Computer Methods and Programs in Biomedicine Update 7 (2025) 100161

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

Validating and updating GRASP: An evidence-based framework for grading and assessment of clinical predictive tools

Mohamed Khalifa [a,b,c,*], Farah Magrabi [b], Blanca Gallego [b,d]

[a] College of Health Sciences, Education Centre of Australia, Sydney, Australia
[b] Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[c] School of Population Health, La Trobe University, Melbourne, Australia
[d] Centre for Big Data Research in Health, Faculty of Medicine, University of New South Wales, Sydney, Australia

ARTICLE INFO

Keywords:
Evidence-based medicine
Clinical prediction
Clinical decision support
Grading and assessment
Validation

ABSTRACT

Background: When selecting clinical predictive tools, clinicians are challenged with an overwhelming and ever-growing number, most of which have never been implemented or evaluated for effectiveness. The authors developed an evidence-based framework for grading and assessment of predictive tools (GRASP). The objective of this study is to refine, validate GRASP, and assess its reliability for consistent application.
Methods: A mixed-methods study was conducted, involving an initial web-based survey for feedback from a wide group of international experts in clinical prediction to refine the GRASP framework, followed by reliability testing with two independent researchers assessing eight predictive tools. The survey involved 81 experts who rated agreement with the framework's criteria on a five-point Likert scale and provided qualitative feedback. The reliability of the GRASP framework was evaluated through interrater reliability testing using Spearman's rank correlation coefficient.
Results: The survey yielded strong agreement of the experts with the framework's evaluation criteria, overall average score: 4.35/5, highlighting the importance of predictive performance, usability, potential effect, and post-implementation impact in grading clinical predictive tools. Qualitative feedback led to significant refinements, including detailed categorisation of evidence levels and clearer representation of evaluation criteria. Interrater reliability testing showed high agreement between researchers and authors (0.994) and among researchers (0.988), indicating strong consistency in tool grading.
Conclusion: The GRASP framework provides a high-level, evidence-based, and comprehensive, yet simple and feasible, approach to evaluate, compare, and select the best clinical predictive tools, with strong expert agreement and high interrater reliability. It assists clinicians in selecting effective tools by grading them on the level of validation of predictive performance before implementation, usability and potential effect during planning for implementation, and post-implementation impact on healthcare processes and clinical outcomes. Future studies should focus on the framework's application in clinical settings and its impact on decision-making and guideline development.

Background

Clinical predictive tools are research-based applications designed to provide clinicians and other healthcare professionals with diagnostic, prognostic, and therapeutic decision support by predicting clinical and other relevant healthcare outcomes [1]. They quantify the contributions of relevant patient characteristics to derive the likelihood of diseases, predict their courses and possible outcomes, or support the decision-making on their management [2]. For example, the Centor Score assesses strep throat likelihood; CHALICE Rule identifies intracranial injury risk in children; Dietrich Rule evaluates appendicitis; LACE Index predicts 30-day readmission or death risk; Manuck Scoring System predicts preterm birth risk; Ottawa Knee Rule determines the need for knee X-rays; PECARN Rule assesses CT scan needs in children [3–10]. Similarly, integrating Internet of Things and Blockchain technologies in healthcare could enhance clinical prediction and decision-making [11]. However, studies discuss that there is an inappropriate but common practice of developing new predictive tools

* Corresponding author at: 1/3 Fitzwilliam St, Parramatta, NSW, 2150, Australia.
E-mail addresses: mohamed.khalifa@chs.edu.au (M. Khalifa), farah.magrabi@mq.edu.au (F. Magrabi), b.gallego@unsw.edu.au (B. Gallego).

https://doi.org/10.1016/j.cmpbup.2024.100161

Available online 20 August 2024
2666-9900/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

instead of validating, updating, or implementing existing ones [12–15]. Facing such an overwhelming and ever-growing number of predictive tools represents a major challenge for clinicians when selecting predictive tools for implementation in clinical practice or for recommendation in clinical guidelines. Moreover, while a few pre-implementation studies compare predictive tools along predictive performance measures, many of these tools have never been implemented or assessed for comparative effectiveness or impact [13–17]. For example, there are 14 tools designed for predicting head injury in children. Only one of them, the PECARN Rule, proved post-implementation effectiveness to minimise missing paediatric traumatic brain injuries in clinical settings [17]. Furthermore, clinicians often rely on personal judgment and anecdotal experience for selecting predictive tools, which can result in inconsistent and unreliable choices that overlook critical factors like predictive performance, usability, and post-implementation impact. This potentially can lead to the adoption of less effective tools in clinical practice. Although some clinicians seek high-quality evidence, such as systematic reviews, the lack of standardized, objective methods to interpret this evidence complicates the identification of the most suitable tools amidst implementation challenges and clinical constraints [13–15,18].

To overcome the challenge of evaluating numerous predictive tools, the authors have developed and published the GRASP framework (Grading and Assessment of Predictive Tools), an innovative, evidence-based system designed to assist clinicians in the evaluation and selection of predictive tools. The framework employs a three-dimensional approach: Phase of Evaluation, Level of Evidence, and Direction of Evidence [17,19]. The Phase of Evaluation categorizes predictive tools based on their stage of development and application, from testing for validity (Phase C), through usability and potential effect testing (Phase B), to final post-implementation effectiveness in clinical practice (Phase A). Within each phase, tools are further graded based on the robustness of evidence supporting their use, with grades ranging from C1 (positive internal validity) to A1 (high-quality experimental studies supporting positive impact post-implementation). The Direction of Evidence assesses the overall positivity or negativity of conclusions drawn from studies on the tool, considering the quality of evidence and how closely studies match the tool's intended use and specifications [17,19]. The GRASP framework assigns letter grades (A, B, C) based on the evaluation phase and numerical scores for evidence level (e.g., C1 for multiple external validity tests), and assesses evidence direction (positive, negative, mixed). For instance, using the GRASP framework, the Ottawa Knee Rule, graded A1, showed positive post-implementation impact, while the LACE Index, graded C1, demonstrated only pre-implementation predictive performance on multiple external validation studies [19]. Moreover, the GRASP framework has been successfully used to evaluate and grade 14 predictive tools for paediatric head injury [17]. This standardized, evidence-based approach helps clinicians select effective tools by critically appraising published evidence, ensuring informed decisions in clinical practice. The GRASP framework's potential impact on clinical decision-making, resource allocation, and guideline development is significant, as it provides a rigorous, transparent method for evaluating tools, ultimately enhancing patient outcomes, optimizing resource use, and supporting the development of evidence-based clinical guidelines [17,19]. In a randomised controlled trial published earlier, the GRASP framework has positively supported and significantly improved evidence-based decision making. It has increased the accuracy and efficiency of selecting predictive tools by expert clinicians [20]. Fig. 1 shows the GRASP framework concept. Table S1 in the Appendix shows the initial GRASP framework detailed report used to assess predictive tools and assign them grades.

This study aims to refine the GRASP framework's evaluation criteria via insights from an extensive panel of clinical prediction international experts, from diverse clinical and prediction backgrounds, and to assess the framework's applicability, verifying the consistency and reliability of outcomes when independent users apply it to grade predictive tools.

Methods

A comprehensive approach of four steps was followed to validate and update the GRASP framework and assess its reliability.

The study design

This foundational phase was divided into two parts. Initially, the

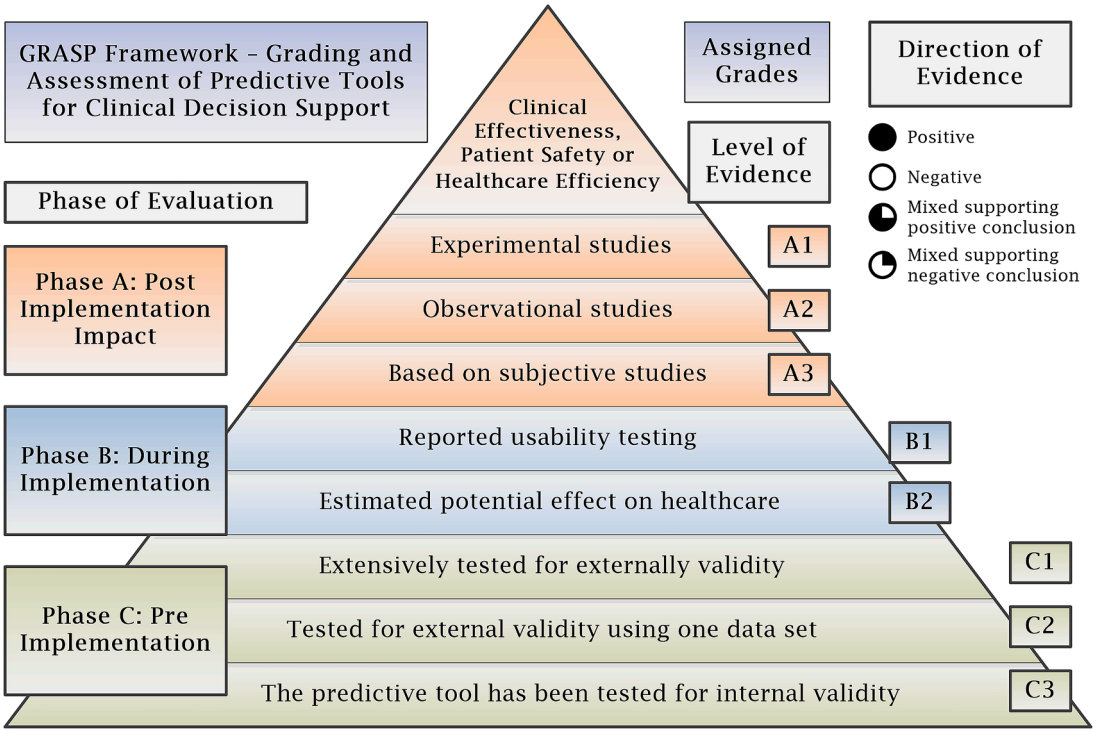


Fig. 1. The published GRASP framework concept initial design [17,19].

effort focused on validating and refining the GRASP framework through expert feedback, collected via a web-based survey. The survey intended to gauge expert agreement with the framework's criteria on a five-point Likert scale, from "Strongly Agree" to "Strongly Disagree", spanning evaluation phases, evidence levels, and evidence directions [21]. Additionally, qualitative feedback was sought for potential criteria modifications. Fifty experts were required as a sample to participate in this process based on similar studies [22–24]. The study design also involved pilot testing to refine the survey before its broader distribution, with the entire process receiving ethical approval from Macquarie University's Human Research Ethics Committee. The detailed survey screenshots are included in the Appendix.

Experts identification

To identify clinical prediction experts, a comprehensive literature search was conducted focusing on recent publications related to the development and evaluation of clinical predictive tools. Experts were defined as researchers who had published at least one paper on these topics in the past five years. The search strategy included prominent databases such as PubMed, Scopus, Web of Science, Embase, and Google Scholar, using keywords like "clinical predictive tools," "development," and "evaluation." This search resulted in 1186 relevant publications. From these publications, 882 unique authors were identified. The inclusion criteria were based on the authors' involvement in developing or evaluating clinical predictive tools. Authors' contact information was extracted, and they were invited to participate in the study via email.

The study survey

The survey, developed on the Qualtrics platform [25], featured eight Likert scale statements and six open-ended questions over seven sections, addressing the GRASP framework for predictive tools assessment. It covered the tools' predictive performance, usability, and post-implementation impact, alongside evaluating evidence direction. Participants provided feedback on the framework's criteria, suggesting additions, removals, or changes, and discussed methods to gauge and capture tools' effectiveness and handling conflicting evidence, aiming to refine and enhance the framework's applicability and accuracy.

