AMSTAR-PF: a critical appraisal tool for systematic reviews of prognostic factor studies

Michael L Henry, doctoral student[1], Neil E O'Connell, professor[2]*, Richard D Riley, professor[3,4], Karel GM Moons, professor[5,6], Beverley J Shea, professor[7,8], Lotty Hooft, professor[5,6], Johanna AA Damen, assistant professor[5,6], Nicole Skoetz, professor[9], Sarah B Wallwork, senior research fellow[1], G Lorimer Moseley, professor[1]

A iliations

[1] IIMPACT in Health, University of South Australia, Kaurna Country, Adelaide, Australia

[2] Centre for Health and Wellbeing Across the Lifecourse, Department of Health Sciences, Brunel University London, Uxbridge, UK

[3] Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom.

[4] National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, United Kingdom.

[5] Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands

[6] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands

[7] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[8] Bruyère Research Institute, Ottawa, ON, Canada

[9] Institute of Public Health, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

*Corresponding Author
N.E. O'Connell, PhD
Centre for Health and Wellbeing Across the Lifecourse
Department of Health Sciences
Brunel University London
Uxbridge UB8 3PH
United Kingdom
E-mail: neil.oconnell@brunel.ac.uk
X: @NeilOConnell . Bluesky: @neiloconnell.bsky.social

The ability to predict the onset or natural history of an illness, or how people may respond to a treatment, guides clinical decision making. These predictions are commonly based on prognostic factors: clinical, patient, or societal variables that are identified as being predictive of a certain future outcome. Prognostic factor research has increased across fields, with a subsequent increase in the number of systematic reviews of prognostic factors studies. Understanding the quality of such prognostic factor reviews is essential for confidence in their findings, but there is no quality appraisal instrument specifically designed for assessing systematic reviews of prognostic factor studies. We developed A MeaSurement Tool to Assess systematic Reviews of Prognostic Factor studies (AMSTAR-PF) to address this gap.

Summary Points

- Research into prognostic factors is vital for many areas of healthcare
- Confidence in the findings of systematic reviews of prognostic factor research can be compromised in a variety of ways
- We developed, refined, and tested A MeaSurement Tool to Assess systematic Reviews of Prognostic Factor studies (AMSTAR-PF), which was based on AMSTAR 2
- AMSTAR-PF uses signalling points and 19 questions over 14 domains to assist in coming to an overall judgement of confidence in the results of the review
- Providing a standardised and reliable tool to appraise review quality will assist users to ascertain confidence in the review's findings

Prognosis research encompasses a range of methods and goals (1, 2), and has been split broadly into four branches, detailed in the Prognosis Research Strategy (PROGRESS) series of papers (3-6). Fundamental, or overall, prognosis research details the average course of health-related conditions in a particular population and setting (3). Prognostic factor (PF) research aims to identify factors that are associated with the risk of certain outcomes occurring (4). Prognostic model research concerns the development, validation, or impact of a prognostic model to estimate an individual's risk of a future outcome occurring (1, 5). Stratified health care research uses prognostic information to tailor treatment to individuals (6).

Knowledge of PFs can form the basis for providing an individual with their likely prognosis, as well as being a building block for prognostic models, and a potential means of stratifying individuals into optimal treatment regimens (7). PFs can be drawn from a range of biopsychosocial domains, including clinical features (eg, tumour grade, blood, imaging or electrophysiology test results, pain levels, omics results), demographics (eg, age, marital status, gender), psychological (eg, comorbid mental health problems, injury beliefs, motivation), and societal factors (eg, geographical location, health care availability, local environment). However, before implementation into healthcare, it is imperative that the evidence to support their use is reliable.

The number of PF studies published each year is increasing (8, 9), many with uncertain or conflicting findings. Consequently, the number of systematic reviews (SRs) seeking to synthesise, pool and consolidate PF knowledge is also increasing (10). Hoffman and colleagues (10) estimated a 20 fold increase in the total number of SRs between 2000 and 2019, with 80

studies being published daily in 2019 across all research types. Between 2000 and 2004, 3.2% of all SRs were classified as addressing aetiology/risk; between 2015 and 2019, this had increased to 22.8% (10). However, significant doubts remain about the quality of published SRs, not only of PFs, but across the full range of research types (11-13). This is important because SRs often form the basis for policy and practice change, which makes the veracity of their findings immediately relevant to population level health outcomes.

Therefore, being able to determine the quality of the methods and content of published PF SRs is critical. PF research has a range of areas where quality can be compromised; some, such as the presence of a pre-registered protocol, clear research question, comprehensive search, non-selective reporting, and transparent and methodologically sound article selection, data extraction, and risk of bias assessment, are common to a range of research types, albeit sometimes with a different focus or methodology. Other important threats to validity that are not captured by other quality appraisal tools are specific to PF research, such as classification of PFs and outcomes, adjustment factors, comparator and other PFs, and appropriate calculation of prognostic effect sizes. PF research directly informs policy and practice, so it is imperative that PF-specific knowledge and methodology are used to systematically appraise research on PFs, in both general areas of potential limitations of quality, as well as the PF specific domains (14). Tools to appraise the quality of primary PF studies exist, for example the Quality In Prognosis Studies (QUIPS) tool (15), but, to our knowledge, there is no dedicated tool to critically appraise the quality of SRs of PF studies. We developed AMSTAR-PF (A MeaSurement Tool to Assess systematic Reviews of Prognostic Factor studies) to address this specific need.

AMSTAR-PF is based on AMSTAR 2(16), which was developed to appraise SRs of interventions. AMSTAR 2 evolved from AMSTAR(17) by incorporating significant feedback about the original AMSTAR, and adapting the tool to incorporate non-randomised, as well as randomised, interventional studies. AMSTAR-PF resembles AMSTAR 2 in several aspects of its content and guidance notes, yet has been comprehensively remodelled to address broad quality issues in a PF-specific way, and with additions to allow systematic appraisal of issues that are unique to PF research. Furthermore, alterations to the answering options allow more nuanced recording of limitations, and align it with a range of other appraisal tools.

