Journal of Clinical Epidemiology 158 (2023) 99−110

Journal of Clinical Epidemiology

REVIEW

Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models

Constanza L. Andaur Navarro[a,b,*], Johanna A.A. Damen[a,b], Toshihiko Takada[a], Steven W.J. Nijman[a], Paula Dhiman[c,d], Jie Ma[c], Gary S. Collins[c,d], Ram Bajpai[e], Richard D. Riley[e], Karel G.M. Moons[a,b], Lotty Hooft[a,b]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[b]*Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[c]*Center for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK*
[d]*NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
[e]*Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK*

Accepted 28 March 2023; Published online 5 April 2023

Abstract

Objectives: We evaluated the presence and frequency of spin practices and poor reporting standards in studies that developed and/or validated clinical prediction models using supervised machine learning techniques.

Study Design and Setting: We systematically searched PubMed from 01/2018 to 12/2019 to identify diagnostic and prognostic prediction model studies using supervised machine learning. No restrictions were placed on data source, outcome, or clinical specialty.

Results: We included 152 studies: 38% reported diagnostic models and 62% prognostic models. When reported, discrimination was described without precision estimates in 53/71 abstracts (74.6% [95% CI 63.4−83.3]) and 53/81 main texts (65.4% [95% CI 54.6−74.9]). Of the 21 abstracts that recommended the model to be used in daily practice, 20 (95.2% [95% CI 77.3−99.8]) lacked any external validation of the developed models. Likewise, 74/133 (55.6% [95% CI 47.2−63.8]) studies made recommendations for clinical use in their main text without any external validation. Reporting guidelines were cited in 13/152 (8.6% [95% CI 5.1−14.1]) studies.

Conclusion: Spin practices and poor reporting standards are also present in studies on prediction models using machine learning techniques. A tailored framework for the identification of spin will enhance the sound reporting of prediction model studies.

Keywords: Diagnosis; Prognosis; Development; Validation; Misinterpretation; Overinterpretation; Overextrapolation; Spin

Systematic review registration: PROSPERO, CRD42019161764.

Funding: There is no specific funding to disclosure for this study. GSC is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is funded by the NIHR Oxford BRC. RB is affiliated to the National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) West Midlands. None of the funding sources had a role in the design, conduct, analyses, or reporting of the study or in the decision to submit the manuscript for publication. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Competing interests: Authors declare no competing interests.

Availability of data, code, and other materials: Data and analytical code is available upon reasonable request to corresponding author.

Declaration of interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Registration and protocol: This review was registered in PROSPERO (CRD42019161764). The study protocol can be accessed in https://doi.org/10.1136/bmjopen-2020-038832.

Author Contributions: Constanza L. Andaur Navarro: Conceptualization, Methodology, Investigation, Data Curation, Formal analysis, Writing - original draft, Writing - review & editing; Johanna A.A. Damen: Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision; Toshihiko Takada: Investigation, Writing - review & editing; Steven WJ Nijman: Investigation, Writing - review & editing; Paula Dhiman: Conceptualization, Methodology, Investigation, Writing - review & editing; Jie Ma: Investigation, Writing - review & editing; Gary S Collins: Conceptualization, Methodology, Writing - review & editing; Ram Bajpai: Investigation, Writing - review & editing; Richard D Riley: Conceptualization, Methodology, Writing - review & editing; Karel GM Moons: Conceptualization, Methodology, Writing - review & editing, Supervision; Lotty Hooft: Conceptualization, Methodology, Writing - review & editing, Supervision.

* Corresponding author. Julius Center for Health Sciences and Primary Care, Universiteitsweg 100, P.O. Box 85500, 3508 GA Utrecht, The Netherlands.

E-mail address: c.l.andaurnavarro@umcutrecht.nl (C.L. Andaur Navarro).

https://doi.org/10.1016/j.jclinepi.2023.03.024

1. Introduction

Prediction models in health care generally use individual data to estimate the probability of the presence of an existing disorder (i.e., a diagnostic model) or of the occurrence of a future outcome (i.e., a prognostic models) [1]. Well-known examples are colonflag to identify colorectal cancer and the Framingham risk score to predict the risk of heart disease within the next 10 years [2,3]. To benefit patients and healthcare providers in clinical practice, studies on clinical prediction models should be conducted following the best available methodological evidence and reported in a transparent and complete manner. However, studies on prediction models are often developed using inappropriate methods and are incompletely reported [4−6]. Inaccurate reporting and misinterpretation of study findings might have consequences for research dissemination and public trust in scientific findings.

To facilitate transparent and complete reporting of study methodology and findings, reporting guidelines are available to authors of biomedical research. However, there is still room for authors to frame or emphasize a particular interpretation of study findings [7,8]. The misuse of language, intentionally or unintentionally, affects the interpretation of study findings and has been described as 'spin' [9−14]. 'Spin' has also been referred to as the discordance between study results and methods, conclusion, or overextrapolation [15]. 'Spin' is prevalent in biomedical literature, and evidence shows that it can have an impact on reader's interpretation and decision-making [15,16].

In recent years, supervised machine learning has gained considerable attention as a flexible suite of data analytic methods for predictions in health care [17]. Machine learning is often described as a set of algorithms that enable computers to learn from data without hard-coded rules, thus coping with the requirements of big data [18]. Neural networks, random forest, and support vector machines are some examples [19]. Nonetheless, studies using machine learning techniques are often questioned about their true effectiveness within the clinical workflow [20,21]. The pressure to publish and the intense commercialization agenda may contribute to the exaggeration of the real benefit of machine learning-based prediction models while underplaying the costs, risks, and limitations. Whether a study applied regression or machine learning techniques, the use of spin and poor reporting practices to describe model development and validation could provide a false impression of the real performance of the model, thus hampering its further independent validation and transportation to daily healthcare settings.