Reliability testing

In the second part of the study, the reliability of the GRASP framework was assessed using the validated and updated version. Two independent researchers with PhDs in health-related disciplines and extensive systematic review experience were trained to grade eight diverse predictive tools. Their task was to independently assess these tools using GRASP. This process aimed to test the framework's interrater reliability, or the consistency of tools grading between independent users. The chosen tools represented a wide spectrum of the GRASP framework's grades [26]. The Spearman's rank correlation coefficient was utilised to measure this reliability, focusing on the ordinal nature of the framework's ratings [27]. Following the grading, researchers provided feedback on the framework's design, usability, and criteria through a brief survey, contributing further to the framework's refinement and practical application.

Analysis and outcomes

The study aimed for three key outcomes. First, quantitatively it sought to refine the GRASP framework criteria based on average expert agreement scores from the survey [28]. Second, qualitatively it planned to incorporate expert feedback into framework updates, using content analysis to identify and categorise suggestions using the NVivo Version 12.3 software [29,30]. Lastly, the study aimed to validate the framework's interrater reliability, comparing the independent assessments of predictive tools to ensure consistent and accurate grading [26]. This multifaceted approach ensures a thorough validation and refinement process, aiming to enhance the GRASP framework's utility and reliability in evaluating clinical predictive tools.

Results

The literature search yielded 1186 relevant publications, from which 882 unique authors were identified, and their emails were extracted. All identified authors were contacted through emails and 81 valid responses were received from international experts. Valid responses were defined as those completing all survey sections and answering all survey questions. Invalid responses were excluded, where participants started but did not complete the survey or omitted some sections or questions.

Experts agreement on GRASP criteria

On average, the 81 experts strongly agreed with the eight closed-ended statements, regarding the evaluation criteria of the GRASP framework, showing an average of 4.35 on a five-points Likert scale. Experts strongly agreed with six of the eight closed-ended agreement statements. They somewhat agreed with one, and were neutral about another, of the eight closed-ended statements. Table 1 shows the

Table 1
Average scores, standard deviations, and confidence intervals of expert experts agreement with the GRASP framework evaluation criteria.

SN	Question	Mean score	Meaning	SD	95 % CI
1	Predictive performance: We should consider the evidence on validating the tool's predictive performance.	4.88	Strongly Agree	0.43	[4.87, 488]
2	Evidence levels on predictive performance: The evidence level could be High (internal + multiple external validation), Medium (internal + external validation once), or Low (internal validation only).	4.44	Strongly Agree	0.87	[4.44, 4.45]
3	Usability: We should consider the evidence on the tool's usability.	4.68	Strongly Agree	0.70	[4.67, 4.68]
4	Potential effect: We should consider the evidence on the tool's potential effect.	4.62	Strongly Agree	0.68	[4.61, 4.62]
5	Usability is higher: The evidence level on tools' usability should be considered higher than the evidence level on tools' potential effect.	2.96	Neither Agree nor Disagree	1.23	[2.95, 2.97]
6	Impact: We should consider the evidence on the tool's impact on healthcare effectiveness, efficiency, or safety.	4.78	Strongly Agree	0.57	[4.78, 4.79]
7	Evidence levels of post-implementation impact: The evidence level could be High (based on experimental studies), Medium (observational studies), or Low (subjective studies).	4.18	Somewhat Agree	1.14	[4.17, 4.19]
8	Evidence direction: Based on the conclusions of published studies, the overall evidence direction could be Positive, Negative or Mixed.	4.25	Strongly Agree	0.78	[4.25, 4.26]
Overall average		**4.35**	**Strongly Agree**	**1.01**	**[4.35, 4.35]**

average agreement score, standard deviation, and the 95 % confidence interval of the experts on each of the eight statements. The 81 experts were from 30 countries, and half of them were from United States, United Kingdom, and Canada. The detailed country distributions of the experts are shown in Table S4 and Figure S1 in the Appendix.

Experts comments, suggestions, and recommendations

A comprehensive analysis of the qualitative feedback of 64 out of 81 experts revealed insights into the GRASP framework's evaluation criteria. Experts provided valuable suggestions on improving the framework, especially in adding, revising, or removing certain criteria based on their expertise. The feedback emphasised the importance of aligning the GRASP criteria with practical and clinical needs, suggesting both positive and negative modifications. Each suggestion was accurately scored for its significance, leading to an aggregate understanding of the most critical areas for enhancement in the GRASP framework.

Predictive performance and performance levels

Most experts who provided qualitative feedback (59/64) advocated for detailed reporting on the validation studies' methodologies, qualities, and types within the GRASP framework. A tool's reliability was linked to its broad validation across diverse healthcare settings and populations, with stability and reliability underscored by consistent predictive performances across multiple external validations. The introduction of a "Strength of Evidence" component was suggested to aid users in selecting tools based on evidence quality and predictive performance, thus facilitating more informed decisions in clinical settings. For example, if two predictive tools were assigned grade C1 (each was externally validated multiple times) but one of them shows strong positive evidence and the other shows a medium or weak positive evidence. It is logic to select the tool with the stronger evidence if both have similar predictive performances for the same tasks. Moreover, experts suggested adding one level below C1, when the internal validity of predictive tools are either not tested or the tools show poor internal validity results.

Usability and potential effect

The feedback highlighted the necessity of reporting on usability and potential effect studies within the GRASP framework, with a particular emphasis on the importance of potential effects on healthcare outcomes over usability. The experts argued for a higher evidence level for potential effect, suggesting that a tool's potential effect on healthcare is paramount, regardless of its usability. The integration of both potential effect and usability was seen as essential for evaluating a tool's overall utility. Accordingly, experts suggested that positive evidence of potential effect the usability together should rank the predictive tool a higher grade than positive evidence on only one of them.

Post-implementation impact and impact levels

Experts recommended detailed reporting on post-implementation impact studies, suggesting the inclusion of a "Strength of Evidence" metric to differentiate between the qualities of observational and experimental studies. This differentiation would help clarify the evidence's reliability and applicability, thereby enhancing the framework's utility in assessing predictive tools' post-implementation impacts.

Direction of evidence

The quality and strength of evidence were highlighted as crucial factors in determining the direction of evidence, especially when faced with conflicting study outcomes. Experts called for a nuanced approach that considers the methodology, population, setting, and other quality metrics of each study to accurately gauge the evidence's direction, ensuring that decisions are based on robust and high-quality evidence. Figure S2 in the Appendix shows the strength of evidence protocol, which is used to decide on the direction of evidence for each level.

Defining and capturing predictive performance

Experts noted that predictive performance evaluations should be tailored to the specific predictive task at hand, emphasising the need for tools to be adjusted according to the clinical conditions, cost-effectiveness, and intended actions based on the tool's outcomes. The distinction between screening and diagnostic tools was underscored, with a call for sensitivity, specificity, and probability/risk estimation to be appropriately applied based on the tool's intended use.

Managing conflicting evidence

Addressing conflicting evidence requires a focus on study quality and evidence strength, with high-quality studies being given precedence in determining the overall evidence direction. The variability in evidence highlights the need for detailed reporting within the GRASP framework, enabling users to make informed decisions based on their specific clinical settings and needs. Figure S3 in the Appendix shows the detailed mixed evidence protocol which is used to sort out and manage conflicting evidence.

Updating the GRASP framework

Incorporating the expert feedback, the GRASP framework was updated to better reflect the nuances of predictive performance evaluation, usability, potential effect, and post-implementation impact. The revisions included more detailed categorisations of evidence levels, a clearer representation of the framework's criteria, and suggestions for an enhanced protocol for assessing the strength of evidence. These updates aim to make the GRASP framework more comprehensive, user-friendly, and applicable to a wide range of clinical predictive tools. For more clarity, the experts recommended that the three levels of internal validation, external validation once, and external validation multiple times, are additionally assigned "Low Evidence", "Medium Evidence", and "High Evidence" labels for predictive performance respectively. Likewise, a fourth level of C0, labelled "No Evidence", is added to reflect that internal validity was either not tested or the tool showed poor internal validity results. Moreover, Phase B: "During Implementation" has been renamed to "Planning for Implementation" to reflect that usability and potential effect should be testing while planning for implementing clinical predictive tools. Furthermore, the Potential Effect is moved to a higher evidence level than Usability and the evidence of both together is higher than any one of them alone, creating three levels, instead of two, within Phase B. Fig. 2 shows the validated and updated GRASP framework concept. Table S2 in the Appendix shows the updated GRASP framework detailed report, adding new information elements and updated levels. Similarly, Table S3 in the Appendix shows the summary of evidence, on each considered published study.

The GRASP framework reliability

Using the updated GRASP framework, two independent researchers evaluated eight predictive tools producing a detailed report on each and assigning grades compared to the authors, as summarised in Table 2. More information on the details of the assigned grades is shown in the Appendix in Table S5. The Spearman's rank correlation coefficients showed a high degree of agreement between researchers and authors (0.994 for both) and between the researchers themselves (0.988), indicating a strong, statistically significant interrater reliability of the GRASP framework. The researchers' feedback, captured through five

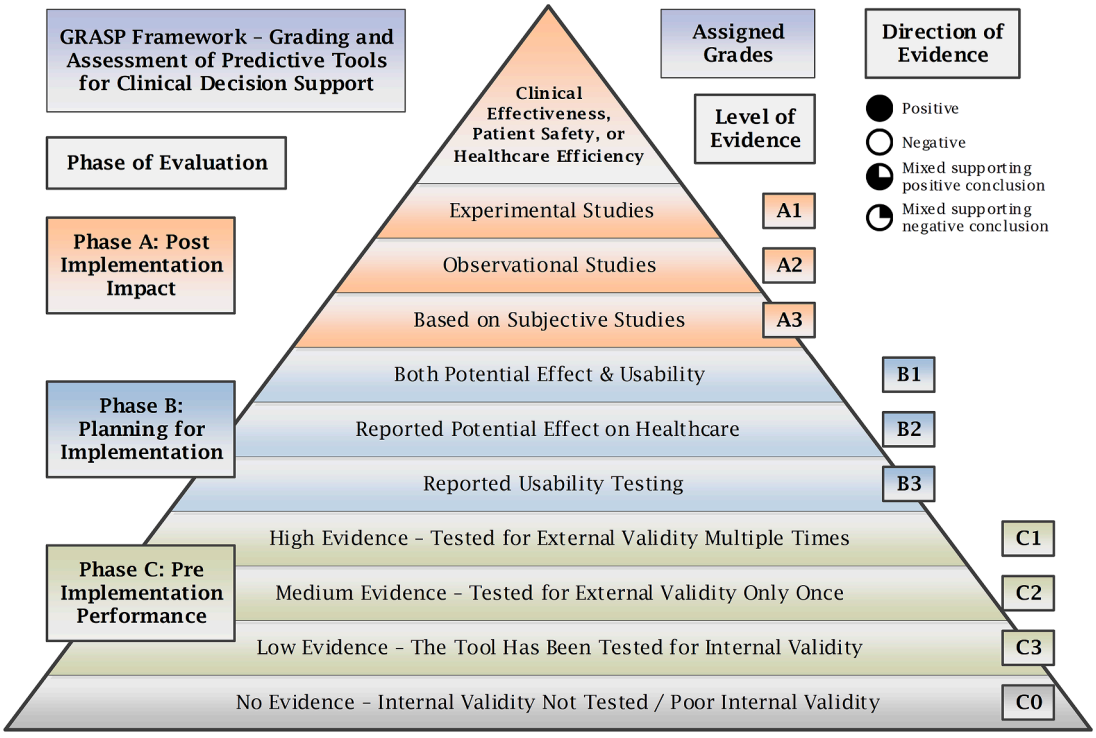


Fig. 2. The validated and updated GRASP framework concept.

Table 2
Grades assigned by the two independent researchers and the authors.