Development of AMSTAR-PF

Initial stages and internal development

An overview of the entire development process is outlined in Figure 1.

In order to optimise efficiency, utility and confidence in the new tool, we gathered a Core Research Group that included experts who had developed tools and guidance for the following: AMSTAR(17) and AMSTAR 2(16) (www.amstar.ca) (BJS), Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I)(18), and - of Exposures (ROBINS-E)(19) (BJS), www.riskofbias.info) Risk Of Bias In Systematic reviews (ROBIS; www.bristol.ac.uk/population-health-sciences/projects/robis/robis-tool)(20) (BJS, KGMM), the PROGRESS partnership(3-6) (RDR, KGMM), Prediction model Risk Of Bias ASsessment Tool (PROBAST), and PROBAST+AI; www.probast.org)(21-23) (KGMM, RDR, LH, JAAD), Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD; www.tripod-statement.org) papers(24-26) (KGMM, RDR, LH), CHecklist for critical Appraisal and data extraction for

systematic Reviews of prediction Modelling Studies (CHARMS)(27) (KGMM), the updated Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2)(28) (KGGM), and guidance for using GRADE (Grading of Recommendations, Assessment, Development, and Evaluations; www.gradeworkinggroup.org) for prognosis(29, 30) (RDR, BJS). We did not include patients or the public in this methodologically-focused project.

The Core Research Group had an initial online meeting in July 2020 to discuss the need and optimal development approach for a quality appraisal tool specifically for SRs of PFs. We considered a range of tools on which to base the new tool. Deepest consideration focussed on AMSTAR 2 (16) and ROBIS (20). AMSTAR 2 is widely used to appraise SRs of interventional studies; ROBIS was designed to appraise risk of bias (RoB) in SRs of a range of study types including prognosis, but was not considered specific or detailed enough to optimally address PF reviews. We first created PF versions of both AMSTAR 2 and ROBIS. For AMSTAR 2, we added or substituted PF-specific terms and items for the interventional items. For ROBIS, we focused on PF-relevant signalling only. We reviewed these two versions, considering their applicability and efficiency and comparing commonalities and differences within the domains. There was significant overlap between these two modified tools. They primarily differed in their focus, AMSTAR 2 being a quality appraisal tool and ROBIS being a RoB tool. Quality appraisal considers RoB, but also the external validity and the reporting quality of the reviewed article(31). Through further discussion, we concluded that AMSTAR 2 had a more user-friendly format and provided a more complete skeleton for appraising overall quality, not just RoB.

The Core Research Group also considered research that compared AMSTAR 2 and ROBIS in interventional studies. One study identified superior agreement amongst users and faster completion when using AMSTAR or AMSTAR 2 for interventional SRs, than when using ROBIS (32). Another showed a preference among healthcare students for AMSTAR 2 over ROBIS, reporting that they found the guidance of the former clearer and easier to follow (33). Similar preferences have been reported by other researchers (34, 35). ROBIS was designed for a range of different review types, and it is possible that its wider generalisability may come at the cost of utility, especially for content-naïve researchers. AMSTAR 2 has been widely used since its publication in 2017, and we predicted that many researchers would be familiar with it, which may aid ease of use and uptake of AMSTAR-PF. We reached 100% consensus within the Core Research Group to use AMSTAR 2 as the basis for our new tool.

Having drafted an initial AMSTAR-PF, replacing AMSTAR 2's interventional focus with PF-specific adaptations, we appraised this draft through extensive group discussion and considered further changes. We considered: the relevance of each AMSTAR 2 domain in turn; additional domains or signalling points that needed to be created; and the structure and format of response options for both the individual domain questions and overall judgement.

Through a comprehensive iterative process that included email collaboration and six online meetings of the Core Research Group between July 2020 and July 2021, drafts of AMSTAR-PF were created, reviewed, and revised. In July 2021, we had a finalised Version 1.0 of the AMSTAR-PF tool, and then developed guidance notes. This Version 1.0 of AMSTAR-PF tool and guidance notes was then sent for external review and feedback.

External review and feedback

The external feedback process and pilot testing were approved by the Human Research Ethics Committee of the University of South Australia (ID 203954, 206951), and participants gave

informed consent to participate. We emailed the draft tool and guidance notes to the core members of the Cochrane Prognostic Methods Group (CPMG; n = 23), because this group was considered to have substantive expertise in prognostic research. The tool and guidance notes were accompanied by an anonymous survey (see Supplementary File 1A) that covered areas of redundancy, missing domains or questions, the relevance and utility of each domain and of the tool as a whole. Responses were free-text boxes. No questions were compulsory. Survey data were collected between 25 October and 27 November 2021 and managed using REDCap (Research Electronic Data Capture) electronic data capture tools hosted at the University of South Australia (36, 37). We received 17 responses; nine responses provided written feedback on the tool or guidance notes or both, and the remaining eight responses contained no information. The Core Research Group then met online to discuss this feedback and made further revisions to the tool and guidance notes, thereby generating Version 2.0 of the AMSTAR-PF tool and guidance notes.

AMSTAR-PF Version 2.0 was then distributed, with the same anonymous REDCap survey described above, to the full mailing list of the CPMG (n =207), who were permitted to share the email with colleagues they believed would have relevant expertise. Feedback was collected between 29 April and 27 May 2022. Twenty-one responses were received; nine provided written feedback on AMSTAR-PF, the guidance notes or both. The authorship team reviewed the feedback and as a result we developed a Microsoft Excel spreadsheet with an auto-populating summary page of question responses, as part of AMSTAR_PF Version 3.0 and guidance notes.

Pilot testing

Three rounds of Pilot testing occurred using 26 different appraisers. Twenty-five of the appraisers were independent of the tool development; they volunteered following an email inviting participation which had been sent to a research group email list, or were asked because they were in the process of completing an umbrella review of PF studies; see Supplementary File 1B for demographic information of the testers and Supplementary File 1C for details of the systematic reviews appraised.

