Journal of Clinical Epidemiology 170 (2024) 111364

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

SPIN-PM: a consensus framework to evaluate the presence of spin in studies on prediction models

Constanza L. Andaur Navarro[a,b,*], Johanna A.A. Damen[a,b], Mona Ghannad[a,b], Paula Dhiman[c,d], Maarten van Smeden[a], Johannes B. Reitsma[a], Gary S. Collins[c,d], 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]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[d]*NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
[e]*Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK*

Accepted 8 April 2024; Published online 15 April 2024

Abstract

Objectives: To develop a framework to identify and evaluate spin practices and its facilitators in studies on clinical prediction model regardless of the modeling technique.

Study Design and Setting: We followed a three-phase consensus process: (1) premeeting literature review to generate items to be included; (2) a series of structured meetings to provide comments discussed and exchanged viewpoints on items to be included with a panel of experienced researchers; and (3) postmeeting review on final list of items and examples to be included. Through this iterative consensus process, a framework was derived after all panel's researchers agreed.

Results: This consensus process involved a panel of eight researchers and resulted in SPIN-Prediction Models which consists of two categories of spin (misleading interpretation and misleading transportability), and within these categories, two forms of spin (spin practices and facilitators of spin). We provide criteria and examples.

Conclusion: We proposed this guidance aiming to facilitate not only the accurate reporting but also an accurate interpretation and extrapolation of clinical prediction models which will likely improve the reporting quality of subsequent research, as well as reduce research waste.

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

Plain language summary

Spin refers to presenting research findings in a way that might mislead readers or make findings seem more significant that they really are. This can be done by showing partial results or by using vague language. Spin can distort the true picture of a study's findings and may influence medical decisions and public understanding. We present SPIN-Prediction Models, a tool for readers to critically evaluate the presence of spin, and for authors to understand their implications and reduce spin practices in studies on prediction models, whether using artificial intelligence or not, making sure findings are described accurately.

Funding: No specific funding was given to 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. The views expressed are those of the authors and not necessarily those of the NHS nor NIHR. 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.

* 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.2024.111364

What is new?

Key findings

- Overoptimistic reporting and misinterpretation of studies on clinical prediction models has received relatively limited attention in biomedical literature. These practices are known as 'spin' and are a wide phenomenon in other study designs.

What this add to what is known?

- SPIN-Prediction Models (SPIN-PM) is a consensus-based framework designed to identify seven common spin practices and 14 facilitators of spin to be avoided when communicating findings in studies on prediction models. We provide an overview of spin in prediction modeling research with further explanation, guidance, and accompanying examples.

What is the implication and what should change now?

- SPIN-PM aims to provide guidance to researchers on better reporting practices and to assist readers, including peer reviewers, journal editors, guideline developers, and policy makers in identifying and evaluating spin practices in studies on prediction models. SPIN-PM was not designed as a scoring tool, or to assess the overall quality of prediction model studies.
- We hope this guidance will further enhance the overall reporting quality and support the critical appraisal of studies on prediction models.

1. Background

Prediction models are often used to complement clinical reasoning and (shared) decision-making typically by estimating the probability of an individual to have a particular outcome (diagnostic model) or to develop an outcome in the future (prognostic model) [1−4]. In recent years, the number of studies on prediction models has increased strongly. This growth is expected to continue given the growing popularity of artificial intelligence (AI) and machine learning (ML), and the increasing availability of larger datasets (eg, routine care). Consequently, there are often multiple prediction models available for the same health outcome or target population, but few are integrated into routine clinical practice [5−7].

Many studies on prediction models, whether AI or non-AI, suffer from shortcomings in the design and statistical analyses, making their findings vulnerable to overinterpretation and exaggerated claims on transportability and clinical usefulness [8−10]. Moreover, particular expectation has grown around AI-based prediction models [11−14]. With a common rhetoric of virtually endless potential and excellent performance in the clinical domain, inadvertent pitfalls in study design and analyses may facilitate an inaccurate interpretation and communication of model performance which consequently might lead to the implementation of suboptimal prediction models in clinical practice [15].

To ultimately improve health outcomes, identification and evaluation of misleading practices when communicating findings is necessary, as the realization of the benefits and harms of prediction models could remain limited, while the investment of resources increases.

1.1. What is spin?

Spin is defined as any (reporting) practice, consciously or unconsciously, that leads to misinterpretation or overinterpretation of study findings, usually emphasizing more favourable findings than the study design, analyses, and actual findings warrant [16,17]. While misinterpretation refers to an inconsistent interpretation of the study findings (ie, incorrect interpretation), overinterpretation refers to when authors take a strong position stemming from their opinion rather than on the study findings. Both practices are closely linked and contribute to the misrepresentation (ie, distorted presentation) of scientific findings. Examples are using exaggerated language, highlighting (only) the findings based on selected subgroups or choosing a particular statistical analysis to shape the impression of their findings to readers [18].