Tools	Grading by researcher 1	Grading by researcher 2	Grading by authors
Centor score [10]	B2	B3	B3
CHALICE rule [9]	B2	B2	B2
Dietrich rule [8]	C0	C0	C0
LACE index [7]	C1	C1	C1
Manuck scoring system [6]	C2	C2	C2
Ottawa knee rule [5]	A1	A2	A1
PECARN Rule [4]	A2	A2	A2
Taylor mortality model [3]	C3	C3	C3

open-ended questions after grading the tools, revealed a unanimous appreciation for the GRASP framework's logical design, clarity, and ease of use. They praised its utility in assessing varying tool qualities and evidence levels. While content with the criteria used for grading, they suggested adding definitions and clarifications to the evaluation criteria, which was incorporated into the framework's update. This feedback underscores the GRASP framework's reliability and user-friendly design.

Discussion

Brief summary

The GRASP framework addresses the challenge clinicians face in evaluating the increasing number of predictive tools for clinical practice and guidelines. Designed to offer an evidence-based, standardised method for assessing these tools, GRASP aids in selecting effective clinical predictive tools which show evidence of high predictive performance, good usability, promising potential effects, and successful implementations. The GRASP framework, went through a journey of initial design, comprehensive development and testing, and implementation and testing and finally extensive validation and updating by a wide group of international experts of clinical prediction [17,19]. This journey improved its concepts, criteria, and reports, with subsequent interrater reliability testing confirming its reliability and consistency for grading predictive tools by independent users.

Predictive performance

Internal validation of a predictive tool's performance is crucial to ensure it predicts accurately as intended, focusing on measures of discrimination and calibration [31,32]. Discrimination refers to the tool's ability to differentiate between patients with and without the outcome, assessed through sensitivity, specificity, and the area under the curve (AUC) [33]. Calibration evaluates the accuracy of predictions against observed outcomes, typically measured by the Hosmer-Lemeshow test or the Brier score [34]. External validation is critical for assessing a tool's reliability and generalizability, with its trustworthiness enhanced by high-quality, extensive external validation across various patient populations and settings [35].

Usability and potential effect

Clinicians value the potential of predictive tools to enhance patient outcomes, efficiency, and safety, focusing on their impact on healthcare processes once implemented [36]. The adoption and success of a Clinical Decision Support (CDS) tool hinge on its ability to improve healthcare processes or clinical outcomes [37]. Usability is also critical; tools must meet specific user objectives within their context [38], as poor usability, and poor integration with other clinical information resources, can lead to failure regardless of performance or potential healthcare benefits [39]. Usability criteria encompass mental effort, user attitude, interaction, ease of use, acceptability, task management effectiveness, resource efficiency, and user satisfaction, including learnability, memorability, and error minimization [40,41].

Post-implementation impact

Clinicians prioritise understanding the post-implementation impact of CDS tools on healthcare aspects, processes, and outcomes. They are particularly interested in the effect size of CDS tools on physicians' performance and patient outcomes [42]. High-quality experimental studies, like randomised controlled trials, are regarded as the most reliable evidence, followed by well-designed observational studies, and finally subjective studies and expert opinions. While experimental methods are traditionally viewed as superior, the importance of high-quality observational studies in comparative effectiveness research is increasingly recognised for their ability to address challenges that experimental methods cannot, highlighting the need for a balanced appreciation of both approaches [43].

Direction of evidence and conflicting conclusions

Encountering conflicting conclusions in the validation and evaluation of predictive tools across different subpopulations or outcomes is common. The determination of what constitutes good predictive performance varies significantly, depending on the clinical condition and the decisions that follow [44]. Systematic reviews and meta-analyses are feasible for homogeneous predictive tools aimed at the same outcomes within similar subpopulations. However, establishing standards for performance and impact across diverse tools and populations presents significant challenges due to the variability in study quality, types, and conditions, complicating the synthesis of data into concise quantitative measures [45].

The GRASP framework overall

The GRASP framework guides the selection of predictive tools for CDS by providing evidence-based grades, but it does not rigidly dictate choices. The preference between tools, such as choosing an A2 tool over an A1 based on patient safety versus cost reduction, depends on the clinician's objectives and priorities. Multiple tools may be recommended in clinical guidelines, each for its unique benefits in predictive performance, potential effect, or post-implementation effects. The GRASP framework serves as a comprehensive yet straightforward method for clinicians to evaluate and select predictive tools, supplemented by detailed reports for in-depth information. It is not intended for direct daily use by clinicians in clinical settings. Instead, it is designed to assist expert clinicians in evaluating and grading predictive tools, which can then be recommended in clinical guidelines. These expert evaluations provide end users, such as clinicians, with the necessary information to select suitable tools for their practice or guidelines, ultimately aiding in the implementation of these tools in their daily work. The GRASP framework ensures objectivity and consistency in evidence grading by incorporating standardized criteria and detailed protocols, such as the Strength of Evidence Protocol and the Mixed Evidence Protocol detailed in the Appendix. These protocols support mitigating variability in study quality and reporting standards, thereby ensuring reliable and consistent tool assessment.

Other methods, approaches, and frameworks

Various frameworks focus on the development, validation, and implementation of clinical predictive tools, with some specialising in performance evaluation and external validation. Steyerberg and Debray have proposed frameworks concentrating on performance measures and external validation interpretation, respectively, while Collins emphasizes on reporting external validation outcomes [31,35,46–50]. The TRIPOD statement and CHARMS checklist aim to enhance reporting standards and critical appraisal of predictive tools [51–53]. Additionally, frameworks by Wallace and Harris, along with Toll's approach, assess post-implementation impacts but lack a unified grading system for tool comparison [54–57]. In contrast, the GRADE framework offers a systematic method to grade evidence quality and recommendation strength [58]. The GRASP framework, compared to TRIPOD, CHARMS, and GRADE, excels in comprehensive evaluation, usability, and post-implementation impact. It uniquely offers a three-dimensional approach, emphasizing practical implementation and real-world effectiveness, making it superior for selecting and grading clinical predictive tools for guidelines and daily practice.

Challenges, limitations, and future work

Analysing expert feedback through open-ended questions presents significant challenges due to the diversity of opinions and experiences, making qualitative content and thematic analysis difficult [59]. The Delphi technique, recommended for developing clinical guidelines and selecting evaluation criteria, typically involves a panel of ten to fifty members to manage the volume of data and analysis effectively [60]. However, limitations in time and resources led to the decision to use a single-round web-based survey for expert feedback, despite anticipating responses from nearly a hundred participants. Out of 882 invited experts only 81 provided valid responses, with a low response rate of 9.2 %. Factors such as lack of incentives and inadequate support from participants' organizations could have contributed to this low response [61]. The survey was designed to be feasible for busy experts, limiting both the number of questions and the completion time to about 20 min, although this constraint may have restricted the depth of feedback obtainable. Evaluating the quality of evidence, a fundamental aspect of the GRASP framework for grading predictive tools, is a challenge. The GRASP framework incorporates various tools for evidence quality assessment, but it could benefit from integrating additional methods, frameworks, and guidelines such as those by Debray, Steyerberg, Collins, the TRIPOD statement, CHARMS checklist, Wallace, Harris, Toll, and the GRADE guidelines to ensure evidence robustness and comprehensiveness. Future research should aim to assess the GRASP framework's impact on clinician decision-making and its application in grading predictive tools in different clinical settings. It is essential to evaluate the framework's effect on improving clinical decision-making. Furthermore, research should apply the GRASP framework to evaluate clinical predictive tools and publish such results to start enhancing evidence-based culture in the field of CDS and predictive tools. To achieve high interrater reliability in broader settings, thorough training on GRASP's structured criteria and protocols is essential. Additionally, the complexity of the GRASP framework necessitates a detailed instruction manual to guide users in accurately and consistently grading and assessing predictive tools. Continuous support and detailed guidelines are necessary to ensure consistent application and understanding, enhancing replicability and maintaining reliability among diverse users. The GRASP framework will evolve through regular feedback from users and experts, periodic reviews, and integration of new evidence and advancements. A dedicated feedback loop, including surveys and expert panels, should ensure the framework remains current, relevant, and effective. This should help to make the framework adaptable to different clinical contexts, specialties, and healthcare settings and can be scaled up for use in large multi-centre studies.

Conclusion

The GRASP framework is an evidence-based approach designed to help clinicians evaluate clinical predictive tools effectively, focusing on predictive performance, potential effect, usability, and post-implementation outcomes. It is suggested to create a web-based platform for clinicians and guideline developers, providing access to detailed information and evidence grades of predictive tools. Updating this system with new evidence poses challenges, emphasising the need for automated methods to maintain current assessments. Expert groups from professional organisations are recommended to grade clinical

predictive tools, ensuring consistency, reliability, and credibility. These organisations should also aid in disseminating evidence-based information, akin to updates on clinical practice guidelines. The GRASP framework will remain in a dynamic and continuous process of development, requiring further validation and refinement based on real-world usage and feedback to confirm its validity and reliability further. Accordingly, future studies are essential to enhance the framework, with its effectiveness dependent on ongoing application and evaluation by expert and end users.

Ethics approval and consent to participate

This study has been approved by the Human Research Ethics Committee, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, Australia, on the 4th of October 2018. Reference No: 5201834324569. Project ID: 3432. Consent to participate was obtained from participants through their agreement to participate in the study, based on the study description and objectives as illustrated in the invitation email sent to them.

Declaration on the use of AI in the writing process

The authors of this manuscript declare that in the writing process of this work, no generative artificial intelligence (AI) or AI-assisted technologies were used to generate content, ideas, or theories. We utilized AI solely for the purpose of enhancing readability and refining language. This use was under strict human oversight and control. After the application of AI technologies, the authors carefully reviewed and edited the manuscript to ensure its accuracy and coherence. The authors understand the potential of AI to generate content that may sound authoritative yet might be incorrect, incomplete, or biased. Considering this, the authors ensured that the manuscript was thoroughly revised by human eyes and judgment. In line with Elsevier's Authorship Policy, the authors confirm that no AI or AI-assisted technologies have been listed as an author or co-author of this manuscript. The authors fully comprehend that authorship comes with responsibilities and tasks that can only be attributed to and performed by humans, and the authors have adhered to these guidelines in the preparation of this manuscript.

CRediT authorship contribution statement

Mohamed Khalifa: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Farah Magrabi:** Writing – review & editing, Validation, Methodology, Investigation. **Blanca Gallego:** Writing – review & editing, Methodology, Investigation.

Declaration of competing interest

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

Acknowledgements

We would like to acknowledge the contributions and thank all the professors, doctors, and researchers who participated in the validation of the GRASP framework including; Abdullah Pandor, Adam Dunn, Alberto Zamora Cervantes, Alex C Spyropoulos, Allyson R Cochran, Alyson Mahar, Anders Granholm, Andrew D MacCormick, Anupam Kharbanda, Ashraf El-Metwally, Beth Devine, Brian Shirts, Carme Carrion, Carrie Ritchie, Cesar Garriga, Christoph U Lehmann, Claudia Gasparini, Claudia Pagliari, Craig Anderson, Douglas P. Gross, Dustin Ballard, Erik Roelofs, Ewout W. Steyerberg, Fabian Jaimes, Felix Zubia-Olaskoaga, Fernando Ferrero, Gary Collins, Gary Maartens, Grégoire Le Gal, Ilkka Kunnamo, Janneke Stalenhoef, Jitendra Jonnagaddala, Julian Brunner, Kent P. Hymel, Kristen Miller, Laura Cowley, Liliana Laranjo da Silva, Luke Daines, Manish Kharche, Maria Lourdes Posadas-Martinez, Mark Ebell, Maryati Mohd. Yusof, Matthias Döring, Matthijs Becker, Maxwell Dalaba, Michael T Weaver, Michelle Ng Gong, Mohamed Hassan Ahmed Fouad, Mowafa Househ, NadÃge Lemeunier, Natalie Edelman, Nathan Dean, Nick van Es, Omar S. Al-Kadi, Oscar Perez Concha, Patrick Vanderstuyft, Peter Dayan, Peter Kent, Pieter Cornu, Rabia Bashir, Reza Khajouei, Robert C Amland, Robert E. Freundlich, Robert Greenes, Rose Galvin, Samina Abidi, Seong Ho Park, Sheila Payne, Sherif Shabana, Simon Adams, Surbhi Leekha, Syed Mustafa Ali, Tero Shemeikka, Thomas Debray, Vassilis Koutkias.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpbup.2024.100161.