'Spin' or overinterpreted scientific findings are a well-established phenomenon in randomized therapeutic intervention trials, observational studies, biomarker studies, diagnostic test accuracy studies, prognostic factor studies, and systematic reviews, however, its form and frequency in prediction model studies are unknown [10,11,22−24]. We conducted a systematic review to estimate the frequency of spin practices and poor reporting standards that might play a role in how the findings of a study are interpreted in studies on prediction models developed using supervised machine learning across clinical domains.

2. Methods

For the reporting of this study, we adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [25].

2.1. Literature search

We aimed to identify primary studies describing the development or validation of prediction models using supervised machine learning techniques across all medical fields published between January 1, 2018 and December 31, 2019. Hence, we searched in PubMed on December 19, 2019 using a comprehensive search strategy that is provided as Supplemental File 1.

2.2. Eligibility criteria

We included studies if they met any of the following criteria: 1) described the development or validation of one or more multivariable prediction models using any supervised machine learning technique aiming for individualized predictions; or 2) reported on the incremental value or model extension aiming to develop a prediction model. A multivariable prediction model was defined as a model aiming to predict a health outcome by using two or more predictor variables. For this study, we considered a study to be an instance of supervised machine learning when it reported any statistical learning technique, except when reporting only models that were strictly regression-based, regardless of whether authors referred to them as machine learning.

We excluded studies if they; 1) investigated a single predictor, test, or biomarker or its causality with an outcome; 2) used machine learning to enhance the reading of images or signals; 3) used as predictors only for genetic traits or molecular markers; and 4) reported on systematic reviews, conference abstracts, or tutorials. The search was restricted to human subjects, English-language articles, and articles available via our institution. Further details about eligibility criteria have been described in the study protocol [26].

2.3. Literature selection

Two independent reviewers screened all titles and abstracts in parallel. One reviewer (CLAN) screened all

What is new?

Key findings

- Spin practices and poor reporting standards are also present in studies on prediction models using machine learning techniques.

What this adds to what was known?

- We systematically reviewed spin practices and poor reporting standards in 152 studies on prediction models, which have not previously been characterized.
- Spin presentation varies per study design, therefore involving specific challenges to its identification and evaluation in studies on prediction models, regardless of modelling approach.

What is the implication, and what should change now?

- Frequent misinterpretation and overinterpretation may lead to an unjustified optimism about the performance of prediction models.
- Establishing a framework for a rigorous identification of spin practices in studies on prediction models will enhance adequate, transparent and sound reporting of prediction model studies.

studies, while the second reviewer came from a group of six (TT, SWJN, PD, JM, RB, and JAAD). Full-text reading of selected articles was performed by one reviewer (CLAN) in combination with one of the other six reviewers (TT, SWJN, PD, JM, RB, and JAAD). In case of disagreement, a third author (JAAD) was involved.

2.4. Data extraction

We defined 'spin practice' as any issue that could make the clinical usefulness of the developed or validated prediction model look more favorable than the study design and results can underpin [10]. We provided examples in Box 1. A previous article about spin in prognostic factor studies already identified several practices, which we modified for our data extraction [22]. We also added spin practices identified in other study designs as well as practices to reduce research waste (e.g., presence of/reference to a study protocol, references to previous evidence), which we grouped under 'poor reporting standards' [11,12,15,31,32]. Reporting practices such as a protocol may aid readers in further check whether the use of certain terms or statements appropriately describes methods, results, and conclusions. For detailed description of extracted items, see Supplemental File 2.

Data extraction was performed in duplicate; one independent reviewer (CLAN) extracted all articles, and the second extraction was carried out by randomly allocating articles to each of the other six reviewers (TT, SWJN, PD, JM, RB, and JAAD). We examined spin practices in abstract and main text separately and across sections (title, introduction, results, and discussion). Discrepancies were discussed between reviewers until agreement was reached.

Our extraction form also included the general characteristics of each study: the aim of the study, type of publication (diagnosis vs. prognosis), year of publication (2018 vs. 2019), journal name, clinical specialty, funding source, disclosure of authors' conflicts of interest (COIs), and mention of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement [7,8]. The extraction form was pilot-tested in five articles and implemented using Research Data Capture (REDCap) [33].

2.5. Synthesis of results

We estimated the frequency of the extracted items with 95% confidence intervals (CIs). We present results separately for abstract, main text, and across sections of the manuscript. Furthermore, we followed the scheme presented in previous articles, wherein spin practices are classified within these three strategies: misleading reporting (i.e., incomplete and selective reporting), misleading interpretation (i.e., unreliable statistical analysis, linguistic spin), and misleading extrapolation (i.e., ignoring uncertainty and claiming irrelevant clinical applicability) [22,31]. We summarized results using descriptive statistics alongside a narrative summary and visual plot. Analyses were carried out using R version 4.0.3 (R Core Team, 2020).

2.5.1. Ethical approval

This study was performed on published studies, thus ethical approval is not required.

3. Results

After our search, we retrieved 24,814 articles. Given time and resource constraints, we randomly sampled 2,482 (10%) studies for screening. After screening, 312 studies were reviewed in full text. A total of 152 studies were found eligible and included in the final analysis. A flowchart of the screening process is provided in Figure 1.