In biomedical research, the concept of spin was introduced by the BMJ in 1995 [19]. Since then, several articles have addressed the implications of spin across different study designs and settings, showing its high prevalence and detrimental effect [18,20−24]. Spin can have serious and undesirable consequences in medical practice (including potential harm to patients), development of clinical guidelines, health policies, funding of subsequent research, and engagement with general audiences.

1.2. Why do we need a spin framework for studies on prediction models?

Spin has been extensively studied for trials, and to a small extent in studies on diagnostic test accuracy and prognostic factors [18,20−22,25−30]. Two recent systematic reviews of spin on prediction model studies, including models using AI/ML, highlight the importance of establishing clearer definitions and guidance to better understand the extent, sources, prevalence, and implications of spin in this research domain [15,31].

Furthermore, evidence suggests that the nature of spin differs depending on the study design [23]. Unlike randomized trials, wherein the most common type on spin can be

found on the estimate of the intervention effect, in prediction model studies it is the interpretation of a model's estimated predictive performance (eg, discrimination and calibration) where the action and consequences of spin must be placed. For example, the predictive performance might be overinterpreted in studies on model development, while in studies on a model's external validation, the estimated performance can be overinterpreted or underinterpreted (eg, to justify the development of a new prediction model). Spin therefore also needs to be considered in the context of the study type (Box 1). Additionally, studies on prediction models are also not typically designed to answer questions on etiology, association, or causality, and are rarely (pre)registered or have publicly available protocols.

Given the differences between study designs and statistical analyses, frameworks on spin previously developed are not directly suitable to identify spin in studies on prediction models. In this article, we present SPIN-Prediction Models (SPIN-PM): a consensus-based framework to provide guidance to researchers on better reporting and communicating prediction model studies and to assist readers, including peer reviewers, journal editors, guideline developers, and policy makers in identifying and evaluating spin practices.

2. Methods

For the reporting of this study, we followed the ACcurate COnsensus Reporting Document reporting guideline for consensus methods in biomedicine [35]. This study was not registered.

2.1. Expert participants

A panel of eight researchers with demonstrable expertise in spin (M.G., J.B.R., G.S.C., K.G.M.M., and L.H.) and prediction model studies (C.L.A.N., J.A.A.D., P.D., G.S.C., J.B.R., Mv.S., K.G.M.M., and L.H.) were invited via e-mail in February 2021 to provide comments and contextualize spin practices during several consensus meetings.

2.2. Procedure

The lead researcher (C.L.A.N.) reviewed key publications on established practices of spin in other study designs and built a preliminary list to be included in SPIN-PM [20,22,23,36]. We defined as 'key publication' an article describing methodological frameworks, critical practices, or relevant context that could be translated to studies on prediction models, serving as background literature for the development of SPIN-PM. The preliminary practices were first discussed with one researcher with expertise in spin in other study design (M.G.).

Box 1 Studies on clinical prediction models

Model development studies

A multivariable model is developed to estimate an outcome probability [32]. This type of study aims to produce a model equation or algorithm (eg, including the identification of the most important predictors and assigning relative weights to each of them) and estimating the model's predictive performance through calibration, discrimination, and potentially clinical utility. Discrimination refers to the measure of how well a prediction model can distinguish individuals with the outcome from those without the outcome (eg, using Area Under the Receiver Operator Curve, c-statistic) [33]. Calibration refers to the measure of agreement between predicted and observed probabilities (eg, calibration plot, calibration curves, observed:expected ratio) [33].

When evaluated in the same data in which it was developed, the performance of a model (called apparent performance) will often be optimistic due to overfitting, notably when development datasets are relatively small (typically with a small number of outcome events). Development studies will therefore include an internal validation to quantify and correct for any 'optimism' in model performance [33]. Examples of internal validation techniques are cross-validation and bootstrapping.

External validation studies

External validation studies consist of assessing the performance of a prediction model in new individuals whose data were not used during model development [34]. There are several types of external validation, most commonly (1) temporal validation: individuals from the same institution as in the development sample, but in a different (usually later) time; (2) geographical validation: individuals from different institutions or countries to the development sample; and (3) domain or setting validation: for example, individuals from secondary care are used to validate a model developed in primary care. Depending on the findings of the external validation, the prediction model may be recalibrated or updated to better fit the population and setting of interest.

We held five online meetings with the panel of researchers, while the sixth was open to all researchers working in one of the affiliated institutions. In addition, the concept of the framework was presented and fine-tuned at the annual epidemiological conference in the Netherlands carried out in 2021. At each meeting and presentation, we

openly discussed clarifications, wording, applicability, and useful examples for each of the items in the preliminary list, with the moderation of the lead researcher (C.L.A.N.). After each meeting, the list was adapted accordingly (C.L.A.N.) and discussed again in the next meeting until all panel's researchers agreed.