References

[1] E.W. Steyerberg, Clinical Prediction models: a Practical Approach to development, validation, and Updating, Springer Science & Business Media, 2008.
[2] E.W. Steyerberg, Clinical Prediction models: a Practical Approach to development, validation, and Updating, 2nd Edition ed., Springer Science & Business Media, 2019.
[3] R.A. Taylor, et al., Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach, Acad. Emerg. Med. 23 (3) (2016) 269–278.
[4] N. Kuppermann, et al., Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study, Lancet 374 (9696) (2009) 1160–1170.
[5] I.G. Stiell, et al., Derivation of a decision rule for the use of radiography in acute knee injuries, Ann. Emerg. Med. 26 (4) (1995) 405–413.
[6] T.A. Manuck, et al., Nonresponse to 17-alpha hydroxyprogesterone caproate for recurrent spontaneous preterm birth prevention: clinical prediction and generation of a risk scoring system, Am. J. Obstetr. Gynecol. 215 (5) (2016), 622. e1-622.e8.
[7] C. van Walraven, et al., Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community, Can. Med. Assoc. J. 182 (6) (2010) 551–557.
[8] A.M. Dietrich, et al., Pediatric head injuries: can clinical factors reliably predict an abnormality on computed tomography? Ann. Emerg. Med. 22 (10) (1993) 1535–1540.
[9] J. Dunning, et al., Derivation of the children's head injury algorithm for the prediction of important clinical events decision rule for head injury in children, Arch. Dis. Child. 91 (11) (2006) 885–891.
[10] R.M. Centor, et al., The diagnosis of strep throat in adults in the emergency room, Med. Decis. Making 1 (3) (1981) 239–246.
[11] K.K. Vaigandla, M.K. Vanteru, M. Siluveru, An extensive examination of the IoT and blockchain technologies in relation to their applications in the healthcare industry, Mesopotam. J. Comput.Sci. 2024 (2024) 1–14.
[12] L. Chen, Overview of clinical prediction models, Ann. Transl. Med. 8 (4) (2020).
[13] E. Christodoulou, et al., A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clin. Epidemiol. 110 (2019) 12–22.
[14] J.A. Damen, et al., Prediction models for cardiovascular disease risk in the general population: systematic review, BMJ 353 (2016).
[15] B.A. Goldstein, et al., Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. JAMIA 24 (1) (2017) 198.
[16] F.E. Babl, et al., Accuracy of PECARN, CATCH, and CHALICE head injury decision rules in children: a prospective cohort study, Lancet 389 (10087) (2017) 2393–2402.
[17] M. Khalifa, B. Gallego, Grading and assessment of clinical predictive tools for paediatric head injury: a new evidence-based approach, BMC Emerg. Med. 19 (1) (2019) 35.
[18] R.T. Sutton, et al., An overview of clinical decision support systems: benefits, risks, and strategies for success, NPJ. Digit. Med. 3 (1) (2020) 17.
[19] M. Khalifa, F. Magrabi, B. Gallego, Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support, BMC Med. Inform. Decis. Mak. 19 (1) (2019) 207.
[20] Khalifa, M., F. Magrabi, and B. Gallego, Evaluating the Impact of Using GRASP Framework on Clinicians and Healthcare Professionals Decisions in Selecting Clinical Predictive Tools. arXiv preprint arXiv:1907.11523, 2019.
[21] K.A. Batterton, K.N. Hale, The Likert scale what it is and how to use it, Phalanx, 50 (2) (2017) 32–39.
[22] R.R. Bond, et al., A usability evaluation of medical software at an expert conference setting, Comput. Methods Programs Biomed. 113 (1) (2014) 383–395.
[23] Lehrer, D. and J. Vasudev, Visualizing information to improve building performance: a study of expert users. 2010.

[24] M. Santiago-Delefosse, et al., Quality of qualitative research in the health sciences: analysis of the common criteria present in 58 assessment guidelines by expert users, Soc. Sci. Med. 148 (2016) 142–151.
[25] Qualtrics experience management solutions, Qualtrics. 2018 [cited 2018 1 January]; Available from: https://www.qualtrics.com/.
[26] S.E. Stemler, A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability, Pract. Assess. Res. Eval. 9 (1) (2019) 4.
[27] C. Croux, C. Dehon, Influence functions of the Spearman and Kendall correlation measures, Stat. Methods Appt. 19 (4) (2010) 497–515.
[28] A.T. Jebb, V. Ng, L. Tay, A review of key Likert scale development advances: 1995–2019, Front. Psychol. 12 (2021) 637547.
[29] K. Dhakal, NVivo, J. Med. Libr. Assoc. JMLA 110 (2) (2022) 270.
[30] B.-M. Lindgren, B. Lundman, U.H. Graneheim, Abstraction and interpretation during the qualitative content analysis process, Int. J. Nurs. Stud. 108 (2020) 103632.
[31] E.W. Steyerberg, F.E. Harrell Jr, Prediction models need appropriate internal, internal-external, and external validation, J. Clin. Epidemiol. 69 (2016) 245.
[32] E.W. Steyerberg, Y. Vergouwe, Towards better clinical prediction models: seven steps for development and an ABCD for validation, Eur. Heart J. 35 (29) (2014) 1925–1931.
[33] K. Hajian-Tilaki, Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, Caspian J. Intern. Med. 4 (2) (2013) 627.
[34] C.H. Schmid, J.L. Griffith, Multivariate classification rules: calibration and discrimination, Encyclop. Biostat. 5 (2005).
[35] G.S. Collins, et al., External validation of multivariable prediction models: a systematic review of methodological conduct and reporting, BMC Med. Res. Methodol. 14 (1) (2014) 40.
[36] A. Bohr, K. Memarzadeh, The Rise of Artificial Intelligence in Healthcare applications, in Artificial Intelligence in Healthcare, Elsevier, 2020, pp. 25–60.
[37] Z. Chen, et al., Harnessing the power of clinical decision support systems: challenges and opportunities, Open Heart 10 (2) (2023) e002432.
[38] M. Broekhuis, et al., Conceptualizing usability for the eHealth context: content analysis of usability problems of eHealth applications, JMIR. Form. Res. 5 (7) (2021) e18198.
[39] V. Sharma, et al., Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records, BMJ Health Care Inform. 28 (1) (2021).
[40] A. Deniz-Garcia, et al., Quality, usability, and effectiveness of mhealth apps and the role of artificial intelligence: current scenario and challenges, J. Med. Internet Res. 25 (2023) e44030.
[41] J.M. Wohlgemut, et al., Methods used to evaluate usability of mobile clinical decision support systems for healthcare emergencies: a systematic review and qualitative synthesis, JAMIa Open. 6 (3) (2023) ooad051.
[42] K.E. Trinkley, et al., Clinician preferences for computerised clinical decision support for medications in primary care: a focus group study, BMJ Health Care Inform. 26 (1) (2019).
[43] Fernainy, P., et al. Rethinking the pros and cons of randomized controlled trials and observational studies in the era of big data and advanced methods: a panel discussion. in BMC proceedings. 2024. Springer.
[44] S.R. Pfohl, et al., A comparison of approaches to improve worst-case predictive model performance over patient subpopulations, Sci. Rep. 12 (1) (2022) 3254.
[45] E. Ahn, H. Kang, Introduction to systematic review and meta-analysis, Korean J. Anesthesiol. 71 (2) (2018) 103–112.
[46] E.W. Steyerberg, et al., Internal and external validation of predictive models: a simulation study of bias and precision in small samples, J. Clin. Epidemiol. 56 (5) (2003) 441–447.
[47] T.P. Debray, et al., A new framework to enhance the interpretation of external validation studies of clinical prediction models, J. Clin. Epidemiol. 68 (3) (2015) 279–289.
[48] T.P. Debray, et al., A guide to systematic review and meta-analysis of prediction model performance, BMJ 356 (2017) i6460.
[49] E.W. Steyerberg, et al., Internal validation of predictive models: efficiency of some procedures for logistic regression analysis, J. Clin. Epidemiol. 54 (8) (2001) 774–781.
[50] E.W. Steyerberg, et al., Assessing the performance of prediction models: a framework for some traditional and novel measures, Epidemiology 21 (1) (2010) 128.
[51] G.S. Collins, et al., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, BMC Med. 13 (1) (2015) 1.
[52] K.G. Moons, et al., Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist, PLoS Med. 11 (10) (2014) e1001744.
[53] K.G. Moons, et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration, Ann. Intern. Med. 162 (1) (2015) W1–W73.
[54] E. Wallace, et al., Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs), BMC Med. Inform. Decis. Mak. 11 (1) (2011) 62.
[55] D. Toll, et al., Validation, updating and impact of clinical prediction rules: a review, J. Clin. Epidemiol. 61 (11) (2008) 1085–1094.
[56] A.H. Harris, Path from predictive analytics to improved patient outcomes: a framework to guide use, implementation, and evaluation of accurate surgical predictive models, Ann. Surg. 265 (3) (2017) 461–463.
[57] E. Wallace, et al., Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review, BMJ Open 6 (3) (2016) e009957.
[58] C.T. Bezerra, et al., Assessment of the strength of recommendation and quality of evidence: GRADE checklist. A descriptive study, Sao Paulo Med. J. 140 (2022) 829–836.
[59] L.S. Nowell, et al., Thematic analysis: striving to meet the trustworthiness criteria, Int. J. Qual. Methods 16 (1) (2017) 1609406917733847.
[60] P. Nasa, R. Jain, D. Juneja, Delphi methodology in healthcare research: how to decide its appropriateness, World J. Methodol. 11 (4) (2021) 116.
[61] M.-J. Wu, K. Zhao, F. Fils-Aime, Response rates of online surveys in published research: a meta-analysis, Comput. Hum. Behav. Rep. 7 (2022) 100206.

The Appendix

The GRASP Framework Detailed Report

Table S1: The GRASP Framework Detailed Report

Name	Name of predictive tool (report tool's creators and year in the absence of a given name)		
Authors/Year	Name of developer, country and year of publication		
Intended use	Predictive task/specific aim/intended use of the predictive tool		
Intended user	Type of practitioner intended to use the tool (e.g. physician or nurse)		
Category	Diagnostic/Therapeutic/Prognostic/Preventive		
Clinical area	Clinical specialty		
Target Population	Target patient population and health care settings in which the tool is applied		
Target Outcome	Event to be predicted (including prediction lead time if needed)		
Action	Recommended action based on tool's output		
Input source	• Clinical (including Diagnostic, Genetic, Vital signs, Pathology) • Non-Clinical (including Healthcare Utilisation)		
Input type	• Objective (Measured input; from electronic systems or clinical examination) • Subjective (Patient reported; history, checklist …etc.)		
Local context	Is the tool developed using location-specific data? (e.g. life expectancy tables)		
Methodology	Type of algorithm (e.g. parametric/non-parametric)		
Endorsement	Organisations endorsing the tool and/or guidelines recommending its utilisation		
Automation Flag	Automation status (manual/automated)		
Phase of Evaluation	**Level of Evidence**	**Grade**	**Evaluation Studies**
Phase C: Before implementation Is it possible?	Internal validation	**C3**	Tested for internally validity (reported calibration & discrimination; sensitivity, specificity, positive and negative predictive values & other performance measures).
	External validation	**C2**	Tested for external validity, using one external dataset.
	External validation multiple times	**C1**	Tested multiple times for external validity, using more than one external dataset.
Phase B: During implementation Is it practicable?	Potential effect	**B2**	Reported estimated potential effect on clinical effectiveness, patient safety or healthcare efficiency.
	Usability	**B1**	Reported usability testing (effectiveness, efficiency, satisfaction, learnability, memorability, and minimizing errors).
Phase A: After implementation: Is it desirable?	Evaluation of post implementation impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency	**A3**	Based on subjective studies; e.g. the opinion of a respected authority, clinical experience, a descriptive study, or a report of an expert committee or panel.
		A2	Based on observational studies; e.g. a well-designed cohort or case-control study.
		A1	Based on experimental studies; properly designed, widely applied randomised/nonrandomised controlled trial.
Final Grade	**Grade ABC,123**	**A1**	**A2** **A3** **B1** **B2** **C1** **C2** **C3**
Direction of Evidence	● Positive Evidence	◕ Mixed Evidence Supporting Positive Conclusion	
	○ Negative Evidence	◔ Mixed Evidence Supporting Negative Conclusion	
Justification	Explains how the final grade is assigned based on evidence; which conclusions were taken into consideration, as positive evidence, and which were considered negative.		
References	Details of studies that support the justification: phase of evaluation, level of evidence, direction of evidence, study type, study settings, methodology, results, findings and conclusions (highlighted according to the colour code).		These two sections are included in the full GRASP report on each tool.
Label/Colour Code	• Positive Findings • Negative Findings	• Important Findings • Less Relevant Findings	