3.1. General characteristics of included studies

Of the 152 articles, 94 (61.8% [95% CI 53.9–69.2]) focused on prognostic models and 58 (38.2% [95% CI 30.8–46.1]) on diagnostic models. Most studies reported the development of prediction models, including internal validation (n = 133/152, 87.5% [95% CI 81.3–91.8]),

Box 1 Examples of spin practices

Examples	Reason	Spin criteria
"The predictive models performed excellent in predicting epithelial ovarian cancer recurrence." [27]	Although reported area unde the curve (AUC) is high, the study has several methodological limitations, so it is likely that the reported AUC is optimistic.	Using overly optimistic or positive words to describe the model or the model's performance. Examples: outperformed, improved, superior, better, novel, unique, etc.
"Classic methods of dealing with missing data such as complete case analysis, …and multiple imputation can potentially bias the estimates of effect of each variable.(ref) …To avoid losing predictive power, the missing data were imputed using the missForest package." [28]	The cited references to support the statement recommend the use of multiple imputations and provide no evidence on missForest imputation.	Using strong affirmative statements to support selected study design and methods
"Although sensitivity was 100% with and without the new biomarker, within the first case, specificity and accuracy were remarkably greater." —This is a fictious example.	The study reports changes on specificity and accuracy but the increase is low	Using strong affirmative statements to describe the model or the model's performance. Examples: clearly shows, strongly recommend, definitely suggest, very important, remarkably greater etc.
"Our finding suggests that random forest model would be best option to implement a system for predicting fatty liver disease patients appropriately and effectively" [29]	Development-only study	Stating a prediction model can be used in routine medical practices without the need for (further) validation and/or clinical impact studies.
"This can be extended to predict other type of ailments which arise from metabolic syndrome" [30]	Development-only study	Stating the use of prediction model in a different outcome, setting or population without stating the need to perform proper evaluation

and 19 (12.5% [95% CI 8.2−18.7]) performed external validation. The clinical specialties with the most publications were oncology (n = 21/152, 14% [95% CI 9.2−20.2]), surgery (n = 20/152, 14% [95% CI 8.7−19.5]), and neurology (n = 20/152, 14% [95% CI 8.7−19.5]). Table 1 shows the characteristics of the included articles. For details on the included articles, see Table S1 (see Supplemental File 3).

Most articles originated in North America (n = 59/152, 38.8% [95% CI 31.4−46.7]) and the first author was often affiliated with a clinical department (n = 85/152, 55.9% [95% CI 48−63.6]). Source of funding was often reported (n = 107/152, 70.4% [95% CI 62.7−77.1]), of which 92/107 (86% [95% CI 78.2−91.3]) were supported by nonprofit organizations. Moreover, 122/152 (80.3% [95% CI 73.2−85.8]) studies were published in journals containing a section for COI, but only 20/122 (16.4% [95% CI 10.9−24]) studies reported at least 1 COI. Reporting guidelines were cited in 13/152 (8.6% [95% CI 5.1−14.1]) studies. Of these 13 studies, 8 (61.5% [95% CI 35.5−82.3]) mentioned TRIPOD [7,8], 3 (23.1% [95% CI 8.2−50.3]) strengthening the reporting of observational studies in epidemiology (STROBE) [34], 2 (15.4% [95% CI 4.3−42.2]) the standards for reporting of diagnostic accuracy studies [35], and 2 (15.4% [95% CI 4.3−42.2]) the *Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research.* [36].

In total, we evaluated 19 spin and poor reporting standards in the abstract (including the title) and 26 in the main text (Table 2). The most frequent poor reporting standards was the absence of a study protocol (n = 150/152, 98.7% [95% CI 95.3−99.6]). Likewise, the most frequent spin practice was the inappropriate comparison to previously developed/validated models. Some examples are provided in Box 1. We found a median of 8 (interquartile range [IQR] 7 to 9) practices in the abstract, as well as in the main text (IQR 7 to 10) (Fig. 2). In Table S2, we provided results stratified by diagnosis vs. prognosis model studies (see Supplemental File 4).

3.2. Misleading reporting

We classified 13 practices as misleading reporting, of which five were assessed in the abstract and eight in the