In parallel, two systematic reviews on spin lead by members of the panel evaluating discrepancies between sections, use of leading words, irrelevant clinical applicability, and unjustified comparisons between models fed-back the feasibility to assess some of the spin practices based on the agreement achieved during data extraction [15,31]. Through this iterative process, SPIN-PM was derived.

3. Results

SPIN-PM proposes two categories of spin through which authors, can consciously or unconsciously, generate spin in studies on prediction models:

a. Misleading interpretation: Claims with overestimation or underestimation of the performance of the developed or validated prediction model. This can be the consequence of either the application of inappropriate methods or statistical analyses, or the use of overly optimistic language to describe the methods or study findings [37,38]. Accordingly, the readers' perception about the quality and quantity of evidence is unsupported by study design, methods, and analyses reported.
b. Misleading transportability: Unjustified claims regarding the applicability, generalizability, or usability of the reported prediction models in routine healthcare practice, or even to other populations, settings, or domains. Without a meaningful external validation (Box 1), inferences on the actual performance or transportability of a prediction model may be overestimated and may cause prediction models to be implemented in settings and populations where there is no robust evidence yet to support this.

Additionally, we proposed two forms of spin: *spin practices* and *facilitators of spin* define as follows:

a. Spin practice: a mismatch between reported information, namely between the actual design and findings, and how these have been interpreted or described by the authors within the manuscript, and that might misdirect the interpretation of readers. A mismatch can occur, for example, when 'positive' words (ie, useful, effective) are overused or misused to describe study findings, while the study design, analysis, and findings do not support such optimistic interpretation. Evaluation of spin, therefore, requires two steps: (1) identify potential spin practices and (2) confirm such mismatch. In case such mismatch cannot directly be verified, the practice may still constitute a *facilitator of spin*.
b. Facilitator of spin: any reporting practices that interferes with the critical appraisal and requires readers to make assumptions. For example, authors reporting performance measures without stating whether these correspond to apparent or internally validated/optimism-corrected measures in studies on model development (Box 1).

Previous classification systems have incorporated selective and incomplete reporting as spin practices [22,36,37]. To objectively identify selective reporting, it is necessary to be able to compare the reported information against a protocol or registration. However, studies on prediction models usually lack publicly available protocols [39]. In this situation, there is no guarantee that because, for example, datasets, subgroups, or thresholds were mentioned in the methods section, these were indeed prespecified in any form of registration or protocol. Similarly, incomplete reporting refers to when authors leave out essential information which hinders the critical appraisal of a study, its findings, interpretation, and conclusions.

The challenge when reporting studies on prediction models in manuscripts with limited word count is that analyses are usually conducted as a funnel—that is, several models may be, for example, developed with different modeling strategies and even different designs or different predicted outcomes, until one sole model is achieved and reported (usually based on the 'best' performing model in the dataset at hand). One cannot assess whether, for example, missed information on other developed or validated models is crucial for the critical appraisal of the reported 'best' performing model. Moreover, current adherence of studies on prediction models to reporting guidelines is deficient in two ways. Authors may be unaware of reporting guidelines (eg, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis [TRIPOD]), while those who are, may still report insufficient information [3,33,40,41]. Hence, for studies on prediction models, we categorized practices related to selective and incomplete reporting as facilitators of spin.

In Tables 1−3, we present each proposed form of spin with their specific criteria for assessment and example. Additionally, we provide extensive explanation for each of these forms of spin in supplemental material 1. In total, the framework consists of two categories of spin, seven spin practices, and 14 facilitators of spin.

4. Discussion

Readers of scientific literature go through an interpretative process which is often influenced by how authors have