Initial testing (Pilot 1) of Version 3.0 of AMSTAR-PF was performed by two PhD students who were independent of the tool development process. They independently appraised a convenience sample of six SRs of PFs for cardiovascular disease, and then provided appraisal results and qualitative written feedback on their experience of using the tool. The PhD students had previously performed quality appraisals as part of systematic and umbrella reviews, including of PF research. This feedback led to very minor changes and generation of AMSTAR-PF Version 4.0 and guidance notes.

A testing protocol for Pilot 2 was pre-registered on the Open Science Framework (OSF) (osf.io/3cgz9). Pilot 2 involved 10 appraisers (PhD students, post-doctoral researchers, and clinicians). One appraiser had been involved in developing AMSTAR-PF (MLH); nine appraisers had not. The appraisers used AMSTAR-PF Version 4.0 to appraise 23 SRs, which investigated PFs for chronic pain (n = 14), brain injury (n = 6), post-traumatic stress disorder (n = 2), and lymphoma (n = 1), and included both narrative syntheses and meta-analyses. All the review articles were appraised independently by two or more appraisers, with 90 completed appraisals in total. All appraisers were invited to provide qualitative feedback on the usability and acceptability of the tool, with eight providing feedback on their experiences. There were deviations from protocol (Supplementary File 1D). Three reviewers were unable to complete their full quota of appraisals. Feedback received during Pilot 2 suggested that further revisions

would be beneficial, so a planned agreement analysis was not undertaken on Pilot 2 in favour of making updates to the tool and guidance notes, then completing another pilot test and calculating agreement on the updated version. Particular difficulty coming to a final conclusion as to the quality of each review was noted, so that became a focus of revision. We had consciously avoided being too prescriptive in our instructions for deciding the overall quality of the review, but the feedback on that suggested that for many appraisers a certain amount of guidance would be helpful as a starting point.

Additionally, two expert researchers, with significant experience in prognostic research, reviewed AMSTAR-PF Version 4.0 and guidance notes and provided feedback. Minor revisions were then made, thereby generating AMSTAR-PF Version 5.0.

Finally, we undertook a third pilot test (Pilot 3), this time on AMSTAR-PF Version 5.0. The protocol for Pilot 3 was pre-registered on OSF (osf.io/acrwf) and is described in detail elsewhere (38). We deviated from the original protocol by our use of Gwet's Agreement Coefficient (AC) rather than kappa scores, as detailed in Supplementary File 1D. Pilot 3 involved eight SRs, two each on PFs for cancer, brain injury, lower back pain, and COVID-19. The articles were appraised by 14 appraisers (11 female, 3 male, ranging in experience from undergraduate student to post-doctoral researcher) using AMSTAR-PF Version 5.0 and guidance notes. No appraisers had been involved in the development of AMSTAR-PF, nor in any earlier feedback or testing rounds. On completion of Pilot 3, these appraisers were also given the opportunity to provide feedback on any aspect of the tool and guidance notes. That feedback, along with analysis of each domain's agreeability scores (38), was used to make final minor changes to the guidance notes, signalling points and wording of the questions, and thus arrive at the AMSTAR-PF tool and guidance notes that are herein presented to the field.

Agreement and usage time

Interrater, inter-pair, and intrapair agreement data from Pilot 3 is summarised in Table 1 and provided in detail elsewhere, alongside individual domain data and information on agreement over time (38). Agreement was calculated twice for the domains, using Gwet's Agreement Coefficient (AC) (39): once with full answering options (Yes, Probably Yes, Probably No, No, and for some questions, N/A) and once with Yes and Probably Yes collapsed into a single positive category, and No and Probably No collapsed into a single negative category. Agreement was moderate, substantial or almost perfect for most domains, according to Gwet's application (40) of Landis and Koch's benchmarks (41). Variability was noted between raters and between pairs (38). Agreement was higher with collapsed answering options, and intrapair agreement was routinely higher than interrater or inter-pair agreement. Mean (SD) time to complete quality appraisal across all articles was 39 (15) minutes. Mean time to reach consensus was 11 (7) minutes (38).

	Interrater	Inter-pair	Intrapair
Domains (all answering options)	0.59 (range 0.21 to 0.90)	0.61 (range 0.24 to 0.91)	0.75 (range 0.45 to 0.95)
Domains (collapsed)	0.72 (range 0.37 to 0.96)	0.73 (range 0.36 to 0.96)	0.83 (range 0.63 to 0.97)
Overall Appraisal	0.46 (95% CI 0.30 to 0.62)	0.46 (95%CI 0.17 to 0.74)	0.68 (range 0.22 to 1.00)

Table 1: Gwet's AC for Interrater, Inter-pair, and Intrapair agreement in Pilot 3 (from (38)). Numbers presented are means (range) for the interrater and inter-pair domains, and mean (range of means) for intrapair. For the overall rating, interrater and inter-pair AC scores are presented with 95% confidence intervals, and the intrapair shows the mean and the range of the intrapair scores. Gwet's AC was calculated unweighted for nominal data (questions 2b, 7c, 9a, 9b, 10, and 12, which include an N/A option), and with linear weighting for ordinal data (the remaining 13 questions, and the overall appraisal).

The AMSTAR-PF tool

AMSTAR-PF contains 19 questions grouped into 14 domains. These domains are listed below, with justification and explanations about their content. The domains are arranged similarly to AMSTAR 2, and broadly follow the regular layout of an SR, from methods through results and discussion. The complete tool and guidance notes are included in the Supplementary Material (2A, 2B, and 2C), and we encourage readers to refer to the guidance notes for more comprehensive detail on the included questions.

AMSTAR-PF Domains

7¡ Research.Question

Systematic reviews of PFs should have a clear, pre-defined research question with appropriate depth of information. PICOTS (7, 27) provides a framework for important information in PF research, and represents the Population, Index prognostic factor/s, Comparator/other prognostic factor/s, Outcome, Timing, and Setting. Having this information clearly detailed allows for a logical step to inclusion/exclusion factors, and for readers to judge the relevance of the review for their own practice or research.