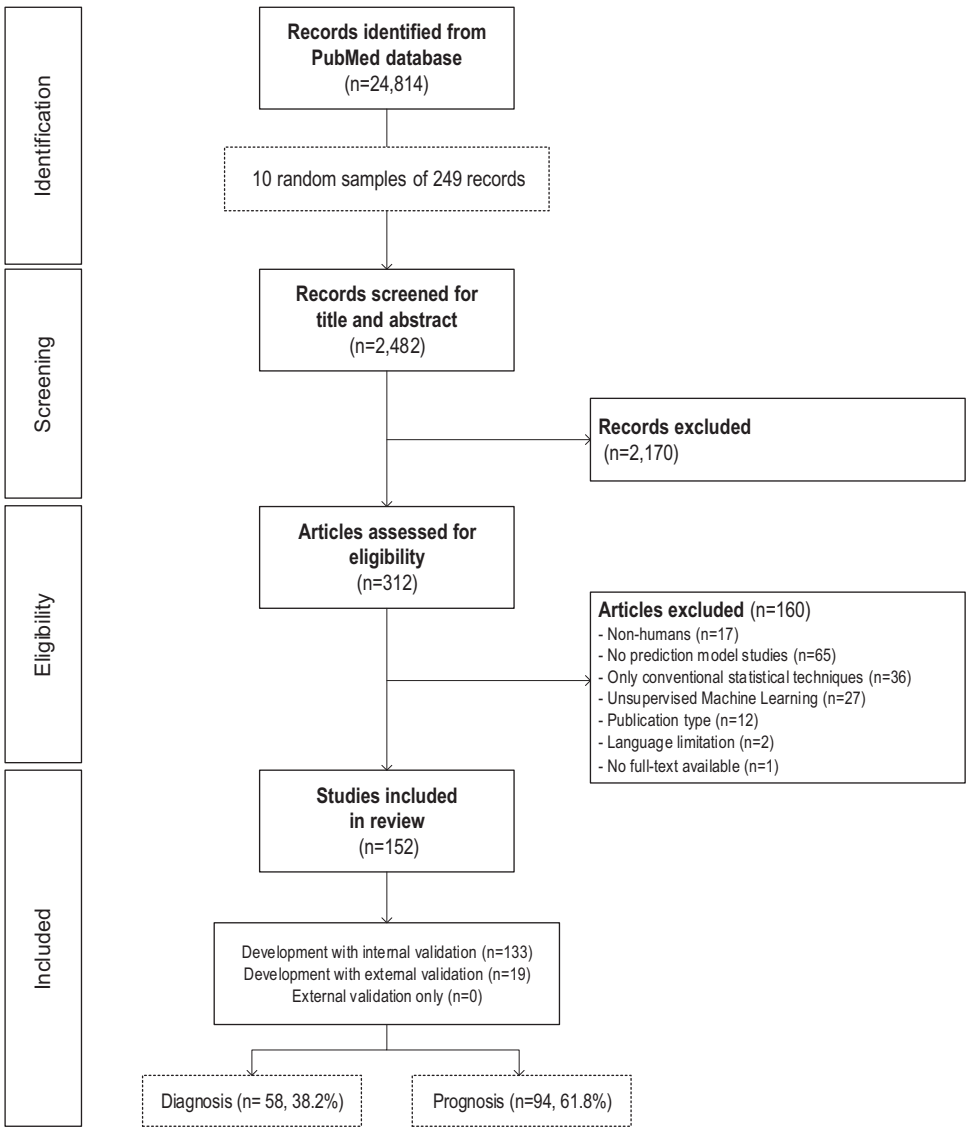

Fig. 1. Flowchart of included articles.

main text (Table 2). Most included studies had four to five misleading reporting practices (n = 118/152, 77.6% [95% CI 70.4−83.5]) in the abstract. In 43/152 (28.3% [95% CI 21.7−35.9]) abstracts, the term 'machine learning' was used rather than the specific term for the algorithm (i.e., support vector machine, k-nearest neighbor). While 81/152 (53.3% [95% CI 21.7−35.9]) abstracts did not report discrimination, 150/152 (98.7% [95% CI 21.7−35.9]) abstracts reported no calibration measure. In 146/152 (96.1% [95% CI 21.7−35.9]) abstracts, no study limitation was mentioned. Likewise, 138/152 (90.8% [95% CI 21.7−35.9]) abstracts did not mention the availability of previous prediction models.

Similarly, most studies had at least four misleading reporting practices (n = 64/152, 42.1% [95% CI 34.5−50.1) in the main text. We identified 7/152 (4.6% [95% CI 2.2−9.2]) studies that reported different performance measures in methods compared to results. In 68/152 (44.7% [95% CI 37.1−52.7]) studies, authors did not provide rationale to support the use of machine learning to address the research question. Similarly, 29/152 (19.1% [95% CI 13.6−26.1]) studies ignored models developed previously. Almost all studies did not provide or reference to a study protocol (n = 150/152, 98.7% [95% CI 95.3−99.6]). Further details can be found in Table 2.

3.3. Misleading interpretation

We classified 21 practices as misleading interpretation, of which eight were identified in the abstract and 13 in the main text (Table 2). Of the 152 studies, 48/152 (31.6% [95% CI 24.7−39.3]) had two misleading

Table 1. General characteristics of included articles (n = 152)

General characteristics	n (%) [95% CI]
Study type	
Diagnosis	58 (38.2) [30.8–46.1]
Prognosis	94 (61.8) [53.9–69.2]
Study aim	
Development only	133 (87.5) [81.3–91.8]
Development with external validation	19 (12.5) [8.2–18.7]
Clinical specialty	
Oncology	21 (14) [9.2–20.2]
Surgery	20 (14) [8.7–19.5]
Neurology	20 (14) [8.7–19.5]
Origin[a]	
Europe	37 (24.3) [18.2–31.7]
North America	59 (38.8) [31.4–46.7]
Asia	46 (30.3) [23.5–38]
Other (Oceania, Latin America)	5 (3.3) [1.4–7.5]
Unclear/Not reported	8 (5.3) [2.7–10]
Affiliation with clinical department[b]	
Yes	85 (55.9) [48–63.6]
No	67 (44.1) [36.4–52]
Conflict of interest	
Yes	20 (13.2) [8.7–19.5]
No	102 (67.1) [59.3–74.1]
Not reported	30 (19.7) [14.2–26.8]
Funding source	
Profit	3 (2) [0.7–5.6]
Nonprofit	92 (60.5) [52.6–67.9]
Both	4 (2.6) [1–6.6]
Unclear	8 (5.3) [2.7–10]
Not reported	45 (29.6) [22.9–37.3]
Reference to reporting guidelines	
Yes	13 (8.6) [5.1–14.1]
No	139 (91.4) [85.9–94.9]

Abbreviations: CI, confidence interval.
[a] Three studies originated in more than one continent.
[b] Reported affiliation of first author.

interpretation practices in the abstract. Out of the 71 abstracts that reported discrimination measures, 53 (74.6% [95% CI 63.4–83.3]) described them without precision estimates. Strong statements to describe model performance were found in 62/152 (40.8% [95% CI 33.3–48.7]) abstracts and 40 (26.3% [95% CI 20–33.8]) used at least one leading word. In 38/152 (25% [95% CI 18.8–32.4]) abstracts, authors emphasize model relevance, while results were not predictive.

Most studies had three to four misleading interpretation practices across sections in the main text (50/152, 32.9%). When reported, discrimination was presented without precision estimates in 53/101 (52.5% [95% CI 42.8–61.9]) studies. Likewise, calibration lacked precision estimates in 7/18 (38.9%) studies. In 59/152 (38.9% [95% CI 20.3–61.4]) studies, we identified strong statements to describe model performance. Further details can be found in Table 2.