Table 1. Proposed framework: spin practices in studies on prediction models

Category of spin	Form of spin	Criteria	Examples
Misleading interpretation			
Unreliable statistical analysis	Ignoring the risk of optimism in model performance	Conclusion or concluding paragraph in *main text* omits statements addressing methodological concern during model development and/or validation that might lead to an overfitted model performance. This can occur if: a. Limited study size b. Limited number of events relative to number of candidate predictors c. Inappropriate dichotomization of continuous predictors d. Traditional stepwise predictor selection strategies	*''Our results suggest that machine learning can predict myopia onset in children. We used the available dataset in our hospital, and we obtained high accuracies''*[a] — *The study has used a very restrictive sample of participants enrolled (and with a few events relative to the number of candidate predictors) in only one location. Limited sample size is not addressed as an important limitation that could have led estimates to be too optimistic.*
Linguistic spin	Unjustified use of strong affirmative statements to support selected study design and methods	*Main text* or *abstract* contains statements about design and selected methods that could be considered *unjustifiable.* An example is 'Given the lack of imputation methods, we carried out complete case-analysis'	*''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''* [42] — The cited references to support the statement recommend the use of multiple imputations and provide no evidence on missForest imputation.
	Unjustified use of strong affirmative statements to describe the model or the model's performance	Conclusion or concluding paragraph in *main text* or *abstract* contains statements using a tone inferring a strong result. Examples: 'clearly shows,' 'strongly recommend,' 'definitely suggest,' 'very important,' 'remarkably greater,' etc.	*''Although sensitivity was 100% with and without the new biomarker, within the first case, specificity and accuracy were remarkably greater''*[a] — The study reports changes on specificity and accuracy but the increase is low.
	Unjustified use of optimistic or positive words to describe the model or model's performance	Statements in *title* and conclusion in *main text* and/or in *abstract* contain terms that positively frame model's predictive performance. Examples: outperformed, improved, superior, better, novel, unique, etc.	*''The predictive models performed excellent in predicting EOC recurrence''* [43] — Although reported AUC is high, the study has several methodological limitations, so it is likely that the reported AUC is optimistic.
Misleading transportability	Stating a prediction model can be used in routine medical practices without the need for (further) external validation	Conclusion or concluding paragraph in *main text* or *abstract* contains statements that claim clinical applicability without stating the need to perform proper validation and/or clinical impact studies.	*''Our finding suggests that random forest model would be best option to implement a system for predicting fatty liver disease patients appropriately and effectively''* [44] — Development-only study.
	Stating any lack of clinical applicability/effectiveness based solely on the poor performance in the specific validation sample	Conclusion or concluding paragraph in *main text* or *abstract* contains statements that limit the applicability of the validated model and thus, supporting the	*''Previous developed models showed low accuracies on our external validation. We therefore developed a better performing model''*[a] — The validation sample

(*Continued*)

Table 1. Continued

Category of spin	Form of spin	Criteria	Examples
		development of a new prediction model.	contains participants with a different case-mix. Instead of recalibrating or adding new predictors, authors have redeveloped the model.
	Stating the use of prediction model for a different outcome, setting, or population without any evaluation	Conclusion in *main text* or *abstract* contains statements that generalize models' performance to different outcome, setting, or population without stating the need to perform proper evaluation.	*"This can be extended to predict other type of ailments which arise from metabolic syndrome"* [45] – Development-only study.

AUC, area under the curve; EOC, epithelial ovarian cancer.
[a] Fictitious example.

framed the findings of a study. To reduce the chance of overvaluing or undervaluing evidence based on a particular framing alone, we developed SPIN-PM, a consensus-based framework to provide guidance to researchers and to assist readers in identifying and evaluating spin practices in studies on prediction models.

Our proposed spin practices and facilitators are consistent with those previously identified in other study designs [22,58]. The difference between some of the proposed spin practices is subtle but they serve slightly different functions in the narrative, for instance between the use of qualifiers and positive words (Table 1, linguistic spin). Qualifiers primarily adjust the precision or certainty of a statement (eg, *clearly shows*), while leading words shape the audience's perception or emotional response to the information conveyed (eg, *groundbreaking*). Some of the facilitators of spin can also be found in TRIPOD and in the PRediction model Risk Of Bias ASsessment Tool (PROBAST), as they were defined as any reporting practices that hinder the critical appraisal of a study due to incomplete or inaccurate reporting practices. We consider facilitators not to be spin practices per se, as they cannot be confirmed within the study at hand.

Our framework focuses on a manuscript's concluding paragraphs because these are especially susceptible to contain misinterpreted or overinterpreted statements as well as unsubstantiated claims of transportability, and readers tend to focus on them to judge the relevance, applicability, and generalizability of the study findings. The interpretation of study findings needs to be framed with the inherent limitations, contextualizing the reported prediction model and its performance. Furthermore, we incorporated a practice in an opposite direction, that is, the use of words to downgrade findings on external validation studies to support the development of a new prediction model (Table 1, misleading transportability). We found this practice equally detrimental, particularly in studies performed by groups of independent researchers. Although extrapolation allows to set the research into 'real-world' context by highlighting the potential final application of the prediction model, we suggest authors to avoid such statements in concluding paragraphs, especially if the study has an explorative rather than applied aim and furthermore, when it is not even supported by their analyses and findings.

SPIN is not per se deliberate malpractice. Authors may overinterpret their findings because of the current academic reward system, methodological illiteracy, and the prioritization of positive and novel studies by journals and funders [59–61]. On the other hand, readers may find overinterpreted articles as consequence of poor peer-review, publication and citation bias, or lack of related expertise [61]. Several more factors and even unconscious ones are likely to play a role in the complex system of interpreting and communicating scientific findings. Previous studies have addressed factors such as conflicts of interest, industry-based research, authorship, affiliation, and journal's impact factor as potential determinants of spin [22,23,62]. SPIN-PM is the first step toward increasing awareness about spin practices in development and validation studies on prediction model.

