None of the investigated 13 trial characteristics were statistically significantly associated with treatment effect estimates in randomized trials of testosterone therapy in men (Figure 2; more detailed results are provided Figures S1-S13 in the Appendix). In addition, there was no statistically significant interaction between type of outcome (objective versus patient-reported) and the magnitude of the ROR for any trial characteristic (see Figure S1-S13 in the Appendix). The direction of the RORs suggested that treatment effects were larger in more recently published randomized trials (i.e. published in the year 2000 or later), smaller trials (N < 50), trials that were registered, trials with a corresponding author based outside of the US, industry-funded trials, trials with an author with a conflict of interest, trials with low risk of bias due to allocation concealment and blinding of participants and personnel, trials with high/unclear risk of bias due to blinding of outcome assessors and missing data and trials with high/unclear overall risk of bias. However, the 95% CIs of all ROR estimates encompassed the null and associations in the opposite direction.

For most of the meta-meta-analyses, there was no or a small amount of between-meta-analysis heterogeneity (see Figures S1-S13 in the Appendix). However, for three meta-meta-analyses, the between meta-analysis heterogeneity was high (for the associations with country of corresponding author [I-squared 63%, tau-squared 0.49], risk of bias due to blinding of participants and personnel [I-squared 61%, tau-squared 0.48], and risk of bias due to missing outcome data [I-squared 65%, tau-squared 0.66]).

Figure 1. Flow diagram of identification, screening and inclusion of meta-analysis.

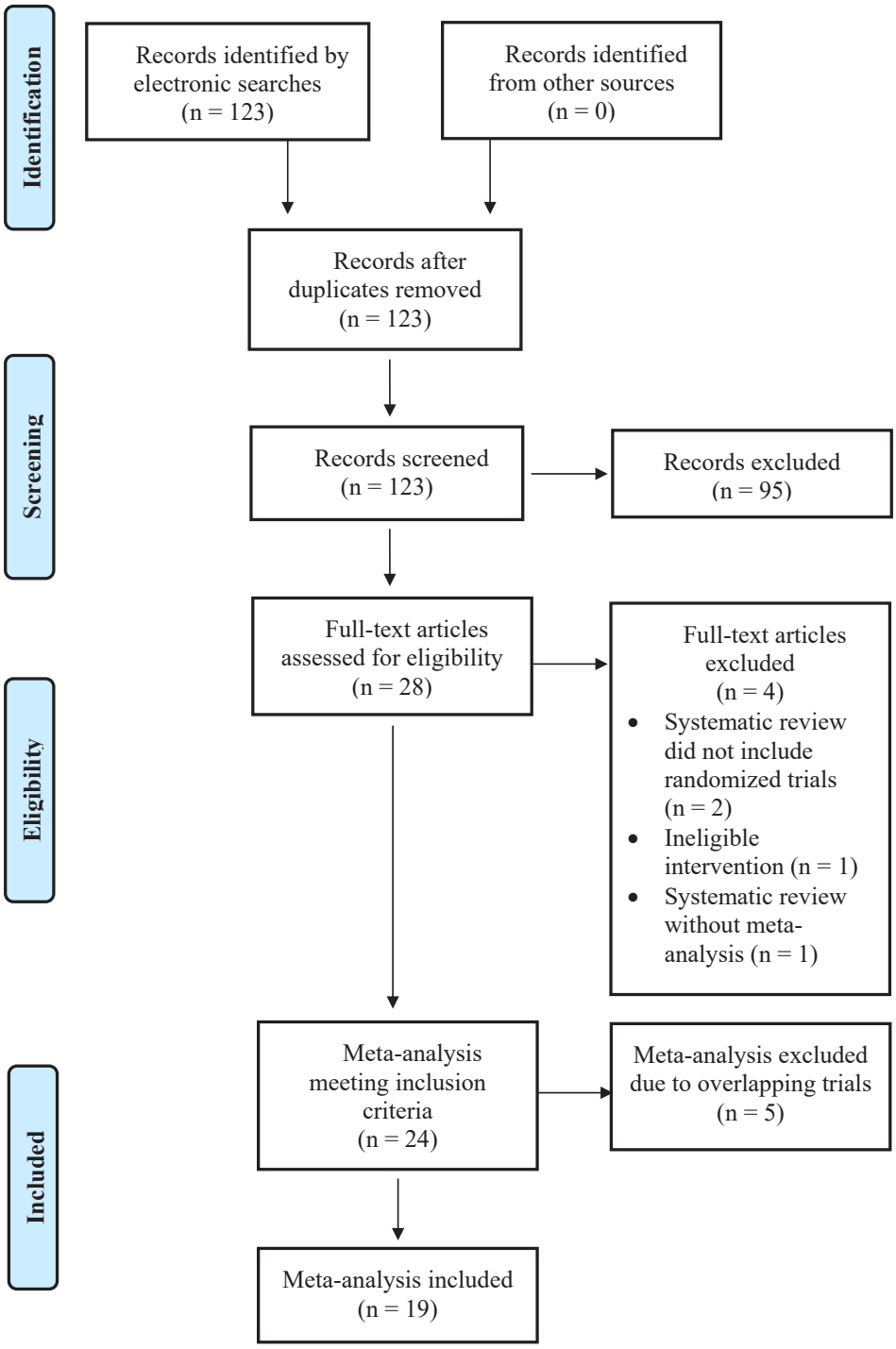


Figure 2: Main results of associations between trial characteristics and treatment effect estimates.

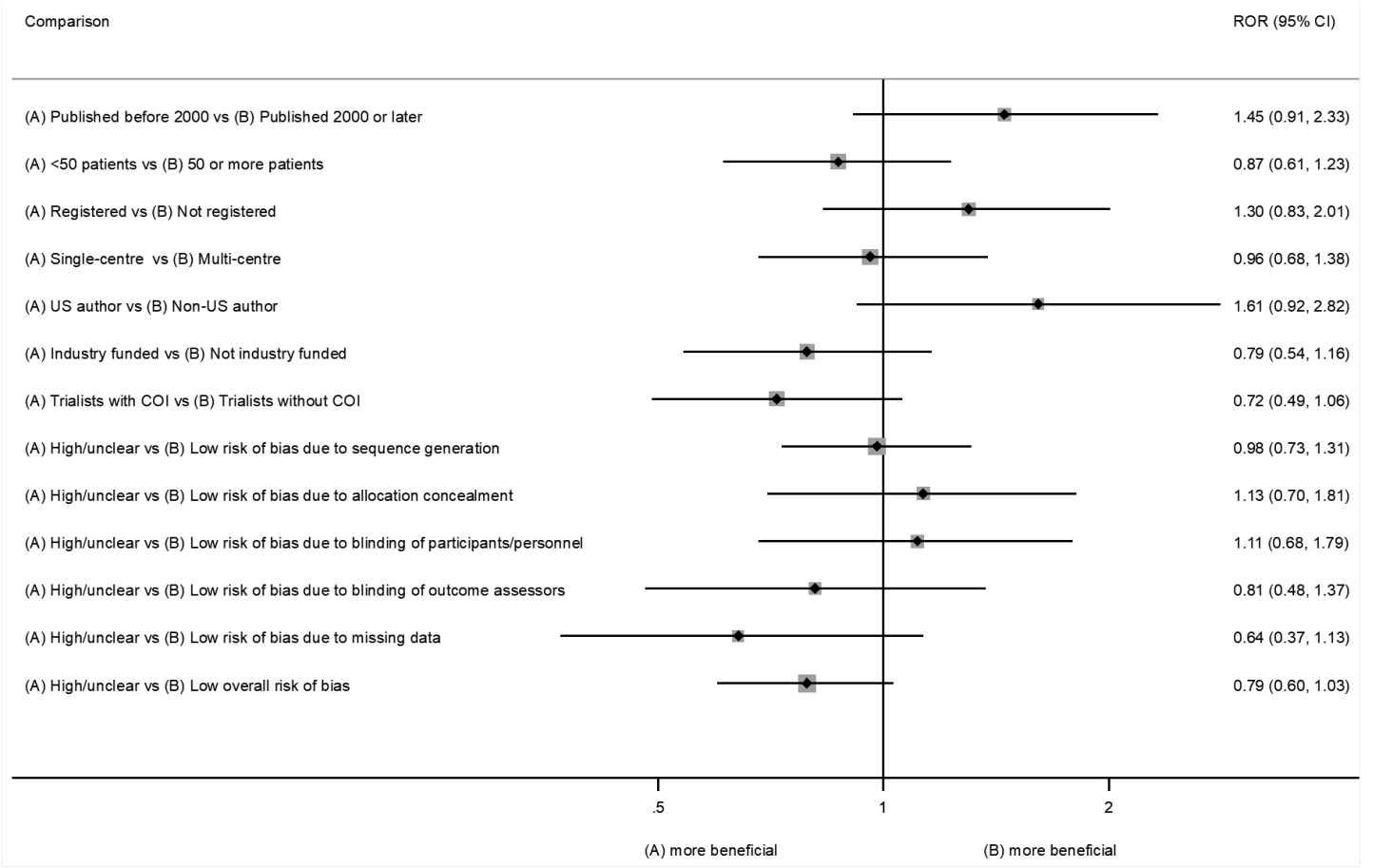


Table 1. Characteristics of included meta-analyses

Meta-analysis ID	Specialty	No. trials	Intervention	Comparator	Outcome	Outcome category	Outcome type
Alexander 2017	Cardiology	5	TRT	Placebo	Composite outcome of myocardial infarction, stroke, and mortality	Objective	Binary
Amanatkar 2014	Psychiatry	8	TRT	Placebo	Mood	Patient-reported	Continuous
Borst 2014	Cardiology	3	TRT	Placebo	Cardiovacular events	Objective	Binary
Corona 2011a	Cardiology	6	TRT	Placebo	Time to 1 mm ST segment depression	Objective	Continuous
Corona 2011b	Endocrinology	4	TRT	Placebo	HbA1c	Objective	Continuous
Corona 2014	Endocrinology	16	TRT	Placebo	Overall erectile function	Patient-reported	Continuous
Corona 2016	Cardiology	14	TRT	Placebo	Body weight	Objective	Continuous
Cui 2013	Urology	5	TRT	Placebo	Prostate-specific antigen levels	Objective	Continuous
Cui 2014	Urology	3	TRT	Placebo	PSA levels' change in the short-term treatment studies	Objective	Continuous
Fernandez-Balsells 2010	Cardiology	7	TRT	Placebo	Composite prostate endpoint	Patient-reported	Binary
Grossmann 2015	Endocrinology	4	TRT	Placebo	HbA1c	Objective	Continuous
Kang 2015	Urology	4	TRT	Placebo	PSA levels	Objective	Continuous

Meta-analysis ID	Specialty	No. trials	Intervention	Comparator	Outcome	Outcome category	Outcome type
Kohn 2016	Urology	11	TRT	Placebo	International Prostate Symptom Score	Patient-reported	Continuous
Neto 2015	Endocrinology	8	TRT	Placebo	Lean body mass	Objective	Continuous
Nian 2017	Urology	5	TRT	Placebo	Psychological subscale score	Patient-reported	Continuous
Price 2012	Cardiology	3	TRT	Placebo	Subjective improvement in symptoms of intermittent claudication	Patient-reported	Binary
Toma 2012	Cardiology	4	TRT	Placebo	Exercise capacity	Objective	Continuous
Xu 2013	Cardiology	15	TRT	Placebo	Cardiovascular-related events	Objective	Binary
Zarrouf 2009	Psychiatry	7	TRT	Placebo	HAM-D response	Patient-reported	Binary

TRT = testosterone replacement therapy

Table 2: Trial characteristics extracted from included randomized-controlled trials of testosterone therapy in men.

Trial characteristic		Number of studies (N=132)
Time effect	study conducted before year 2000 study conducted year 2000 or later	30 (23%) 102 (77%)
Sample size	>200 101-200 50-100 <50	12 (9%) 21 (16%) 35 (27%) 64 (48%)
Trial registration	registered not registered	104 (79%) 28 (21%)
Centre status	single centre randomized trial multi-centre randomized trial not reported	55 (42%) 31 (23%) 46 (35%)
Regionality	US author non-US author	59 (45%) 73 (55%)
Funding source	industry funding non-industry funding both no funding not reported	34 (26%) 50 (38%) 20 (15%) 1 (<1%) 27 (20%)
COI statement	at least one trialist declares having a COI no trialist declares having a COI COIs not declared	32 (24%) 21 (16%) 79 (60%)
Risk of bias (RoB) characteristics	random sequence generation	3 = high risk (2%) 73 = low risk (56%) 56 = unclear risk (42%)
	allocation concealment	3 (2%) 47 (36%) 82 (62%)
	blinding of participants and personnel	9 (7%) 69 (52%) 54 (41%)
	blinding of outcome assessors	2 (1%) 104 (79%) 26 (20%)
	incomplete outcome data	25 (19%) 84 (64%) 23 (17%)
	overall risk of bias	35 (27%) 23 (17%) 74 (56%)

DISCUSSION

To our knowledge, this is the first meta-epidemiological study systematically investigating the influence of trial characteristics in randomized trials of testosterone therapy in men. We found no clear evidence that any trial characteristic was associated with treatment effects in the trials; all of the associations were statistically non-significant, with wide confidence intervals making it impossible to rule out a positive, negative or null association. Further, for the most part we observed little or no between-meta-analysis heterogeneity in the ROR estimates, except in three cases, where the heterogeneity appeared to be driven by the results of a single meta-analysis (by Neto et al.; we were unable to determine anything unique about this meta-analysis that would explain why it had very different results).

The direction of several of the ROR associations was comparable to the direction observed in previous meta-epidemiological studies. For example, we observed larger treatment effect estimates in randomized trials with less than 50 participants compared with trials with 50 or more participants, which is in line with previous studies reporting stronger effect estimates in small to moderately sized trials [116]. Also, trials with high/unclear risk of bias due to blinding of outcome assessors had larger effect estimates than trials rated at low risk of bias, a finding similar to that observed by others [24, 117]. Furthermore, our findings suggest but do not confirm support for a potential *industry bias* (i.e. larger treatment effects when at least one investigator reported a conflict of interest), a finding comparable with other meta-epidemiological studies of industry ties with outcomes [118].

Our findings differ from that observed previously in several ways. The postulated *decline effect*, an association between year of study publication and reported effect size, was not statistically significant in our analysis. Despite the majority of testosterone therapy trials performed after the year 2000 (N= 102 vs. 30) and the tripled testosterone use in the US from 2001 through 2011 [119], we did not detect a field-specific *decline effect* over time. Similarly, we did not detect the previously suggested *US effect* of overestimated effect sizes from authors working in the United States [120]. Also previously reported associations between treatment effect magnitude and trial registration [121] or centre status [122] were not confirmed in the present analyses.

In our analyses, a substantial number of studies were rated at “high/unclear risk of bias” (ranging from 17% to 62% across the domains). This finding confirms the persistent high prevalence of incomplete reporting (previous research suggests that 89% of published randomized trials include at least one “unclear” risk of bias domain [123]) and stresses the need for improved reporting in trials.

A potential explanation for the statistically non-significant findings in our meta-epidemiological study is the relatively small number of included meta-analyses and randomized trials. Despite the fact that we systematically included all available, non-overlapping meta-analyses of testosterone therapy in men in the present study, the potential risk of being underpowered cannot be ruled out. Thus, it is possible that a larger meta-epidemiological study including more testosterone therapy trials will be able to detect different or stronger bias associations between trial characteristics and treatment effect estimates.

Alternatively, the small magnitude of the investigated characteristics might also reflect their minor discipline- and topic-specific impact in the field of testosterone research.

A key strength of our study includes the use of systematic methods to minimise error in the identification and selection of meta-analyses, and collection of data from meta-analyses and trials. In addition, unlike many previous meta-epidemiological studies, we investigated the influence of a large number of (n=13) of trial characteristics, several of which have been underexplored. However, our findings must be considered in light of some limitations. We did not contact the trialists for clarification about any missing or unclear information in the trial reports. Therefore, the associations between trial characteristics and treatment effects reflects what was *reported* in the trial report, not necessarily what was *done* by the trialists.

CONCLUSION

The present meta-epidemiological study underlines the necessity for complete reporting to assess the safety and efficacy of testosterone therapy in men. Additionally, authors of systematic reviews and meta-analysis of testosterone trials should carefully consider potential characteristics that may bias the results of the included studies. Given the unquestionable importance of well-designed and -conducted randomized trials for the production of high-quality evidence, future trials on testosterone therapy, should not only be adequately performed but also transparently reported.

Chapter 4

A randomised trial of an editorial intervention to reduce spin in the abstract's conclusion of manuscripts showed no significant effect

Mona Ghannad

Bada Yang

Mariska Leeflang

Adrian Aldcroft

Patrick M. Bossuyt

Sara Schroter

Isabelle Boutron

Journal of Clinical Epidemiology 2020;130:69-77

Supplementary data to this article can be found online at:
https://doi.org/10.1016/j.jclinepi.2020.10.014

ABSTRACT

Objective To estimate the effect of an intervention compared to the usual peer-review process on reducing spin in the abstract's conclusion of biomedical study reports.

Study Design and Setting We conducted a two-arm, parallel-group RCT in a sample of primary research manuscripts submitted to *BMJ Open*. Authors received short instructions alongside the peer reviewers' comments in the intervention group. We assessed presence of spin (primary outcome), types of spin, and wording change in the revised abstract's conclusion. Outcome assessors were blinded to the intervention assignment.

Results Of the 184 manuscripts randomised, 108 (54 intervention, 54 control) were selected for revision and could be evaluated for the presence of spin. The proportion of manuscripts with spin was 6% lower (95% CI: 24% lower to 13% higher) in the intervention group (57%, 31/54) than in the control group (63%, 34/54). Wording of the revised abstract's conclusion was changed in 34/54 (63%) manuscripts in the intervention group and 26/54 (48%) in the control group. The four pre-specified types of spin involved: (i) selective reporting (12 in the intervention group versus 8 in the control group); (ii) including information not supported by evidence (9 versus 9); and (iii) interpretation not consistent with study results (14 versus 18); and (iv) unjustified recommendations for practice (5 versus 11).

Conclusion These short instructions to authors did not have a statistically significant effect on reducing spin in revised abstract conclusions and, based on the confidence interval, the existence of a large effect can be excluded. Other interventions to reduce spin in reports of original research should be evaluated.

Study registration osf.io/xnuyt

INTRODUCTION

Ethically, research findings should be disseminated completely and accurately [124]. However, authors may intentionally or non-intentionally misrepresent or overinterpret their results, which is referred to as 'spin' [10, 12]. Through "spin", the effectiveness of interventions is typically presented in a more favourable way than is justified by the study findings.

Several studies have documented a high prevalence of spin in the biomedical literature [10, 12, 13, 15-17, 125-127]. A recent systematic review of 35 reports evaluated the prevalence of spin in clinical trials, observational studies, diagnostic accuracy tests, systematic reviews and meta-analyses [23]. The median prevalence of spin was 67% (range: 10% - 84%), with the highest prevalence of spin found in trials [23, 128].

A consequence of biased representation of results in scientific reports is that the published literature may suggest stronger evidence than is justified by the study findings, thus contributing to the increase in false discovery rate [2]. This justifies efforts to prevent or reduce spin due to selective, biased and incomplete reporting in biomedical research. Researchers, authors, peer reviewers and journal editors undoubtedly share responsibility.

Meta-research – research on research – can generate solid evidence to inform editorial policies and interventions [128, 129]. Many studies have been published on peer review and the publishing process, and some interventions have been developed and implemented to improve the quality of peer-review [129, 130]. However, to date, there has been no intervention designed to mitigate or reduce the prevalence of spin in biomedical literature. Thus, we developed a specific editorial intervention to reduce spin and conducted a randomised controlled trial to evaluate its impact.

Given that the abstract and its conclusions are often the most widely read part of scientific article [14], we considered a concise intervention (i.e., a set of short instructions for authors) to reduce spin in the abstract conclusion of primary research and research synthesis manuscripts that are submitted for publication. The intervention was developed in collaboration with editors, epidemiologists with expertise in the field of spin, and patient representatives. We aimed to obtain an initial estimate of the effectiveness of this intervention, compared to the usual peer-review process, at a large general medical journal, in a randomised trial.

Methods

We conducted this study in accordance with the study protocol registered on the Open Science Framework in August 2019 (osf.io/xnuyt).

Design and setting

This was a two-arm parallel-group randomised controlled trial of research manuscripts submitted for publication to BMJ Open, a general medical journal.

