Validating and Updating GRASP: An Evidence-Based Framework for Grading and Assessment of Clinical Predictive Tools

Mohamed Khalifa (dr.m.khalifa@gmail.com)
Macquarie University Faculty of Medicine and Health Sciences  https://orcid.org/0000-0002-5919-8352

Farah Magrabi
Macquarie University

Blanca Gallego
University of New South Wales

Research article

Keywords: Evidence-Based Medicine, Clinical Decision Support, Clinical Prediction, Grading and Assessment, Validation

Posted Date: August 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-15929/v2

License: ☑️ 📧 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background: When selecting predictive tools, clinicians are challenged with an overwhelming and ever-growing number, most of which have never been implemented or evaluated for comparative effectiveness. The authors developed an evidence-based framework for grading and assessment of predictive tools (GRASP). The objective of this study is to update GRASP and evaluate its reliability.

Methods: A web-based survey was developed to collect responses of a wide international group of experts, who published studies on clinical prediction tools. Experts were invited via email and their responses were quantitatively and qualitatively analysed using NVivo software. The interrater reliability of the framework, to assign grades to eight predictive tools by two independent users, was evaluated.

Results: We received 81 valid responses. On five-points Likert scale, experts overall strongly agreed with GRASP evaluation criteria=4.35/5, SD=1.01, 95%CI [4.349, 4.354]. Experts strongly agreed with six criteria: predictive performance=4.88/5, SD=0.43, 95%CI [4.87, 4.88] and evidence levels of predictive performance=4.44/5, SD=0.87, 95%CI [4.44, 4.45], usability=4.68/5, SD=0.70, 95%CI [4.67, 4.68] and potential effect=4.62/5, SD=0.68, 95%CI [4.61, 4.62], post-implementation impact=4.78/5, SD=0.57, 95%CI [4.78, 4.79] and evidence direction=4.25/5, SD=0.78, 95%CI [4.25, 4.26]. Experts somewhat agreed with one criterion: post-implementation impact levels=4.18/5, SD=1.14, 95%CI [4.17, 4.19]. Experts were neutral about one criterion; usability is higher than potential effect=2.96/5, SD=1.23, 95%CI [2.95, 2.97]. Sixty-four respondents provided recommendations to six open-ended questions regarding updating evaluation criteria. Forty-three suggested potential effect is higher than usability. Experts highlighted the importance of quality of studies and strength of evidence. Accordingly, GRASP concept and its detailed report were updated. The framework’s interrater reliability was tested, and two independent reviewers produced accurate and consistent results in grading eight predictive tools using the framework.

Conclusion: Before implementation, internal and external validation of predictive performance of tools is essential in evaluating sensitivity and specificity. During planning for implementation, potential effect is more important that usability to evaluate acceptability of tools by users. Post-implementation, it is crucial to evaluate tools’ impact on healthcare processes and clinical outcomes. The GRASP framework aims to provide clinicians with a high-level, evidence-based, and comprehensive, yet simple and feasible, approach to evaluate, compare, and select predictive tools.

Background

Clinical predictive tools are research-based applications designed to provide clinicians and other healthcare professionals with diagnostic, prognostic, or therapeutic decision support through predicting clinical and other related healthcare outcomes [1, 2]. 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 [3, 4]. Many studies discuss that there is an inappropriate but common practice of developing new predictive tools instead of validating or updating
existing ones [2, 5-8]. This represents a major challenge for clinicians, when selecting predictive tools for implementation in their clinical practice or for recommendation in clinical practice guidelines, facing such overwhelming and ever-growing number of tools. Moreover, while a few pre-implementation studies compare similar predictive tools along some predictive performance measures, we find that many of these tools have never been implemented or assessed for comparative performance or impact [9-13].

When developing clinical practice guidelines and selecting predictive tools to be recommended for use, some clinicians rely on their previous experience, subjective evaluation, or recent exposure. Objective methods and evidence based approached are not always used in such decisions [14, 15]. Other clinicians, when developing practice guidelines, search the literature for best available published evidence. Commonly they look for research studies that describe the development, implementation, or evaluation of predictive tools. More specifically, some clinicians look for systematic reviews on predictive tools, comparing their predictive performance or development methods. However, there are no available approaches to objectively summarise or interpret such evidence, to identify the most appropriate predictive tool(s) for a given application while considering various implementation challenges and clinical settings constraints [16-18].