Spin evaluation requires background knowledge about studies on prediction models and to be weighted in the light of the context at hand. Authors and readers are still required to judge how detrimental the spin practice is within the context of their particular research question, study type (model development only, development with external validation, external validation only) or publication type (preprint, peer-reviewed, proceedings). Similarly, they need to determine whether the use of qualifiers (ie, *very, clearly*) or 'hedging' (ie, *may, could*) relativizes the certainty of a statement based on the findings that have been reported. Readers might still disagree regarding the likely effect of certain criteria; thus, comprehensive evaluation of spin practices will remain partially subjective. Efforts to measure and improve the inter-rater reliability should be considered when using SPIN-PM for systematic evaluation.

Table 2. Examples of facilitators of spin in studies on prediction models

Category of spin	Facilitators	Criteria	Example
Misleading interpretation	Study aim is unclear or not reported.	Study aim is partially described or unspecified within *abstract* or *main text.*	*"This paper proposes an importance-driven approach to identify key markers/ features for the detection of early Parkinson's disease."* [46] — The study later reports findings of a classification model based on key predictors.
	Key details of dataset are partially reported or unreported.	Information about origin and/or collection of data, OR enrollment of participants is not provided within the manuscript, supplementary material, and/or referenced. This can occur when using, for example: a. Online open data repositories b. Biobank data c. Population cohorts d. Well-known randomized controlled trials	*"Patients diagnosed with COVID-19 from March 4th to April 5th from eight large University Hospitals were eligible if they had positive reverse transcription polymerase chain reaction (PCR-RT) and signs of COVID-19 pneumonia on unenhanced chest CT."* [47] — Insufficient description of data sources. Further references are not provided.
	Citation of the original article that describes the development of the prediction model being validated is missed.	Information about the development of the model being validated is not properly provided within the manuscript or referenced. It applies to validation studies only.	*"The CHADS2 score ranging from 0 to 6 was calculated for each patient as congestive heart failure or left ventricular ejection fraction _75 years (1 point); and history of stroke, transient ischemic attack (TIA), or systemic embolism (2 points)."* [48] — Reference of the original study is not provided.
	Inappropriate exclusion of participants from the analysis.	Discussion and/or Conclusion in *main text* lack addressing potential risk of using an unrepresentative sample of participants during analysis. Examples where this can occur include: a. Inappropriate handling of missing values b. Excluding participants with incomplete follow-up.	*"The results were robust to inclusion of participants with known risk factors for cardiovascular disease…"* [49] —Participants with fatal myocardial infarction were excluded in this study. Participants are thus a lower-risk sample of the original population at risk. Limitations regarding the unrepresentativeness of the sample are not further discussed.
	Additional complexities in the analysis are ignored (if applicable).	Discussion and/or Conclusion in *main text* lack addressing potential limitation due to unaddressed complexities in the data, such as long-term outcomes, competing risks, or clustering. An example is death in elderly patients before a second event of interest (competing risk).	*"The main finding of our study is the high specificity for accidental fall prediction reported in older inpatients."* [50] — The aim is to predict fall risk in older patients; however, the outcome is treated dichotomously (ie, fallers have one or more falls), while data were collected accounting for all fall events per patient. Potential limitations are not further discussed.
	Inappropriate method for internal validation is used.	Discussion and/or Conclusion in *main text* lack addressing potential limitations due to splitting the original data to obtain a dataset for internal validation (ie, testing).	*"We randomly split the original dataset into 70% of patients for a training subset and 30% for a testing or validation subset."* [51] — The total sample size was relatively small. No further concerns are mentioned in discussion.
	Reported results are not in accordance with study aim and methods.	Statements in Results describe findings based solely on analysis that go beyond the aim and methods reported in *main text*. This can occur when reporting on, for example: a. Prognostic factors b. Unplanned subgroups c. Results based on an analysis that was not preplanned.	*"To develop and validate a prognostic model applicable to high, middle, low income countries".* [52] — Validation was done using data from high-income countries alone. However, this is highlighted later as limitation.