All research manuscripts consecutively submitted between June 1st and October 9th 2019 to *BMJ Open*, and sent for peer review, were screened for eligibility until the planned sample size was achieved.

Research manuscripts sent for peer review were randomly allocated to the intervention or control group. An investigator (MG) indicated which of the manuscripts were assigned to the intervention group using a flag and an additional note in the journal's electronic tracking system (ScholarOne). This informed the handling editors which manuscripts needed to include additional instructions in the decision emails sent to authors alongside the peer reviewers' reports. In the control group, the handling editors sent recommended revisions and reviewers' comments to authors in their usual manner. In the (flagged) intervention group, editors sent additional instructions to authors as part of the decision letter alongside the peer reviewers' comments, inviting the authors to check for and remove spin in the abstract of their revised manuscript.

The intervention was only applied if the editorial decision was 'revise'; manuscripts that were rejected after initial peer review were excluded. Manuscripts without an editorial decision by February 12th 2020, the timeframe that we monitored the decision status, were also excluded. All manuscripts selected for revision were resubmitted by the authors within 100 days of original submission, in compliance with the BMJ Open time requirement for authors.

Eligibility criteria

Quantitative research manuscripts (primary research or research synthesis) submitted to BMJ Open were eligible. Narrative reviews, protocols, qualitative studies and mixed methods studies with a qualitative component were not eligible, as there are currently no reference studies characterising spin in these publications.

We considered manuscripts that had passed the initial editorial assessment and had been sent for peer review. Only manuscripts that were not rejected after peer review and that were resubmitted by the authors within 100 days of original submission were ultimately included in the analysis.

The intervention: Instructions to reduce spin

The intervention was developed in close collaboration with editors, epidemiologists with expertise in the field of spin, along with input from patient representatives.

To develop the intervention, we considered different factors. We defined spin as reporting that contains 'misrepresentation, over-interpretation, or inappropriate extrapolation of results' in the abstract's conclusion [10, 23]. We decided that the best time point to intervene was at the revision stage of the manuscript as the author may be more likely to follow the request of editors and peer reviewers to get their manuscript accepted. We chose to add the instructions under the heading "Editorial Requirement" in the decision letter email sent to authors with the peer reviewers' comments. We hoped authors would be more likely to follow editor's requests knowing it is the editor who makes the final decision. We intended the instructions to authors to be concise and easy to understand, applicable to all types of studies, to facilitate future wider implementation, and to target prevalent types of "spin" that are seen in abstract's conclusions. It is important to note that spin may also have been present in the main text conclusions, but our instructions were not targeted at removing spin in that part of the manuscript.

We developed the initial list of instructions for authors (Appendix A), based on a classification of types of spin in a recent systematic review by Chiu et al. [23]. The selected statements targeted the most prevalent types of spin identified generic to all empirical study designs that we expected to see in a general medical journal.

Patient and public involvement

To include the patient's perspective in the research process, we shared the initial draft of the instructions to authors with three of The BMJ's patient and public reviewers. These reviewers were invited to share their thoughts on the importance of the items proposed, and any additional spin practices and phrases they frequently observe in abstract's conclusions that can mislead readers and patients. We then revised the instructions based on their feedback and piloted it with three PhD students in the clinical department at the University of Amsterdam to generate the final version of the instructions to authors (Box 1).

Box 1: Instructions for authors to reduce spin

With your revision, please **carefully check and confirm that your abstract conclusion**:

- ☐ Focuses on the primary aim and primary outcome(s) of your study (e.g. avoids selective reporting or focusing on subgroup, secondary or exploratory analysis)
- ☐ Includes only information supported by evidence and in accordance with the results of your study (e.g., avoids claims of significant effect with non-statistically significant outcomes)
- ☐ Avoids any interpretation that is not consistent with your study design or the results (e.g., avoids causal claims in studies with non-randomised designs, or inappropriate extrapolation of results to different populations or outcomes than investigated in your study)
- ☐ Avoids recommendations for practice not justified by study findings (e.g. calls for action that have not been evaluated)

Control group: usual practice

Manuscripts allocated to the control group received the usual decision letter email without the additional instructions (Appendix B).

Random assignment and allocation concealment

Eligible manuscripts were assigned with a unique individual identifier (ID) by one researcher (MG), and allocated to the assigned group at three timepoints, until the planned sample size was reached. The random assignment sequence was generated by researchers not involved in the selection or assessment of the manuscripts (SS, PB). The researchers (SS, PB) kept a record of the randomised list of manuscript IDs for the purpose of verification, and the lead investigator (MG) was given access to the copy.

The sample of eligible manuscripts was prepared by MG in July and December 2019. The initial set of eligible manuscripts allocated (n=154) was selected from a consecutive sample submitted to BMJ Open between June and July 2019. The subsequent set of eligible manuscripts allocated (n= 30) was selected from a consecutive sample submitted between September and October 2019. Decisions about eligibility were made by MG, prior to and without knowledge of the group allocation.

Random allocation of the first 154 manuscripts was performed in a single block as follows. A first block of 150 eligible manuscripts was allocated by SS to assigned groups (allocation ratio 1:1) on July 22nd, 2019. In the first 150 manuscripts assigned, two manuscripts allocated to the intervention group did not receive the additional spin instructions in the decision letter, thus resulting in a balance of 73 to the intervention and 75 to the control group. To account for this, on July 30th, 2019, 4 additional manuscripts selected consecutively from the initial list of eligible manuscripts were allocated by PB to assigned groups (3 to the intervention group and 1 to the control group) for a balance of 76 manuscripts per group.

We monitored the editorial decision status of manuscripts in the first sample allocated until mid-December, 2019. Of the 154 manuscripts randomised, 93 manuscripts were selected for revision. To reach our planned sample size, 30 additional manuscripts were allocated in a second single block by SS to assigned groups (15 manuscripts per group) on December 20th, 2019. Of the 30 additional manuscripts randomised, 15 manuscripts had been selected for revision by February 12th, 2020.

In the first sample of 154 manuscripts, we assessed all eligible study designs sent for peer review, regardless of whether any of the peer reviewers had already completed their review of the manuscript. In an effort to decrease the waiting time during the editorial decision period, we only assessed eligible study designs sent for peer review in the second sample of 30 manuscripts if at least one peer reviewer had already completed their review.

Blinding

To minimise bias, outcome assessors (BY, ML) were blinded to group allocation. Outcome assessors were also blinded to authors' names and affiliations, and the editorial decision letter

with the peer reviewers' comments. The editors at BMJ Open sent the intervention emails, and therefore could not be blinded to group allocation.

After submitting a manuscript, all authors received an automated email confirming it had been received and that BMJ Open is striving to improve its peer review process and their manuscript may be included in a study. However, authors were not informed of the purpose of this specific study.

Outcome assessment

For each included manuscript, two investigators (BY, ML), trained in the field of epidemiology, with expertise in systematic review methodology and critical appraisal, independently assessed spin in the abstracts' conclusions of the revised manuscripts. As previously stated, we defined spin as reporting that contains 'misrepresentation, over-interpretation, or inappropriate extrapolation of results' in the abstract conclusion [10, 23]. BY and ML formally assessed the revised abstracts' conclusions against each of the four types of spin, using the data extraction form presented in Box 2. This assessment was primarily based on the abstract alone, but if further information was deemed necessary the full-text manuscript was consulted. To ensure consistency of coding, we conducted a pilot of 10 manuscripts prior to the start of the study. The agreement for the pilot between the two investigators was 80% or higher for each of the four spin items. The level of agreement between the two investigators (BY, ML) for scoring the primary and secondary outcomes are outlined in Appendix C.

All manuscript versions were available on ScholarOne. MG provided BY and ML the de-identified abstracts of the submitted manuscripts and the de-identified abstracts and full texts of the revised manuscripts following peer review. After data extraction, BY and ML discussed all discrepancies, to obtain consensus.

Box 2: Outcome assessment form

Spin items to look for in abstract conclusion	Present?
Selective reporting or focus on outcomes or analyses other than the primary outcomes and analysis. e.g., selective reporting or focus on subgroup, secondary or exploratory analysis.	☐
Information that is not supported by evidence or in accordance with the study results. e.g., Claiming a significant effect with non-statistically significant outcomes.	☐
Interpretation that is not consistent with the study design or the results. e.g., causal claims in non-randomised study designs (without any efforts to improve exchangeability/comparability), or inappropriate extrapolation of results to different populations than were investigated in the study.	☐
Recommendations for clinical practice not justified by study findings (e.g. calls for action that have not been evaluated).	☐

Outcomes

Primary outcome

The primary outcome was the presence of spin in the revised abstract conclusion (yes/no). If an abstract contained one or more of the four pre-specified types of spin, we classified the abstract as 'spin present'.

Secondary outcomes

The secondary outcomes were: the presence of each type of spin in the revised abstract conclusion (i.e., 4 secondary outcomes), and a change in wording in the revised abstract conclusion from what was originally submitted (yes/no). It is important to note that if the wording of the abstract was changed by one or two words, in a manner that did not alter the conclusion, we scored it as unchanged.

Ethics and trial registration

This study was exempt from an evaluation by the local ethics committee at Amsterdam UMC as it did not recruit patients. We additionally asked BMJ for any ethical concerns, whereupon The BMJ confirmed that the work is part of quality improvement of its processes. All submitting authors at BMJ Open were notified that BMJ has a programme of research into peer review and that their paper may be entered into a study. MG was given access to confidential manuscript data under a confidentiality agreement. The study protocol was registered at Open Science Framework (OSF) prior to recruitment.

Statistical analysis

Sample size justification

We allowed for detecting a 15% absolute difference in the proportion of articles with spin (primary outcome) between the intervention and control group (a reduction from 30% to 15%) with a power of 80% and a two-sided alpha of 5%. This required a sample size of at least 98 manuscripts. The estimate of 30% for the prevalence of spin was based on a recent systematic review of spin by Chiu et al. [23].

To allow for manuscripts not invited for revision and not resubmitted within 3 months of the editorial decision letter, we randomised 184 manuscripts in our study, in order to achieve a target sample size of 108 manuscripts.

Analysis

We expressed the effect of the intervention as the absolute difference in the proportion of manuscripts with spin, with a 95% confidence interval based on the normal distribution approximation. All manuscripts resubmitted after invitation for revision were analysed in the group to which they were allocated. The statistical analysis was performed using R software (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Overall, 108/184 (59%) of allocated manuscripts were selected for revision, and could be evaluated for spin (Figure 1). Five manuscripts allocated to the intervention group did not receive the additional instructions to reduce spin but were included in the intention-to-treat analysis.

Of the 108 evaluated manuscripts, 84 reported an observational study, 17 reported a systematic review, meta-analysis or overview of systematic reviews, while 5 reported a randomised controlled trial. Two manuscripts did not specify a study design in the abstract. The baseline characteristics of the evaluated manuscripts are shown in Table 1.

Primary outcome

The proportion of manuscripts with spin in the revised abstract conclusion was 6% lower in the intervention group (31/54, 57%) than in the control group (34/54, 63%; 95% CI: 24% lower to 13% higher); the difference was not statistically significant (Table 2).

Secondary outcomes

Two of the four types of spin were more often observed in the control group: 'interpretation not consistent with study results', and 'unjustified recommendations for practice'. However, 'selective reporting' was more frequently observed in the intervention group; there was no difference between the groups in the proportion of manuscripts 'including information not supported by evidence' (Table 2).

The wording of the revised abstract conclusion was changed by authors in 63% (34/54) of manuscripts in the intervention group compared with 48% (26/54) of manuscripts in the control group, a difference of 15% more in the intervention group, [95% CI: 4% lower to 33% higher] (Table 2).

Although we did not formally assess spin in the initially submitted manuscript, it was possible to identify if a specific type of spin was present in the submitted version. In a number of cases, authors introduced slight modifiers but failed to remove spin. As an example, in one meta-analysis, the authors revised the conclusion from *"The meta-analysis found a higher prevalence of (…) compared to the general population, with substantial regional difference."* to *"The meta-analysis found a relatively high prevalence of (…) although there was significant heterogeneity between gender and across regions.*" However, despite the change in wording, the two types of spin detected in the revised abstract conclusion (selective reporting and unjustified recommendations for practice) remained unchanged from the initial submitted version.

Figure 1. Spin intervention flow diagram.

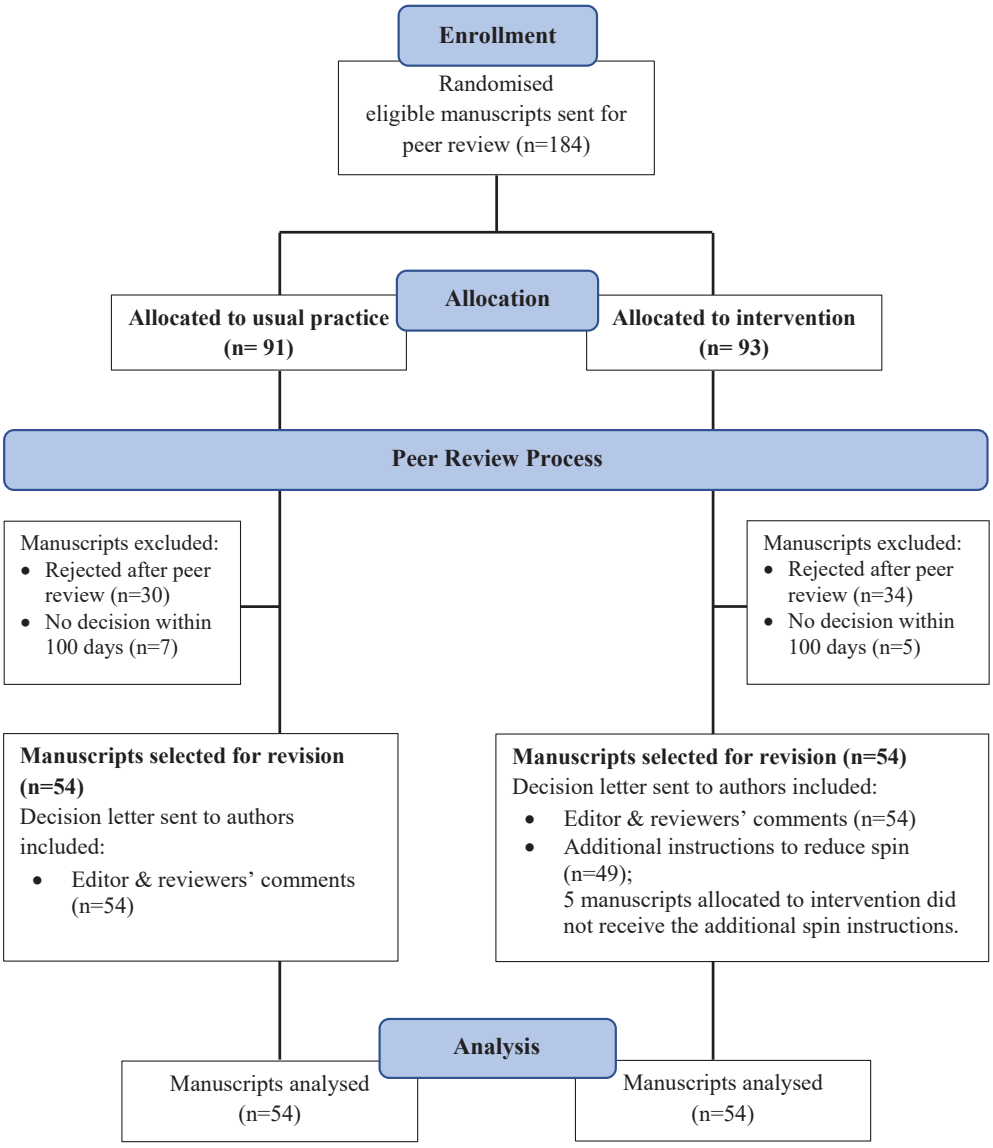


Table 1. Baseline characteristics of the evaluated manuscripts.

	Intervention, n=54 n (%)	**Usual practice, n=54 n (%)**
Study design		
Observational studies	42 (78%)	42 (78%)
Evidence synthesis	8 (15%)	9 (17%)
RCT	2 (4%)	3 (6%)
Not reported	2 (4%)	0 (0%)
Conflict of interest		
No	43 (80%)	46 (85%)
Yes	10 (19%)	5 (9%)
Not reported	1 (2%)	3 (6%)
Funding		
Nonprofit	38 (70%)	35 (65%)
No funding	15 (28%)	16 (30%)
For profit	0 (0%)	2 (4%)
Mix (nonprofit and for profit)	1 (2%)	0 (0%)
Not reported	0 (0%)	1 (2%)
Primary subject		
Clinical research	21 (39%)	32 (59%)
Public & global health, epidemiology, research methods	18 (33%)	9 (17%)
Health services research	12 (22%)	10 (19%)
Nutrition and metabolism, sports and exercise medicine, occupational and environmental medicine	3 (6%)	3 (6%)

Table 2. Frequency of overall spin, types of spin, and changes to wording in the abstract's conclusion of revised manuscripts.

	Intervention, n=54 n (%)	**Usual practice, n=54 n (%)**	**Risk difference (%) [95% CI]**
Overall spin present	31 (57%)	34 (63%)	-6% [-24, 13]
Selective reporting	12 (22%)	8 (15%)	7% [-7, 22]
Including information not supported by evidence	9 (17%)	9 (17%)	0% [-14, 14]
Interpretation not consistent with study results	14 (26%)	18 (33%)	-7% [-25, 10]
Unjustified recommendations for practice	5 (9%)	11 (20%)	-11% [-24, 2]
Wording change	34 (63%)	26 (48%)	15% [-4, 33]

DISCUSSION

The editorial instructions for reducing spin in the abstract's conclusion of research manuscripts tested in this study may have led authors to revise the wording of their conclusions, but the changes in wording did not lead to a significant decrease in the prevalence of spin. In a number of cases, the changes to wording led to the introduction of modifiers but did not remove the presence of spin.

To our knowledge, this is the first randomised controlled trial of an editorial intervention to reduce spin. Despite the fundamental role of peer review, current editorial practices are not preventing the high prevalence rate of reporting biases and spin [12] [15] [23]. Various efforts have been undertaken to mitigate the issue of bad reporting as it directly impacts patient care [131, 132]. The EQUATOR Network, launched in 2008, holds a library of more than 400 reporting guidelines to help authors, peer reviewers, editors, and other stakeholders improve the reporting of published articles [132, 133]. However, despite initial evidence of effectiveness for some of the reporting guidelines [134, 135], there are also associated challenges with their implementation that are being addressed [132]. Simply disseminating instructions or guidelines may not be sufficiently effective. In a recent study evaluating the appropriate use of reporting guidelines of health research, Caulley and colleagues showed that major reporting guidelines are frequently used inappropriately, indicating that authors may need additional education [136]. Automated tools aiming to improve reporting are starting to appear to help authors and editors [137] [132].

The validity of our study may have been affected by potential limitations. First, we only evaluated the first resubmitted version after review and further changes may have been made to other versions before acceptance. Second, the authors were asked but not required to address the instructions and these could have been overlooked by authors. Third, we only evaluated the presence of spin in the revised abstract conclusion. Although the randomization process would guarantee that all different types of articles with spin in both groups only differ based on chance, in a relatively small trial such as ours an assessment of spin before and after the revision could have been informative for evaluating the effects of the editorial intervention. Fourth, for feasibility purposes, the editors could not be blinded to allocation of the intervention. As editors aim to keep the journal to a high standard of reporting, they may have also addressed spin practices in their decision letter in the control group, thus potentially diluting the effect of the intervention. We designed our intervention to be concise and applicable to all types of study designs, which may have rendered the instructions less specific and clear for authors in identifying and removing spin in their abstract conclusion. Our pre-defined primary outcome was the complete absence of spin, whereas it is possible that some authors made word changes to reduce the level of spin and this was not assessed. Therefore, prior to completely dismissing the intervention based solely on our small study, the effectiveness of this intervention in reducing the *level* of spin should be tested in a larger study, particularly given the relative ease of implementation across a large sample of manuscripts and that we demonstrated its feasibility.

The success of any intervention aiming to improve the quality of research articles depends on their effectiveness, feasibility, adoption and appropriate use [2]. Future research could also

evaluate the reasons why authors fail to follow editorial instructions such as those included in our intervention. For instance, it is possible that authors are not fully aware of what spin is, and why it is important to avoid it. Van der Steen and colleagues theorise that "*the motivations and subsequent behaviour leading to reporting bias, may result from a natural tendency to publicise our successes*" [138]. In addition, authors may have been trained in a culture where spin practices are omnipresent, and regarded as a necessary element in the reporting [138]. Recognition of spin could be facilitated if instructions for authors could contain specific examples of typical wording to be avoided when writing conclusions. Qualitative studies to better understand the perceptions of authors when receiving the intervention could also be helpful to improve its content. Based on a better understanding of possible reasons, we can develop more effective interventions to restore balance between study results generated by research finding and conclusions made by study authors. There are other stakeholders in the publication process; studies could also evaluate the effect of peer reviewers being explicitly instructed to check for spin and/or editors highlighting spin in a standalone section of feedback.