The elements of PICOTS should be described in a way that makes the review question clear and reproducible, whether the authors choose to strictly define the parameters of a certain element or leave it more open and inclusive.

8¡ Pre_registered.Protocol

It is recommended that an SR is clearly planned, and that this process is documented in a protocol prior to commencing the SR. This helps as a safeguard to ensure that important methodological or analytical decisions are not based on findings made during the conduct of the review, and hence is an important step to minimise bias. The protocol should be registered and made publicly available by the time the SR is published, to enable readers to compare the protocol to the review conducted. Any deviations from the protocol should be documented and justified in the SR.

The AMSTAR-PF tool divides this question into two sub-sections, to clearly address both the presence of a publicly available pre-registered protocol, and how any deviations from the protocol are handled.

9¡ Included.Study.Designs.and.Types

It is important for authors to state which study designs are eligible for inclusion in the review. Prospective and retrospective cohort studies, case control studies, registry data and data from randomised trials may be included, because all can give information on PFs, though with potential different methodological issues to consider. Prospective cohort studies are seen as the ideal design for PF research, as researchers have maximum control over timing and types of PF and outcome measurements. Other study designs are more likely to have weaknesses,

though in many cases are easier and cheaper to run, and so are likely to be encountered in the reported research.

As well as different study designs, studies with different primary objectives may be included. Most commonly, studies of PFs will be the source of evidence in an SR of PFs, however in certain situations prognostic modelling studies (ie, those that aim to develop a model for individual risk prediction) or aetiological research (ie, those aiming to identify causal factors) can provide information about PFs(1).

Decisions made by review authors on the types and designs of studies eligible for inclusion in a review can have repercussions on bias and heterogeneity and may necessitate separate synthesis when calculating PF effect sizes.

0¡ Search.Strategy

A comprehensive search strategy is a vital component of an SR of PFs(42, 43). The search strategy should be wide-ranging enough to ensure all relevant studies are captured and allow for a complete synthesis of available evidence. An SR founded on an inadequate search may be missing important data and hence give an unreliable picture of the current state of evidence on a topic.

What constitutes a comprehensive search strategy will be different in different topic areas, as well as for different PFs and outcomes. Appraisers are encouraged to consider the signalling points listed as well as any other potential sources of data relevant to the SR topic and aims.

❶ Process.for.evaluating.and.including.studies

It is recommended that at least two review authors independently determine the eligibility of each study found in the search for inclusion into the SR. This commonly involves a two-stage process, whereby an initial screen is performed on title and abstract, followed by a secondary screening of the full text of all potentially relevant articles. While the process may differ in different reviews, at each point of decision-making it is recommended that at least two independent reviewers are involved in the decision-making process. A plan for resolution of disagreements should be described.

❷ Excluded.studies

It is recommended that all potentially relevant studies that were read at the full text stage are listed with a justification for exclusion of each. This is important to allow readers to judge the validity of the SR and applicability to their own research or practice.

❸ Data.Extraction

At least two assessors should independently extract the necessary data from the included studies, and a plan should be followed to resolve any disagreements. Having two independent extractors helps to minimise the risk of errors in this process.

The data extracted from the included studies is likely to vary depending on the aims of the PF review(27). However, all included studies should be described in adequate detail to allow readers to assess the appropriateness of each study's inclusion and relevance.

PF effect estimates and their precision from primary studies may be derived in different ways and presented in different formats, which can make extraction for the review challenging. In certain situations, authors of SRs may have needed to calculate PF effect estimates and their precision from the results provided in the primary studies(7). Where this has been done, the methods used should be clearly reported and appropriate. A major topic is whether unadjusted

or adjusted PF results were extracted and, if the latter, the adjustment factors of interest. Adjusted results examine the contribution of a PF over and above (ie, after adjusting for) other variables, which should typically refer to well-known existing PFs. However, extracting or deriving estimates adjusted for the same set of factors in each study is very challenging, and often impossible. How this issue was dealt with should be explained.

For clarity, AMSTAR-PF subdivides the data extraction question into three sub-questions, dealing with a) the process of data extraction, b) the detail of report of extracted data, and c) appropriateness of calculating effect sizes from reported data, where relevant.

❹¡ Assessing.Risk.of.Bias

Two assessors should independently assess the risk of bias of each included article in the SR and have a plan in place for resolving any disagreements during the consensus process.

The method chosen for assessing risk of bias should be pre-planned and clear. This will often involve the use of a recognised risk of bias tool; preferably one designed for PF studies such as the Quality In Prognosis Studies (QUIPS) (15).

This question is subdivided into two sections in AMSTAR-PF, to cover both the process (8a) and technique used (8b) to complete risk of bias assessment.

❺¡ Data.Synthesis

The review protocol should have stated the principles on which review authors based their decision to synthesise data from included studies, and how it is planned to do this. There are many areas in the included studies where differences may arise (eg, different lengths of follow-up; different cut-points used to dichotomise continuous factors; different adjustment factors), and PICOTS (7, 27) provides a good framework for exploring many of these. Additionally, different ways of calculating and presenting prognostic effect sizes and their precision provides another area in which differences can arise, and compatibility of studies decrease. Even in non-quantitative syntheses, the principles behind ensuring interpretability of results remain, and separate summaries may need to be presented (eg, by unadjusted and adjusted results; for each cut-point) when included studies differ in key areas.

If meta-analysis was performed, it is expected that authors used appropriate methods, such as a random-effects model to account for unexplained heterogeneity in PF effects. Authors should have also pre-planned how they will investigate heterogeneity, and quantify it appropriately. As well as statistical heterogeneity, an examination of potential clinical or methodological reasons for heterogeneity may be necessary.

This question is subdivided into two sections in the AMSTAR-PF tool, with 9a focussing on the approach taken to ensure interpretability of results, and 9b dedicated to analysis methods (in the case meta-analysis was performed).