**The GRASP Framework**

To overcome this major challenge, the authors have developed a new evidence-based framework for grading and assessment of predictive tools (The GRASP Framework) [19]. This framework aims to provide clinicians with standardised objective information on predictive tools to support their search for and selection of effective tools for their tasks. Based on the critical appraisal of the published evidence on predictive tools, the GRASP framework uses three dimensions to grade predictive tools: 1) Phase of Evaluation, 2) Level of Evidence and 3) Direction of Evidence.

**Phase of Evaluation:** Assigns A, B, or C based on the highest phase of evaluation. If a tool's predictive performance, as reported in the literature, has been tested for validity, it is assigned phase C. Predictive performance is defined as the capability of the tool to use clinical and other relevant patient characteristics to predict an outcome to support diagnostic, prognostic or therapeutic decisions [1, 2]. If a tool's usability and/or potential effect have been tested, it is assigned phase B. Usability is defined as the degree to which a tool can be used by specific users to achieve specific and quantifiable objectives in a specific context of use [20, 21]. While the potential effect, of a predictive tool, is defined as the expected, estimated or calculated impact of using the tool on different healthcare aspects, processes or outcomes, assuming the tool has been successfully implemented and is used in the clinical practice, as designed by its developers [22, 23]. Finally, if a tool has been implemented in clinical practice, and there is published evidence evaluating its impact, it is assigned phase A. The post-implementation impact is defined as the achieved change or influence, of predictive tools, on different healthcare aspects, processes or outcomes, after the tools have been successfully implemented and used in the clinical practice [24, 25].
**Level of Evidence:** A numerical score, within each phase, is assigned based on the level of evidence associated with each tool. A tool is assigned grade C1 if it has been tested for external validity multiple times; C2 if it has been tested for external validity only once; and C3 if it has been tested only for internal validity. C0 means that the tool did not show sufficient internal validity to be used in clinical practice. Similarly, B1 is assigned to a predictive tool that has been evaluated during implementation for its usability; while if it has been studied for its potential effect on clinical effectiveness, patient safety or healthcare efficiency, it is assigned B2. Finally, if a predictive tool had been implemented then evaluated after implementation for its impact, on clinical effectiveness, patient safety or healthcare efficiency, it is assigned score A1 if there is at least one experimental study of good quality evaluating its impact, A2 if there are observational studies evaluating its impact and A3 if the impact has been evaluated through subjective studies, such as expert panel reports.

**Direction of Evidence:** For each phase and level of evidence, a direction of evidence is assigned based on the collective conclusions reported in the studies. The evidence is considered positive if all studies about a predictive tool reported positive conclusions and negative if all studies reported negative or equivocal conclusions. The evidence is considered mixed if some studies reported positive and some reported either negative or equivocal conclusions. To decide an overall direction of evidence, a protocol is used to sort the mixed evidence into supporting an overall positive conclusion or supporting an overall negative conclusion. The protocol is based on two main criteria; 1) The degree of matching between the evaluation study of the predictive tool and the original specifications of the predictive tool, in terms of the intended use of the predictive tool, predictive task and objectives, intended user, clinical area and specialty, target population including age group, gender, or disease, clinical settings, and type of practice, and 2) The quality of the evaluation study. Studies evaluating predictive tools in closely matching conditions to the tool specifications and providing high quality evidence are considered first for their conclusions in deciding the overall direction of evidence. The mixed evidence protocol is detailed and illustrated in Figure 7 in the Appendix.

The final grade assigned to a 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. It is true that the design of the framework looks hierarchical in the concept, however, it is not a simple linear rating system. A clinical predictive tool does not need to fulfill all lower grades first to move to higher grades. We only require that a tool be internally validated before we can use the GRASP framework to grade and assess this tool and assign a final grade based on the published evidence. A tool could have been only internally validated, fulfilling Grade C3, then could have been implemented with some experimental post-implementation studies showing its positive effectiveness, so it is now assigned Grade A1. In such case, the tool does not have to be evaluated for external validation, usability, or potential effect before being assigned post-implementation effectiveness grade. At the same time, the visual presentation of assigning the final grade passes through the three dimensions: phases of evaluation, level of evidence, and direction of evidence. The GRASP framework concept is shown in Figure 1, the various grading and assessment levels are shown in Table 1, and the GRASP framework detailed report is presented in Table 4 in the Appendix [19].
Table 1: The Various Grading & Assessment Levels of the GRASP framework