CONCLUSIONS

Spin in biomedical research is a major contributor to avoidable waste in research and, through exaggerated claims, may endanger the health of patients or encourage the introduction of ineffective interventions [10, 16]. Despite the study being designed to detect a decrease in spin, we found no evidence that it did. Thus, we need to consider other methods to inform authors of the manifestations and impact of spin.

ACKNOWLEDGEMENT

We would like to thank the patient and public reviewers at *The BMJ* (Rebecca Harmston, Andrew Demaine, and Melissa Hicks) for taking the time to share their thoughts and perspectives on the initial author instructions. We would like to thank the PhD researchers (Stijn Jonge, Frits Mulder, and Victorine Roos) for piloting the author instructions and for their comments.

Chapter 5

Publications with high Altmetric scores

Mona Ghannad
Reza Ramezan
Patrick M. Bossuyt
Jeffrey K Aronson
Elizabeth Wager
Jon Brassey
Carl Heneghan

Submitted

ABSTRACT

Background: Alternative metrics have been developed to measure the attention publications receive from social news media, and blogs.

Objectives: We aimed to discover which types of studies reported in recent research articles in medical journals receive the highest Altmetric scores, among those generating attention in Altmetric data sources.

Methods: We identified 679 primary research articles through a weekly search of PubMed, exploring the "big five" medical journals, and a daily email from EvidenceUpdates of suggested references, from those published between October 2018 and September 2019. The Altmetric score was manually recorded for each of the articles once a month. We limited our selection to articles with a manually recorded Altmetric score of more than 50. For each article we extracted study design, intervention type, journal, journal impact factor, journal category, and direction of conclusion. We developed a model for the growth of the Altmetric score of an article over time. We performed analysis of variance to evaluate the association of high online media attention with intervention type, adjusting for journal category, study design, and journal impact factor.

Results: We included 324 primary research articles, with a median Altmetric score of 184 [Interquartile Range, 111 – 378]. Journal category and impact factor were significantly associated with adjusted Altmetric score ($P < 0.00001$). Nutritional intervention (median = 4.69) accumulated significantly higher adjusted median Altmetric score compared to lifestyle and environmental (1.47), pharmacological (1.05), and other interventions (0.82).

Conclusions: Of the publications that generated an Altmetric score of 50 or more, reports of evaluations of nutritional interventions are mentioned more often than other types of interventions in news media, social media, and other online sources covered by Altmetric. This seems to indicate that interventions that a wide range of readers can apply in their daily life attract more attention and are discussed more often than other interventions in medicine and public health.

BACKGROUND AND RATIONALE

Alternative metrics are metrics and qualitative data that were developed to be complementary to traditional, citation-based metrics. The Altmetric system attempts to measure how often journal articles and other scholarly outputs are discussed and used around the world. This includes the attention an article receives from social and news media, such as Twitter and Facebook, and in research blogs.[26, 139] Altmetric scores are widely consulted by journal editors, researchers, and potential readers to identify articles that are generating interest. However, there is a debate about how well a score reflects the quality of an article.[140, 141]

It is worrisome when biased interpretations of findings are naively accepted as facts, without careful scrutiny of the methods and results, misleading other researchers and the public. Diet studies are particularly vulnerable to subjective interpretation and biased misrepresentation.[141, 142] Most often, cohort study designs do not allow inferences about causality and are potentially subject to confounding.[141, 142] In addition, the effect sizes in studies of diet and health outcomes are often small.[141, 142]

This is of particular concern, since lifestyle factors and their association with health and longevity have always been of great public interest, and still generate significant attention from social and news media.[26, 27] A recent example was a dietary guideline on red meat published in the *Annals of Internal Medicine*, which within 18 days of publication had an Altmetric score of 3705.[143] Widespread interest in human health and disease is highlighted by the annual published lists of the 100 articles with the highest Altmetric scores.[26] Analysis of the top 100 articles in 2015 showed that the most popular subject was medical and health science (n = 36).[26] Further inspection of these articles showed that the most frequent theme was diet (11/36).[26]

Colquhoun and Plested contested the usefulness of Altmetric scores through the example of an article with the second highest Altmetric score of all articles published in 2013 in the *New England Journal of Medicine*, titled “Primary Prevention of Cardiovascular Disease with a Mediterranean Diet”.[141, 144] The journal’s press release promoted the article in a tweet, as follows “Our new post focuses on trial that shows Mediterranean diet results in less cardiovascular events than low-fat diet”.[141] However, both the title of the paper and the press release misinterpreted the actual findings of the research, which were, according to Colquhoun and Plested, that “the diets had no detectable effect on myocardial infarction, no effect on death from cardiovascular causes, and no effect on death from any cause”.[141] [144] The only effect was that the diet reduced the risk of stroke for the groups assigned to a Mediterranean diet with extra-virgin olive oil or to a Mediterranean diet with nuts.[144] The article, originally published in 2013 [144] was retracted in 2018, because of methodology (i.e., “irregularities in the randomization procedures)”, and an updated report was published in 2018.[145]

We wondered whether the high level of interest in dietary interventions and differences is a persisting phenomenon, and performed a new analysis of the Altmetric scores of nutritional studies, relative to other interventions. For this purpose, we evaluated articles published in medical journals in 2019 with an Altmetric score of more than 50.

METHODS

Dataset compilation strategy

Our initial sampling dataset was compiled by tracking the Altmetric scores of primary research articles identified through a weekly search of PubMed, exploring the "big five" medical journals (*New England Journal of Medicine*, *Lancet*, *Annals of Internal Medicine*, *BMJ*, and *JAMA*), and a daily email from EvidenceUpdates of suggested references, published in English between 4 October 2018 and 3 September 2019. The Altmetric score was manually recorded for each article on a tracker sheet once a month. The initial dataset was compiled by one researcher (JB) and included 679 articles.

Study inclusion

From this initial dataset, we subsequently preselected for inclusion 491 articles with an Altmetric score of more than 50. We then downloaded the online publication date (ePub), the Altmetric score, and the historic Altmetric score by using the PubMed unique identifier (PMID) for the 491 included articles (Altmetric search date 8 November 2019) using R software (R Foundation for Statistical Computing, Vienna, Austria). The historic Altmetric record is available for 1 – 7 days, 1 month, 3 months, 6 months, and 1 year from the Altmetric search date (8 November 2019). As we could only retrieve the historic Altmetric records for articles published after 8 November 2018, we chose to exclude articles published in 2018. Two studies were further excluded, as their Altmetric score was not available to download. Three other studies had to be excluded, as the downloaded Altmetric score turned out to be less than 50. We analysed the remaining 324 articles, published between January and September 2019.

Data collection

We developed and piloted our data extraction form on a sample of 150 articles. The following data were collected: study design, intervention type, journal, journal impact factor, journal category (general medical, specialty journal), and direction of conclusion.

Journal impact factor was downloaded and retrieved from Journal Citation Report in 2019 by Clarivate.

Two researchers (JB, MG) independently assessed study design and direction of conclusion for all articles included in the analysis. Disagreements were resolved by a third researcher (JKA). We defined "general medical journals" according to the definition provided in Wikipedia: "A general medical journal is an academic journal dedicated to medicine in general, rather than a specific field of medicine."[146] Likewise, "specialty journals" were defined as journals dedicated to a specific field of medicine.

Assessment of direction of conclusion

The direction of the conclusion was assessed by scoring the abstract conclusion or summary statements followed by presentation of the results. We used the abstract to evaluate the conclusion section, as it was considered representative of the authors' main conclusions. Conclusions were categorized as positive, negative, mixed, or not applicable. Conclusions were

scored as "positive" if the summary statements referred to the evaluated intervention using words such as effective, beneficial, or impactful, or asserted that they were associated with events or outcomes of interest, with no evidence of any or significant harm (e.g., "the study strongly supports" or "X was significantly associated with Y"). Conclusions were scored as "negative" in the absence of any observed beneficial effects or associations (i,e., neutral outcomes, with neither benefits nor harms), or in the presence of any adverse effects or harms for the intervention evaluated (e.g. "use of X is not associated with mortality benefits" or "X provided no important benefit compared with Y, and probably carries a small risk of serious harms"). Conclusions were scored as "mixed" if the summary statement contained both positive and negative associations or effects of an intervention, as previously described, or stated a positive association or effect for one of the evaluated interventions but no association or effect for other evaluated interventions, or concluded a very small association or effect of the evaluated intervention that did not exceed currently established benefits. Conclusions were scored as "not applicable" if the summary statement only restated the results without any (positive or negative) remarks on the association or effect of the evaluated intervention.

Altmetric score adjustment

The Altmetric Attention score is derived from the source data, representing a weighted count of the amount of attention received by a research article (e.g. news reports, blogs, Twitter, Facebook, Reddit, YouTube, policy documents, patents, Wikipedia, Peer review from Publons and Pubpeer, F1000, Open Syllabus, and Stack Overflow).[147] The Altmetric score is not normalized and does not have a scale, although a score of 0 indicates that a publication has not attracted any attention [148] and a score of 20 or more corresponds to articles in the top 5%.[149]

The attention that an article receives is measured from the date of publication.[26] Time influences media attention, and articles generally receive the most online attention during the first few months of publication.[139] As the historic Altmetric record is available only retroactively from the Altmetric search date and at long time intervals (at 1 month, 3 months, 6 months, and 1 year after 8 November 2019), it is not possible to obtain the Altmetric score for the same post-publication time. Potentially, articles published earlier may have accumulated higher Altmetric scores simply because of having more exposure time than articles published later. To account for differences in post-publication exposure time, we fitted a model to adjust for the time between the date of publication and the Altmetric score (see below).

Statistical analyses

We used ANOVA to compare reports describing nutritional, lifestyle and environmental, pharmacological, and other types of intervention. To account for confounding and to increase precision, we adjusted our analyses for the following pre-specified variables: (1) study design, (2) intervention type, (3) journal, (4) journal impact factor, (5) journal category, and (6) direction of conclusion. Log transformation of the adjusted Altmetric scores was required to better approximate normality and the constant variance assumptions of the ANOVA model.

To investigate the Altmetric score plateau trend statistically, we fitted a regression model of the form

$$y_t = a - \frac{b}{t}$$

where t is the time since publication and y_t is the Altmetric score after t days from publication. For a positive value of b, the shape of the function $y_t = a - \frac{b}{t}$ illustrates an increasing function with a plateauing shape. Out of the 324 articles in this study, there were enough datapoints in 312 publications to fit the model. The remaining 12 publications either had a constant Altmetric scores over the measured period or were too recent, i.e. not enough time had passed since their publication at the time of data collection. The aforementioned model provided a good fit to the data; the average and median R^2 were, respectively, 0.90 and 0.97. We have also noticed that 90% of the R^2 values were above 0.70 (Appendix B).

The statistical analysis was carried out with R software (version 3.6.2, R foundation for Statistical Computing).

RESULTS

Characteristics of included articles

The 324 included articles were published in 47 journals. One quarter (n=80, 25%) of the articles were published in the *New England Journal of Medicine (NEJM);* 43% of the journals were specialty journals and the rest were general medical journals. The median impact factor of the journals was 28 [interquartile range 19 to 59].

In all, 201 (62%) articles reported on randomized controlled trials, 62 (19%) presented evidence syntheses, and 61 (19%) were observational studies. Most of the interventions were pharmacological (n=159, 49%), and in 207 (64%) of the articles assessed, the abstract conclusion was in favour of the study intervention. Table 1 provides key characteristics of the included studies. Appendix A shows the journals with the included number of articles.

Table 1. Characteristics of included articles

Article characteristics	Total (n=324)
Study design	
Randomized controlled trial (RCT)	201 (62%)
Evidence synthesis (i.e., systematic review/meta-analysis, pooled individual patient data, guidelines)	62 (20%)
Observational studies	61 (19%)
Intervention type	
Lifestyle & Environmental	17 (5%)
Nutritional	24 (8%)
Other (risk factors, surgery, health services research, devices, pharmacological & devices combined, vaccines, education, non-drug, procedural, program)	124 (38%)
Pharmacological	159 (49%)
Journal category	
Specialty	138 (43%)
General medical	186 (57%)
Direction of conclusion	
Positive	207 (64%)
Negative	67 (21%)
Mixed	45 (14%)
N/A	5 (1%)

Analysis of Altmetric scores

The electronically retrieved Altmetric scores of the 324 articles in our dataset ranged from 51 to 9208, with a median of 184 [interquartile range 111 to 378]. Figure 1 describes the overall distribution of Altmetric scores for the 324 articles.

The adjusted Altmetric scores ranged from 0.1 to 30.5, with a median of 0.8 [Q1–Q3, 0.4–1.6], and a mean of 1.6. Figure 1 and Figure 2 show, respectively, the overall distribution of Altmetric scores and a box-and-whiskers plot for the 324 articles. Nutritional interventions accumulated higher median Altmetric scores (median=4.69) compared to other types of interventions: lifestyle and environmental (1.47), Pharmacological (1.05), and other interventions (0.82). Figure 3 plots the means along with 95% bootstrap confidence intervals across the four different intervention categories.

Figure 1. Altmetric score distribution for articles (n=324) [Inset graph represents articles with an Altmetric score below 1000]

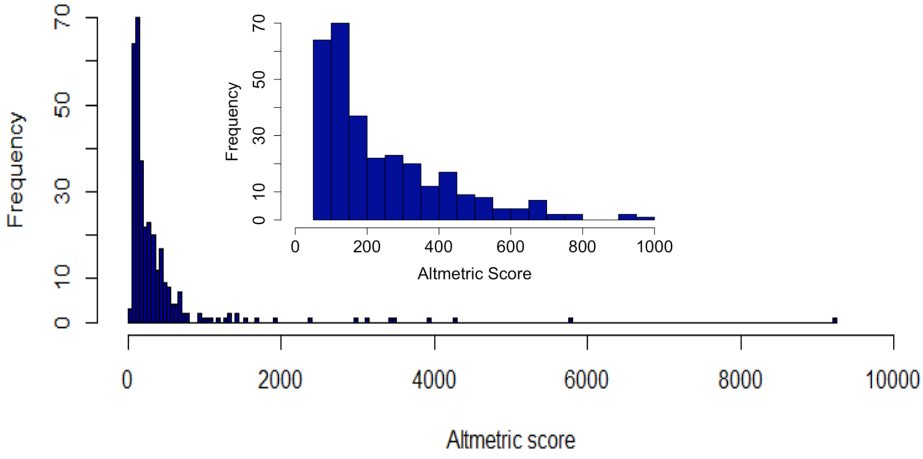


Figure 2. Altmetric score by intervention type.

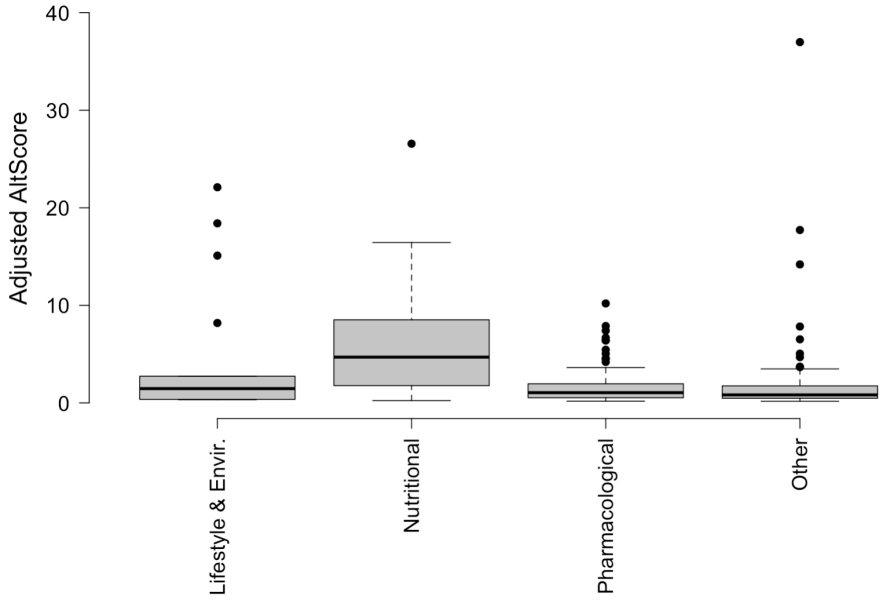
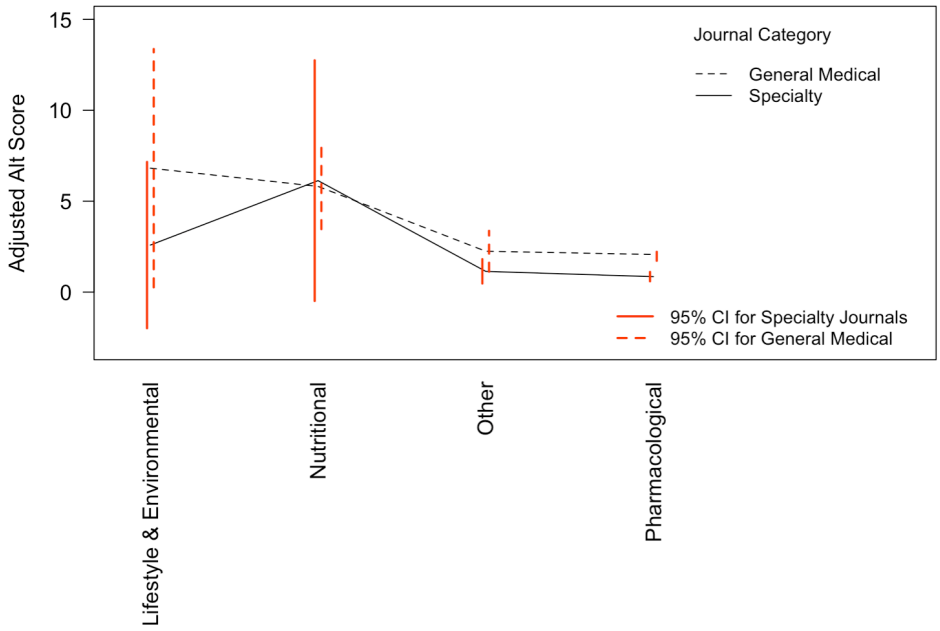


Figure 3. 95% confidence intervals for the mean of adjusted Atmetric score grouped by the levels of intervention and journal category.



Factors associated with Altmetric score

Intervention, journal category, and impact factor all showed statistically significant (P <0.00001) association with Altmetric score through an ANOVA model (Table 2). No significant effect modifier (interaction term) was identified.

Table 2. 3-way Analysis of Variance (n=324)

	DF	SS	MS	F-value	p-value
Intervention type	3	38.364	12.788	19.594	<0.00001
Journal Category	1	55.090	55.090	84.412	<0.00001
Impact Factor	1	29.241	29.241	44.0805	<0.00001
Residuals	317	206.885	0.653		

DISCUSSION

In our dataset of 324 studies with high Altmetric scores of 50 or above, published in 2019, studies describing a nutritional intervention had higher Altmetric scores than articles describing other types of interventions; this association persisted even after accounting for several potentially confounding variables. This confirms previous analyses and suggestions: nutritional intervention receive a lot of attention. These results show an association but not causation, particularly since the data were not collected by random sampling in an experimental design.

Our study had limitations. Since we analysed only articles with high Altmetric scores, our ability to identify factors that drive high Altmetric scores is limited. To account for differences in post-publication exposure time of articles in our dataset, we modelled adjusted Altmetric scores, which reflected the actual pattern of Almtetric score development over time after publication. However, we did not evaluate how the various Altmetric data sources contribute to the Altmetric score. Some previous studies have analysed several of the broad categories of Altmetric indicators and have noted important differences in data coverage across diverse Altmetric data.[150] These studies have also indicated uneven distributions of publication and article-level metrics across various research topics.[150]

Several of these studies highlighted intrinsic differences between the different Altmetric data sources and their relation to data sources outside social media (policy documents and peer review platforms).[150] We need to select datasets carefully before making any generalizable claims about drivers of the Altmetric score.[150]

We focused on compiling articles with Altmetric score over 50. As we do not know the prevalence of subjects published in medical journals, we cannot infer widespread preferential publication of nutritional interventions. Our study was not designed to determine whether journals preferentially select for publication studies that may be characterized as *hot research topics* (i.e., "current excitement about a topic")[150, 151] from an Altmetric perspective. Additionally, as we took a manual approach to our initial data set compilation strategy, rather than an automated approach, we cannot be sure that we have captured all eligible publications.