76¡ Small.study.effects

Small study effects is the phenomenon that occurs when studies with smaller sample sizes demonstrate systematically different PF effects than studies with larger sample sizes. This might be due to publication bias, such that smaller PF studies are may likely to published when they provide statistically significant results, compared to non-significant results. However, heterogeneity may also cause small study effects, for example if smaller studies have a shorter length of follow-up than larger studies, and a PF effect genuinely varies from short to longer term. Hence, PF reviews investigating asymmetry in funnel plots should refer to the examination

of small study effects, rather than publication bias, because it is not possible to be certain that publication bias is the reason for any asymmetry observed, particularly in the absence of formal registries for PF studies.

Commonly, statistical tests for asymmetry may accompany a funnel plot, however these tests often have low power, so they are not a panacea.

If funnel plots or asymmetry tests do suggest small study effects may be present, it is expected that this forms part of the discussion about the review's findings and conclusions.

77¡ Discussing.Risk.of.Bias

The potential effects of bias on the review's results and conclusions are an important issue. In cases where only studies adjudged to have a low risk of bias were included in the review, there may be relatively little discussion.

If studies of varying quality were included (which is often the situation with PF reviews) then it is expected that this is explored. If a meta-analysis was performed, then evaluation of how effects vary by study quality may be undertaken with techniques such as subgroup analysis, meta-regression analysis, or sensitivity analysis. Even if these aren't possible, we recommend authors provide some commentary as to possible impact of bias on the individual included studies, and the results of the review.

78¡ Discussing.Heterogeneity

There are many potential causes of heterogeneity in PF research; some of these are explored as sources of bias in Question 8 and accounting for heterogeneity is also touched upon in Question 9. If a meta-analysis was performed, it is generally expected that heterogeneity will be accounted for in the meta-analysis and quantified, for example using tau-squared (the estimate of between-study variance in PF effects) and prediction intervals (for a PF effect in a new study). In situations where only a small number of studies are included, then the estimate of heterogeneity will be very uncertain (and prediction intervals wide), but acknowledgment of this issue should be expected.

Irrespective of whether meta-analysis is undertaken and estimates of heterogeneity are obtained or not, the review authors should still consider and discuss potential sources of clinical and methodological heterogeneity, and the possible impact on the results, conclusions, and recommendations of the SR.

79¡ Conflicts.of.Interest

It is important that review authors document any funding sources or other potential conflicts of interest explicitly in the manuscript. Furthermore, it is recommended that any conflicts of interest in the included studies also be documented. This can help readers better assess the quality of evidence and any potential conflicts in the included studies.

70¡ Certainty.in.Results

It is recommended that authors of the review address the level of certainty around their key findings. The GRADE guidelines (44) are a commonly used framework for authors, however review authors may use other methods for addressing this issue. It is noted that at the time of writing, there is not a full range of GRADE guidelines covering all the range of PF studies, so authors may be required to modify existing resources (for example existing GRADE guidelines for prognosis (30, 45)) or develop their own methods using similar principles.

Overall.confidence.in.the.results.of.the.review

AMSTAR-PF is not designed to give an overall score, as a high score could mask critical failings in one or more key areas. Rather, each domain should be considered potentially vital to the overall confidence in the results of the review, and any errors or oversights noted appraised with this in mind.

We have suggested four categories for overall confidence in the results of the review: High?. Moderate?Low?and.Critically.Low. The attached tool and guidance notes provide further detail about our suggested approach to classifying a review; however, we stress that appraisers can, and should, make decisions given their own topic and methodological knowledge.

Applying AMSTAR-PF

Many elements of AMSTAR-PF are open to varying interpretations and may have different levels of importance in different fields or in different reviews. We recommend that team members planning to use the tool meet prior to undertaking appraisal to ensure consistency in interpretation and application, and to discuss areas of topic or methodological importance that may need clarification or standardisation. We envisage that a common reason for appraising SR quality may be as part of an umbrella review. If so, it may be important to clarify areas of quality that are particularly important to the umbrella review question and outcomes sought, particularly if the findings will directly influence policy or practice.

AMSTAR-PF has 19 questions over 14 domains, with some of the questions containing signalling points to assist with answering the question. The response options for each of the questions are Yes (Y), Probably Yes (PY), Probably No (PN), No (N), and for some questions, Not Applicable (N/A). The signalling points have the same response options. For simplicity, all questions and signalling points are worded such that 'Yes' indicates higher quality, and 'No' indicates lower quality. We recommend that Yes and No options are used when there is clear evidence in the review for or against the signalling point or question, and the Probably Yes and Probably No options when the evidence is less clear, or assumptions need to be made when answering. The recommendations for using this answering system are consistent with other quality appraisal and RoB tools, eg, ROBINS-I (18), PROBAST (21, 22), ROB 2 (46), and ROB-ME (47). Similar to other appraisal tools (eg (18, 19, 23, 46, 47)) responses of 'Yes' and 'Probably Yes' have similar implications for the overall appraisal of quality of the review, and as do 'No' and 'Probably No'.

Deciding which answering option to choose requires a certain amount of judgement, alongside topical knowledge and methodological knowledge. The "probably" option is appropriate where there is imperfect information available and indicates that the reviewers have had to make a judgement because of that. Taking question 1 (Did the review clearly define the research question, including the relevant components of PICOTS?) as an example, the first step would be to assess the extent to which the signalling questions have been addressed in the systematic review, and mark these accordingly. Using those responses as a guide, appraisers should then consider the question as a whole, and to what extent it was answered. If appraisers feel that the review has clearly defined all relevant aspects of the review question, then "Yes" is an appropriate response. If they feel that the question is adequately defined, but missing an element of PICOTS or specificity around a certain signalling point or aspect of the review question, then "Probably Yes" may be preferred, as it indicates that there is a lack of clarity and a judgment was made. Conversely, we recommend that a review with a poorly defined, or ambiguous research question is given a "No" response, whereas a review with a few well-defined elements, but still

overall lacking clarity and appropriate definitions of the elements, may receive a “Probably No” response. The use of the different response options may differ in different reviews, in different questions, and across different reviewer teams and fields.