<table>
<thead>
<tr>
<th>Phase of Evaluation</th>
<th>Level of Evidence</th>
<th>Assigned Grade</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase C: Before implementation</td>
<td>Insufficient internal validation</td>
<td>C0</td>
<td>The tool was either insufficiently internally validated or validation was insufficiently reported.</td>
</tr>
<tr>
<td></td>
<td>Internal validation</td>
<td>C3</td>
<td>The tool was positively tested for internally validity (predictive performance measures).</td>
</tr>
<tr>
<td></td>
<td>External validation</td>
<td>C2</td>
<td>Tested once for external validity, using one external dataset.</td>
</tr>
<tr>
<td></td>
<td>External validation multiple times</td>
<td>C1</td>
<td>Tested multiple times for external validity, using more than one external dataset.</td>
</tr>
<tr>
<td>Phase B: During implementation</td>
<td>Potential effect</td>
<td>B2</td>
<td>Reported estimated potential effect on clinical effectiveness, patient safety or healthcare efficiency.</td>
</tr>
<tr>
<td></td>
<td>Usability</td>
<td>B1</td>
<td>Reported usability testing (effectiveness, efficiency, satisfaction, learnability, memorability, and minimizing errors).</td>
</tr>
<tr>
<td>Phase A: After implementation</td>
<td>Evaluation of post implementation impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency</td>
<td>A3</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2</td>
<td>Based on observational studies, e.g. a well-designed cohort or case-control study.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1</td>
<td>Based on experimental studies; properly designed, widely applied randomised/nonrandomised controlled trial.</td>
</tr>
</tbody>
</table>

Study Objectives

The primary objective of this study is to update the evaluation criteria of the GRASP framework through the feedback of a wide international group of experts. The secondary objective is to evaluate the initial reliability of the preliminary version of the GRASP framework to ensure that the outcomes produced by independent users, when grading predictive tools using the GRASP framework, are consistent and reliable.

Methods

The Study Design

The study is composed of two parts. The first part includes validating and updating the GRASP framework and the second part includes evaluating the framework reliability. For the first part, a survey was designed to solicit the feedback of experts on the criteria used by the GRASP framework for grading and assessment of predictive tools. The main outcome of this part is to measure the degree of agreement of each individual expert, using a Likert scale, with a series of statements regarding the GRASP framework evaluation criteria. The analysis includes calculating the average agreement of all the experts
with the framework evaluation criteria, including the three dimensions; phases of evaluation: before, during and after implementation, levels of evidence, and directions of evidence within each phase. In addition, experts’ qualitative feedback on adding, removing, or updating any of the criteria, used to grade predictive tools, and their further open-ended responses, suggestions, and recommendations will also be analysed and considered in updating the design and content of the GRASP framework. Figure 2 shows a flow diagram of the study design and the screenshots at the end of the Appendix show the details of the survey statements and questions.

Based on similar studies, validating and updating systems through surveying expert users, it was estimated that the required sample size for this study is around fifty experts [30-33]. Experts were identified as healthcare researchers who have published at least one paper on developing, implementing, or evaluating predictive tools and clinical decision support systems. Expert healthcare researchers were not required to be practicing clinicians, rather they should have experience in evidence-based methods and clinical decision support development and evaluation approaches. To search for such researchers and publications, the concepts of Clinical Decision Support, Clinical Prediction, Developing, Validating, Implementing, Evaluating, Comparing, Reviewing, Tools, Rules, Models, Algorithms, Systems, and Pathways were used. The search engines used included MEDLINE, EMBASE, CINAHL and Google Scholar. For the purpose of emails currency, the search was restricted to the last three years.

The 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. The authors expected the distribution of the survey to take two weeks, and the collection of the feedback to take another four weeks. The authors expected the response rate to be around 10%. Before the deployment of the survey a pilot testing was conducted through asking ten experts to take the survey. The feedback of the pilot testing was used to improve the survey design and content; some statements and questions were rephrased, some were rearranged, and some were supported by definitions and clarifications. Experts who participated in the pilot testing were excluded from the participation in the final survey. An invitation email, introducing the study objectives, the expected survey completion time, which was estimated at 20 minutes, and a participation consent was submitted to the identified experts with the link to the web-based survey. A reminder email, in two weeks, was sent to the experts who have not responded or completed the survey.