The results of our study are in agreement with the findings of Fang et al.[150] In a study of nearly 12.3 million Web of Science publications published between 2012 and 2018, they found that most Altmetric events are garnered by the fields of biomedical and health sciences.[150] Furthermore, of the 1796 micro-topics the study authors considered, 75 (4%) received more attention and were identified as '*hot research topics'*.[150] They detected the following as hot research topics: daily health keeping (e.g. low carbohydrate diet, longevity), global infectious diseases (e.g. Ebola virus), lifestyle diseases (e.g. obesity), emerging biomedical technologies (e.g. mobile health), and medical issues caused by some social activities (e.g. Brexit, public involvement in the medical system).[150]

Our analysis of the top 100 discussed articles, ranked by the Altmetric Attention Score, published between 15 November 2018 and 2019, also highlighted "modern life" and "healthy living" among the themes garnering high attention in 2019 (data not shown). According to the classification of subjects provided by Altmetric, medical and health sciences represented 54 of the top 100 articles. The most common topic in the medical and health science subjects was

nutrition (17 articles), followed by lifestyle and environmental (16) topics, featuring 61% of the medical and health science subjects represented.

This seemingly persistent pattern of nutritional and lifestyle factors garnering high media attention has important implications. As scholarly publishing has evolved through the digital transition, mentions of scientific outputs in social web tools, such as blogs, news, and social media (Facebook, Twitter, etc), may reflect the potential appeal of topics and interventions that are of higher interest to media outlets or that the general public can apply in their lives, rather than how rigorous the methods of the studies are.[139, 152] [153] As indicated by several other studies, online media attention is not necessarily a proxy measure of high research quality. [139, 152, 153]

All this raises the question "*what is the value of alternative metrics?*".[152] It has been hypothesized that Altmetric may serve as an early surrogate and faster measure of scientific impact, and may predict citations.[150, 152, 154] Previous studies have suggested that dissemination of medical research in the media can affect the behaviour of patients, clinicians, other healthcare providers, researchers, and the public.[139] [155] A positive correlation between some alternative metrics and citations has also previously been observed, although most often the correlation is weak to moderate.

Several studies have shown a positive correlation between an article's Altmetric score before retraction and the probability of the document's eventual retraction.[152, 156, 157] Thus, alternative metrics may in fact be capturing different kinds of hidden impacts, which could be explored by future research.[152]

We believe we should be cautious when positively linking the single quantitative value of an Altmetric score to research quality or scientific impact. Quantitative measures of social media attention disregard the intrinsic reasons for the attention, and whether it is due to controversy, criticism, or commendation.[152] The extent of the problem is evident when considering contradictory information about health and nutrition in the media. [158, 159] Conflicting information on the health benefits and harms of dietary consumption may adversely influence public opinion.[158, 159] It may lead to 'confusion about what foods are best to eat and the belief that nutrition scientists keep changing their minds', and to discrediting valid nutrition and health recommendations, such as fruit/vegetable consumption and exercise.[159] Given that high online media attention to an article can indicate broad societal impact[139, 152], it is important to recognize and mitigate the limitations of Altmetric scores.

CONCLUSION

Consistent with the results of previous studies, we observed that evaluations of nutritional interventions published in 2019 are mentioned more often in news media, social media, and other online sources covered by Altmetric, compared with other types of studies that generate an Altmetric score of 50 or more. This suggests that interventions that a wide range of readers can apply in their daily life attract more attention and are discussed more often than other interventions in medicine and public health.

APPENDICES

Appendix A. List of journals of included studies, if over 2%.

Journal	Number of articles published (%, if over 2%)
New England Journal of Medicine (NEJM)	80 (24.7)
Lancet	37 (11.4)
JAMA	30 (9.3)
BMJ	26 (8.0)
Annals of Internal Medicine	25 (7.7)
Journal of the American College of Cardiology	13 (4.0)
Journal of Clinical Oncology	10 (3.1)
Lancet Diabetes & Endocrinology	10 (3.1)
Lancet Oncology	9 (2.8)
JAMA Internal Medicine	8 (2.5)
Cochrane Database of Systematic Reviews	7 (2.2)
Circulation	6 (1.9)
Canadian Medical Association Journal (CMAJ)	4 (1.2)
Journal of the American Geriatrics Society	4 (1.2)
JAMA Psychiatry	4 (1.2)
PLOS Medicine	4 (1.2)
Annals of the Rheumatic Diseases	3 (0.9)
European Heart Journal	3 (0.9)
European Urology	3 (0.9)
Gastroenterology	3 (0.9)
Journal of the American Society of Nephrology	3 (0.9)
Neurology	3 (0.9)
European Respiratory Journal	2 (0.6)
JAMA Pediatrics	2 (0.6)
Lancet Neurology	2 (0.6)
Thorax	2 (0.6)
Total	**303**

Appendix B. Modelling the Adjusted Altmetric score

We hypothesized that the Altmetric score plateau over time, allowing us to estimate an adjusted Altmetric score for each article by the number of days after publication. To test our hypothesis, we plotted the historic Altmetric scores of the 324 research articles in our dataset (Figure B.1). Each set of connected points represents one article. We observed that most of the articles had already reached a plateau. However, for some publications the Altmetric score was still rising. To better visualize this, we plotted the subset of articles (n=38) that had an increase of at least 50 points in the Altmetric score between the 3rd and 6th months Altmetric history time interval after our search date, 8 November 2019 (Figure B.2).

Since the model b=0 provides a horizontal fit for Altmetric scores that have already reached a plateau, we only considered the cases (n=38) that depicted a higher rate of increase, reminiscent to the rise in Altmetric score initially after publication. The minimum R^2 among the 38 cases depicting a quick rise of the model was 87%, confirming the fit to our model. We further evaluated the performance of the model, by fitting it to the entire dataset of 324 articles. We observed that the model fitted the data well; (i) there were enough data points in 312 out of 324 research studies to fit the model, (ii) the average R^2 was 0.90 in 312 fits of the model, and (iii) 90% of the fits had an R^2 of at least 0.71. Additionally, we fitted the model to four randomly chosen articles, once from each intervention category (Figure B.3), further illustrating its validity.

Figure B.1. Historic Altmetric scores of included articles (n=324).

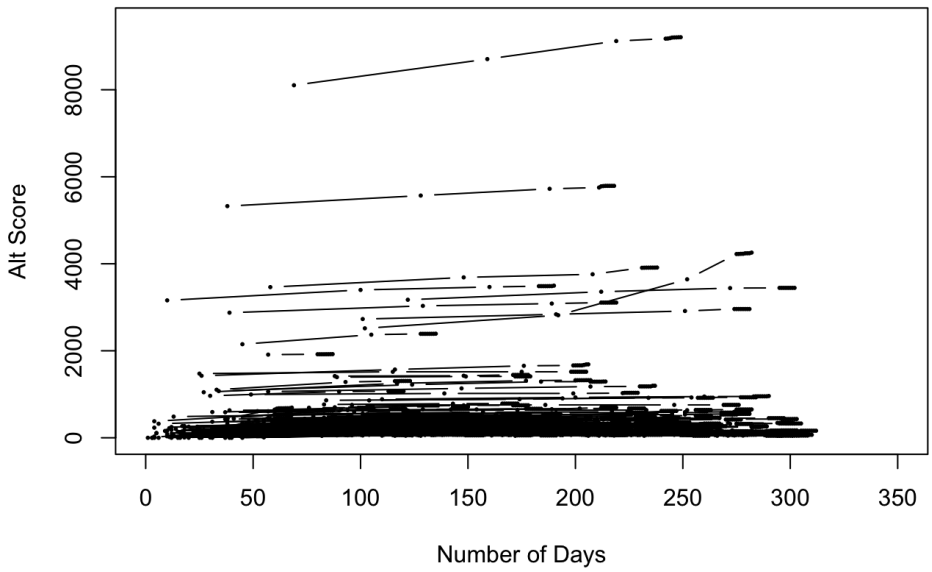
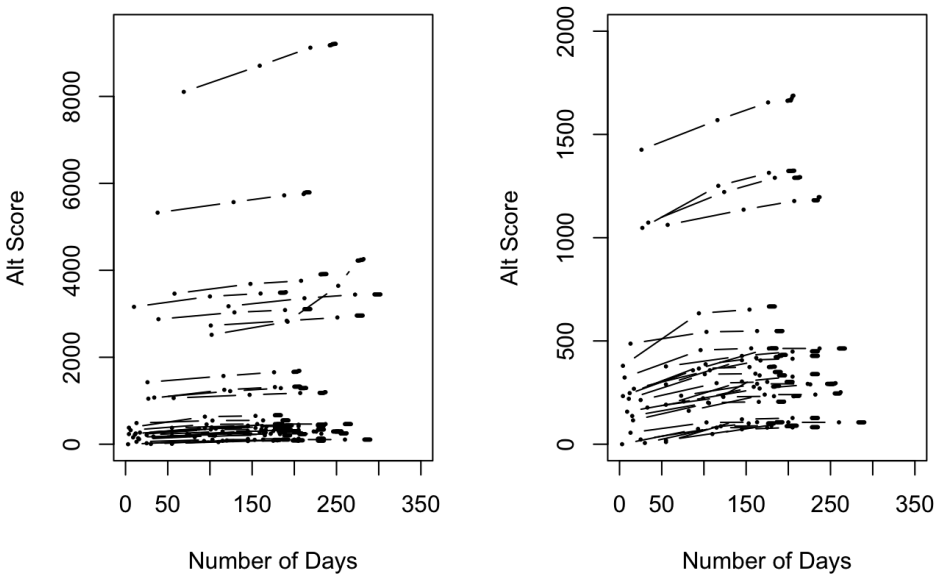
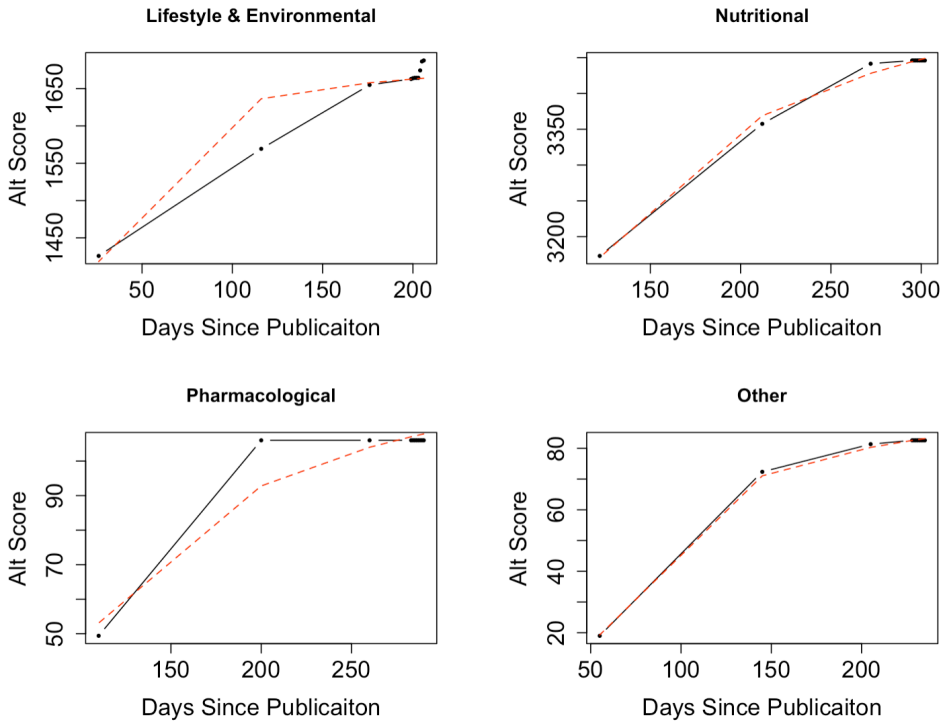


Figure B.2. Altmetric score trend over time.



The left panel shows all cases and the panel on the right-hand-side shows those cases that had an increase of over 50 points in the Altmetric score between the first recording (6 months before 8 November) and the second recording three months later. There are 38 such cases in the data-set.

Figure B.3. Examples of the trends in Altmetric scores and the predicted trends from the model.



The Altmetric data are in black and the fits to the model in red.

Chapter 6

Stop this waste of people, animals and money

David Moher
Larissa Shamseer
Kelly D Cobey
Manoj M Lalu
James Galipeau
Marc T Avey
Nadera Ahmadzai
Mostafa Alabousi
Pauline Barbeau
Andrew Beck
Raymond Daniel
Robert Frank
Mona Ghannad
Candyce Hamel
Mona Hersi
Brian Hutton
Inga Isupov
Trevor A McGrath
Matthew DF McInnes
Matthew J Page
Misty Pratt
Kusala Pussegoda
Beverley Shea
Anubhav Srivastava
Adrienne Stevens
Kednapa Thavorn
Sasha van Katwyk
Roxanne Ward
Dianna Wolfe
Fatemeh Yazdi
Ashley M Yu
Hedyeh Ziai

Nature 2017; 549 (7670): 23–25

Note: The published article representing this chapter is a news-style summary of the work, in line with style of the publishing journal. The main text of this chapter appeared as a supplement to the published article. It contains a detailed account of the methods and results of the study.

Details on access to both the article and supplement are below:

Link to published article:
https://www.nature.com/news/stop-this-waste-of-people-animals-and-money-1.22554

Original research article:
https://www.nature.com/news/polopoly_fs/7.46207.1504703578!/suppinfoFile/549023a_S1.pdf

COMMENT

ILLUSTRATION BY DAVID PARKINS

Stop this waste of people, animals and money

Predatory journals have shoddy reporting and include papers from wealthy nations, find **David Moher**, **Larissa Shamseer**, **Kelly Cobey** and colleagues.

Predatory journals are easy to please. They seem to accept papers with little regard for quality, at a fraction of the cost charged by mainstream open-access journals. These supposedly scholarly publishing entities are murky operations, making money by collecting fees while failing to deliver on their claims of being open access and failing to provide services such as peer review and archiving.

Despite abundant evidence that the bar is low, not much is known about who publishes in this shady realm, and what the papers are like. Common wisdom assumes that the hazard of predatory publishing is restricted mainly to the developing world. In one famous sting, a journalist for *Science* sent a purposely flawed paper to 140 presumed predatory titles (and to a roughly equal number of other open-access titles), pretending to be a biologist based in African capital cities[1]. At least two earlier, smaller surveys found that most authors were in India or elsewhere in Asia[2,3]. A campaign to warn scholars about predatory journals has concentrated its efforts in Africa, China, India, the Middle East and Russia. Frequent, aggressive solicitations from predatory publishers are generally considered merely a nuisance for scientists from rich countries, not a threat to scholarly integrity.

Our evidence disputes this view. We spent 12 months rigorously characterizing nearly 2,000 biomedical articles from more than 200 journals thought likely to be predatory. More than half of the corresponding authors hailed from high- and upper-middle-income countries as defined by the World Bank.

Of the 17% of sampled articles that reported a funding source, the most frequently named funder was the US National Institutes of Health (NIH). The United States produced more articles in our sample than all other countries save India. Harvard University (with 9 articles) in Cambridge, Massachusetts, and the University of Texas (with ▶

11 articles across all campuses) were among the eight institutions with the most articles. It is easy to imagine other, similar institutions coming up in a different sample. The point is, the problem of predatory journals is more urgent than many realize.

Articles in our sample consistently failed to report key information necessary for readers to assess, reproduce and build on the findings. Fewer than 10% of studies claiming to be randomized controlled trials described how patients were allocated to treatment groups; where blinding was possible, fewer than one-quarter noted whether patients and outcome assessors were blinded to group assignment.

Whether authors are being duped or are overzealously seeking to lengthen their publication lists, this represents enormous waste. Just the subset of articles that we examined contained data from more than 2 million individuals and over 8,000 animals. By extrapolation, we estimate that at least 18,000 funded biomedical-research studies are tucked away in poorly indexed, scientifically questionable journals. Little of this work will advance science. It is too dodgily reported (and possibly badly conducted) and too hard to find.

In our view, publishing in predatory journals is unethical. Individuals who agree to be studied expect that their participation could benefit future patients. Use of animals in biomedical research is rationalized on the assumption that experiments will contribute valuable information. Even assuming authors are publishing more than one paper from their study (and some are), they should be held to a higher standard of disclosure. Publishers, funders and research institutions must join together to prevent research from ending up in predatory journals.

WHAT WE DID

We drew our sample from the journals and publishers whose status as predatory was deemed as "potential, possible, or probable" by librarian Jeffrey Beall of the University of Colorado, Denver. (These controversial compilations were taken offline early in 2017, but remain available in web archives and are one of the few tools that researchers have to investigate illegitimate journals.)

We took a random subset of 185 publishers and obtained lists of their journals. At least two people, working independently, assessed whether each journal met Medline's selection criteria as having a biomedical scope. We randomly selected 200 journals from this list; we also included 45 biomedical standalone journals from a set that had been developed similarly for another study[4].

In February 2016, we visited each journal's website and downloaded copies of up to 25 articles, starting with the ones most recently published. The total number of articles we obtained came to 3,702, because many journals listed fewer than 25 articles. Of those, 1,907 reported primary biomedical research or systematic reviews and so were included in our analysis. This left us with 41 single-journal publishers and 179 titles from 51 multi-journal publishers (see Supplementary information and https://osf.io/r2gj6/).

SLOPPY WORK

We examined each paper in light of reporting guidelines relevant for each type of study. For example, for randomized controlled trials, we cross-checked articles using a modified Consolidated Standards of Reporting Trials (CONSORT) checklist.

Although adherence to and enforcement of guidelines is patchy even in mainstream publications, reporting quality in our sample was much worse. Articles were particularly deficient in descriptions of study methods, results and — for clinical trials and systematic reviews — study registration. Of the 94 randomized controlled trials that we examined, most items in CONSORT were reported 40% of the time. Fewer than 14% of trials gave a registration number or registry name. Yet, a study of mainstream journals in the Netherlands found registration information in at least 60% of trials[5].

"In our view, publishing in predatory journals is unethical."

Of the 21 systematic reviews in our sample, only two reported assessing the risk of bias. Yet 70% did so in an evaluation of 300 Medline-indexed systematic reviews[6]. Even for animal studies, for which reporting in mainstream journals is remarkably poor, we found that performance of predatory journals was much worse; for instance, just 3% of 201 relevant predatory articles reported blinding. A separate study found that blinding is recorded in 20% of articles in PLoS journals and in 21% of articles in Nature journals[7].

Of the articles in our sample that evaluated humans or whole animals, only 40% noted that they received approval from an ethics committee. Previous studies show that ethics-committee approval is mentioned in more than 90% of animal[7] and 70% of human[8] studies published in mainstream journals.

GLOBAL PROBLEM

Nearly three-quarters (1,397) of the publications we examined did not report information about funding; 10% stated that they were not funded. The remaining 323 articles named 345 different funders, mainly academic institutions (124) and government agencies (122).

For 1,907 papers, corresponding authors came from 103 countries, including India (27%), the United States (15%), Nigeria (5%), Iran (4%) and Japan (4%) (see 'Global predation'). These figures should be interpreted in the context of total scientific output per nation. According to tallies in the academic databases Scopus and PubMed, the United States produced about 5 times as many biomedical articles as India last year, and 80 times as many as Nigeria. An analysis of general academic articles from 2013 to 2015 in Scopus found that 10% or more from India and Nigeria were in predatory journals, as compared to less than 1% from Japan and the United States[9].

Researchers at universities in Indiana also looked at who has published in predatory pharmaceutical journals in 2013 (ref. 2). They compared authors who published in seven predatory titles on Beall's list with those in five open-access journals that rejected the journalist's bait in the sting operation[1]. Sixty-five per cent of authors of predatory-journal articles had never published before, as compared to 19% of those in vetted open-access journals. In that study, 75% of all authors in predatory journals were from South Asia (mainly India), 14% from Africa (mainly Nigeria) and only 3% from North America. By contrast, 57% of authors in our sample are from higher-income or upper-middle-income countries.

This could be because our sampling looked mainly at corresponding authors rather than all authors, or because it considered a broader swath of articles, titles and years. It is possible that our sample included some journals particularly popular with high-income-country authors. Or, perhaps, predatory journals have stepped up their aggressive e-mails in richer countries.