The AMSTAR-PF guidance notes (see Supplementary File 2B) provide more detailed information about each of AMSTAR-PF’s domains and application, and we recommend that users of AMSTAR-PF review these and refer to them when using the tool. We stress that given the variety of topic areas, methodologies, review aims, and expertise of appraisers on a team, these notes should be seen as guidance only; teams may wish, or need, to further operationalise their application. Furthermore, teams may find it useful to pre-plan a meeting after appraisers have completed appraisal of 3-6 reviews, in order to confirm their interpretations, compare appraisals and identify any unexpected inconsistencies in interpretation or application (38, 48). This is especially true if appraisers have different experience levels or knowledge bases, or need more guidance in certain areas.

Discussion

A MeaSurement Tool to Assess systematic Reviews of Prognostic Factor studies - AMSTAR-PF - was specifically designed to appraise the quality of PF SRs. It is, to our knowledge, the first such tool and fills an important gap for people assessing or undertaking SRs of PFs, which have many unique challenges to consider (7). The development process involved a wide range of researchers, was comprehensive and iterative, and several pilot trials were undertaken before arriving at the final version presented here.

AMSTAR-PF and guidance notes were based on AMSTAR 2, and therefore share commonalities. There are, however, differences in the domains and how they have been presented, with some of AMSTAR-PF’s domains divided into sub questions in order to highlight important elements of that domain. Other modifications were a change in the order of some domains, and changes in guidance to reflect current ideas and recommendations for best practice that have evolved since AMSTAR 2 was published in 2017. For instance, we recommend that it is not enough to state there was a protocol, but rather that it should be publicly available (Q2a). Furthermore, we added a question about the process used in performing RoB (Q8a), one about the interpretability of results (Q9a), and one dealing with the certainty around key findings (Q14). Other changes were necessitated to accommodate the different focuses of PF research; changes were made to the wording of questions, and additional questions (eg, Q7c, around PF effect estimates) were added.

The categories for the overall rating of the appraised SR remain the same in AMSTAR-PF as they are in AMSTAR 2, but answering options for the domains have been changed. AMSTAR-PF differs from AMSTAR 2 in its inclusion of “probably yes” and “probably no” responses for all questions, whereas a “partial yes” option was only available in some AMSTAR 2 questions, and there was no “probably no” option in any question. AMSTAR-PF also uses these same options (Y/PY/PN/N, +/- N/A) for each of the signalling points, which diverges from the checkboxes used in AMSTAR 2. We considered a range of options for answering the questions and signalling points. Some other tools have a ‘no information’ or ‘unclear’ option, or force a binary Y or N decision for questions and/or signalling points. The change to include the additional “probably” options aims to enhance the useability of the tool, especially given judgements will need to be made where there is inherent uncertainty, and the extra options provide added detail that may be useful

when two appraisers meet to reach consensus. We acknowledge that such responses can add complexity when coming to a final decision on each domain and on the appraisal as a whole, but we consider that the overall benefit of the added detail makes this change worthwhile, especially given recognised deficits in quality in reporting of prognosis research (www.tripod-statement.org) (7, 8, 25, 49). This answering structure aligns AMSTAR-PF with many other tools currently being used, eg, ROBINS-I (18), ROBINS-E (19), ROB2 (46), ROBIS (20), ROB-ME (47) and PROBAST (21-23).

AMSTAR 2 uses the concept of "critical domains" – pre-specified domains considered to be of integral importance to a review's quality – to assist in coming to a final judgement on the quality of the review under appraisal. We consider that such critical domains may be less consistent in PF research than in interventional research, and that it is more beneficial to instead have appraisers approach each and every domain as potentially critical to the outcome and findings of the review. We therefore chose to diverge from AMSTAR 2 in this respect. Broadly, however, AMSTAR-PF follows a similar format and style to AMSTAR 2. This resemblance and familiarity was deliberate so as to assist users already familiar with AMSTAR 2. A summary of the main differences between AMSTAR 2 and AMSTAR-PF is shown in Box 1.

AMSTAR-PF is designed to facilitate a comprehensive appraisal of the quality of PF SRs. Whiting and colleagues (31) describe three components of quality; internal validity (risk of bias), external validity (applicability/variability) and reporting quality. AMSTAR-PF incorporates considerations of each of these components in the included domains, and all should be considered when coming to a final judgement of the quality of the review in question. Like AMSTAR 2, the highlighting of different elements of quality may make AMSTAR-PF useful as a brief checklist for people conducting an SR of PFs, or as a teaching aid. It is not, however, a substitute for more comprehensive methodology and rationale in conducting SRs, nor a replacement for detailed RoB guidance.

AMSTAR-PF was developed for SRs of PF studies. PF research, however, is variably defined; some authors (49), for instance, highlight causal PFs as part of this broad subgroup, and differentiate these as "prognostic determinants (causal)" as opposed to "prognostic markers (non-causal)". AMSTAR-PF does not differentiate between causal and non-causal PFs, and is suitable for either or both. Specific guidance on appraising the quality of the methods used to establish the presence and/or degree of causation is, however, outside the scope of this tool, as is aetiological research more generally. Areas of research outside of PF reviews may need their own specific guidance, whether from a distinct tool or tailored by modifying existing resources, such as AMSTAR-PF.

There are a number of strengths to the development of AMSTAR-PF and the resulting tool. We developed AMSTAR-PF using a rigorous, multi-stage design and validation process, which aligned with recommendations for developing quality assessment tools (31), and had input from diverse international experts within and external to the Core Research Group. Responses from the surveys to members of the Cochrane Prognostic Methods Group confirmed a need for a quality appraisal tool for SRs of PF studies, and feedback was generally positive for the included items and AMSTAR-PF as a whole. Where it was not, changes were made to improve the tool. Three pilot trials to test agreeability and usability, involving twenty-five independent testers with diverse backgrounds and research expertise, contributed to the final version.