The Updated GRASP Framework Detailed Report

Table S2: The Updated GRASP Framework Detailed Report

Name	Name of predictive tool (report tool's creators and year in the absence of a given name)		
Author	Name of developer (first author or researcher)		These are now three separate fields instead of one field reporting the three information.
Country	Country of development		
Year	Year of development		
Category	Diagnostic/Therapeutic/Prognostic/Preventive		
Intended Use	Predictive task/specific aim/intended use of the predictive tool		
Intended User	Type of practitioner intended to use the tool (e.g. physician or nurse)		
Clinical Area	Clinical specialty		
Target Population	Target patient population and health care settings in which the tool is applied		
Target Outcome	Event to be predicted (including prediction lead time if needed)		
Action	Recommended action based on tool's output		
Input Source	• Clinical (including Diagnostic, Genetic, Vital signs, Pathology) • Non-Clinical (including Healthcare Utilisation)		
Input Type	• Objective (Measured input; from electronic systems or clinical examination) • Subjective (Patient reported; history, checklist …etc.)		
Local Context	Is the tool developed using location-specific data? (e.g. life expectancy tables)		
Methodology	Type of algorithm used for developing the tool (e.g. parametric/non-parametric)		
Endorsement	Organisations endorsing the tool and/or clinical guidelines recommending its utilisation		
Automation Flag	Automation status (manual/automated)		
Internal Validation	Method of internal validation		These are new fields.
Dedicated Support	Name of the supporting/funding research networks, programs, or professional groups		
Tool Citations	Total citations of the tool		
Studies	Number of studies reporting the tool		
Authors No	Number of authors		
Sample Size	Size of patient/record sample used in the development of the tool		
Journal Name	Name of the journal that published the tool's primary development study		
Journal Rank	Impact factor of the journal		
Citation Index	Calculated as: Average Annual Citations = number of citations/age of primary publication		
Publication Index	Calculated as: Average Annual Studies = number of studies/age of primary publication		
Literature Index	Calculated as: Citations and Publications = number of citations X number of studies		
Phase of Evaluation	**Level of Evidence**	**Grade**	**Evaluation Studies**
Phase C: **Before implementation**	Insufficient internal validation	**C0**	Not tested for internal validity, insufficiently internally validated, or internal validation was insufficiently reported.
	Internal validation	**C3**	Tested for internally validity (reported calibration & discrimination; sensitivity, specificity, positive and negative predictive values & other performance measures).

Is it possible?	External validation	**C2**	Tested for external validity, using one external dataset.	
	External validation multiple times	**C1**	Tested multiple times for external validity, using more than one external dataset.	
Phase B: **Planning for implementation (Renamed from: During Implementation)** **Is it practicable?**	Usability	**B3**	Reported usability testing (tool effectiveness, efficiency, satisfaction, learnability, memorability, and minimizing errors).	Instead of two levels in Phase B, we have now three levels. Potential Effect is higher than Usability, and both together is higher than any of them alone.
	Potential effect	**B2**	Reported estimated potential effect on clinical effectiveness, patient safety or healthcare efficiency.	
	Potential effect & Usability	**B1**	Both potential effect and usability are reported.	
Phase A: **After implementation:** **Is it desirable?**	Evaluation of post implementation impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency	**A3**	Based on subjective studies; e.g. the opinion of a respected authority, clinical experience, a descriptive study, or a report of an expert committee or panel.	
		A2	Based on observational studies; e.g. a well-designed cohort or case-control study.	
		A1	Based on experimental studies; properly designed, widely applied randomised/nonrandomised controlled trial.	

Final Grade	**Grade ABC/123**	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**

Tool Label	One-word description of the most prominent prediction, potential effect or impact on processes or outcomes. E.g. "Grade A2 – Efficiency" (the tool improves efficiency by saving money, resources or time, proved through observational post-implementation impact studies).		New field
Direction of Evidence	● Positive Evidence	◕ Mixed Evidence Supporting Positive Conclusion	
	○ Negative Evidence	◔ Mixed Evidence Supporting Negative Conclusion	
Justification	Explains how the final grade is assigned based on evidence; which conclusions were taken into consideration, as positive evidence, and which were considered negative.		
Evidence Summary (Renamed from References)	Details of studies; using the Evidence Summary, to support the justification, where comparative predictive performance and effectiveness studies are highlighted.		
Findings Codes (Renamed from Colour Code)	Positive Findings / Negative Findings / Important Findings		

The Evidence Summary

Table S3: The Evidence Summary

Study	The published study (According to Reference Style)
Country	Country of study
Year	Year of study
Phase	Before Implementation, planning for implementation, after Implementation
Type	Development / Internal Validation / External Validation / Usability / Potential Effect / Post-Implementation Impact.
Tools	Single Tool vs Comparative Study (comparing multiple tools or one tool vs clinical practice).
Intended use [1]	Predictive task/specific aim/intended use of the predictive tool
Intended user [1]	Type of practitioner intended to use the tool (e.g. physician or nurse)
Clinical Area [1]	Clinical specialty
Target Population [1]	Patients (age group, gender group, clinical specifications, e.g. cardiac population). Providers (age group, gender, clinical specifications, e.g. specialty).
Settings [1]	Inpatient, outpatient, intensive care … etc.
Practice [1]	Clinical vs non-clinical practice.
Methods [2]	Tool development methods: recursive partitioning, multivariate logistic regression … etc. Internal validation methods: out-of-sample, bootstrapping, cross validation, split sample … etc. External validation methods: national, international … etc. Usability: acceptance, satisfaction, adoption … etc. Potential Effect: feasibility, cost-effectiveness, economic analysis … etc. Impact: experimental (randomised, non-randomised, controlled, quasi-experimental), observational (cohort studies, case-control, cross-sectional), subjective (expert opinion, reports) … etc.
Sample Size [2]	Number of patients/records/users recruited in the study
Data Collection [2]	Prospective/retrospective data
Outcomes [2]	Reported outcome measures: Development/Validation: reported calibration/discrimination; sensitivity, specificity, positive & negative predictive values & other performance measures. Usability: acceptance, satisfaction … etc. Potential Effect: feasibility, cost-effectiveness, economic analysis … etc. Impact: effect size, duration of implementation … etc.
Institute [2]	Name and type of hospital (Multiple hospitals, single hospital, tertiary care … etc).
Support [2]	Dedicated support of research networks, programs or groups.
Authors [2]	Number of researchers.
Journal [2]	Journal name and impact factor.
Direction of Evidence	Positive / Equivocal / Negative (Based on study findings and conclusions).
Matching of Evidence	Considering fields [1] (Matching/Non-Matching to the tool's original specifications)
Quality of Evidence	Considering fields [2] (High Quality/Low Quality of the study)
Strength of Evidence	Based on Evidence Matching and Quality: Strong Evidence / Medium Evidence / Weak Evidence
Label	Effectiveness / Efficiency / Safety / Workflow / Processes (one or more).
Notes	Special important study information.
Fields 1 (Matching of Evidence): Intended Use, Intended User, Clinical Area, Target Population, Settings, Practice. Fields 2 (Quality of Evidence): Methods, Sample Size, Data Collection, Outcomes, Institute, Support, Authors, Journal.	

Experts' Country Distributions

Table S4: Country Distributions of Expert Respondents

Country	Responses	Percent	Cumulative
United States	20	24.7%	24.7%
United Kingdom	12	14.8%	39.5%
Canada	8	9.9%	49.4%
Netherlands	5	6.2%	55.6%
Spain	5	6.2%	61.7%
Australia	4	4.9%	66.7%
Argentina	2	2.5%	69.1%
Belgium	2	2.5%	71.6%
Germany	2	2.5%	74.1%
Austria	1	1.2%	75.3%
China	1	1.2%	76.5%
Colombia	1	1.2%	77.8%
Croatia	1	1.2%	79.0%
Denmark	1	1.2%	80.2%
Finland	1	1.2%	81.5%
France	1	1.2%	82.7%
Ghana	1	1.2%	84.0%
Greece	1	1.2%	85.2%
Iran	1	1.2%	86.4%
Japan	1	1.2%	87.7%
Jordan	1	1.2%	88.9%
Korea	1	1.2%	90.1%
Malaysia	1	1.2%	91.4%
Mexico	1	1.2%	92.6%
New Zealand	1	1.2%	93.8%
Norway	1	1.2%	95.1%
Pakistan	1	1.2%	96.3%
South Africa	1	1.2%	97.5%
Sweden	1	1.2%	98.8%
Taiwan	1	1.2%	100.0%
Total	**81**	**100%**	