Corresponding authors in our survey named 1,291 institutions as primary affiliations; 15 did not name any. We contacted 16 vice-presidents (or the senior administrative person) of research at some of the top institutions whose researchers were publishing in predatory journals. Our e-mail to Bangalore Medical College and Research Institute bounced back. Three institutions provided feedback; one (Manipal University, India, 15 papers) detailed an intervention launched earlier this year, and provided data that the effort reduced the number of articles published in presumed predatory journals.

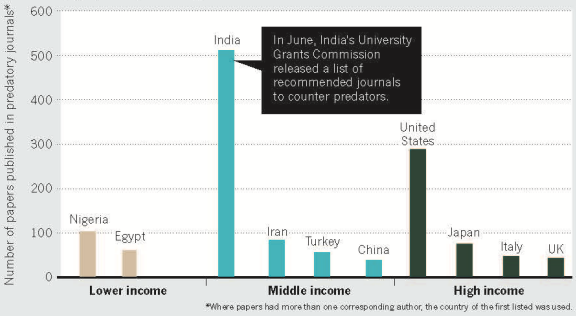
Responses from the University of Benin in Nigeria (8 papers) and Menoufia University in Egypt (8 papers) said that they warned against or made lists of illegitimate publishers for researchers to consult. The Mayo Clinic in Rochester, Minnesota (7 papers), sent a note to *Nature* that articles in predatory journals are not considered for academic advancement. D. Y. Patil University in India, which, with 20 papers, had the most in our sample, did not reply. Nor did the University of Tehran, which, with 14 papers from 14 authors, tied with D. Y. Patil University for the most unique authors.

We also attempted to contact corresponding authors at some of the leading institutions. Fifteen articles — including all 9 at Bangalore

GLOBAL PREDATION

A sample of 1,907 papers in more than 200 supposed predatory journals found that most of the articles come from India. Surprisingly, however, more than half of the papers have authors from higher-income or upper-middle-income countries.

Predatory papers by country and income



SOURCE: STUDY REPORTING IN PREDATORY JOURNALS GROUP

Medical College and Research Institute — did not include author e-mails. In total, we sent e-mails to 87 authors (who had collectively written 119 of the articles in our sample). Three bounced back. Of the 18 replies we received, 3 asked for educational material about predatory journals. Only two said that they were aware of the journal's categorization as potentially predatory, and only four were aware of Beall's list. Similarly, 90% of 1,088 Italian authors surveyed in 2016 stated that they were unaware of Beall's list (see go.nature.com/2ixzvqm).

Only three people who responded to us had submitted their manuscript to other publications before its acceptance in the predatory journal. Seven indicated receiving guidance of some form on where to submit; an equal number said that their work had been cited. This suggests that authors are not using these titles as a last resort: the scientific community needs a better understanding of what precisely makes predatory journals attractive.

Although, collectively, the articles we examined were atrocious in terms of reporting, we did not examine the performance of individual titles. We should note that several publishers have protested against their inclusion on Beall's list. Also, the term 'predator' in particular is questioned, because on occasion it is hard to tell whether a journal is simply inept or disdaining research integrity and scientific robustness to pursue profit. And rather than being prey, some authors may purposely seek out low-barrier ways to publish.

Some may argue that authors in higher-income countries publish in the 'best' predatory journals. Our study was not set up to resolve this question; we did examine reporting of randomization and allocation across the 94 clinical trials, and found no statistical differences between rich-world corresponding authors and those from elsewhere.

It is nearly impossible for prospective authors to differentiate predatory journals by metrics. For example, the *Journal of Surgery* from Avens Publishing Group — the title most favoured by US authors in our sample — does not have easily identifiable metrics that distinguish it from non-predatory journals. We contacted the editor-in-chief of the journal, who replied that he hadn't heard of predatory journals but that the journal sends manuscripts to peer reviewers and has rejected manuscripts on the basis of their merit. (Note that at least two other questionable journals carry the same title.)

Our experience with these journals is that they provide both poor vetting and poor access. Their websites and archiving systems are unstable. Although some articles appear in PubMed (often after a delay), the titles are not indexed by Medline and are difficult to find. Indeed, some titles in our sample, such as the *International Journal of Pharmaceutical and Medical Research*, ask authors to assign copyright to the journal, which is against the mores of open access.

Even without Beall's list, savvy authors should know when to suspect that a journal is predatory. Our research group has identified 13 characteristics of predatory journals; these include low article-processing fees (less than US$150); spelling and grammar errors on the website; an overly broad scope; language that targets authors rather than readers; promises of rapid publication; and a lack of information about retraction policies, manuscript handling or digital preservation. Manuscript submissions by e-mail and the inclusion of distorted images are also common[4].

However, predatory journals are becoming increasingly adept at appearing legitimate, and little is being done to warn authors away from them. Just one of the ten most common funders reported in our study, the University Grants Commission, India, provides guidance about journal selection on its website.

STOP THE ROT

We believe that publishers, research institutions and funders should issue explicit warnings against illegitimate publishers and develop cohesive recommendations on publication integrity together.

Funders and research institutions should increase the funds that they make available towards open-access publication; prohibit the use of funds to support predatory journal publications; make sure that researchers are trained in how to select appropriate journals when submitting their work; and audit where grantees, faculty members and research staff publish. When seeking promotion or funding, researchers should include a declaration that their CV is free of predatory publications. Publication lists could be checked against lists such as the Directory of Open Access Journals (DOAJ) or the *Journal Citation Reports*. Developing automated tools to facilitate the proposed audits would also be valuable.

Before approving a study, ethics committees should ask researchers to declare in writing their willingness to work with their institutional resources, such as librarians, to ensure they do not submit to any journals without reviewing evidence-based criteria for avoiding these titles.

If not, predatory journals will continue to erode the integrity of scientific scholarship. Substandard publications have permeated authentic electronic databases. A problem largely unknown a decade ago, there are now a roughly estimated 8,000 predatory titles that collectively 'publish' more than 400,000 items a year[3]. We need to cut off the supply of manuscripts to these illegitimate outfits. ■

David Moher *is a clinical epidemiologist at the Ottawa Hospital Research Institute, Ontario, Canada, and part of the Study Reporting in Predatory Journals Group.*
e-mail: dmoher@ohri.ca

1. Bohannon, J. *Science* **342,** 60–65 (2013).
2. Xia, J. *et al. J. Assoc. Infor. Sci. Tech.* **66,** 1406–1417 (2015).
3. Shen, C. & Björk, B.-C. *BMC Med.* **13,** 230 (2015).
4. Shamseer, L. *et al. BMC Med.* **15,** 28 (2017).
5. van de Wetering, F. T., Scholten, R. J. P. M., Haring, T. & Hooft, L. *PLoS ONE* **7,** e49599 (2012).
6. Page, M. J. *et al. PLOS Med.* **13,** e1002028 (2016).
7. Baker, D., Lidster, K., Sottomayor, A. & Amor, S. *PLoS Biol.* **12,** e1001756 (2014).
8. Taljaard, M. *et al. Br. Med. J.* **342,** d2496 (2011).
9. Machacek, V. & Srholec, M. *Predatory journals in Scopus* (IDEA, 2017); available at http://go.nature.com/2wd3es7

Supplementary information and a full list of authors accompanies this article online: see go.nature.com/2vlnvmw

Chapter 7

Summary and future perspectives

Summary

Through the research work presented in this thesis, we explored the ramifications of (1) suboptimal reporting practices in context of diagnostic/prognostic biomarker studies and randomized trials, (2) a proposed strategy aimed at improving biased reporting, and (3) other challenges in publication and dissemination of biomedical research. Our work builds on a growing number of research publications on research misconduct and misbehaviours, commonly named "questionable research practices".

Responsible research practices and fair reporting is an element of research integrity. Articles published in *The Lancet* illustrated the problem of waste during various stages of research encompassing design, conduct and reporting. [8, 9] Given that much of this waste is avoidable, there is a need to develop and implement remedies. [8] Of these, accurate interpretation and presentation of results in published data is essential in order to avoid producing misleading studies and waste valuable resources.

In **Chapter 1**, we reported a descriptive systematic review of 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. 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 biased reporting and poor study design, have been attributed to the lack of success in identifying clinically relevant biomarkers. We investigated biased reporting and interpretation in published articles as a potential contributing factor, which had 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.

Our analysis of 200 recent evaluations of ovarian cancer biomarkers confirmed that 140 (70%) contained 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. 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%). This review confirmed that biased reporting and interpretation is prevalent in recent clinical evaluations of biomarkers in ovarian cancer. These results indicated a need for strategies to minimize biased reporting and interpretation.

In **Chapter 2**, we found that practices facilitating spin, such as suboptimal design features and inadequate reporting of methods, were also prevalent in biomarker evaluations. In the sample of studies described in Chapter 1, 93 (47%) had acquired samples and data from a single centre. The median sample size was 156 patients (ranging from 13 to 50,078), with only 5 (3%) of studies justifying their sample size. Studies often recruited patients in disjoint groups (33%), (i.e., groups of comparison not originating from one single study group), sometimes with extreme phenotypic contrasts; 46 (23%) included healthy controls and 5 (3%) exclusively included advanced stage cases. Eligibility criteria and sampling methods were rarely reported.

The reporting of methods was incomplete; few studies reported eligibility criteria (10%) and sampling methods (10%). This review showed that inadequate study designs were frequent in clinical evaluations of ovarian cancer, which could lead to biased and premature conclusions about the performance of the marker in clinical applications.

In the study reported in **Chapter 3**, we performed a meta-epidemiological study to identify potential trial characteristics associated with reported treatment effect estimates in randomized trials of testosterone therapy in adult men. Of 132 randomized trials, 19 were meta-analyses, comprising data from 10,725 participants. None of the investigated design characteristics, including year of publication, sample size, trial registration status, centre status, regionality, funding source, and conflict of interest, were statistically significantly associated with reported treatment effects of testosterone therapy in men.

In our analyses, a substantial number of studies were rated at "high/unclear risk of bias" (ranging from 17% to 62% across the domains). Previous research suggests that 89% of published randomized trials include at least one "unclear" risk of bias domain [123]. Our results confirm the persistent high prevalence of incomplete reporting in trials.

This review showed no clear evidence that trial characteristics are associated with treatment effects in randomized trials of testosterone therapy in men. However, the associations between trial characteristics and treatment effects reflects what was *reported* in the trial report, not necessarily what was *done* by the trialists. Given the importance of well-designed and well-conducted randomized trials for the production of high-quality evidence, future trials on testosterone therapy should not only be adequately performed but also transparently reported to assess the safety and efficacy of testosterone therapy in men.

To date, there has been no intervention designed to mitigate or reduce the prevalence of spin in biomedical literature. In the study reported in **Chapter 4**, we developed a specific editorial intervention to reduce spin and conducted a two-arm parallel-group randomised controlled trial to evaluate its impact. We conducted this study in collaboration with BMJ Open, a general medical journal. Our primary outcome was the presence of spin; secondary outcomes were types of spin and wording change in the revised abstract's conclusion. Outcome assessors were blinded to the intervention assignment.

Of the 184 manuscripts randomised, 108 (54 intervention, 54 control) were selected for revision and could be evaluated for the presence of spin. The proportion of manuscripts with spin was 6% lower (95% CI: 24% lower to 13% higher) in the intervention group (57%, 31/54) than in the control group (63%, 34/54), a non-significant difference. Wording of the revised abstract's conclusion was changed in 34/54 (63%) manuscripts in the intervention group and 26/54 (48%) in the control group. The four pre-specified types of spin involved: (i) selective reporting (12 in the intervention group versus 8 in the control group); (ii) including information not supported by evidence (9 versus 9); and (iii) interpretation not consistent with study results (14 versus 18); and (iv) unjustified recommendations for practice (5 versus 11).

Our short editorial instructions to authors may have led authors to revise the wording of their conclusions, but it did not have a statistically significant effect on reducing spin in revised abstract conclusions and, based on the confidence interval, the existence of a large effect can

be excluded. Thus, other interventions to reduce spin in reports of original research should be evaluated.

Alternative metrics have been developed to measure the attention publications receive from social news media and blogs. In the study reported in **Chapter 5**, we aimed to discover which types of studies reported in recent research articles in medical journals receive the highest Altmetric scores, among those generating attention in Altmetric data sources. In the 324 primary research articles included in our study, the median Altmetric score was 184 [Interquartile Range, 111 – 378]. Journal category and impact factor were significantly associated with adjusted Altmetric score ($P < 0.00001$). Nutritional interventions (median = 4.69) accumulated significantly higher adjusted Altmetric scores compared to lifestyle and environmental (1.47), Pharmacological (1.05), and other interventions (0.82).

We observed that, in publications that generated an Altmetric score of 50 or more, reports of evaluations of nutritional interventions were mentioned more often than other types of interventions in news media, social media, and other online sources covered by Altmetric. This seems to indicate that interventions that a wide range of readers can apply in their daily life attract more attention and are discussed more often than other interventions in medicine and public health.

Entities that have become known as 'predatory' journals and publishers are permeating the world of scholarly publishing, yet little is known about the papers they publish. For the project summarized in **Chapter 6**, we examined a cross-section of 1907 human and animal biomedical studies, recording their study designs, epidemiological and reporting characteristics. In our sample more than two million humans and over eight thousand animals were included in predatory publications. Only 40% of studies report having ethics approval. Of the 17% of articles reporting their funding source, the US National Institutes of Health was most frequently named. Corresponding authors were most often from India (511/1907, 26.8%) and the US (288/1907, 15.1%). The reporting quality of work reported in our sample was poor and worse than contemporaneous samples from the legitimate literature. Many studies were missing key methodological details and findings. Our results raise important ethical concerns since research in predatory journals is difficult to identify and not indexed in scientifically curated biomedical databases. Funders and academic institutions need to develop explicit policies to drive grantees and prospective authors away from these entities.

Future perspectives

Research misconduct is a broad term that encompasses different areas. For the widely accepted definition (e.g., falsification, fabrication, and plagiarism (FFP)), many regulatory and ethical policies are in check. However, an important challenge remains: beyond FFP, there is currently a lack of consensus about what types of behaviour constitute research misconduct [160], which subsequently impedes strategies for detection and limitation of such misbehaviours.

'Spin' practices are commonly perceived acceptable amongst researchers and not regarded as detrimental research practices, despite previous evidence documenting some of its negative consequences.[10] This perception among researchers may in large be due to lack of awareness. Continued efforts to document and characterise specific strategies of spin in emerging fields are needed to further build on previous evidence, and to improve our understanding of this concept within each research field. We could also extend on the evidence by documenting negative consequences of spin in research reports, as it may further encourage researchers and stakeholders to regard spin as detrimental research practice. 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.

Simply characterising spin and documenting its negative consequences will probably not be sufficient to change both perceptions and practices. Awareness of current behaviour is only one step towards changing behavior and research practices. Active interventions are needed to develop strategies that facilitate change. Since we observed a limited effect of our editorial intervention, we should look at other strategies to improve reporting. These can include but should not be limited to pre-registration of the study design, primary outcome(s), and analysis plan as a highly effective form of blinding, given the data do not exist and the outcomes are not yet known at the time of study initiation. Another strategy to consider may be assembling diverse and multidisciplinary teams that include statisticians in research teams, or including a statistician in the editorial board of the journal, to help ensure the rigorous and robust conduct and interpretation of research methodology. Thereby, limiting the possibility of spin in the findings and conclusions, reducing bias and improving transparency in medical research.

The role of funders and academic institutions is integral in disseminating research integrity and best research practices. Currently, most university promotion and tenure committees, emphasis and reward perceived impact of research with (albeit narrow) criteria focused on publications and associated metrics (i.e., impact factors, citations, or Altmetrics) rather than rigor. [161] Issues related to publication practices, such as 'predatory' journals, arise as a symptom of this flawed 'publish or perish' system. To foster research integrity and increase the value of research, it is imperative to further current initiatives that place more emphasis on rigorous research and other positive publication practices with greater societal value.

Trust in science can be eroded by the frequent use of suboptimal reporting practices and inadequate methodology. 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, funders and academic institutions undoubtedly share responsibility.

Samenvatting

In het in dit proefschrift gepresenteerde onderzoek hebben we het volgende onderzocht (1) suboptimale rapportagepraktijken bij de evaluatie van diagnostische en prognostische biomarkers en in gerandomiseerd onderzoek, (2) een strategie voor het verminderen van 'spin' in rapportages van onderzoek en (3) andere uitdagingen bij de publicatie en verspreiding van biomedisch onderzoek. Ons werk bouwt voort op een geleidelijk groeiend aantal publicaties over wangedrag en verwijtbaar gedrag in onderzoek, gewoonlijk "dubieuze onderzoekspraktijken" genoemd.

Verantwoorde onderzoekspraktijken en eerlijke rapportage maken beide deel uit van integriteit in onderzoek. Artikelen gepubliceerd in The Lancet illustreren het probleem van vermijdbare verspilling ('waste') in de verschillende stadia van onderzoek, waaronder ontwerp, uitvoering en rapportage. [8, 9] Aangezien een groot deel van deze verspilling vermijdbaar is, is er behoefte aan de ontwikkeling en toepassing van remedies. [8] Een nauwkeurige interpretatie en presentatie van de resultaten uit onderzoek is van essentieel belang om te voorkomen dat misleidende conclusies worden getrokken en kostbare middelen worden verspild.

In **Hoofdstuk 1** rapporteerden we een systematisch literatuuronderzoek naar de aanwezigheid van spin in recente klinische studies die de prestatie van biomarkers bij eierstokkanker evalueren. Spin werd verder ingedeeld als een verkeerde voorstelling dan wel overinterpretatie van studiebevindingen. Er is veel geïnvesteerd in de ontdekking van biomarkers voor eierstokkanker, maar tot nu toe zijn weinig van die merkers met succes in de klinische praktijk geïntroduceerd. Factoren als bevooroordeelde rapportage en gebrekkige studieopzetten worden deels verantwoordelijk geacht voor dit gebrek aan succes. Een verkeerde voorstelling of overinterpretatie van studiebevindingen kan leiden tot ongerechtvaardigd optimisme over de prestaties van vermeende biomarkers.

Onze analyse van 200 recente evaluaties van biomarkers voor eierstokkanker bevestigde dat 140 (70%) ten minste één vorm van spin bevatten in de titel, de samenvatting of de conclusie van het artikel, waarbij de prestaties van de biomarker werden overdreven. De meest voorkomende vormen van spin waren: (1) uitspraken over ander gebruik van de biomarker dan onderzocht (65; 32,5%); (2) mismatch tussen beoogd doel en conclusie (57; 28,5%); en (3) onjuiste presentatie van resultaten (40; 20%). Onze review bevestigde dat een bevooroordeelde rapportage en interpretatie veelvuldig voorkomen in recente klinische evaluaties van biomarkers bij eierstokkanker. Deze resultaten geven aan dat er behoefte is aan strategieën om vertekening in rapportage en interpretatie te beperken.

In **Hoofdstuk 2** zagen we dat sommige praktijken die spin in de hand werken, zoals een suboptimaal design en een inadequate rapportage van methoden, vaak voorkwamen in evaluaties van biomarkers. In de steekproef van studies voor Hoofdstuk 1 werden in 93 gevallen (47%) monsters en gegevens in één enkel centrum verzameld. De mediane steekproefgrootte bedroeg 156 patiënten (variërend van 13 tot 50.078), met slechts bij 5 (3%) een rechtvaardiging van de steekproefgrootte. Studies rekruteerden vaak patiënten in disjuncte groepen (33%); dit zijn groepen die niet afkomstig waren uit één enkele reks. Dat gebeurde

soms met extreme fenotypische contrasten; 46 (23%) includeerden bij voorbeeld gezonde controles en 5 (3%) onderzochten uitsluitend patiënten in een gevorderd stadium van ziekte. Criteria om deel te kunnen nemen aan een studie en methoden van monsterneming werden zelden gerapporteerd. De rapportage van methoden was onvolledig; weinig studies vermeldden inclusiecriteria (10%) en bemonsteringsmethoden (10%). Dit overzicht toont aan dat inadequate studie-opzetten frequent voorkomen in klinische evaluaties van eierstokkanker, wat zou kunnen leiden tot vertekende en voorbarige conclusies over het nut van de marker in klinische toepassingen.