Consideration of the results of our testing of AMSTAR-PF should recognise the participants in our pilot trials may not be representative of all researchers. In particular, most were not

experienced in PF research, nor were they always knowledgeable about the topic area of the sample reviews. We consider that this holds ecological validity – critical appraisal of SRs may often be undertaken by graduate students who are new to both prognostic research and the health condition of interest. The agreeability scores obtained in our testing showed 'slight' to 'substantial' inter-rater and inter-pair agreement and 'fair' to 'almost perfect' intra-pair agreement, which is similar or better than other published RoB/quality appraisal tools (38). The fact that intra-pair agreement scores were routinely higher than interrater and inter-pair scores may point to a benefit of discussing interpretation and understanding of the tool prior to using it, or after a small number of initial appraisals, in order to better operationalise the application of the AMSTAR-PF questions within appraisers' specific areas. Additionally, the guidance notes were modified to provide more thorough guidance following results and feedback from the testing, which may have enhanced agreeability and useability of the final version presented here.

We faced several challenges in developing this tool. We attempted to canvas a wide range of opinions and feedback on AMSTAR-PF, however participation rates from those invited to review it were sometimes low. This is highlighted in the survey to the CPMG, where the response rate was around 10% (assuming all emails were still monitored, and not duplicates). This is a low participation rate, although within published ranges for emailed surveys (50, 51). The low response may reflect the time commitment needed to comprehensively review and provide feedback on the document, as well as difficulties with recruiting busy academics, and a lack of incentive for completing it. Many of the concepts in appraising PF reviews are complex, and determining whether something has been addressed adequately is inherently subjective. This is a recognised challenge with all such tools, and the final version of AMSTAR-PF represents answering options and decision-making guidance that we deemed to balance the potential benefits of prescriptive, standardised scoring with the need to allow appraisers' judgements and expertise to influence the appraisal process. We have sought to provide additional guidance in coming to judgements, however acknowledge that our guidance is general and may not be detailed enough for all situations; it is recommended that any review team includes members with appropriate methodological and topical knowledge. In certain situations, appraisers may feel that it is beneficial to modify or amend certain elements to better suit their purposes. This comes with risks, however, so we recommend appraisers carefully consider the potential drawbacks of making alterations, particularly as non-standard tools present difficulties in interpretation and comparability. When modifications are deemed necessary, clear documentation of what changes were made and why they were made will be critical.

Conclusion

AMSTAR-PF, A MeaSurement Tool to Assess systematic Reviews of Prognostic Factor studies, represents the first quality appraisal tool specifically developed for SRs of PFs. We undertook a rigorous iterative process with field experts to develop both the tool and the guidance notes. We obtained input from a wide group of experienced prognosis researchers, and we tested the tool for agreement, usability, and acceptability in three separate cohorts of largely inexperienced researchers. We present the final tool to the field for immediate application.

References

1. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
2. Moons KG, Hooft L, Williams K, Hayden JA, Damen JA, Riley RD. Implementing systematic reviews of prognosis studies in Cochrane. Cochrane Database Syst Rev. 2018;10:ED000129.
3. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595.
4. Riley RD, Hayden JA, Steyerberg EW, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10:e1001380.
5. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10:e1001381.
6. Hingorani AD, Windt DA, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ. 2013;346:e5793.
7. Riley RD, Moons KGM, Snell KIE, et al. A guide to systematic review and meta-analysis of prognostic factor studies. BMJ. 2019;364:k4597.
8. Riley RD, van der Windt DA, Croft P, Moons KGM. Prognosis research in healthcare: concepts, methods, and impact. Oxford: Oxford University Press; 2019. 376 p.
9. Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. Diagn Progn Res. 2019;3:13.
10. Hoffmann F, Allers K, Rombey T, et al. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol. 2021;138:1-11.
11. Ioannidis JPA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q. 2016;94:485-514.
12. Page MJ, Shamseer L, Altman DG, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med. 2016;13:e1002028.
13. Almeida MO, Yamato TP, Parreira PdCS, Costa LOP, Kamper S, Saragiotto BT. Overall confidence in the results of systematic reviews on exercise therapy for chronic low back pain: a cross-sectional analysis using the Assessing the Methodological Quality of Systematic Reviews (AMSTAR) 2 tool. Braz J Phys Ther. 2020;24:103-17.
14. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med. 2006;144:427-37.
15. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158:280-6.
16. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.
17. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10.
18. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.
19. Higgins JPT, Morgan RL, Rooney AA, et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). Environ Int. 2024;186:108602.
20. Whiting P, Savovic J, Higgins JP, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225-34.
21. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170:51-8.
22. Moons KGM, Wolff RF, Riley RD, 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.

23. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.
24. Moons KG, Altman DG, Reitsma JB, 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.
25. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015;13:1-10.
26. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
27. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11:e1001744.
28. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529-36.
29. Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ. 2015;350:h870.
30. Foroutan F, Guyatt G, Zuk V, et al. GRADE Guidelines 28: Use of GRADE for the assessment of evidence about prognostic factors: rating certainty in identification of groups of patients with different absolute risks. J Clin Epidemiol. 2020;121:62-70.
31. Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. Syst Rev. 2017;6:204.
32. Gates M, Gates A, Duarte G, et al. Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. J Clin Epidemiol. 2020;125:9-15.
33. Lee SWH. What tool do undergraduate pharmacy students prefer when grading systematic review evidence: AMSTAR-2 or ROBIS? Cochrane Evid Synth Methods. 2023;1:e12023.
34. Pieper D, Puljak L, González-Lorenzo M, Minozzi S. Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies. J Clin Epidemiol. 2019;108:26-33.
35. Perry R, Whitmarsh A, Leach V, Davies P. A comparison of two assessment tools used in overviews of systematic reviews: ROBIS versus AMSTAR-2. Syst Rev. 2021;10:273.
36. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019;95:103208.
37. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42:377-81.
38. Henry M, O'Connell N, Riley R, et al. Agreeability testing of AMSTAR-PF, a tool for quality appraisal of systematic reviews of prognostic factor studies. medRxiv [Preprint]. 2025 Apr 14: 2025.04.10.25325555. Available from: https://doi.org/10.1101/2025.04.10.25325555.
39. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61:29-48.
40. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. 4th ed. Gaithersburg, MD: Advanced Analytics, LLC; 2014. 428 p.
41. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.
42. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7:e32844.