Table 3. Examples of facilitators of spin in studies on prediction models

Category of spin	Facilitators	Criteria	Example
Misleading interpretation	Inaccurate reporting of performance measures in development studies	Results in *main text* are described in tables or text without stated if reported performance is apparent or optimism corrected.	*''Additional prediction metrics (*eg, *recall and precision) are shown in* Table 2.'' [53] — Unclear in main text whether reported measures are apparent, or optimism corrected. No further details are provided in Table 2 either.
	Measures of performance are partially reported.	Results in *main text* are partially reported. Reporting only discrimination might mislead reader into appraising a model as ''good'' without knowing if the model provides accurate individual probabilities (calibration). Both calibration and discrimination should be reported when developing risk prediction models.	*''In general, NN-based models show better performance when predicting readmission, except for CHF (where GBM outperforms NN).''* [54] — The aim was to identify patients at high risk of readmission; however, calibration measures are not reported. Judgment of ''better'' performance is based solely in discrimination ability.
	Performance measures without confidence intervals are reported.	Results in *main text* are reported without confidence intervals to indicate the level of precision of reported performance measures.	*''GLMN outperforms the other MLTs among those implemented with CC analysis, with higher values of all the measures used to compare the algorithms.''* [55] — Table 2 does not report confidence intervals for any of the performance measures presented.
	Inappropriate presentation of plots	Receiver operating characteristics curves and calibration plots are presented with their axes squashed or truncated.	''Fig.5: R*OC Curve for classification of Gaussian K-Base NB Classifier''* [56] — Left plot presents the y-axis squashed. Also, there is no clear definition of what plot A or B represent in the main text.
	Unsubstantiated claims of superiority of one modeling approach over another are stated	Statements in *main text* and/or in *abstract* claim superiority of one modeling approach over another.	*''The performance of most of the ML-based models was significantly better than that of conventional methods using a single clinical feature, Knosp grade, which is commonly used to predict TTS response.''* [57]
	Unfair comparison between models	Statements in main text and/or in abstract compare models based solely on their predictive performance ignoring methodological differences, changes on patient's characteristics, or clinical context.	*''In previous research, accuracies of up to 94% were reported. Our model achieved 95% and is therefore better.''*[a] *— Comparison is made based on 1% increase on model performance. Remains unknown whether this difference was due to design, methods, or conduct applied to during models' development.*
Misleading transportability	Unsubstantiated claims of clinical usefulness are reported.	Model performance measures to determine relevant threshold to support clinical decisions based on prediction models are not reported in *main text.* Examples are net benefit (NB), decision curve analysis (DCA), and net reclassification improvement (NRI).	*''The results demonstrate that the proposed technique is suitable with optimal discrimination … producing accurate, specific, and decision-oriented rules to facilitate physician and make informed choices about their management and improve health condition.''* [45] — Measures to assess clinical usefulness are not reported throughout the manuscript.

[a] Fictitious example.

4.1. Strengths and limitations

We conducted an iterative process to develop SPIN-PM by engaging in multiple consensus meetings with methodologists and statisticians with related expertise. Future research may expand the panel by including other type of stakeholders (eg, policy makers), by increasing the number of participants for further consensus and by choosing a more robust consensus methodology (eg, Delphi method, nominal group technique).

Our aim was two-fold: to establish clear definitions and to enhance the understanding of spin within studies on prediction models by providing examples and contextualizing their consequences. We broadened our criteria to enhance the applicability of the framework. By doing so, we have ensured that SPIN-PM is suitable for a wide range of contexts, including not only development and external validation studies but also both non-AI−based and AI-based prediction models.

Despite this stepwise process, some practices and facilitators are likely to have been overlooked. Furthermore, our SPIN-PM framework does not provide means to discern whether spin practices are the result of inexperience, undeliberate or deliberate misconduct, or both. Additionally, the proposed practices do not determine the optimal degree of proper framing for the communication of prediction models. Instead, we provide guidance on how to identify and be cautious about practices that could potentially adversely impact readers' interpretations.

4.2. Implication for researchers, reviewers, and readers

Spin practices are embedded in the process of writing, reviewing, and publishing scientific literature, in which different players share a collective responsibility. All authors and editors commonly use language to emphasize or 'spin' the certainty of the results. Behavior that is often encouraged by today's volume of research publication in which results will hardly speak by themselves [17]. The consequence is a biased representation of science that will almost always suggest robust solutions to healthcare problems. On the other hand, authors of well-conducted studies with scientific novelty and importance may appropriately use spin to frame their research finding and to one extent, it might be necessary to stand out and allow further research.

Inexperience regarding studies on prediction models, lack of guidance, and language barriers may explain the presence of spin. Guidance on 'what to write' is available through reporting guidelines, instead guidance on 'how to write' is scarce [3,33,63]. Both TRIPOD and SPIN-PM serve as tools to improve reporting quality. While adhering to TRIPOD and the upcoming TRIPOD + AI checklist (www.tripod-statement.org) can reduce the chances of omitting essential information to evaluate a prediction model, SPIN-PM provides guidance on how such omissions may lead to misinterpretation of results in other sections of the manuscript [3,33,64,65].