Hoofdstuk 3 bevat een verslag van een meta-epidemiologische onderzoek naar trialkarakteristieken die geassocieerd zouden kunnen zijn met effectschattingen in gerandomiseerde trials van testosterontherapie bij volwassen mannen. Van de 132 gerandomiseerde trials waren er 19 meta-analyses, met gegevens van 10.725 deelnemers. Geen van de onderzochte designkarakteristieken - inclusief jaar van publicatie, steekproefgrootte, trialregistratiestatus, centrumstatus, regio, financieringsbron, of belangenverstrengeling - waren statistisch significant geassocieerd met de gerapporteerde behandelingseffecten van testosterontherapie bij mannen. In onze analyses werd een aanzienlijk aantal studies beoordeeld met een "hoog/onduidelijk risico op vertekening" (variërend van 17% tot 62% over de domeinen). Eerder onderzoek suggereert dat 89% van de gepubliceerde gerandomiseerde trials ten minste één "onduidelijk" *risk of bias* domein bevatten [123]. Onze resultaten bevestigen de aanhoudend hoge prevalentie van onvolledige rapportage in trials.

Dit review bood hiermee geen duidelijk bewijs dat trialkarakteristieken geassocieerd zijn met behandelingseffecten in gerandomiseerde trials van testosterontherapie bij mannen. Echter, de associaties tussen trialkarakteristieken en behandelingseffecten weerspiegelen enkel wat er in het verslag staat, niet noodzakelijkerwijs wat er door de onderzoekers is gedaan. Gezien het belang van goed opgezette en goed uitgevoerde gerandomiseerde trials moeten toekomstige trials over testosterontherapie dan ook niet alleen adequaat worden uitgevoerd, maar ook transparant worden gerapporteerd, om de veiligheid en werkzaamheid van testosterontherapie bij mannen te kunnen beoordelen.

Tot op heden is er nog geen interventie ontwikkeld om de prevalentie van spin in biomedische literatuur te verminderen. Voor het project waarover we in **hoofdstuk 4** verslag uitbrengen hebben we een redactionele interventie ontwikkeld om spin tegen te gaan. We hebben vervolgens een twee-armige gerandomiseerde gecontroleerde trial met parallelle groepen uitgevoerd om het effect van onze interventie te evalueren. We voerden deze studie uit in samenwerking met BMJ Open, een algemeen medisch tijdschrift. Onze primaire uitkomst was de aanwezigheid van spin; secundaire uitkomsten waren soorten spin en verandering van formulering in de conclusie van de herziene samenvatting. Beoordelaars waren geblindeerd voor de toewijzing van de interventie.

Van de 184 gerandomiseerde manuscripten werden er 108 (54 interventie, 54 controle) door de tijdschriftredactie geselecteerd voor revisie; deze konden worden beoordeeld op de aanwezigheid van spin. Het aandeel manuscripten met spin was 6% lager (95% CI: 24% lager tot 13% hoger) in de interventiegroep (57%, 31/54) dan in de controlegroep (63%, 34/54), een niet-significant verschil. De formulering van de conclusie in het abstract werd gewijzigd in

34/54 (63%) herziene manuscripten in de interventiegroep en 26/54 (48%) in de controlegroep. De vier vooraf gespecificeerde types van spin betroffen: (i) selectieve rapportage (12 in de interventiegroep versus 8 in de controlegroep); (ii) het opnemen van informatie die niet wordt ondersteund door de data (9 versus 9); en (iii) interpretatie die niet consistent is met de studieresultaten (14 versus 18); en (iv) ongerechtvaardigde aanbevelingen voor de praktijk (5 versus 11).

De korte redactionele instructies aan de auteurs hebben dus ertoe geleid dat auteurs de formulering van hun conclusies soms herzagen, maar dit had geen statistisch significant effect op het verminderen van spin in conclusies. Op basis van het betrouwbaarheidsinterval kan het bestaan van een groot effect van onze interventie worden uitgesloten. Andere interventies om spin in verslagen van oorspronkelijk onderzoek te verminderen, moeten dus worden ontwikkeld en geëvalueerd.

Er zijn alternatieve maten ontwikkeld om de aandacht te meten die publicaties krijgen in social media en blogs. In het onderzoek waarover in **hoofdstuk 5** verslag wordt uitgebracht, hebben we geprobeerd te achterhalen welke typen studies de hoogste Altmetric-scores krijgen, binnen alle studies die enige vorm van aandacht genereren in de bronnen die voor Altmetric worden gebruikt. In de 324 onderzoeksartikelen die in onze studie werden opgenomen, was de mediaan van de Altmetric-score 184 [Interkwartielafstand, 111 - 378]. Tijdschriftcategorie en impact factor waren significant geassocieerd met de Altmetric score (P <0.00001). Voedingsinterventies (mediaan = 4,69) behaalden significant hogere scores in vergelijking met leefstijl en milieu (1,47), farmacologische (1,05), en andere interventies (0,82).

Zo zagen we dat, in publicaties die een Altmetric score van 50 of meer genereerden, evaluaties van voedingsinterventies vaker dan andere soorten interventies worden genoemd in nieuwsmedia, social media en in andere online bronnen die door Altmetric worden bestreken. Dit lijkt erop te wijzen dat interventies die door een breed scala van lezers in hun dagelijks leven kunnen worden toegepast, meer aandacht krijgen en vaker worden besproken dan andere interventies in de geneeskunde of op het terrein van de volksgezondheid.

Sommige tijdschriften staan bekend als "roofzuchtig" (*predatory journals).* Er is echter weinig bekend over de artikelen die zij publiceren. Voor het project dat in **hoofdstuk 6** wordt samengevat, onderzochten we een steekproef van 1907 biomedische studies, bij mens en dier, waarbij we studieopzet, epidemiologische kenmerken en rapportagekenmerken vastlegden. Onze steekproef omvatte meer dan twee miljoen mensen en meer dan achtduizend dieren.

Slechts 40% van de studies meldde een ethische goedkeuring. Van de 17% van de artikelen die hun financieringsbron vermeldden, werd de US National Institutes of Health het vaakst genoemd. Corresponderende auteurs waren het vaakst afkomstig uit India (511/1907, 26,8%) en de VS (288/1907, 15,1%). De kwaliteit van de rapportage in onze steekproef was slechter dan die van andere steekproeven uit de “gewone” literatuur. In veel studies ontbraken belangrijke methodologische details en bevindingen.

Onze resultaten doen belangrijke ethische vragen rijzen, aangezien onderzoek in rooftijdschriften moeilijk te identificeren is en niet als dusdanig geïndexeerd is in wetenschappelijk databanken. Geldschieters en academische instellingen moeten een expliciet

beleid ontwikkelen om subsidieverleners en potentiële auteurs weg te houden van deze rooftijdschriften.

Toekomstperspectieven

Wangedrag bij onderzoek is een brede term, die verschillende gebieden bestrijkt. Voor het tegengaan van de praktijken in de basisdefinitie - vervalsing, verzinsel en plagiaat - zijn al veel maatregelen genomen. Een belangrijke uitdaging blijft echter bestaan: buiten FFP is er momenteel een gebrek aan consensus over de vormen van gedrag die moeten worden bestempeld als te vermijden gedrag bij onderzoek [160]. Die afwezigheid belemmert de ontwikkeling van strategieën voor het opsporen en beperken van dergelijk wangedrag.

"Spin" wordt door veel onderzoekers als aanvaardbaar beschouwd en als niet schadelijk, ondanks dat eerder onderzoek de negatieve gevolgen ervan heeft laten zien. [10] Deze perceptie onder onderzoekers kan grotendeels te wijten zijn aan een gebrek aan bewustzijn. Verdere inspanningen om specifieke vormen van spin te documenteren en te karakteriseren zijn dan ook nodig, ook om ons begrip van dit concept te verbeteren. We zouden de negatieve gevolgen van spin in onderzoeksrapporten beter kunnen documenteren, omdat dit onderzoekers en belanghebbenden ertoe kan aanzetten om spin als een schadelijke onderzoekspraktijk te beschouwen. Bestaande onderwijsprogramma's voor beginnende onderzoekers kunnen hiermee worden verrijkt, om zo aan toekomstige auteurs duidelijk te maken wat de verschillende facetten van ‘spin’ zijn, en wat hun impact is.

Het simpelweg karakteriseren van spin en het documenteren van de negatieve gevolgen zal waarschijnlijk echter niet voldoende zijn om de praktijk te veranderen. Bewustmaking vormt slechts één stap op de weg naar gedragsverandering. Actieve interventies zijn daarnaast nodig, om die verandering ondersteunen. Aangezien we maar een beperkt effect van onze eigen redactionele interventie hebben waargenomen, moeten we uitkijken naar andere manieren om de rapportage te verbeteren. Deze kunnen bestaan uit, maar hoeven niet beperkt te zijn tot, het vooraf registreren van de onderzoeksopzet, van de primaire uitkomst(en) en van het analyseplan. Een andere te overwegen strategie is het samenstellen van diverse en multidisciplinaire teams, met statistici en sceptici, in onderzoeksteams, of het opnemen van een statisticus in de redactieraad van het tijdschrift, om een rigoureuze en robuuste uitvoering én interpretatie van het onderzoek te bevorderen. Zo kan de kans op spin in de bevindingen en conclusies worden beperkt, vooringenomenheid verkleind, en wordt de transparantie van medisch onderzoek verder verbeterd.

De rol van financiers en academische instellingen bij het bevorderen van goed gedrag in onderzoek is cruciaal. Momenteel leggen de meeste promotie- en aanstellingscommissies van universiteiten de nadruk op publicaties, en leunen zij vooral op de bijbehorende meetmethoden, zoals impactfactoren en aantallen publicaties en citaties. [161] Sommige problemen rond verwijtbare publicatiepraktijken en 'roofzuchtige' tijdschriften ontstaan onzes inziens als symptoom van dit 'publiceer of verdwijn'-systeem. Om de integriteit van onderzoek te bevorderen en de waarde van onderzoek te verhogen is het dan ook absoluut noodzakelijk om initiatieven te versterken die meer nadruk leggen het belang van op rigoureus onderzoek en positieve publicatiepraktijken, en op onderzoek met een grotere maatschappelijke waarde.

Vertrouwen in de wetenschap kan wel degelijk worden ondermijnd, door suboptimale rapportagepraktijken en inadequate methodologie. Inspanningen om bevooroordeelde en onvolledige rapportage in biomedisch onderzoek te voorkomen of te reduceren moeten daarom worden gesteund. Dit is een gezamenlijke verantwoordelijkheid, van onderzoekers en auteurs, van peer reviewers en tijdschriftredacteuren, en van financiers en academische instellingen.

Le résumé

L'interprétation des données est subjective et peut conduire à des biais

L'élément humain dans le processus d'interprétation en science est subjectif et sujet à des préjugés [1]. Dans son article sur l'effet du biais interprétatif sur les preuves de la recherche, Kaptchuk soutient que la bonne science est incarnée dans la "tension entre l'empirisme des données concrètes et le rationalisme des convictions profondes" [1]. L'interprétation peut être fondée sur un bon jugement ou une erreur, mais la distinction ne peut être observée que rétrospectivement. Ioannidis et ses collègues notent également qu'un défi majeur pour les scientifiques est d'équilibrer la capacité à voir des modèles nouveaux et inattendus dans les données, tout en évitant simultanément l'apophonie - la tendance à voir des structures ou des modèles dans des données aléatoires [2]. La combinaison de l'apophénie et des biais d'interprétation peut facilement nous conduire à de fausses conclusions [2].

Resch et ses collègues ont documenté un exemple de biais de confirmation dans une étude contrôlée randomisée, dans laquelle 398 chercheurs ont été randomisés sans le savoir pour évaluer des rapports fictifs de traitement de l'obésité pour un journal respecté. Les rapports ne différaient que dans leur description de l'intervention de traitement : un traitement non prouvé mais crédible ou un traitement non conventionnel. Les examinateurs ont fait preuve d'un biais significatif en faveur du traitement crédible, défavorisant un rapport techniquement bon mais non conventionnel [3].

Les résultats expérimentaux sont généralement jugés en fonction des attentes, et les preuves qui ne sont pas conformes à des principes bien confirmés peuvent être écartées en trouvant de manière sélective des failles dans la conception ou la conduite de l'étude [1]. Lorsque les premiers essais contrôlés randomisés sur l'hormonothérapie substitutive (HTS) n'ont pas montré de réduction du risque de maladie coronarienne [4], les défenseurs de cette approche ont fait valoir que la maladie était bien trop avancée dans la population étudiée pour bénéficier du traitement, estimant qu'il était encore utile pour la prévention primaire [1].

Les premières preuves négatives en faveur de l'hormonothérapie substitutive auraient peut-être été plus facilement acceptées si le mécanisme physiopathologique n'avait pas créé une forte attente que le système cardiovasculaire bénéficie des oestrogènes [5].

Des biais potentiels peuvent également se produire avant la collecte des données. Le fait d'être convaincu de l'hypothèse peut affecter la collecte des données, entraînant ainsi un biais d'orientation. Des étudiants diplômés en psychologie ont découvert que les rats élevés spécialement pour la luminosité des labyrinthes avaient des performances supérieures à ceux élevés pour la grisaille des labyrinthes, bien que les deux groupes soient des rats de laboratoire standard attribués au hasard [6].

Des articles publiés dans *The Lancet ont* illustré le problème des déchets de la recherche au cours des différentes étapes de la recherche, qui englobent la conception, la conduite et le compte rendu. [7, 8] Étant donné qu'une grande partie de ces déchets sont évitables, il est nécessaire de développer et de mettre en œuvre des solutions. [7] Parmi ceux-ci, une interprétation et une présentation précises des résultats dans les données publiées sont

essentielles pour éviter de produire des études trompeuses et de gaspiller des ressources précieuses.

Contexte et objectifs

Le "Spin" est un concept standard dans les relations publiques et la politique, obtenu en fournissant une interprétation biaisée d'un événement afin de tromper l'opinion publique (https://en.wikipedia.org/wiki/Spin_(propagande)). Par exemple, la façon dont les nouvelles sont rapportées peut contenir des biais et des distorsions, et donc modifier la perception d'un événement, par des tactiques telles que la présentation sélective de faits spécifiques (c'est-à-dire "le choix des cerises"), ou la sous-estimation d'informations potentiellement négatives.

Le concept de "spin" a également été étudié dans les communications scientifiques. Les auteurs disposent d'une grande latitude pour interpréter et rapporter leurs résultats. [9] Le "spin" a été défini comme l'utilisation de pratiques de compte-rendu, pas nécessairement intentionnelles, "qui ne reflètent pas fidèlement la nature et la portée des résultats et qui pourraient affecter l'impression que les résultats produisent chez les lecteurs, une façon de déformer le compte-rendu scientifique sans pour autant mentir". [10] Plusieurs études ont montré que les auteurs d'études cliniques peuvent présenter et interpréter les résultats de leurs recherches avec une forme de manipulation. [9, 11-16] Le "spin", ou la représentation biaisée des résultats dans les rapports scientifiques, peut nuire aux patients et constitue une source de gaspillage évitable dans la recherche. [2, 7]

L'objectif principal de ce projet de doctorat était d'identifier et de documenter les pratiques sous-optimales de reporting dans les rapports publiés et de suggérer des stratégies privilégiées pour les surmonter. Nous nous sommes concentrés sur trois sujets clés: (1) étudier les pratiques de compte rendu sous-optimales, telles que la fausse représentation et la sur-interprétation des résultats des études, également appelées "spin", et les plans ou méthodes d'étude inadéquats, dans les études de biomarqueurs diagnostiques/pronostiques et les essais randomisés (chapitres 1-3); (2) a mis au point une intervention visant à réduire le "spin" et a évalué la faisabilité de la stratégie proposée, en réalisant un essai sur le terrain en collaboration avec le groupe de publication du BMJ (Londres, Royaume-Uni) (chapitre 4); et (3) a examiné d'autres aspects des pratiques de déclaration sous-optimales conduisant à des biais et à des gaspillages dans les publications scientifiques (chapitres 5-6). La discussion met en évidence des stratégies potentielles pour éviter ces problèmes et ces lacunes dans le processus de publication, dans le but ultime d'accroître la confiance et la valeur des rapports publiés sur la recherche clinique.

Chapitre 1: Une étude systématique révèle que les biais de spin ou d'interprétation sont nombreux dans les évaluations des biomarqueurs du cancer de l'ovaire

Ghannad M, Olsen M, Boutron I, Bossuyt PMM.
Journal of Clinical Epidemiology 2019;116:9-17

Dans cette étude, nous avons effectué une revue systématique descriptive de la présence de spin, encore classée comme une fausse représentation et une surinterprétation des résultats de l'étude, dans des études cliniques récentes évaluant la performance des biomarqueurs dans le cancer des ovaires. De nombreuses recherches ont été consacrées à la découverte de

biomarqueurs du cancer de l'ovaire, mais peu sont introduites avec succès dans les soins cliniques. Un certain nombre de facteurs, tels que les rapports biaisés et la mauvaise conception des études, ont été attribués au manque de succès dans l'identification de biomarqueurs cliniquement pertinents. Nous avons étudié les rapports et l'interprétation biaisés dans les articles publiés comme un facteur contributif potentiel, qui n'a pas été caractérisé auparavant dans les biomarqueurs du cancer de l'ovaire. La pratique de la fausse représentation ou de la surinterprétation fréquente des résultats des études peut conduire à un optimisme déséquilibré et injustifié dans l'interprétation des résultats des études sur la performance des biomarqueurs présumés.

Notre analyse de 200 évaluations récentes de biomarqueurs du cancer de l'ovaire a confirmé que 140 (70%) contenaient au moins une forme de spin (c'est-à-dire une fausse représentation ou une surinterprétation des résultats de l'étude) dans le titre, le résumé ou la conclusion du texte principal, exagérant la performance du biomarqueur. Les formes de spin les plus fréquentes identifiées sont les suivantes: (1) autres objectifs du biomarqueur déclaré non investigué (65; 32,5%); (2) décalage entre l'objectif visé et la conclusion (57; 28,5%); et (3) présentation incorrecte des résultats (40; 20%). Cet examen a confirmé que la présentation et l'interprétation biaisées sont courantes dans les récentes évaluations cliniques des biomarqueurs du cancer de l'ovaire. Ces résultats ont montré la nécessité de stratégies visant à réduire au minimum les biais dans la présentation et l'interprétation des résultats.

Chapitre 2: Lacunes dans les évaluations des biomarqueurs du cancer de l'ovaire : une étude systématique

Olsen M, **Ghannad M**, Lok C, Bossuyt PMM.
Chimie clinique et médecine de laboratoire 2019;58(1):3-10

Dans cette étude, nous avons constaté que les pratiques facilitant le spin, telles que les caractéristiques de conception sous-optimales et la communication inadéquate des méthodes, étaient également répandues dans les évaluations de biomarqueurs. Dans l'échantillon d'études décrit au chapitre 1, 93 (47 %) avaient acquis des échantillons et des données auprès d'un seul centre. La taille médiane de l'échantillon était de 156 patients (allant de 13 à 50 078), 5 (3 %) études seulement justifiant leur taille d'échantillon. Les études ont souvent recruté des patients dans des groupes disjoints 66 (33%), (c'est-à-dire des groupes de comparaison provenant d'un seul groupe d'étude), parfois avec des contrastes phénotypiques extrêmes ; 46 (23%) comprenaient des témoins sains et 5 (3%) comprenaient exclusivement des cas à un stade avancé. Les critères d'éligibilité et les méthodes d'échantillonnage ont rarement été signalés. Les rapports sur les méthodes étaient incomplets ; peu d'études ont fait état de critères d'admissibilité (10 %) et de méthodes d'échantillonnage (10 %). Cet examen a montré que les plans d'étude inadéquats étaient fréquents dans les évaluations cliniques du cancer de l'ovaire, ce qui pouvait conduire à des conclusions biaisées et prématurées sur la performance du marqueur dans les applications cliniques.

Chapitre 3: Aucune preuve trouvée d'une association entre les caractéristiques de l'essai et les effets du traitement dans les essais randomisés de la thérapie à la testostérone chez les hommes: étude méta-épidémiologique

Harin R, **Ghannad M**, Bertizzolo L, Page MJ.
Journal of Clinical Epidemiology 2020;122:12-19

Nous avons réalisé une étude méta-épidémiologique afin d'identifier les caractéristiques potentielles des essais associées aux estimations des effets de traitement rapportés dans les essais randomisés de la thérapie à la testostérone chez les hommes adultes. Sur 132 essais randomisés, 19 étaient des méta-analyses, comprenant les données de 10 725 participants. Aucune des caractéristiques de conception étudiées, y compris l'année de publication, la taille de l'échantillon, le statut d'enregistrement de l'essai, le statut du centre, la région, la source de financement et le conflit d'intérêt, n'a été associée de manière statistiquement significative aux effets de traitement signalés de la thérapie à la testostérone chez les hommes.