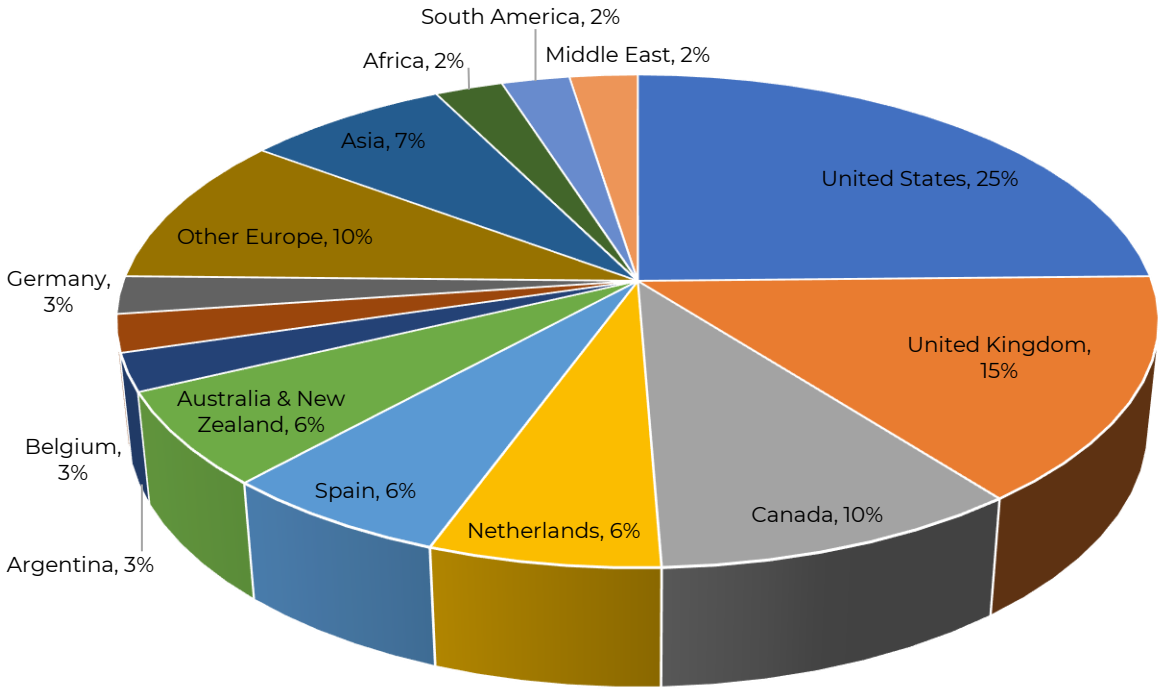


Figure S1: Country Distributions of Expert Respondents

The Strength of Evidence Protocol

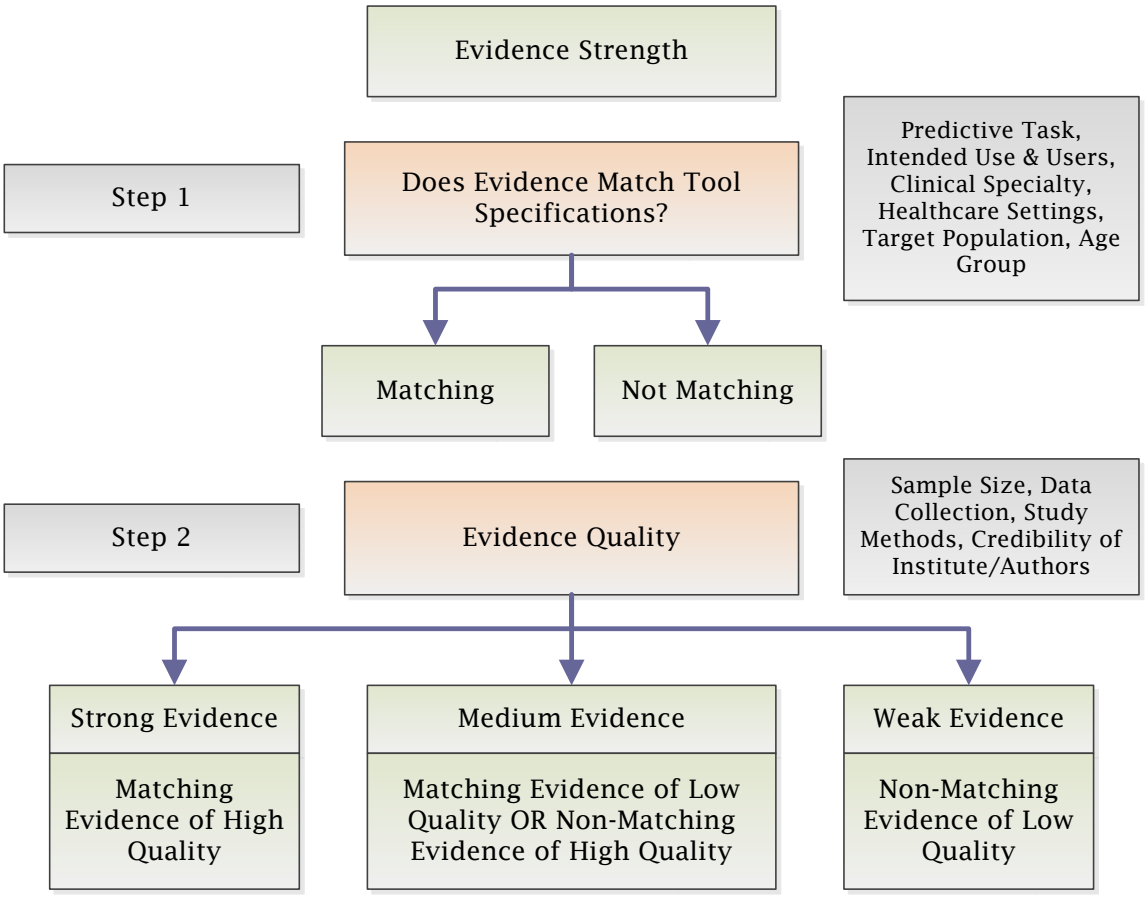


Figure S2: The Strength of Evidence Protocol

The strength of the evidence protocol considers two main criteria of the published studies. Firstly, it considers the degree of matching between the published study conditions, through which the tools is being evaluated, and the original tool specifications, in terms of the predictive task, target outcomes, intended use and users, clinical specialty, healthcare settings, target population, and age group. Secondly, it considers the quality of the study, in terms of the sample size, data collection, study methods, and credibility of institute and authors. Based on these two criteria, the strength of evidence is classified into 1) Strong Evidence: matching evidence of high quality, 2) Medium Evidence: matching evidence of low quality or non-matching evidence of high quality, and 3) Weak Evidence: non-matching evidence of low quality.

The Mixed Evidence Protocol

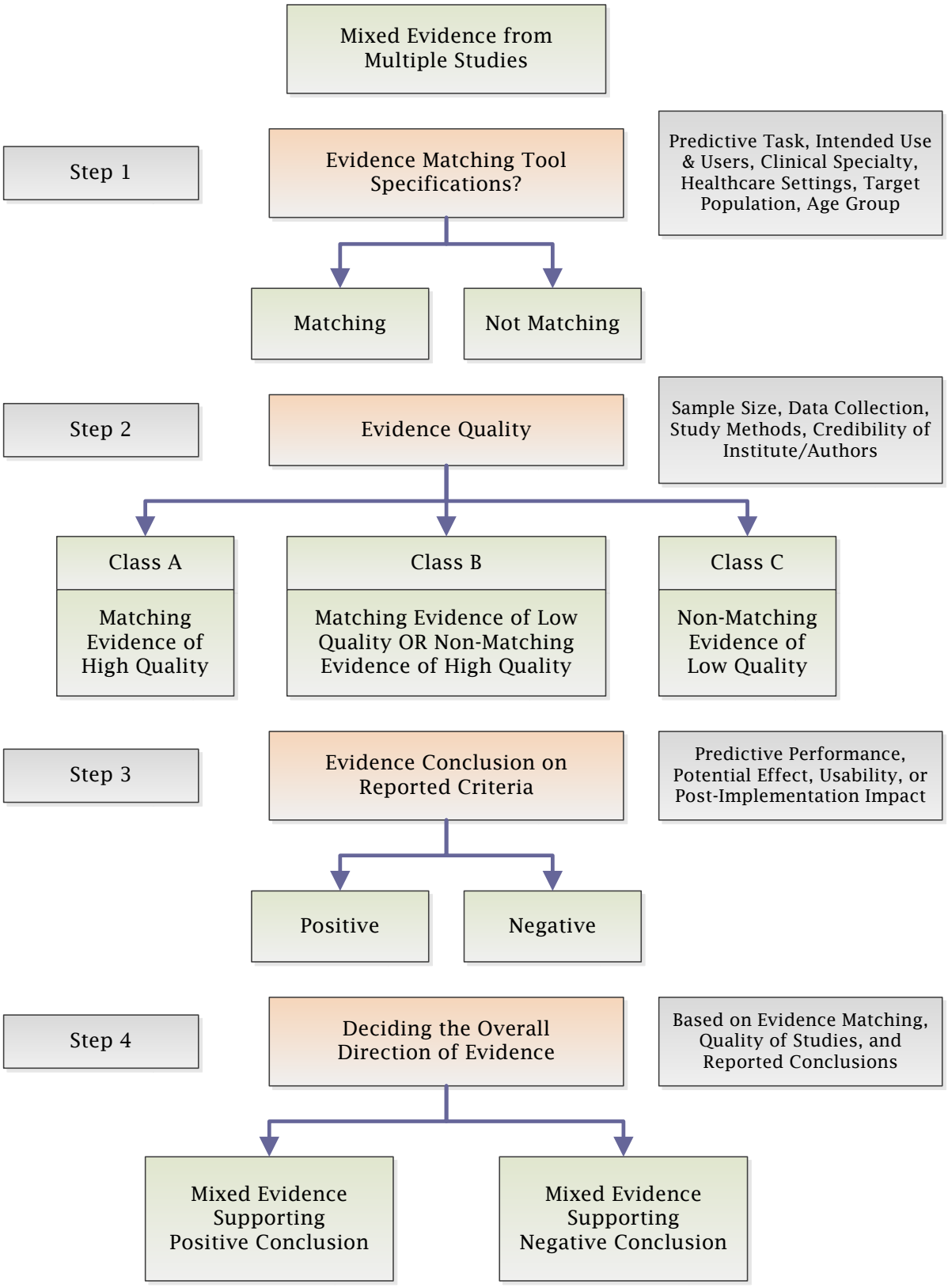


Figure S3: The Mixed Evidence Protocol

The mixed evidence protocol is based on four steps. Firstly, it considers the degree of matching between the evaluation study conditions and the original tool specifications, in terms of the predictive task, outcome, intended use and users, clinical specialty, healthcare settings, target population, and age group. Secondly, it considers the quality of the study, in terms of sample size, data collection, study methods, and credibility of institute or authors. Based on these two criteria, the studies in the mixed evidence on the tool are classified into 1) Class A: matching evidence of high quality, 2) Class B: matching evidence of low quality or non-matching evidence of high quality, and 3) Class C: non-matching evidence of low quality. Thirdly, it considers the evidence conclusion on the reported evaluation criteria, the predictive performance, potential effect, usability, and post-implementation impact. In the fourth step, studies evaluating predictive tools in closely matching conditions to the tool specifications and providing high quality evidence, Class A, are considered first; taking into account their conclusions on the evaluation criteria in deciding the overall direction of evidence. On the other hand, studies evaluating predictive tools in different conditions to the tool specifications and providing low quality evidence, Class C, are considered last. The conclusion of one study in Class A is considered a stronger evidence than the conflicting conclusions of any number of studies in Class B or C, and the overall direction of the evidence is decided towards the conclusion of the study of Class A. When multiple studies of the same class; for example, Class A, report conflicting conclusions, then we compare the number of studies reporting positive conclusions to those reporting negative conclusions and the overall direction of the evidence is decided towards the conclusion of the larger group. If the two groups are of the same size, then we check if there are more studies in other classes, if not then we examine the reported evaluation criteria and their values in the two groups of studies.

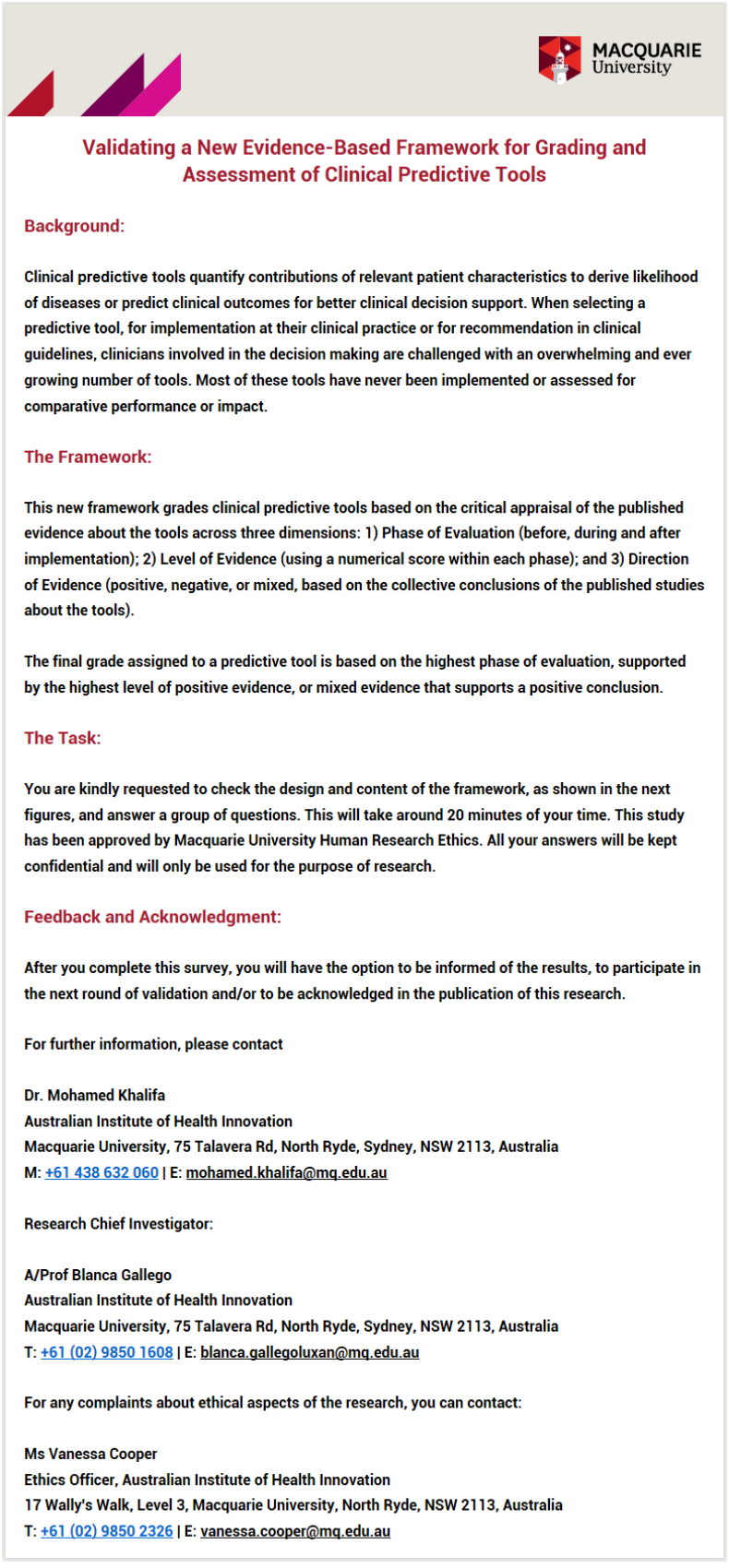
Interrater Reliability Detailed Results

Table S5: Grading the Predictive Tools by the Independent Researcher vs the Authors