4.3. Future research

This work should be seen as starting point to identify and discuss spin practices in prediction research. Further research could improve the consensus methodology, identify further facilitators of spin, or develop a severity score based on the likelihood to distort reader's interpretation of each practice (low, moderate, high, unclear) [36]. Similarly, an overall 'spin-measure' per study could contribute to the critical appraisal when conducting systematic reviews of prediction models. Moreover, there is an urgent need to implement effective long-term interventions to reduce spin practices across all study designs [66,67].

4.4. When and how should SPIN-PM be used?

SPIN-PM is primarily intended for researchers reporting prediction models and to assist readers in identifying and evaluating spin practices. We anticipate SPIN-PM can be used by decision-makers when assessing potential utility of prediction models for routine clinical care. However, the use of SPIN-PM should be limited to flag and appraise the quality of reporting of a study rather than the overall methodological quality or conduct. For this, we recommend using PROBAST and PROBAST + AI (www.probast.org) [4]. We stress that SPIN-PM is not a scoring tool or an instrument for assessing overall quality or overall presence of spin. Rather, it serves as a set of definitions designed to assist any reader in identifying misleading practices, facilitating critical appraisal, and potentially advancing research on spin in studies on prediction models. We encourage authors, peer reviewers, and editors to provide further feedback on how we can improve SPIN-PM or suggest opportunities to expand the framework's items in collaboration with the authors of this article.

5. Conclusion

Spin is a widely recognized phenomenon in the biomedical literature. We call researchers who develop and validate prediction models, either non-AI−based or AI-based, into publishing results that are both robust and consistent, as this can improve the quality of future research. With SPIN-PM, we aim to reduce vague and biased reporting of findings. We believe by doing so, the uptake of prediction models in clinical practice will be facilitated.

Ethical approval

Ethical approval was not required for this study because solely relied on publicly available data and did not involve any interventions on human participants. Thus, it falls outside the scope of institutional requirements for ethical approval.

Contributors expertise

CLAN had the idea for the article and led the development of SPIN-PM. JAAD has extensive experience in systematic reviews of prediction models studies. MG has expertise in developing intervention strategies to reduce spin. PD has experience in developing and validating prediction models. MvS, JBR, GSC, RDR, KGMM, and LH have extensive experience on methodological aspects of

prediction model development and validation. GSC and KGMM co-lead TRIPOD, an international collaboration for the development of consensus reporting guidelines for prediction model studies. JBR and LH have participated in the development of a classification scheme for spin in diagnostic accuracy test studies, while GSC in prognostic factor studies. CLAN wrote this manuscript with substantial contribution and revisions from all the authors.

CRediT authorship contribution statement

Constanza L. Andaur Navarro: Writing − review & editing, Writing − original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Johanna A.A. Damen:** Writing − review & editing, Supervision, Methodology, Investigation, Conceptualization. **Mona Ghannad:** Methodology, Investigation. **Paula Dhiman:** Writing − review & editing, Methodology, Investigation, Conceptualization. **Maarten van Smeden:** Writing − review & editing, Methodology, Investigation, Conceptualization. **Johannes B. Reitsma:** Writing − review & editing, Methodology, Investigation, Conceptualization. **Gary S. Collins:** Writing − review & editing, Methodology, Investigation, Conceptualization. **Richard D. Riley:** Writing − review & editing, Methodology, Investigation, Conceptualization. **Karel G.M. Moons:** Writing − review & editing, Supervision, Methodology, Investigation, Conceptualization. **Lotty Hooft:** Writing − review & editing, Supervision, Methodology, Investigation, Conceptualization.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgments

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.2024.111364.

References

[1] Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381.
[2] Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009; 338(7707):1373−7.
[3] 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.
[4] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1−33.
[5] Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416.
[6] Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak 2006;6:1−10.
[7] Van Dieren S, Beulens JWJ, Kengne AP, Peelen LM, Rutten GEHM, Woodward M, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart 2012;98:360−9.
[8] Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011;9(1):103.
[9] 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.
[10] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021;375:n2281.
[11] Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2020;2(12):e677−80.
[12] 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.
[13] Chen JH, Asch SM. Machine learning and prediction in medicine-beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507−9.
[14] 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.
[15] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models [Internet]. J Clin Epidemiol 2023;158(11):99−110.
[16] Fletcher RH, Black B. Spin in scientific writing: scientific mischief and legal jeopardy. Med Law 2007;26(3):511−25.
[17] Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A 2018;115:2613−9.
[18] Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MMG. Misreporting of diagnostic accuracy studies : evidence of "spin.". Radiology 2013;267:581−8.
[19] Horton R. The rhetoric of research. BMJ 1995;310:985.
[20] 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.
[21] 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.