Dans nos analyses, un nombre important d'études ont été classées dans la catégorie "risque élevé/non clair de biais" (allant de 17% à 62% pour l'ensemble des domaines). Ce résultat confirme la prévalence élevée et persistante des rapports incomplets (des recherches antérieures suggèrent que 89 % des essais randomisés publiés comportent au moins un domaine de risque de biais "peu clair" [17]) et souligne la nécessité d'améliorer les rapports sur les essais.

Cet examen n'a pas montré de preuves claires que les caractéristiques des essais sont associées aux effets du traitement dans les essais randomisés de la thérapie à la testostérone chez les hommes. Cependant, les associations entre les caractéristiques des essais et les effets du traitement reflètent ce qui a été *rapporté* dans le rapport sur les essais, et pas nécessairement ce qui a été *fait* par les chercheurs. Ainsi, la présente étude méta- épidémiologique souligne la nécessité d'un rapport complet pour évaluer la sécurité et l'efficacité de la thérapie à la testostérone chez les hommes. Étant donné l'importance d'essais randomisés bien conçus et bien conduits pour la production de preuves de haute qualité, les futurs essais sur la thérapie à la testostérone devraient non seulement être réalisés de manière adéquate mais aussi faire l'objet d'un rapport transparent.

Chapitre 4: Un essai randomisé d'une intervention éditoriale visant à réduire le spin dans la conclusion des manuscrits du résumé n'a montré aucun effet significatif

Ghannad M, Yang B, Leeflang M, Aldcroft A, Bossuyt PMM, Schroter S, Boutron I.
Journal of Clinical Epidemiology 2020;130:69-77

À ce jour, la littérature biomédicale ne fait état d'aucune intervention visant à atténuer ou à réduire la prévalence du spin. Dans cette étude, nous avons développé une intervention éditoriale spécifique pour réduire le spin et avons mené un essai contrôlé randomisé à deux bras en groupes parallèles pour évaluer son impact. Nous avons mené cette étude en collaboration avec BMJ Open, une revue médicale générale. Notre résultat principal était la présence de spin; nos résultats secondaires étaient les types de spin et le changement de formulation dans la conclusion du résumé révisé. Les évaluateurs des résultats ont été aveuglés par la mission de l'intervention.

Sur les 184 manuscrits randomisés, 108 (54 interventions, 54 contrôles) ont été sélectionnés pour la révision et ont pu être évalués pour la présence de spin. La proportion de manuscrits présentant un effet de spirale était de 6 % inférieure (IC 95 % : 24 % inférieur à 13 % supérieur)

dans le groupe d'intervention (57 %, 31/54) que dans le groupe de contrôle (63 %, 34/54). La formulation de la conclusion du résumé révisé a été modifiée dans 34/54 (63 %) des manuscrits du groupe d'intervention et dans 26/54 (48 %) du groupe de contrôle. Les quatre types d'effets préétablis étaient concernés : (i) rapport sélectif (12 dans le groupe d'intervention contre 8 dans le groupe de contrôle) ; (ii) inclusion d'informations non étayées par des preuves (9 contre 9) ; et (iii) interprétation non conforme aux résultats de l'étude (14 contre 18) ; et (iv) recommandations injustifiées pour la pratique (5 contre 11).

Nos brèves instructions éditoriales aux auteurs ont peut-être amené ces derniers à revoir la formulation de leurs conclusions, mais cela n'a pas eu d'effet statistiquement significatif sur la réduction de la rotation dans les conclusions révisées des résumés et, sur la base de l'intervalle de confiance, l'existence d'un effet important peut être exclue. Ainsi, d'autres interventions visant à réduire le spin dans les rapports de recherche originaux devraient être évaluées.

Chapitre 5: Publications avec des scores Altmetrics élevés

Ghannad M, Ramezan R, Bossuyt PMM, Aronson JK, Wager E, Brassey J, Heneghan C
Soumettre à *Journal of Clinical Epidemiology*

Des mesures alternatives ont été développées pour mesurer l'attention que les publications reçoivent des médias d'information sociaux et des blogs. Dans cette étude, nous avons cherché à découvrir quels types d'études rapportées dans des articles de recherche récents publiés dans des revues médicales reçoivent les scores Altmetric les plus élevés, parmi ceux qui génèrent de l'attention dans les sources de données Altmetric. Dans les 324 articles de recherche primaire inclus dans notre étude, le score Altmetric médian était de 184 [Interquartile Range, 111 - 378]. La catégorie de revue et le facteur d'impact ont été associés de manière significative au score Altmetric ajusté (P <0,00001). L'intervention nutritionnelle (médiane = 4,69) a accumulé des scores Altmetric ajustés significativement plus élevés que le mode de vie et l'environnement (1,47), la pharmacologie (1,05) et d'autres interventions (0,82).
Nous avons observé que, dans les publications qui ont généré un score Altmetric de 50 ou plus, les rapports d'évaluation des interventions nutritionnelles étaient mentionnés plus souvent que les autres types d'interventions dans les médias d'information, les médias sociaux et les autres sources en ligne couvertes par Altmetric. Cela semble indiquer que les interventions qu'un large éventail de lecteurs peuvent appliquer dans leur vie quotidienne attirent davantage l'attention et sont discutées plus souvent que d'autres interventions en médecine et en santé publique.

Chapitre 6: Évaluer le contenu scientifique des journaux de prédateurs

Moher D, Shamseer L, Cobey KD, Lalu MM, Galipeau J, Avey MT, Ahmadzai N, Alabousi M, Barbeau P, Beck A, Daniel R. Frank R, **Ghannad M,** Hamel C, Hersi M, Hutton B, Isupov I, McGrathTA, McInnes MDF, Page MJ, Pratt M, Pussegoa K, Shea B, Srivastava A, Stevens A, Thavorn K, van Katwyck S, Ward R, Wolfe D, Yazdi F, Yu AM, Ziai H.
Nature News. 2017;549(7670):23

Les entités connues sous le nom de revues et d'éditeurs "prédateurs" imprègnent le monde de l'édition scientifique, mais on sait peu de choses sur les articles qu'elles publient. Nous avons examiné un échantillon de 1907 études biomédicales humaines et animales, en enregistrant

leurs modèles d'étude, leurs caractéristiques épidémiologiques et leurs rapports. Dans notre échantillon, plus de deux millions d'humains et plus de huit mille animaux ont été inclus dans des publications prédatrices. Seulement 40 % des études déclarent avoir reçu une approbation éthique. Sur les 17 % d'articles mentionnant leur source de financement, les Instituts nationaux de la santé des États-Unis ont été les plus souvent cités. Les auteurs correspondants étaient le plus souvent originaires d'Inde (511/1907, 26,8 %) et des États-Unis (288/1907, 15,1 %). La qualité des travaux signalés dans notre échantillon était médiocre et pire que celle des échantillons contemporains de la littérature légitime. Dans de nombreuses études, il manquait des détails méthodologiques et des résultats clés. Nos résultats soulèvent d'importantes questions éthiques, car les recherches effectuées dans des revues prédatrices sont difficiles à identifier et ne sont pas indexées dans des bases de données biomédicales scientifiquement établies. Les bailleurs de fonds et les institutions universitaires doivent élaborer des politiques explicites pour éloigner les bénéficiaires de subventions et les auteurs potentiels de ces entités.

Perspectives d'avenir et remarques finales

L'inconduite en matière de recherche est un terme général qui englobe différents domaines. Pour la définition largement acceptée (par exemple, la falsification, la fabrication et le plagiat (FFP)), de nombreuses politiques réglementaires et éthiques sont en échec. Toutefois, un défi important demeure : au-delà de la FFP, il y a actuellement un manque de consensus sur les types de comportement qui constituent une inconduite en matière de recherche [160], ce qui entrave par la suite les stratégies de détection et de limitation de ces inconduites.

Les pratiques "tournantes" sont généralement perçues comme acceptables par les chercheurs et ne sont pas considérées comme des pratiques de recherche préjudiciables, malgré des preuves antérieures documentant certaines de leurs conséquences négatives [10]. Cette perception parmi les chercheurs peut être due en grande partie à un manque de sensibilisation. Des efforts continus pour documenter et caractériser les stratégies spécifiques de "spin" dans les domaines émergents sont nécessaires pour continuer à s'appuyer sur les preuves précédentes et pour améliorer notre compréhension de ce concept dans chaque domaine de recherche. Nous pourrions également approfondir les preuves en documentant les conséquences négatives de la rotation dans les rapports de recherche, car cela pourrait encourager davantage les chercheurs et les parties prenantes à considérer la rotation comme une pratique de recherche préjudiciable. Les programmes éducatifs existants pour les chercheurs en début de carrière peuvent être enrichis par la mise en œuvre d'initiatives de mentorat et de formation, en sensibilisant les auteurs aux formes et aux facilitateurs du spin et à son impact.

La simple caractérisation du spin et la documentation de ses conséquences négatives ne suffiront probablement pas à changer les perceptions et les pratiques. La prise de conscience des comportements actuels n'est qu'une étape vers le changement des comportements et des pratiques de recherche. Des interventions actives sont nécessaires pour développer des stratégies qui facilitent le changement. Puisque nous avons observé un effet limité de notre intervention éditoriale, nous devrions envisager d'autres stratégies pour améliorer le reportage. Celles-ci peuvent inclure, sans s'y limiter, le pré-enregistrement du plan d'étude, du ou des résultats primaires et du plan d'analyse comme une forme très efficace d'aveuglement, étant donné que les données n'existent pas et que les résultats ne sont pas encore connus au moment

du lancement de l'étude. Une autre stratégie à envisager peut consister à réunir des équipes diverses et multidisciplinaires qui incluent des statisticiens dans les équipes de recherche, ou à inclure un statisticien dans le comité de rédaction de la revue, afin d'aider à garantir la conduite et l'interprétation rigoureuses et solides de la méthodologie de recherche. Ainsi, on limitera la possibilité de faire tourner les résultats et les conclusions, on réduira les biais et on améliorera la transparence de la recherche médicale.

Le rôle des bailleurs de fonds et des institutions universitaires est essentiel pour diffuser l'intégrité de la recherche et les meilleures pratiques de recherche. Actuellement, la plupart des comités de promotion et de titularisation des universités mettent l'accent et récompensent l'impact perçu de la recherche avec des critères (bien qu'étroits) axés sur les publications et les mesures associées (c'est-à-dire les facteurs d'impact, les citations ou Altmetrics) plutôt que sur la rigueur. Les questions liées aux pratiques de publication, telles que les revues "prédatrices", sont un symptôme de ce système imparfait de "publier ou périr". Pour favoriser l'intégrité de la recherche et accroître la valeur de la recherche, il est impératif de poursuivre les initiatives actuelles qui mettent davantage l'accent sur la recherche rigoureuse et d'autres pratiques de publication positives ayant une plus grande valeur pour la société.

La confiance dans la science peut être érodée par le recours fréquent à des pratiques de publication sous-optimales et à une méthodologie inadéquate. Les efforts visant à prévenir ou à réduire les rapports biaisés et incomplets dans la recherche biomédicale doivent être entrepris avec vigueur et à l'unisson, étant donné la complexité complexe qui implique de multiples acteurs. Les chercheurs et les auteurs, les pairs évaluateurs et les rédacteurs de revues, les bailleurs de fonds et les institutions universitaires partagent sans aucun doute la responsabilité.

References

[1] Kaptchuk TJ. Effect of interpretive bias on research evidence. BMJ 2003;326(7404):1453-5.

[2] Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature Human Behaviour 2017;1:0021.

[3] Herper M. Data show panic and disorganization dominate the study of Covid-19 drugs, https://www.statnews.com/2020/07/06/data-show-panic-and-disorganization-dominate-the-study-of-covid-19-drugs/; 2020 2020].

[4] Resch KI, Ernst E, Garrow J. A randomized controlled study of reviewer bias against an unconventional therapy. J R Soc Med 2000;93(4):164-7.

[5] Herrington DM, Reboussin DM, Brosnihan KB, Sharp PC, Shumaker SA, Snyder TE, et al. Effects of estrogen replacement on the progression of coronary-artery atherosclerosis. N Engl J Med 2000;343(8):522-9.

[6] Nabel EG. Coronary heart disease in women--an ounce of prevention. N Engl J Med 2000;343(8):572-4.

[7] Rosenthal R, Persinger GW, Kline LV, Mulry RC. The Role of the Research Assistant in the Mediation of Experimenter Bias. J Pers 1963;31:313-35.

[8] Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383(9912):101-4.

[9] Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374(9683):86-9.

[10] Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A 2018;115(11):2613-9.

[11] Bailar JC. How to Distort the Scientific Record without Actually Lying: Truth and the Arts of Science. European Journal of Oncology 2006;11(4):217-24.

[12] 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.

[13] 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.

[14] Mathieu S, Giraudeau B, Soubrier M, Ravaud P. Misleading abstract conclusions in randomized controlled trials in rheumatology: comparison of the abstract conclusions and the results section. Joint Bone Spine 2012;79(3):262-7.

[15] 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.

[16] 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(36):4120-6.

[17] 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(2):581-8.
[18] Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and Interpretation of Randomized Controlled Trials With Statistically Nonsignificant Results for Primary Outcomes. Jama-J Am Med Assoc 2010;303(20):2058-64.
[19] 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.
[20] 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(36):4120-U346.
[21] 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.
[22] 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(8):1353-62.
[23] Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: A methodological systematic review. PLoS Biol 2017;15(9):e2002173.
[24] Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PloS one 2016;11(7):e0159267.
[25] Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol 2016;77:24-37.
[26] Warren HR, Raison N, Dasgupta P. The Rise of Altmetrics. JAMA 2017;317(2):131-2.
[27] Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. Am J Clin Nutr 2013;97(1):127-34.
[28] Ioannidis JPA, Bossuyt PMM. Waste, Leaks, and Failures in the Biomarker Pipeline. Clin Chem 2017;63(5):963-72.
[29] Ioannidis JP. Biomarker failures. Clin Chem 2013;59(1):202-4.
[30] 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.
[31] Pepe MS, Feng ZD. Improving Biomarker Identification with Better Designs and Reporting. Clinical Chemistry 2011;57(8):1093-5.
[32] 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.

[33] Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383(9912):101-4.
[34] 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.
[35] Pavlou MP, Diamandis EP, Blasutig IM. The long journey of cancer biomarkers from the bench to the clinic. Clin Chem 2013;59(1):147-57.
[36] 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.
[37] 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(3):560-70.
[38] 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.
[39] 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.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. PLoS One 2012;7(4):e35621.
[45] 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(4):387-91.
[46] 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.

[47] Liao CI, Chow S, Chen Lm, Kapp DS, Mann A, Chan JK. Trends in the incidence of serous fallopian tube, ovarian, and peritoneal cancer in the US. Gynecologic Oncology 2018;149(2):318-23.
[48] Timmermans M, Sonke GS, Van de Vijver KK, van der Aa MA, Kruitwagen RFPM. No improvement in long-term survival for epithelial ovarian cancer patients: A population-based study between 1989 and 2014 in the Netherlands. European Journal of Cancer 2018;88:31-7.
[49] Diamandis EP. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? 10. 2012.
[50] Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: The long and uncertain path to clinical utility. 24. 2006:971-83.
[51] Pavlou MP, Diamandis EP, Blasutig IM. The long journey of cancer biomarkers from the bench to the clinic. 59. 2013:147-57.
[52] 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. Clinica Chimica Acta 2014;427:49-57.
[53] Duffy MJ, Sturgeon CM, Sölétormos G, Barak V, Molina R, Hayes DF, et al. Validation of new cancer biomarkers: A position statement from the european group on tumor markers. Clinical Chemistry 2015;61(6):809-20.
[54] Ioannidis JPA, Bossuyt PMM. Waste, leaks, and failures in the biomarker pipeline. 63. American Association for Clinical Chemistry Inc.; 2017:963-72.
[55] Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, Van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(4):469-76.
[56] Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: Standards for study design. 100. 2008:1432-8.
[57] Pepe MS, Feng Z. Improving biomarker identification with better designs and reporting. 57. 2011:1093-5.
[58] Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. 5. 2005:142-9.
[59] Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. 60. Elsevier USA; 2007:1205-19.
[60] Diamandis EP. Cancer biomarkers: Can we turn recent failures into success? 102. Oxford University Press; 2010:1462-7.
[61] Ioannidis JPA. Biomarker failures. 59. 2013:202-4.
[62] Leung F, Diamandis EP, Kulasingam V. Ovarian cancer biomarkers: Current state and future implications from high-throughput technologies. Academic Press Inc.; 2014, p. 25-77.
[63] Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: A systematic review and classification. 19. 2013:4578-88.
[64] Strimbu K, Tavel JA. What are biomarkers? 5. 2010:463-6.

[65] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van Der Meulen JHP, et al. Empirical evidence of design-related bias in studies of diagnostic tests. Journal of the American Medical Association 1999;282(11):1061-6.
[66] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: The cross-sectional study. Journal of Clinical Epidemiology 2003;56(11):1118-28.
[67] Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. 155. American College of Physicians; 2011:529-36.
[68] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015;12(10):e1001885.
[69] Pearl ML, Dong H, Tulley S, Zhao Q, Golightly M, Zucker S, et al. Treatment monitoring of patients with epithelial ovarian cancer using invasive circulating tumor cells (iCTCs). Gynecologic Oncology 2015;137(2):229-38.
[70] Chen X, Paranjape T, Stahlhut C, McVeigh T, Keane F, Nallur S, et al. Targeted resequencing of the microRNAome and 3′UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer. 34. Nature Publishing Group; 2014:2125-37.
[71] Ransohoff DF, Gourlay ML. Sources of bias in specimens for research about molecular markers for cancer. Journal of Clinical Oncology 2010;28(4):698-704.
[72] Furukawa TA, Guyatt GH. Sources of bias in diagnostic accuracy studies and the diagnostic process. 174. 2006:481-2.
[73] Vandenbroucke JP, Von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. 18. 2007:805-35.
[74] Moore HM, Kelly AB, Jewell SD, McShane LM, Clark DP, Greenspan R, et al. Biospecimen reporting for improved study quality (BRISQ). Cancer Cytopathology 2011;119(2):92-102.
[75] Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. PLoS Med 2012;9(5):e1001216.
[76] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Annals of Internal Medicine 2015;162(1):55-63.
[77] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11):e012799.
[78] Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. 8. 2010.
[79] Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. 383. Lancet Publishing Group; 2014:267-76.