Tool	Two Researchers (R1 & R2) vs Paper Authors (A)	Assigned Grade	Impact After Implementation			Planning for Implementation			Performance Before Implementation		
			Experimental Studies	Observational Studies	Subjective Studies	Potential Effect & Usability	Potential Effect	Usability	External Validation Multiple Times	External Validation Only Once	Internal Validation
			A1	A2	A3	B1	B2	B3	C1	C2	C3
Centor Score [1]	R1	B2	○				◕		●		●
	R2	B3	◔				○	●	●		●
	A	B3	◔					●	●		●
CHALICE Rule [2]	R1	B2					◕		◕		●
	R2	B2					◕		●		●
	A	B2					◕		●		●
Dietrich Rule [3]	R1	C0									○
	R2	C0									○
	A	C0									○
LACE Index [4]	R1	C1							◕		●
	R2	C1							◕		●
	A	C1							◕		●
Manuck Scoring System [5]	R1	C2								●	●
	R2	C2								●	●
	A	C2								●	●
Ottawa Knee Rule [6]	R1	A1	●				●		●		●
	R2	A2		●					●		●
	A	A1	●						●		●
PECARN Rule [7]	R1	A2		●				●	◕		●
	R2	A2		◕		●	●		●		●
	A	A2		◕			●		●		●
Taylor Mortality Model [8]	R1	C3									●
	R2	C3									●
	A	C3									●

Direction of Evidence		
	● Positive Evidence	◕ Mixed Evidence Supporting Positive Conclusion
	○ Negative Evidence	◔ Mixed Evidence Supporting Negative Conclusion

The Survey Screenshots



Validating a New Evidence-Based Framework for Grading and Assessment of Clinical Predictive Tools

Background:

Clinical predictive tools quantify contributions of relevant patient characteristics to derive likelihood of diseases or predict clinical outcomes for better clinical decision support. When selecting a predictive tool, for implementation at their clinical practice or for recommendation in clinical guidelines, clinicians involved in the decision making are challenged with an overwhelming and ever growing number of tools. Most of these tools have never been implemented or assessed for comparative performance or impact.

The Framework:

This new framework grades clinical predictive tools based on the critical appraisal of the published evidence about the tools across three dimensions: 1) Phase of Evaluation (before, during and after implementation); 2) Level of Evidence (using a numerical score within each phase); and 3) Direction of Evidence (positive, negative, or mixed, based on the collective conclusions of the published studies about the tools).

The final grade assigned to a predictive tool is based on the highest phase of evaluation, supported by the highest level of positive evidence, or mixed evidence that supports a positive conclusion.

The Task:

You are kindly requested to check the design and content of the framework, as shown in the next figures, and answer a group of questions. This will take around 20 minutes of your time. This study has been approved by Macquarie University Human Research Ethics. All your answers will be kept confidential and will only be used for the purpose of research.

Feedback and Acknowledgment:

After you complete this survey, you will have the option to be informed of the results, to participate in the next round of validation and/or to be acknowledged in the publication of this research.

For further information, please contact

Dr. Mohamed Khalifa
Australian Institute of Health Innovation
Macquarie University, 75 Talavera Rd, North Ryde, Sydney, NSW 2113, Australia
M: +61 438 632 060 | E: mohamed.khalifa@mq.edu.au

Research Chief Investigator:

A/Prof Blanca Gallego
Australian Institute of Health Innovation
Macquarie University, 75 Talavera Rd, North Ryde, Sydney, NSW 2113, Australia
T: +61 (02) 9850 1608 | E: blanca.gallegoluxan@mq.edu.au

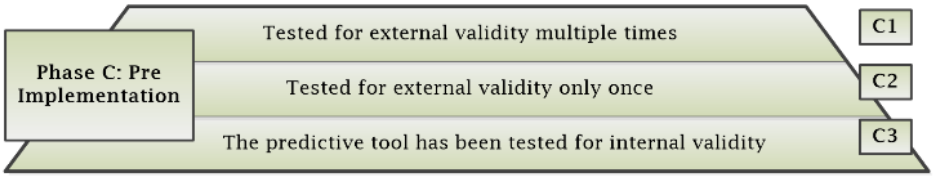
For any complaints about ethical aspects of the research, you can contact:

Ms Vanessa Cooper
Ethics Officer, Australian Institute of Health Innovation
17 Wally's Walk, Level 3, Macquarie University, North Ryde, NSW 2113, Australia
T: +61 (02) 9850 2326 | E: vanessa.cooper@mq.edu.au

Section 1: The survey introduction

MACQUARIE University

Published Evidence on Evaluating the Tools Before Implementation



How much do you agree with the following?

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Q1) We should consider the evidence on validating the tool's predictive performance.	○	○	○	○	○
Q2) The evidence level could be High (internal + multiple external validation), Medium (internal + external validation once), or Low (internal validation only).	○	○	○	○	○

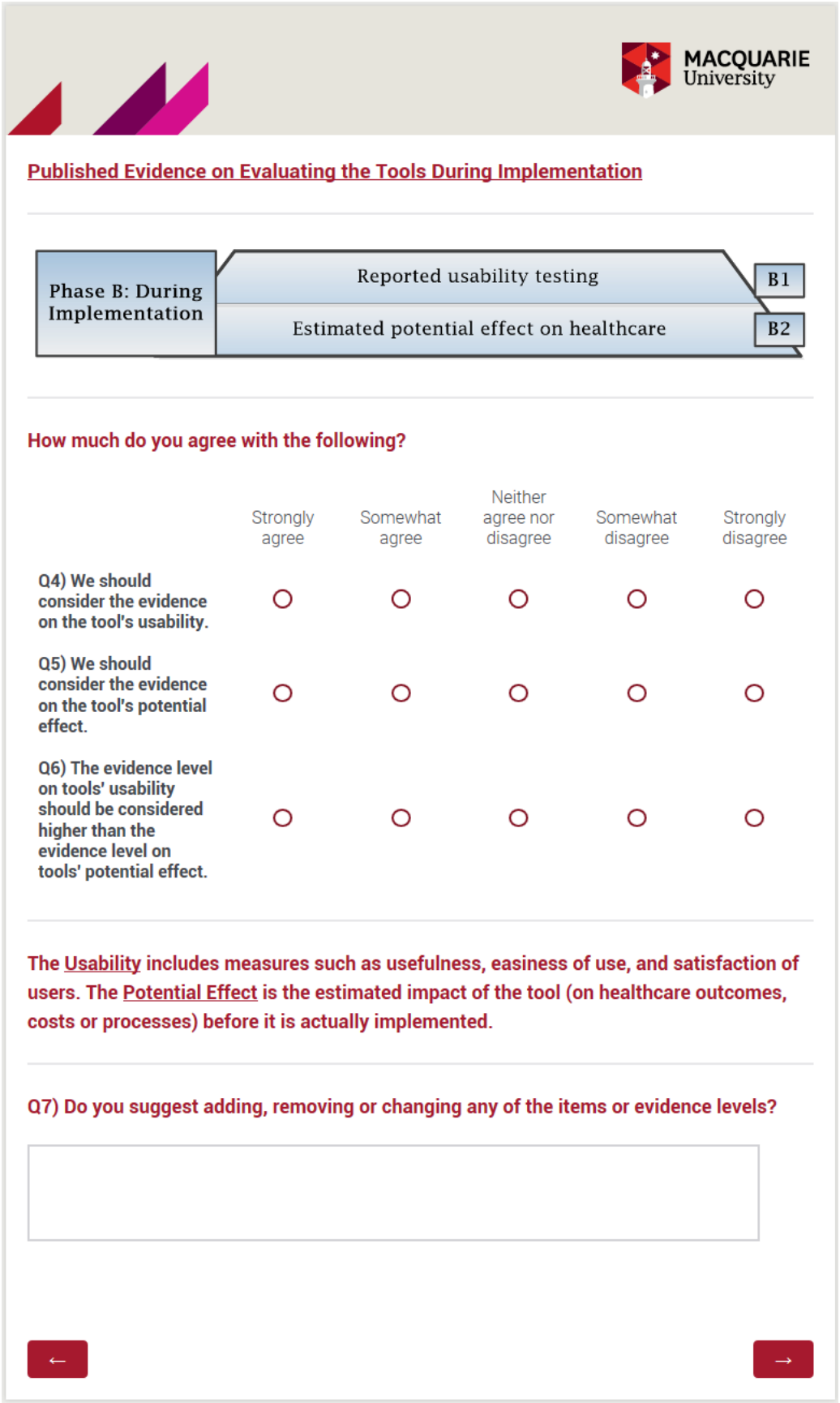
Validating the predictive performance includes measures such as sensitivity and specificity of tools. Internal Validation includes testing the performance of the tool on the same data that was used for its development, while External Validation includes testing the performance on new data, different from that used for development.

Q3) Do you suggest adding, removing or changing any of the items or evidence levels?

←

→

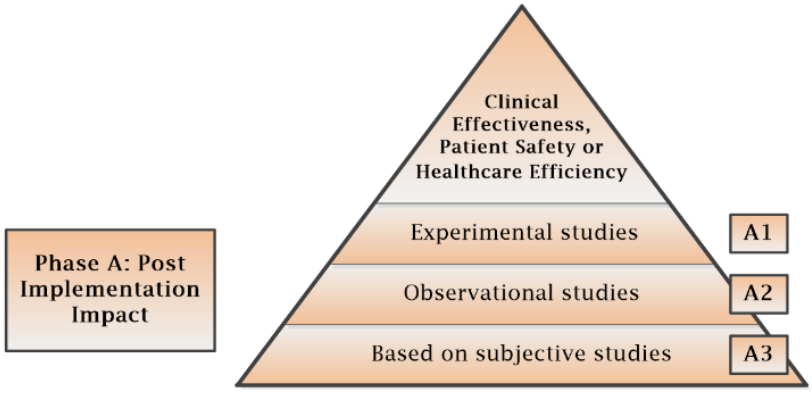
Section 2: Criteria of evaluating tools' predictive performance before implementation



Section 3: Criteria of evaluating tools' usability and estimated potential effect

MACQUARIE University

Published Evidence on Evaluating the Tools After Implementation



How much do you agree with the following?