[22] 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.
[23] Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: a methodological systematic review. PLoS Biol 2017;15(9):1−16.
[24] 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.
[25] 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.
[26] 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(1):1.
[27] 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.
[28] Dwan K, Altman DG, Clarke M, Gamble C, Higgins JPT, Sterne JAC, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. PLoS Med 2014;11(6):1−22.
[29] Won J, Kim S, Bae I, Lee H. Trial registration as a safeguard against outcome reporting bias and spin ? A case study of randomized controlled trials of acupuncture. PLoS One 2019;10:1−19.
[30] Ioannidis JPA. Spin, bias, and clinical utility in systematic reviews of diagnostic studies. Clin Chem 2020;66:863−5.
[31] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. J Clin Epidemiol 2023;157:120−33.
[32] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683−90.
[33] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, 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.
[34] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691−8.
[35] Gattrell WT, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med 2024;21(1):e1004326.
[36] 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.
[37] 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.
[38] Krishnamurti T, Woloshin S, Schwartz LM, Fischhoff B. A randomized trial testing US food and drug administration "breakthrough" language. JAMA Intern Med 2015;175(11):1856−8.
[39] Peat G, Riley R, Croft P, Morley KI, Kyzas PA, Moons KGM, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med 2014;11(7):e1001671.
[40] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022;22:12.
[41] Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med 2018;16(1):1−12.
[42] 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.
[43] 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.
[44] Wu CC, Yeh WC, Hsu WD, Islam MM, Nguyen PA, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed 2019;170:23−9.
[45] 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.
[46] Xiao C, Liu Y, Feng DD, Wang X. Key marker selection for the detection of early Parkinson's disease using importance-driven models. Annu Int Conf IEEE Eng Med Biol Soc 2018;2018:6100−3.
[47] Chassagnon G, Vakalopoulou M, Battistella E, Christodoulidis S, Hoang-Thi TN, Dangeard S, et al. AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. Med Image Anal 2021;67:101860.
[48] Caro-Codón J, Lip GYH, Rey JR, Iniesta AM, Rosillo SO, Castrejon-Castrejon S, et al. Prediction of thromboembolic events and mortality by the CHADS2and the CHA2DS2-VASc in COVID-19. Europace 2021;23(6):937−47.
[49] Aslibekyan S, Campos H, Loucks EB, Linkletter CD, Ordovas JM, Baylin A. Development of a cardiovascular risk score for use in low- and middle-income countries. J Nutr 2011;141(7):1375−80.
[50] Beauchet O, Noublanche F, Simon R, Sekhon H, Chabot J, Levinoff E, et al. Falls risk prediction for older inpatients in acute care medical wards: is there an interest to combine an early nurse assessment and the artificial neural network analysis? J Nutr Health Aging 2018;22(1):131−7.
[51] Sanchez Fernandez I, Sansevere AJ, Gainza-Lein M, Kapur K, Loddenkemper T. Machine learning for outcome prediction in electroencephalograph (EEG)-Monitored children in the intensive care unit. J Child Neurol 2018;33(8):546−53.
[52] Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, et al. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ 2012;345:1−12.
[53] Hunter-Zinck HS, Peck JS, Strout TD, Gaehde SA. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. J Am Med Inf Assoc 2019;26:1427−36.
[54] Garcia-Arce A, Rico F, Zayas-Castro JL. Comparison of machine learning algorithms for the prediction of preventable hospital readmissions. J Healthc Qual 2018;40(3):129−38.
[55] Lorenzoni G, Sabato SS, Lanera C, Bottigliengo D, Minto C, Ocagli H, et al. Comparison of machine learning techniques for prediction of hospitalization in heart failure patients. J Clin Med 2019;8(9):1298.
[56] Kaviarasi R, Gandhi Raj R. Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed Gaussian classifier system. J Med Syst 2019;43(7):201.
[57] Fan Y, Li Y, Li Y, Feng S, Bao X, Feng M, et al. Development and assessment of machine learning algorithms for predicting remission after transsphenoidal surgery among patients with acromegaly. Endocrine 2020;67(2):412−22.
[58] 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.
[59] Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific

standards are a disservice to patients and society. J Clin Epidemiol 2021;138:219−26.
[60] Koletsi D, Karagianni A, Pandis N, Makou M, Polychronopoulou A, Eliades T. Are studies reporting significant results more likely to be published? Am J Orthod Dentofacial Orthop 2009;136(5):632.e1−5.
[61] Young NS, Ioannidis JPA, Al-Ubaydli O. Why current publication practices may distort science. PLoS Med 2008;5(10):e201.
[62] Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68:25−34.
[63] Boers M. Graphics and statistics for cardiology: designing effective tables for presentation and publication. Heart 2018;104:192−200.
[64] Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open 2019;9(4):e025611.
[65] 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.
[66] 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.
[67] Blanco D, Schroter S, Aldcroft A, Moher D, Boutron I, Kirkham JJ, et al. Effect of an editorial intervention to improve the completeness of reporting of randomised trials: a randomised controlled trial. BMJ Open 2020;10(5):e036799.