[80] Korevaar DA, Wang J, Van Enst WA, Leeflang MM, Hooft L, Smidt N, et al. Reporting diagnostic accuracy studies: Some improvements after 10 years of STARD. Radiology 2015;274(3):781-9.
[81] Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. 101. 2009:1446-52.
[82] Henry NL, Hayes DF. Cancer biomarkers. 6. John Wiley and Sons Ltd; 2012:140-6.
[83] The Ovarian Tumor Tissue Analysis consortium (OTTA), https://ottaconsortium.org.
[84] Ovarian Cancer Association Consortium (OCAC), http://apps.ccge.medschl.cam.ac.uk/consortia/ocac//aims/aims.html.
[85] Monaghan PJ, Lord SJ, St John A, Sandberg S, Cobbaert CM, Lennartz L, et al. Biomarker development targeting unmet clinical needs. 460. Elsevier B.V.; 2016:211-9.
[86] Nieschlag E, Nieschlag S. ENDOCRINE HISTORY: The history of discovery, synthesis and development of testosterone for clinical use. European journal of endocrinology / European Federation of Endocrine Societies 2019;180(6):R201-R12.
[87] Snyder PJ, Bhasin S, Cunningham GR, Matsumoto AM, Stephens-Shields AJ, Cauley JA, et al. Effects of Testosterone Treatment in Older Men. The New England journal of medicine 2016;374(7):611-24.
[88] Walther A, Breidenstein J, Miller R. Association of Testosterone Treatment With Alleviation of Depressive Symptoms in Men: A Systematic Review and Meta-analysis. JAMA psychiatry 2019;76(1):31-40.
[89] Bhasin S, Brito JP, Cunningham GR, Hayes FJ, Hodis HN, Matsumoto AM, et al. Testosterone Therapy in Men With Hypogonadism: An Endocrine Society Clinical Practice Guideline. The Journal of clinical endocrinology and metabolism 2018;103(5):1715-44.
[90] Savovic J, Harris RJ, Wood L, Beynon R, Altman D, Als-Nielsen B, et al. Development of a combined database for meta-epidemiological research. Research synthesis methods 2010;1(3-4):212-25.
[91] Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ (Clinical research ed 2011;343:d5928.
[92] Page MJ, Higgins JP. Rethinking the assessment of risk of bias due to selective reporting: a cross-sectional study. Systematic reviews 2016;5(1):108.
[93] Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Statistics in medicine 2002;21(11):1513-24.
[94] Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Statistics in medicine 2000;19(22):3127-31.
[95] Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, Bryant DM, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ (Clinical research ed 2007;334(7597):786.
[96] Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in

randomised controlled trials: meta-epidemiological study. BMJ (Clinical research ed 2013;346:f457.
[97] Alexander GC, Iyer G, Lucas E, Lin D, Singh S. Cardiovascular Risks of Exogenous Testosterone Use Among Men: A Systematic Review and Meta-Analysis. The American journal of medicine 2017;130(3):293-305.
[98] Nian Y, Ding M, Hu S, He H, Cheng S, Yi L, et al. Testosterone replacement therapy improves health-related quality of life for patients with late-onset hypogonadism: a meta-analysis of randomized controlled trials. Andrologia 2017;49(4).
[99] Kohn TP, Mata DA, Ramasamy R, Lipshultz LI. Effects of Testosterone Replacement Therapy on Lower Urinary Tract Symptoms: A Systematic Review and Meta-analysis. Eur Urol 2016;69(6):1083-90.
[100] Corona G, Giagulli VA, Maseroli E, Vignozzi L, Aversa A, Zitzmann M, et al. THERAPY OF ENDOCRINE DISEASE: Testosterone supplementation and body composition: results from a meta-analysis study. European journal of endocrinology / European Federation of Endocrine Societies 2016;174(3):R99-116.
[101] Neto WK, Gama EF, Rocha LY, Ramos CC, Taets W, Scapini KB, et al. Effects of testosterone on lean mass gain in elderly men: systematic review with meta-analysis of controlled and randomized studies. Age (Dordr) 2015;37(1):9742.
[102] Kang DY, Li HJ. The effect of testosterone replacement therapy on prostate-specific antigen (PSA) levels in men being treated for hypogonadism: a systematic review and meta-analysis. Medicine 2015;94(3):e410.
[103] Grossmann M, Hoermann R, Wittert G, Yeap BB. Effects of testosterone treatment on glucose metabolism and symptoms in men with type 2 diabetes and the metabolic syndrome: a systematic review and meta-analysis of randomized controlled clinical trials. Clin Endocrinol (Oxf) 2015;83(3):344-51.
[104] Borst SE, Shuster JJ, Zou B, Ye F, Jia H, Wokhlu A, et al. Cardiovascular risks and elevation of serum DHT vary by route of testosterone administration: a systematic review and meta-analysis. BMC medicine 2014;12:211.
[105] Corona G, Isidori AM, Buvat J, Aversa A, Rastrelli G, Hackett G, et al. Testosterone supplementation and sexual function: a meta-analysis study. J Sex Med 2014;11(6):1577-92.
[106] Amanatkar HR, Chibnall JT, Seo BW, Manepalli JN, Grossberg GT. Impact of exogenous testosterone on mood: a systematic review and meta-analysis of randomized placebo-controlled trials. Annals of clinical psychiatry : official journal of the American Academy of Clinical Psychiatrists 2014;26(1):19-32.
[107] Cui Y, Zong H, Yan H, Zhang Y. The effect of testosterone replacement therapy on prostate cancer: a systematic review and meta-analysis. Prostate Cancer Prostatic Dis 2014;17(2):132-43.
[108] Xu L, Freeman G, Cowling BJ, Schooling CM. Testosterone therapy and cardiovascular events among men: a systematic review and meta-analysis of placebo-controlled randomized trials. BMC medicine 2013;11:108.
[109] Cui Y, Zhang Y. The effect of androgen-replacement therapy on prostate growth: a systematic review and meta-analysis. Eur Urol 2013;64(5):811-22.

[110] Price J, Leng GC. Steroid sex hormones for lower limb atherosclerosis. Cochrane database of systematic reviews (Online) 2012;10:CD000188.
[111] Toma M, McAlister FA, Coglianese EE, Vidi V, Vasaiwala S, Bakal JA, et al. Testosterone supplementation in heart failure: a meta-analysis. Circulation Heart failure 2012;5(3):315-21.
[112] Corona G, Rastrelli G, Monami M, Guay A, Buvat J, Sforza A, et al. Hypogonadism as a risk factor for cardiovascular mortality in men: a meta-analytic study. European journal of endocrinology / European Federation of Endocrine Societies 2011;165(5):687-701.
[113] Corona G, Monami M, Rastrelli G, Aversa A, Sforza A, Lenzi A, et al. Type 2 diabetes mellitus and testosterone: a meta-analysis study. International journal of andrology 2011;34(6 Pt 1):528-40.
[114] Fernandez-Balsells MM, Murad MH, Lane M, Lampropulos JF, Albuquerque F, Mullan RJ, et al. Clinical review 1: Adverse effects of testosterone therapy in adult men: a systematic review and meta-analysis. The Journal of clinical endocrinology and metabolism 2010;95(6):2560-75.
[115] Zarrouf FA, Artz S, Griffith J, Sirbu C, Kommor M. Testosterone and depression: systematic review and meta-analysis. Journal of psychiatric practice 2009;15(4):289-305.
[116] Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ (Clinical research ed 2013;346:f2304.
[117] Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ (Clinical research ed 2012;344:e1119.
[118] Chartres N, Fabbri A, McDonald S, Turton J, Allman-Farinelli M, McKenzie J, et al. Association of industry ties with outcomes of studies examining the effect of wholegrain foods on cardiovascular disease and mortality: systematic review and meta-analysis. BMJ open 2019;9(5):e022912.
[119] Baillargeon J, Kuo YF, Westra JR, Urban RJ, Goodwin JS. Testosterone Prescribing in the United States, 2002-2016. JAMA 2018;320(2):200-2.
[120] Baillargeon J, Urban RJ, Ottenbacher KJ, Pierson KS, Goodwin JS. Trends in androgen prescribing in the United States, 2001 to 2011. JAMA internal medicine 2013;173(15):1465-6.
[121] Odutayo A, Emdin CA, Hsiao AJ, Shakir M, Copsey B, Dutton S, et al. Association between trial registration and positive study findings: cross sectional study (Epidemiological Study of Randomized Trials-ESORT). BMJ (Clinical research ed 2017;356:j917.
[122] Unverzagt S, Prondzinsky R, Peinemann F. Single-center trials tend to provide larger treatment effects than multicenter trials: a systematic review. J Clin Epidemiol 2013;66(11):1271-80.
[123] Jorgensen L, Paludan-Muller AS, Laursen DR, Savovic J, Boutron I, Sterne JA, et al. Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials:

overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. Systematic reviews 2016;5:80.
[124] World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191-4.
[125] 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.
[126] Ghannad M, Olsen M, Boutron I, Bossuyt PM. A systematic review finds that spin or interpretation bias is abundant in evaluations of ovarian cancer biomarkers. J Clin Epidemiol 2019;116:9-17.
[127] Reynolds-Vaughn V, Riddle J, Brown J, Schiesel M, Wayant C, Vassar M. Evaluation of Spin in the Abstracts of Emergency Medicine Randomized Controlled Trials. Ann Emerg Med 2019:423-31.
[128] Bero L. Meta-research matters: Meta-spin cycles, the blindness of bias, and rebuilding trust. PLoS Biol 2018;16(4):e2005972.
[129] Schroter S, Loder E, Godlee F. Research on peer review and biomedical publication. BMJ 2020;368:m661.
[130] Bruce R, Chauvin A, Trinquart L, Ravaud P, Boutron I. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. BMC Med 2016;14(1):85.
[131] Dancey JE. From quality of publication to quality of care: translating trials to practice. J Natl Cancer Inst 2010;102(10):670-1.
[132] Moher D. Reporting guidelines: doing better for readers. BMC Med 2018;16(1):233.
[133] Enhancing the QUAlity and Transparency Of health Research, http://www.equator-network.org/; [accessed 29 May.2020].
[134] Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database Syst Rev 2012;11:MR000030.
[135] Stevens A, Shamseer L, Weinstein E, Yazdi F, Turner L, Thielman J, et al. Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. BMJ 2014;348:g3804.
[136] Caulley L, Catala-Lopez F, Whelan J, Khoury M, Ferraro J, Cheng W, et al. Reporting guidelines of health research studies are frequently used inappropriately. J Clin Epidemiol 2020;122:87-94.
[137] Hopewell S, Boutron I, Altman DG, Barbour G, Moher D, Montori V, et al. Impact of a web-based tool (WebCONSORT) to improve the reporting of randomised trials: results of a randomised controlled trial. BMC Med 2016;14(1):199.
[138] van der Steen JT, Ter Riet G, van den Bogert CA, Bouter LM. Causes of reporting bias: a theoretical framework. F1000Res 2019;8:280.
[139] Haneef R, Ravaud P, Baron G, Ghosn L, Boutron I. Factors associated with online media attention to research: a cohort study of articles evaluating cancer treatments. Res Integr Peer Rev 2017;2:9.

[140] Bornmann L, Haunschild R. Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. PLoS One 2018;13(5):e0197133.
[141] David Colquhoun AP. Why Altmetrics is bad for science—and healthcare. *The BMJ Opinion.* The BMJ; 2014.
[142] Ioannidis JPA. Neglecting Major Health Problems and Broadcasting Minor, Uncertain Issues in Lifestyle Science. JAMA 2019:1-2.
[143] Johnston BC, Zeraatkar D, Han MA, Vernooij RWM, Valli C, El Dib R, et al. Unprocessed Red Meat and Processed Meat Consumption: Dietary Guideline Recommendations From the Nutritional Recommendations (NutriRECS) Consortium. Ann Intern Med 2019.
[144] Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella D, Aros F, et al. Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med 2013;368(14):1279-90.
[145] Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella D, Aros F, et al. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. N Engl J Med 2018;378(25):e34.
[146] Wikipedia. General medical journal, https://en.wikipedia.org/wiki/General_medical_journal#:~:text=A%20general%20medical%20journal%20is,a%20specific%20field%20of%20medicine.; 2020 [accessed April.2020].
[147] Altmetric. How is the Altmetric Attention Score calculated?, https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated-; 2019 [accessed April.2020].
[148] Altmetric. Putting the Altmetric Attention Score in context, https://help.altmetric.com/support/solutions/articles/6000060970-putting-the-altmetric-attention-score-in-context; 2019 [accessed April.2020].
[149] Serghiou S, Ioannidis JPA. Altmetric Scores, Citations, and Publication of Studies Posted as Preprints. JAMA 2018;319(4):402-4.
[150] Fang Z, Costas R, Tian W, Wang X, Wouters P. An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics. Scientometrics 2020:1-31.
[151] Catalogue of Bias Collaboration AJ, Bankhead C, Nunan D. Hot stuff bias, https://catalogofbias.org/biases/hot-stuff-bias/; 2017 [accessed September.2020].
[152] Copiello S. Other than detecting impact in advance, alternative metrics could act as early warning signs of retractions: tentative findings of a study into the papers retracted by PLoS ONE. Scientometrics 2020.
[153] Patel RB, Vaduganathan M, Bhatt DL, Bonow RO. Characterizing High-Performing Articles by Altmetric Score in Major Cardiovascular Journals. JAMA Cardiol 2018;3(12):1249-51.
[154] Sathianathen NJ, Lane Iii R, Murphy DG, Loeb S, Bakker C, Lamb AD, et al. Social Media Coverage of Scientific Articles Immediately After Publication Predicts Subsequent Citations - #SoME_Impact Score: Observational Analysis. J Med Internet Res 2020;22(4):e12288.

[155] Piwowar H. Altmetrics: Value all research products. Nature 2013;493(7431):159.
[156] Furman JL, Jensen K, Murray F. Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. Research Policy 2012;41(2):276-90.
[157] Shema H, Hahn O, Mazarakis A, Peters I. Retractions from altmetric and bibliometric perspectives. Information - Wissenschaft & Praxis 2019;70(2-3):98.
[158] Lee CJ, Nagler RH, Wang N. Source-specific Exposure to Contradictory Nutrition Information: Documenting Prevalence and Effects on Adverse Cognitive and Behavioral Outcomes. Health Commun 2018;33(4):453-61.
[159] Nagler RH. Adverse outcomes associated with media exposure to contradictory nutrition messages. J Health Commun 2014;19(1):24-40.
[160] Dal-Ré R, Bouter LM, Cuijpers P, Gluud C, Holm S. Should research misconduct be criminalized? Research Ethics 2020;16(1-2):1-12.
[161] Moher D, Bouter L, Kleinert S, Glasziou P, Sham MH, Barbour V, et al. The Hong Kong Principles for assessing researchers: Fostering research integrity. PLoS Biol 2020;18(7):e3000737.

Publications

PUBLICATIONS INCLUDED IN THIS THESIS

Ghannad M, Olsen M, Boutron I, Bossuyt PMM. A systematic review finds that spin or interpretation bias is abundant in evaluations of ovarian cancer biomarkers. J Clin Epidemiol. 2019 Dec;116:9-17. doi: 10.1016/j.jclinepi.2019.07.011.

Olsen M, **Ghannad M**, Lok C, Bossuyt PMM. Shortcomings in the evaluation of biomarkers in ovarian cancer: a systematic review. Clin Chem Lab Med. 2019 Dec 18;58(1):3-10. doi: 10.1515/cclm-2019-0038.

Haring R, **Ghannad M**, Bertizzolo L, Page MJ. No evidence found for an association between trial characteristics and treatment effects in randomized trials of testosterone therapy in men: a meta-epidemiological study. J Clin Epidemiol. 2020 Jun;122:12-19. doi: 10.1016/j.jclinepi.2020.02.004. Epub 2020 Feb 24.

Ghannad M, Yang B, Leeflang M, Aldcroft A, Bossuyt PMM, Schroter S, Boutron I. A randomised trial of an editorial intervention to reduce spin in the abstract's conclusion of manuscripts showed no significant effect. J Clin Epidemiol. 2020 Oct 20:S0895-4356(20)31153-7. doi: 10.1016/j.jclinepi.2020.10.014.

Moher D, Shamseer L, Cobey KD, Lalu MM, Galipeau J, Avey MT, Ahmadzai N, Alabousi M, Barbeau P, Beck A, Daniel R, Frank R, **Ghannad M**, et al. Stop this waste of people, animals and money. Nature. 2017 Sep 6;549(7670):23-25. doi: 10.1038/549023a.

Ghannad M, Ramezan R, Bossuyt PMM, Wager E, Aronson JK, Brassey J, Heneghan C. Publications with high Altmetrics scores. (Submitted)

OTHER PUBLICATIONS

Ghannad M, Dennehy M, la Porte CJL, Seguin I, Tardiff D, Mallick R, Sabri E, Zhang GJ; Kanji S, Cameron DW. (2019) A drug interaction study investigating the effect of Rifabutin on the pharmacokinetics of Maraviroc in healthy subjects. PLOS ONE 14(10): e0223969. doi: 10.1371/journal.pone.0223969

Fergusson D, Monfaredi Z, Pussegoda K, Garritty C, Lyddiatt A, Shea B, Duffett L, **Ghannad M**, … Yazdi F. (2018) The prevalence of patient engagement in published trials: a systematic review. Research involvement and engagement 4(17). doi: 10.1186/s40900-018-0099-x

Hamel C, **Ghannad M**, McInnes MDF, Marshall J, Earnshaw J, Ward R, Skidmore B, Garritty C. (2017) Potential benefits and harms of offering ultrasound surveillance to men aged 65 years and older with a subaneurysmal (2.5-2.9 cm) infrarenal aorta. Journal of Vascular Surgery 67(4): 1298-1307. doi: 10.1016/j.jvs.2017.11.074

Augustine S, Avey M T, Harrison B, Locke T, **Ghannad M**, Moher D, Thébaud B. (2017) Mesenchymal Stromal Cell Therapy in Bronchopulmonary Dysplasia: Systematic Review and Meta-Analysis of Preclinical Studies. Stem Cells Translational Medicine 6(12): 2079-2093. doi: 10.1002/sctm.17-0126

PhD portfolio

Promotors: Patrick M Bossuyt and Isabelle Boutron
PhD Period: October 2016 – September 2020

Courses at the AMC Graduate School	**Year**
AMC world of science	2017
Clinical Epidemiology 1: Randomized Controlled Trials	2017
Clinical Epidemiology 2: Observational Clinical Epidemiology – Effects and Effectiveness	2017
Clinical Epidemiology 3: Evaluation of medical tests	2018
Clinical Epidemiology 4: Systematic Reviews	2018
Genetic Epidemiology	2018
Computing in R	2020

Courses at other institutions	**Year**
Methods in diagnostic tests, biomarkers, and screening evaluation (Université de Paris)	2016
32nd Residential 3-week Summer Course in Epidemiology, 9 ECTs, (Florence, Italy)	2019
Catalogue of Bias workshop organized by the Centre for Evidence-Based Medicine (CEBM), Stratford Upon Avon, Warwickshire (UK)	2019
Dutch language course (University of Amsterdam)	2020

MiRoR PhD Training	**Year**
Ghent, Belgium (3 days)	
Introduction to research and waste in research and reproducibility, protocol writing, data management plan, communication of research results, review of basic statistical concepts	2016
Liverpool, UK (3 days)	
Introduction to qualitative research methods, core outcome measures and reporting bias, patient and public involvement in research, evidence-based guideline development, statistical analysis plan, and data sharing	2017
Amsterdam, Netherlands (3 days)	
Scientific communication and writing, bibliometrics, research integrity, higher education institutions and responsible research and innovation, fundamentals of natural language processing, short introductions to supervised machine learning	2017
Split, Croatia (3 days)	2018
Writing research grant proposals, research ethics in H2020,	
Barcelona, Spain (3 days)	2019
Peer-review, career planning and job applications, introduction to Python	

Conferences, symposiums and scientific meetings	Year
Presentation of PhD project at the METHODS team meeting, Inserm / Université de Paris, Paris	2017
Oral presentation at the International Congress on Peer Review and Scientific Publication, Chicago	2017
Presentation link: https://www.youtube.com/watch?v=5MkpZDLqXIY	
Poster presentation at the Amsterdam Public Health Meeting – **awarded best poster for Methodology**	2017
Oral presentation at EBM Live, Oxford	2018
Poster presentation at MEMTAB (Methods for Evaluation of medical prediction Models, Tests And Biomarkers), Amsterdam	2018
Oral presentation at Cochrane Colloquium, Edinburgh	2018
Oral presentation at the 6th World Conference on Research Integrity Doctoral Forum, Hong Kong	2019
Oral presentation at the Centre for Evidence-Based Medicine (CEBM), Oxford	2019
Poster presentation and contribution to the booklet for the "Meta-research for transforming clinical research" conference organized by the MiRoR consortium at the Académie Nationale de Médecine, Paris	2019
Oral presentation at the REWARD-EQUATOR conference, Berlin (February 2020) – **selected as one of the top four scoring abstracts from more than 150 submitted abstracts**	2020
Oral presentation at the 2nd PEERE International Conference on Peer Review, Valencia (online)	2020

Other activities	Year
Master class with prof JPA Ioannidis	2016
MiRoR newsletter	2017
Conducted and published an interview with Jacques Demotes-Mainard, Director at ECRIN, Director General, ECRIN for the 2nd issue of MiRoR newsletter.	
MiRoR Common Project:	2017-2018
The MiRoR common project went beyond each individual project and involved the whole group in an innovative project, focusing on Questionable Research Practices (QRP).	
Assisting a Cochrane workshop in ME-methods, Edinburgh	2018
MiRoR Challenge Project	2018-2019
The MiRoR Challenge Project aimed to initiate entrepreneurial and innovative skills among students. Divided into 4 groups, the objective was to propose a scientific project considering all aspects (development of the intervention, evaluation of the intervention, ethics, financial plan, project management, communication of results). Each team presented their project to the consortium and to a group of 4 external experts who rated the projects.	
Junior board member, Amsterdam Public Health (APH) Methodology	2019

Awards	Year
Full doctoral fellowship in a Marie Curie joint doctorate training program, Methods in Research on Research Innovative Training Network (http://miror-ejd.eu)	2016
AMC Young talent fund (3.687,50 euro) The AMC Young Talent Fund provides young excellent AMC PhD candidates the unique opportunity to follow a course or an internship at one of the top international research institutes. Grants are awarded after an assessment of the applications by a selection committee. The AMC Young Talent Fund is made possible by private donations to the AMC Foundation. Dutch language course (University of Amsterdam)	2019


PARIOS
SCIENCE