3.4. Misleading extrapolation

We classified 11 practices as misleading extrapolation, of which six were assessed in Abstract and five in main text (Table 2). Across abstracts, recommendation to use the model in clinical practice was provided in 21 studies; however, 20 (95.2% [95% CI 77.3–99.8]) of them lacked any form of external validation despite their small sample size. Likewise, recommendations to use the model in different settings of population were given in nine studies, all of which lacked external validation in the same study.

In the main text, 86/152 (56.6% [95% CI 48.6–64.2]) studies made recommendations to use the model in clinical practice, however, 74/86 (86% [95% CI 77.2–91.8]) lacked external validation in the same article. Out of the 13/152 (8.6% [95% CI 5.1–14.1]) studies that recommended the use of the model in a different setting or population, 11/13 (84.6% [95% CI 57.8–95.7]) studies lacked external validation. Finally, qualifiers (such as ''very''and ''may'') were used frequently to describe findings in the main text (n = 64/152, 42.1% [95% CI 34.5–50.1]). Further details can be found in Table 2.

3.5. Extent of spin practices across sections

Most articles contained no spin practice in title (n = 132/152, 86.8% [95% CI 80.5–91.3]), three spin practices in results (n = 61/152, 40.1% [95% CI 32.7–48.1]), and three in discussion (n = 61/152, 40.1% [95% CI 32.7–48.1]). Regarding the main text, articles contained two spin practices in results (n = 48/152, 31.6% [95% CI 24.7–39.3]), four in the discussion (n = 36/152, 23.7% [95% CI 17.6–31]), and one in another section (n = 69/152, 45.4% [95% CI 37.7–53.3]). We showed the extent of occurrence of spin per sections in Figure 3.

4. Discussion

We systematically assessed how often spin practices and poor reporting standards occurred in 152 prediction model studies using supervised machine learning. Our study revealed that both were widely present in studies on prediction models developed using supervised machine learning.

4.1. Principal findings

The most frequent poor reporting standard was the absence of a predefined protocol or registration. Moreover, the use of reporting guidelines was scarce. Although infrequent, we also observed a few discrepancies between methods and results in the main text, as well as

Table 2. Frequency of 'spin' practices and poor reporting standards in title, abstract, and main text

Spin practices and poor reporting standards		No. (%) [95% CI of percentage]	
		Abstract (n = 152)	Main text (n = 152)
Misleading reporting			
Results section			
Machine learning techniques used are unreported	Poor reporting standard	43 (28.3) [21.7–35.9]	NE
Differences between performance measures prespecified in methods and reported in results section	Spin	NA	7 (4.6) [2.2–9.2]
Discrimination is not reported	Poor reporting standard	81 (53.3) [45.4–61]	51 (33.6) [26.5–41.4]
Calibration is not reported	Poor reporting standard	150 (98.7) [95.3–99.6]	134 (88.2) [82.1–92.4]
Discussion and conclusion section			
Limitations are not reported	Poor reporting standards	146 (96.1) [91.7–98.2]	28 (18.4) [13.1–25.3]
Other sections			
Rationale to use machine learning techniques to address the objective in introduction is unavailable	Poor reporting standard	NA	68 (44.7) [37.1–52.7]
No references to existing models	Poor reporting standard	138 (90.8) [85.1–94.4]	29 (19.1) [13.6–26.1]
Main results are reported as supplemental file	Poor reporting standard	NA	14 (9.2) [5.6–14.9]
The study protocol is unavailable	Poor reporting standard	NA	150 (98.7) [95.3–99.6]
Misleading interpretation			
Title			
Title is inconsistent with the study results	Spin	6 (3.9) [1.8–8.3]	NA
Use of leading words	Spin	18 (11.8) [7.6–17.9]	NA
Novel		3 (2) [0.7–5.6]	
Excellent		0	
Accurate		1 (0.7) [0–3.6]	
Optimal		0	
Perfect		0	
Significant		0	
Improved		7 (4.6) [2.2–9.2]	
Other[b]		7 (4.6) [2.2–9.2]	
Results section			
Discrimination is reported without precision estimates	Poor reporting standards	53 (74.6) [63.4–83.3][a]	53 (52.5) [42.8–61.9][a]
Calibration is reported without precision estimates	Poor reporting standards	2 (100) [34.2–100][a]	7 (38.9) [20.3–61.4][a]
Use of strong statements to describe the model and/or model performance/accuracy/ effectiveness	Spin	62 (40.8) [33.3–48.7]	59 (38.8) [31.4–46.7]
Use of leading words	Spin	40 (26.3) [20–33.8]	59 (38.8) [31.4–46.7]
Novel		4 (2.6) [1–6.6]	0
Excellent		1 (0.7) [0–3.6]	5 (3.3) [1.4–7.5]
Accurate		13 (8.6) [5.1–14.1]	7 (4.6) [2.2–9.2]
Optimal		3 (2) [0.7–5.6]	0
Perfect		0	1 (0.7) [0–3.6]
Significant		10 (6.6) [3.6–11.7]	30 (19.7) [14.2–26.8]
Promising		6 (3.9) [1.8–8.3]	0
Improved		4 (2.6) [1–6.6]	4 (2.6) [1–6.6]
Outperform		4 (2.6) [1–6.6]	12 (7.9) [4.6–13.3]
Other[c]		26 (17.1) [11.9–23.9]	14 (9.2) [5.6–14.9]
Spin in tables or figures		NA	10 (6.6) [3.6–11.7]

(*Continued*)