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Q8) We should consider the evidence on the tool's impact on healthcare effectiveness, efficiency or safety.	○	○	○	○	○
Q9) The evidence level could be High (based on experimental studies), Medium (observational studies), or Low (subjective studies).	○	○	○	○	○

Effectiveness includes improving clinical outcomes, Efficiency includes reducing costs, while Safety includes minimising patient complications. We are assuming that all study types (experimental, observational & subjective) have the same high quality.

Q10) Do you suggest adding, removing or changing any of the items or evidence levels?

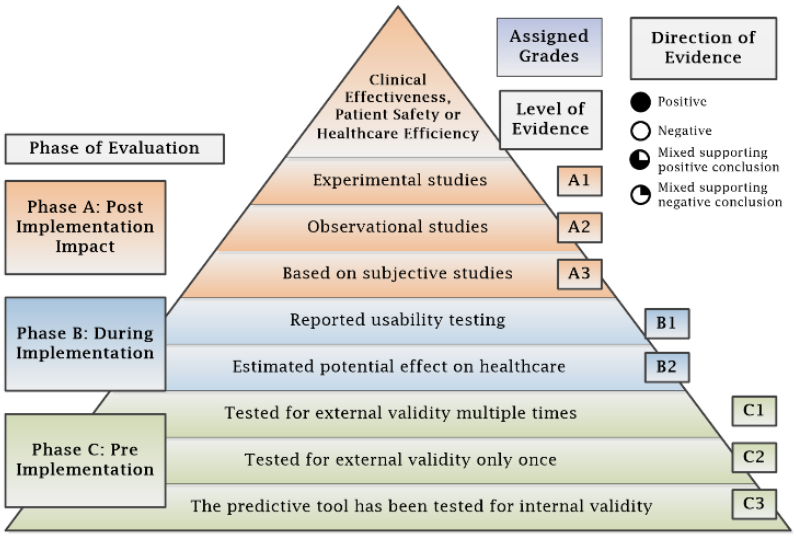
←

→

Section 4: Criteria of evaluating tools' impact post-implementation

MACQUARIE University

The Framework Three Dimensions



In addition to the phases of evaluation (before, during and after implementation) and the related levels of evidence (with each phase), how much do you agree with the following?

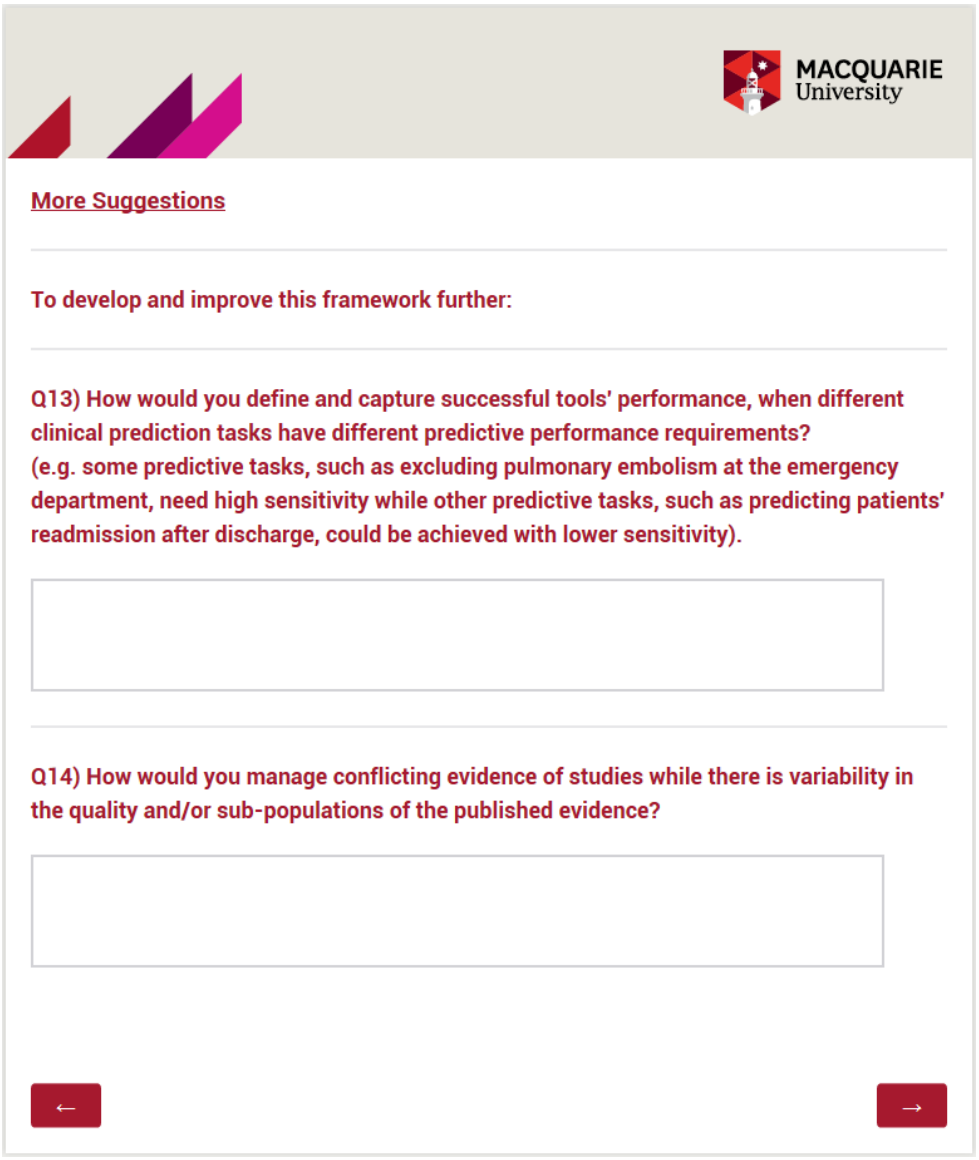
	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Q11) Based on the conclusions of published studies, the overall evidence direction could be Positive, Negative or Mixed.	○	○	○	○	○

Evidence direction is positive if all studies reported positive conclusions about the tool, mixed if some studies reported positive and some reported negative, and negative if all studies reported negative conclusions.

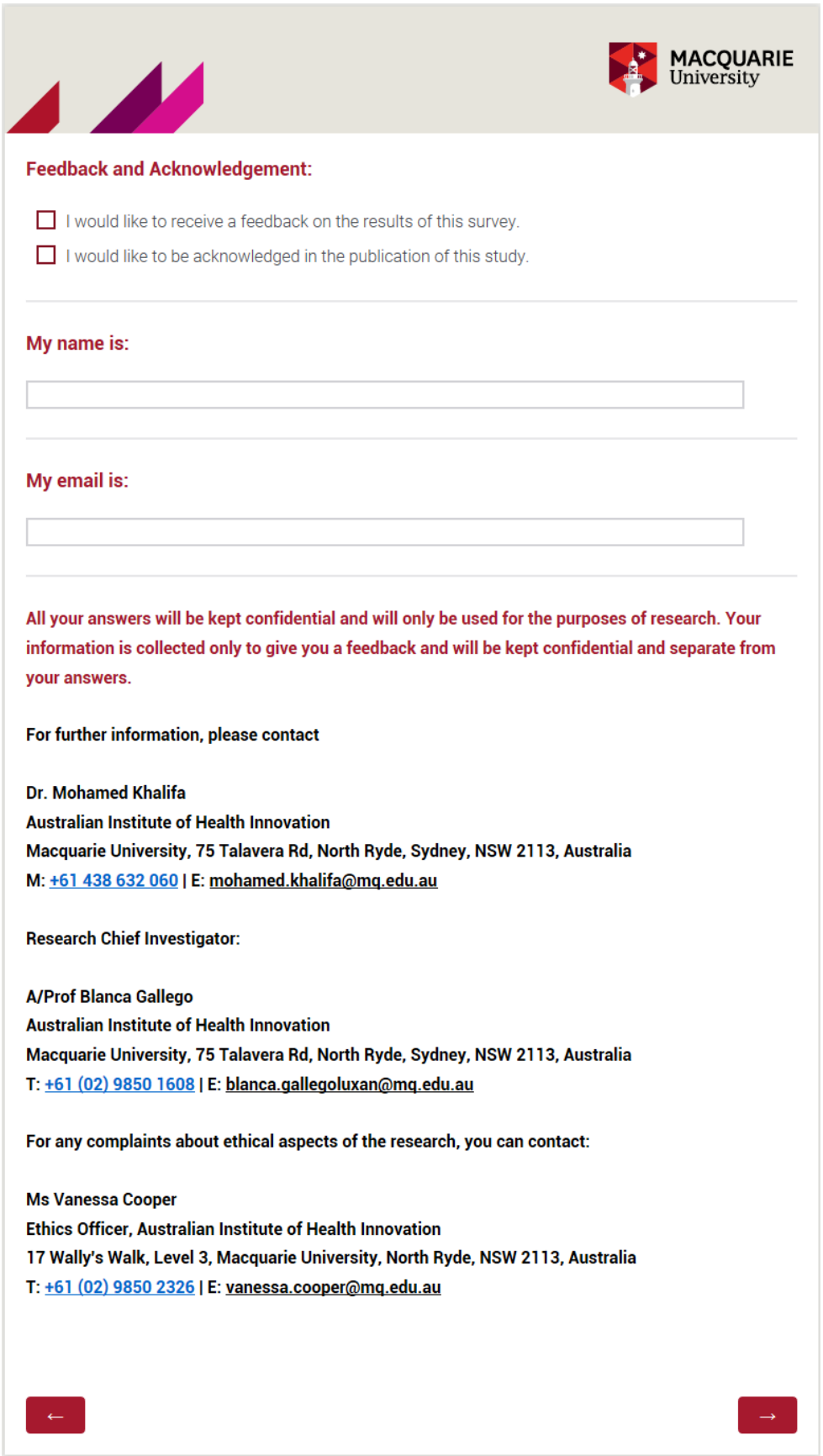
Q12) Do you suggest adding, removing or changing any of the phases of evaluation, levels of evidence, or directions of evidence?

← →

Section 5: Criteria of evaluating direction of published evidence



Section 6: Defining successful predictive performance and managing conflicting evidence.



Section 7: providing contacts to request feedback and acknowledgment

The Survey

Do you find the framework logical?

Do you find the framework useful?

Do you find the framework easy to use?

Please provide your views about the criteria used by GRASP to grade predictive tools:

Please list any criteria you wish to add/remove/change in GRASP:

The interrater reliability post-task questionnaire given to the independent reviewers

Appendix References

1. Centor, R.M., et al., *The diagnosis of strep throat in adults in the emergency room*. Medical Decision Making, 1981. **1**(3): p. 239-246.
2. Dunning, J., et al., *Derivation of the children's head injury algorithm for the prediction of important clinical events decision rule for head injury in children*. Archives of disease in childhood, 2006. **91**(11): p. 885-891.
3. Dietrich, A.M., et al., *Pediatric head injuries: can clinical factors reliably predict an abnormality on computed tomography?* Annals of emergency medicine, 1993. **22**(10): p. 1535-1540.
4. van Walraven, C., et al., *Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community*. Canadian Medical Association Journal, 2010. **182**(6): p. 551-557.
5. Manuck, T.A., et al., *Nonresponse to 17-alpha hydroxyprogesterone caproate for recurrent spontaneous preterm birth prevention: clinical prediction and generation of a risk scoring system*. American Journal of Obstetrics & Gynecology, 2016. **215**(5): p. 622. e1-622. e8.
6. Stiell, I.G., et al., *Derivation of a decision rule for the use of radiography in acute knee injuries*. Annals of emergency medicine, 1995. **26**(4): p. 405-413.
7. Kuppermann, N., et al., *Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study*. The Lancet, 2009. **374**(9696): p. 1160-1170.
8. Taylor, R.A., et al., *Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach*. Academic emergency medicine, 2016. **23**(3): p. 269-278.