43. Stallings E, Gaetano-Gil A, Alvarez-Diaz N, et al. Development and evaluation of a search filter to identify prognostic factor studies in Ovid MEDLINE. BMC Med Res Methodol. 2022;22:107.
44. Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from: https://gdt.gradepro.org/app/handbook/handbook.html.
45. Huguet A, Hayden JA, Stinson J, et al. Judging the quality of evidence in reviews of prognostic factor research: adapting the GRADE framework. Syst Rev. 2013;2:71.
46. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.
47. Page MJ, Sterne JAC, Boutron I, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. BMJ. 2023;383:e076754.
48. Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies [last updated August 2022]. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5. Cochrane; 2024. Available from: www.training.cochrane.org/handbook.
49. Kent P, Cancelliere C, Boyle E, Cassidy JD, Kongsted A. A conceptual framework for prognostic research. BMC Med Res Methodol. 2020;20:172.
50. Shih T-H, Fan X. Comparing response rates in e-mail and paper surveys: A meta-analysis. Educ Res Rev. 2009;4:26-40.
51. Wu M-J, Zhao K, Fils-Aime F. Response rates of online surveys in published research: A meta-analysis. Comput Hum Behav Rep. 2022;7:100206.

Contribution statement

MLH, NEO'C and GLM conceived the original idea. MLH, NEO'C, RDR, KGM, BJS, LH developed the initial versions of the tool and guidance notes. All authors were involved in subsequent revisions. MLH wrote the first draft of the manuscript. All authors critically revised the manuscript and approved the final version.

MLH is the guarantor of the article. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding

This work received no specific funding.

MLH was supported by an Australian Government Research Training Scholarship

RDR was supported by funding from the MRC Better Methods Better Research panel (grant reference: MR/V038168/1) and by the National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham.

GLM was supported by a Leadership Investigator Grant from the National Health & Medical Research Council of Australia to GLM (ID 1178444).

The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests

All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/disclosure-of-interest/ and declare: no support from any organisation for the submitted work.

Between 2020 and 2023 NEO'C was Coordinating Editor of the Cochrane Pain, Palliative and Supportive Care group, whose activities were funded by The UK National Institute of Health and Care Research (NIHR). He is the Chair of the International Association for the Study of Pain (IASP) Methodology, Evidence Synthesis and Implementation Special Interest Group and has held a grant from the ERA-NET Neuron Co-Fund.

RDR is an NIHR Senior Investigator. RDR receives royalties from two textbooks ('Prognosis Research in Healthcare' and 'IPD Meta-Analysis') and consultancy fees as a Statistical Editor for the BMJ.

GLM has received support, unrelated to the current work, from: Reality Health, ConnectHealth UK, Institutes of Health California, AIA Australia, Workers' Compensation Boards and professional sporting organisations in Australia, Europe, South and North America. Professional and scientific bodies have reimbursed him for travel costs related to presentation of research on pain and pain education at scientific conferences/symposia. He has received speaker fees for lectures on pain, pain education and rehabilitation. He receives royalties for books on pain and pain education. He is non-paid CEO of the non-profit Pain Revolution. There are no disclosures immediately relevant to this work;

There are no other relationships or activities that could appear to have influenced the submitted work.

Data

Data related to agreement testing can be found in the referenced paper, or on reasonable request from the corresponding author at neil.oconnell@brunel.ac.uk. The draft versions of the tool and guidance notes generated during the development process are available upon reasonable request, from the corresponding author.

Transparency

Transparency: The lead author (MLH) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and registered) have been explained.

Dissemination plans

There are no specific patient and public dissemination plans for this paper. The work is methodologically focussed and primarily intended for the research and clinical community, and did not involve patients or public during its development.

Provenance and peer review

Provenance and peer review: Not commissioned; externally peer reviewed.

Acknowledgements

The authors would like to thank all those who provided feedback and suggestions on earlier iterations of this work, as well as those who piloted the tool, including: Ruth Appiah, Carolyn Berryman, Chloe Blacket, Aidan Cashin, Sophie Crouch, Grace Ferencz, Michael Ferraro, Bethany Gower, Ashley Grant, Katherine Henry, Aleksandra Herman, Emma Karran, Indika Koralegedera, Hayley Leake, Erin MacIntyre, Brendan Mouatt, Karma Phuentsho, Chandan Poddar, Daniel Van Der Laan, Ellana Welsby, Louise Wiles, Erica Wilkinson, Marelle Wilson, Monique Wilson, and members of the Cochrane Prognostic Methods Group. The authors would also like to acknowledge the writers of AMSTAR and AMSTAR 2, whose comprehensive and user-friendly tool and guidance notes formed a basis for the current work.

ORCID

MLH: 0000-0003-2871-6695

NEO'C: 0000-0003-1989-4537

RDR: 0000-0001-8699-0735

KGMM: 0000-0003-2118-004X

LH: 0000-0002-7950-2980

JAAD: 0000-0001-7401-4593

NS: 0000-0003-4744-6192

Figure and Box Legends

Figure 1: Diagram of the steps to developing AMSTAR-PF.
We used feedback received at each step of the process to progressively re-draft the AMSTAR-PF tool and guidance notes.
CPMG; Cochrane Prognostic Methods Group.

Box 1: Summary of key differences between AMSTAR 2 (A-2) and AMSTAR-PF (A-PF), with select examples.
PICOTS, Population, Index prognostic factor, Comparator prognostic factors, Outcome, Timing, Setting (7); PICO, Patient, Population or Problem, Intervention, Comparison, Outcome; PF, Prognostic Factor; Q, Question; QUIPS, Quality In Prognosis Studies (15); RoB, Risk of Bias