Table 2. Continued

Spin practices and poor reporting standards		No. (%) [95% CI of percentage]	
		Abstract (n = 152)	Main text (n = 152)
Discussion and conclusion section			
Use of strong statements to describe model and/or model performance/ accuracy/effectiveness	Spin	NA	64 (42.1) [34.5–50.1]
Use of leading words	Spin	NA	64 (42.1) [34.5–50.1]
Novel			3 (2) [0.7–5.6]
Excellent			8 (5.3) [2.7–10]
Accurate			10 (6.6) [3.6–11.7]
Optimal			2 (1.3) [0.4–4.7]
Perfect			1 (0.7) [0–3.6]
Significant			15 (9.9) [6.1–15.6]
Superior			7 (4.6) [2.2–9.2]
Outperform			4 (2.6) [1–6.6]
Other[d]			26 (17.1)[11.9–23.9]
Invalid comparison of results to previous development and/or validation studies is given	Spin	NA	99 (65.1) [57.3–72.2]
The comparison in favour of similar prediction models			52 (52.5) [42.8–62.1][a]
Some outcomes in favour and not in favour for other			22 (22.2) [15.2–31.4][a]
Unclear			11 (11.1) [6.3–18.8][a]
Nonrelevant models are not discussed	Spin	NA	28 (23) [16.4–31.2][a]
Authors make use of leading words to reject those nonrelevant models	Spin		10 (38.5) [22.4–57.5][a]
Emphasis on model relevance while results are not predictive	Spin	38 (25) [18.8–32.4]	NE
Discrepancy between full text and abstract explanation of the study findings	Spin	7 (4.6) [2.2–9.2]	NA
Misleading extrapolation			
Discussion and conclusion section			
Recommendation to use the model in clinical practice without external validation in same study	Spin	20 (95.2) [77.3–99.8][a]	74 (86) [77.2–91.8][a]
Recommendation to use the model in different setting or population without external validation in same study	Spin	9 (100) [70.1–100][a]	11 (84.6) [57.8–95.7][a]
No recommendation for further studies	Poor reporting standard	15 (9.9) [6.1–15.6]	38 (25) [18.8–32.4]
Qualifiers are used	Spin	50 (32.9) [25.9–40.7]	64 (42.1) [34.5–50.1]
Other benefits not prespecified in Methods are addressed	Spin	NA	10 (6.6) [3.6–11.7]
Conclusions are inconsistent with the reported study results	Spin	23 (16.4) [11.2–23.4][a]	NE
Conclusion focuses solely on significant results	Spin	85 (55.9) [48–63.6]	NE

Abbreviations: CI, confidence interval; NA, not applicable; NE, not extracted.
[a] Valid percentage: with respect to the articles which reported the information.
[b] Predictive, well-calibrated, promote, intelligent, outperform, improved.
[c] Best, efficient, superior, satisfactory, greater, substantial, well, effective.
[d] Remarkable, substantial, better, robust, satisfied, superior, huge.

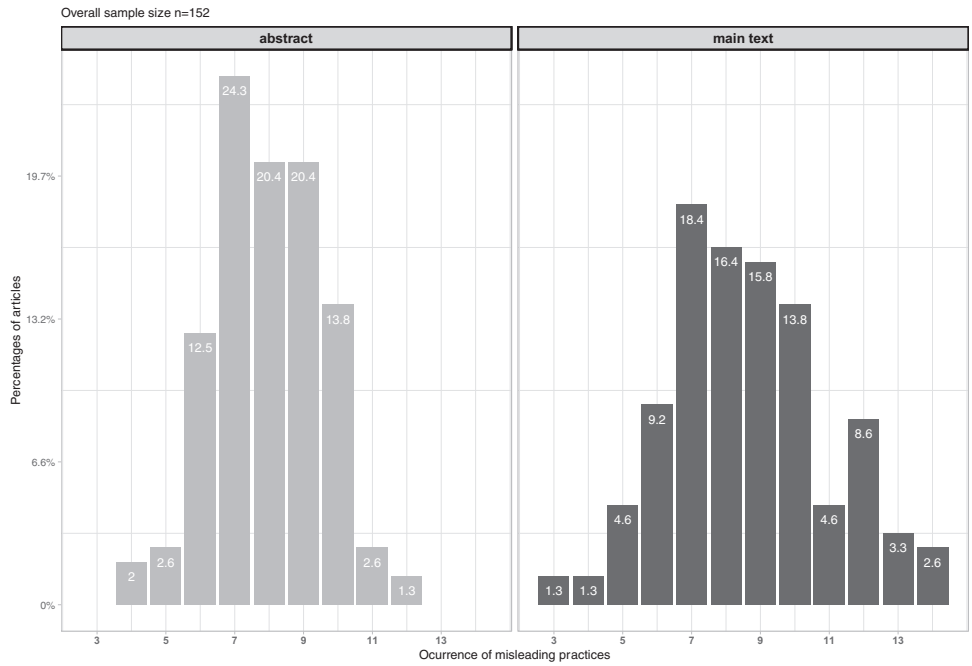

Fig. 2. Distribution of misleading practices.

discrepancies between the abstract and main text conclusions. However, a protocol and the use of reporting guidelines could reduce selective and incomplete reporting. TRIPOD, the reporting guideline for studies on prediction models, also includes a version for proper reporting of Abstracts [7,8,37].

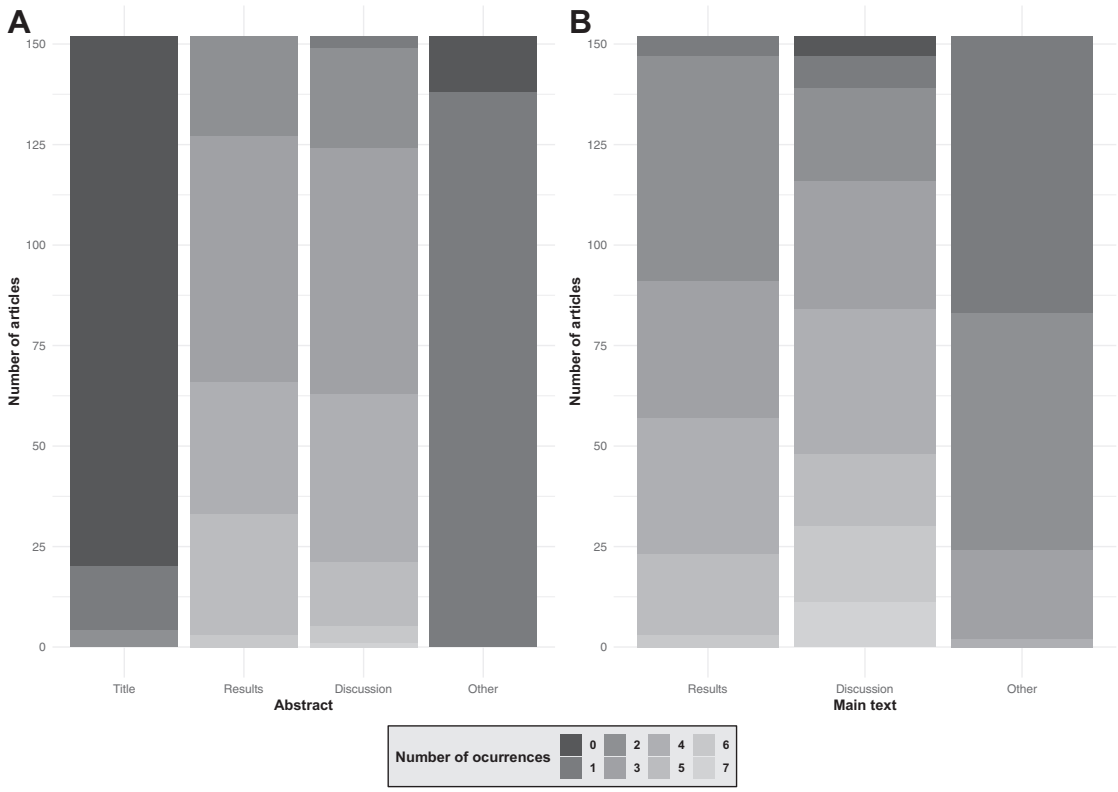

Fig. 3. Extent of occurrences per section. (A) Occurrence of spin in Abstract's sections. (B) Occurrence of spin in Main text's sections.

We found that the occurrence of spin practices was similar between the abstract and main text. We found that studies often made inappropriate recommendations to use the prediction model in daily clinical practice, ignored limitations, and reported performance measures without precision estimates in the abstract. Previous research on nonrandomized studies has identified that abstracts are the most frequent section with spin [31]. As a primary source of dissemination, the content of an abstract must be accurate and useful, not only for evidence users but also for the general audience. Furthermore, research shows that the main factor associated with spin in a press release was the presence of spin in the abstract [14]. Spin in abstracts could partially be explained by the limited word count and the need to attract potential readers; however, any recommendation in concluding statements in an abstract should be consistent with the study design, findings, and limitations to avoid misleading the readers, especially those who can only access this information.

We further noticed that a considerable number of studies neither report their limitations nor their findings within the context of previously developed models. Given the high number of developed models already available in the biomedical literature, researchers should focus on carrying out systematic reviews and validating the most promising models to avoid further research waste [4,32].

4.2. Strengths and limitations

To our knowledge, there has been no systematic review about spin practices and poor reporting standards in prediction model studies and, particularly not in studies on machine learning based prediction models. We appraised a sample of articles covering a wide range of outcomes and clinical domains. In addition, we evaluated spin in the title, abstract, and across several sections of the abstract and main text.

However, several limitations are worth highlighting. In our study, we modified the pre-existing tool used in prognostic factor studies as such, but faced certain challenges during data extraction. Although this tool enabled us to capture several practices, it failed to identify aspects particularly related to prediction model studies (i.e., selection of predictors, categorization of continuous predictors, threshold definition). Furthermore, we focused on the use of leading words (i.e., linguistic spin) rather than allowing certain degree of rhetoric and evaluating it within its specific context. Similarly, we could not determine if the use of qualifiers was detrimental because we only counted the occurrence rather than to evaluate its use to show uncertainty. The appraisal of spin practices relied mostly on the subjective judgement of reviewers; thus, it is possible that others will interpret the authors' statements differently as we did, especially the linguistic spin. Although we reduced interpretation bias by resolving any discrepancies through discussion, reviewers were not blinded to authors, funding source, or journal. Likewise, this appraisal depended on what was reported in articles thus, some of our findings might be the consequence of poor reporting quality rather than misleading practices.

As we did not cover the full range of potential spin practices, our findings should be interpreted bearing this in mind. Furthermore, our review does not provide a comparison group and as such, we avoided drawing inferences, associations, or causality between studies characteristic and spin practices and poor reporting standards. Our aim was to bring attention to practices indicative of spin in prediction model studies based on a descriptive analysis. Despite these limitations, we still provided exploratory evidence about the presence of spin and poor reporting standards in prediction model studies.

4.3. Comparison with other studies

A systematic review including 35 publications assessing misleading practices showed that spin evaluation varies per study design [15]. Unfortunately, no study assessing spin practices in prediction model studies was found in this review. Within prognostic research, studies on prognostic factors in oncology frequently overinterpret their findings, hampering clinical applicability [22].

4.4. Unanswered questions, recommendations, and future research

There are several obstacles to ensuring accurate interpretation and dissemination of research. The reward system within academia and the increasing amount of published research make spin in research, to some extent, necessary and therefore more frequent. As authors, we naturally want our studies to be published and will consciously and subconsciously use language to increase credibility and readability of our findings. For example, authors reporting exploratory analysis, such as studies describing model development only and not any form of evaluation, might allow themselves to overinterpret and extrapolate their results, as it might not be expected to become available in daily clinical practice. But a growing concern is that spin in primary studies is linked to inappropriate reporting of press releases and news media [14,38]. The reach of spin in biomedical research therefore also extends to general audiences, potentially biasing behaviour and jeopardizing public trust. Authors should make every effort to avoid distortion and hype and should focus on overall quality, transparency, and further research.

Spin is prevalent in all biomedical literature, and therefore, further evaluation of spin practices and poor reporting standards will benefit those who rely upon biomedical research findings and evidence. However, to some extent, it requires subjective judgement. There is a need to develop an instrument or classification scheme with clear definitions tailored to identify and evaluate spin practices in studies on

prediction models based on the consensus between experts. Likewise, further guidance and interventions on how to write study findings and reduce spin could also be helpful [39]. This can guide junior researchers, peer-reviewers, and journal editors to be cautious on how study findings are written, while achieving transparency, accuracy, and conciseness. Similarly, readers should be aware of practices that can mislead their interpretation of findings before deploying models into daily health care settings. We theorized that spin practices in prediction model studies might have a larger impact on clinical guidelines and research funds, especially given the rise of machine learning and artificial intelligence in health care applications [16,40−42]. The effects on reader's interpretation, role of peer-reviewers, number of citations, and assignments of research funds still needs to be assessed within studies on prediction models [11,13,14,23].

Spin practices and its association with methodological quality and risks of bias still needs to be systematically assessed to provide evidence of its effect on overall quality of biomedical evidence. A severity scale for spin in prediction models still need to be developed.

5. Conclusion

Authors have several opportunities to frame the impression their findings will produce in readers. We provide a description of the existence of spin practices and poor reporting standards in studies on prediction models and indicate the need for strategies to improve how study results are portrayed to increase prediction model validations and uptake in daily clinical practice.

Acknowledgments

The authors would like to acknowledge René Spijker for his assistance in developing the search strategy. We thank the peer-reviewers for critically reading the manuscript and suggesting substantial improvements.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2023.03.024.

References

[1] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:1317−20.

[2] Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med 2017;6(10):2453−60.

[3] Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837−47.

[4] Damen JAAG, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416.

[5] Collins GS, De Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40.

[6] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.

[7] Moons KGM, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1−73.

[8] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55.

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

[10] 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.

[11] 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.

[12] 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.

[13] Boutron I, Haneef R, Yavchitz A, Baron G, Novack J, Oransky I, et al. Three randomized controlled trials evaluating the impact of "spin" in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit. BMC Med 2019;17(1):1−10.

[14] Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med 2012;9(9):e1001308.

[15] Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: a methodological systematic review. PLoS Biol 2017;15(9):1−16.

[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;32:4120−6.

[17] Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019;188:2222−39.

[18] Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med 2016;46:2455−65.

[19] Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019;19(1):281.

[20] Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:1−12.

[21] Morley J, Floridi L, Goldacre B. The poor performance of apps assessing skin cancer risk. BMJ 2020;368:m428.

[22] Kempf E, de Beyer JA, Cook J, Holmes J, Mohammed S, Nguyên TL, et al. Overinterpretation and misreporting of prognostic factor studies in oncology: a systematic review. Br J Cancer 2018;119:1288−96.

[23] Haneef R, Lazarus C, Ravaud P, Yavchitz A, Boutron I. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. PLoS One 2015;10:1−15.
[24] McGrath TA, Bowdridge JC, Prager R, Frank RA, Treanor L, Dehmoobad Sharifabadi A, et al. Overinterpretation of research findings: evaluation of "spin" in systematic reviews of diagnostic accuracy studies in high-impact factor journals. Clin Chem 2020;66:915−24.
[25] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
[26] Andaur Navarro CL, Damen JAAG, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. BMJ Open 2020;10(11):1−6.
[27] Zhang F, Zhang Y, Ke C, Li A, Wang W, Yang K, et al. Predicting ovarian cancer recurrence by plasma metabolic profiles before and after surgery. Metabolomics 2018;14(5):1−9.
[28] Chen D, Goyal G, Go RS, Parikh SA, Ngufor CG. Improved interpretability of machine learning model using unsupervised clustering: predicting time to first treatment in chronic lymphocytic leukemia. JCO Clin Cancer Inform 2019;3:1−11.
[29] Wu CC, Yeh WC, Hsu WD, Islam MM, Nguyen PAA, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed 2019;170:23−9.
[30] Perveen S, Shahbaz M, Keshavjee K, Guergachi A. A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression. Sci Rep 2018;8(1):1−12.
[31] 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;15:1−8.
[32] Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet 2014;383:267−76.
[33] Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208.
[34] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2008;335:806−8.
[35] 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):1−17.
[36] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323.
[37] Heus P, Reitsma JB, Collins GS, Damen JAAG, Scholten RJPM, Altman DG, et al. Transparent reporting of multivariable prediction models in journal and conference abstracts: TRIPOD for abstracts. Ann Intern Med 2020;173:43.
[38] Adams RC, Challenger A, Bratton L, Boivin J, Bott L, Powell G, et al. Claims of causality in health news: a randomised trial. BMC Med 2019;17(1):1−11.
[39] Ghannad M, Yang B, Leeflang M, Aldcroft A, Bossuyt PM, Schroter S, et al. A randomized trial of an editorial intervention to reduce spin in the abstract's conclusion of manuscripts showed no significant effect. J Clin Epidemiol 2021;130:69−77.
[40] el Hechi M, Ward TM, An GC, Maurer LR, El Moheb M, Tsoulfas G, et al. Artificial intelligence, machine learning, and surgical science: reality versus hype. J Surg Res 2021;264:A1−9.
[41] Manlhiot C. Machine learning for predictive analytics in medicine: real opportunity or overblown hype? Eur Heart J Cardiovasc Imaging 2018;19(7):727−8.
[42] Modine T, Overtchouk P. Machine learning is No magic: a plea for critical appraisal during periods of hype. JACC Cardiovasc Interv 2019;12(14):1339−41.