**The Study Survey**

The web-based survey was developed using Qualtrics experience management platform [34]. The survey, illustrated through the screenshots in the Appendix, included eight five-points Likert scale closed-ended agreement statements and six open-ended suggestions and recommendations questions arranged over seven sections. The first introduced the participants to the study objectives and the requested task. The second focused on the tools’ predictive performance before implementation, such as internal and external validation. The third focused on tools’ usability and/or potential effect during implementation. The fourth
focused on tools’ post-implementation impact. The fifth focused on the evaluation of the direction of the published evidence. Experts were also requested to provide their free-text feedback on adding, removing, or changing any of the criteria used for the assessment of phases of evaluation, levels of evidence, or directions of evidence. The sixth section requested the experts’ their free-text feedback on the best methods suggested to define and capture successful tools’ predictive performance. Experts were also asked about managing conflicting evidence of studies.

Reliability Testing

The second part of this study; evaluating the framework reliability, followed the completion of the first part and used the validated and updated version of the GRASP framework. Two independent and experienced researchers were trained, for four hours each by the authors, on using the framework to grade predictive tools. The two researchers hold PhD degrees in closely relevant health disciplines, have several relevant publications, and have extensive experience in conducting systematic reviews and using evidence-based methods. The researchers were asked to grade eight different predictive tools independently, using the validated and updated version of the GRASP framework, the full text studies describing the development of the tools, and the comprehensive list of all the published evidence on each tool along with the full text of each study.

The objective of this part of the study was to measure the reliability of using the framework, by independent users, to grade predictive tools. Since the tested function of the GRASP framework here is grading tools, the interrater reliability was the best measure to evaluate its reliability. The eight predictive tools were selected, before conducting the reliability testing and after examining and assigning grades to thirty candidate predictive tools, by the authors of this research, using the validated and updated version of the GRASP framework. The grading and assessment of the candidate predictive tools was done by the first author, MK, and revised by the second author, FM. Disagreements between the first and the second authors were resolved by the third author, BG. The eight predictive tools were selected because of their diversity, as they cover a wide range of GRASP framework grades; covering eight different grades of the designed ten grades of the framework; from Grade C0 up to Grade A1, which should support the validity of the interrater reliability evaluation. The interrater reliability, also called interrater agreement, interrater concordance, and interobserver reliability, is the degree of agreement or the score of how much homogeneity, or consensus, there is in the ratings given by independent reviewers [35].

Since the target ratings of the GRASP framework are ordinal, based on the available collected data, the correlation testing is an appropriate method for showing the interrater reliability. The Spearman's rank correlation coefficient is a nonparametric correlation estimator. It is widely used in the applied sciences and reported to be a robust measure of correlation [36]. This correlation coefficient is a test for a monotonic relationship, while in practice we may also be interested to capture concealed statistical relationships. Therefore some researchers prefer to add some divergence measures, which was not feasible using the available collected data [37]. After grading the tools, the two independent researchers
were asked to provide their open-ended feedback. Through a short five questions survey, they were asked if the GRASP framework design was logical, if they found it useful, easy to use, their opinion in the criteria used for grading, and if they wish to add, remove, or change any of them.

Analysis and Outcomes

Three major outcomes were planned. Firstly, through the eight closed-ended agreement statements of the survey, the average scores, and distributions of experts’ opinions on the different criteria used by the GRASP framework to grade and assess the predictive tools should help to improve such criteria. A five-points Likert scale ranging from strongly agree to strongly disagree was used, where the first was assigned the score of five and the last was assigned the score of one, to translate qualitative values into quantitative measures for the sake of the analysis [38, 39]. Secondly, the six open-ended free text questions should provide experts with the opportunity to suggest adding, removing, or updating any of the framework criteria. A qualitative content analysis was planned to search for common patterns, quantify the provided free-text responses, and measure the frequency and significance of repeated concept, ideas, and suggestions [40]. The qualitative content analysis should help categorise exerts feedback suggestions and recommendations into specific information and should also support updating the framework design and detailed content. The qualitative data analysis was conducted using the NVivo Version 12.3 software package [41].

To quantify the feedback of experts, their comments, suggestions, and recommendations were coded according to the relevant GRASP framework evaluation criteria. These included the predictive performance, usability, potential effect, and post-implementation impact, within the phase of the evaluation, level of evidence, and direction of evidence. An additional code was given for each feedback as positive or negative, then a score from 1 to 10, based the strength of the recommendation of each feedback. Thirdly, an interrater reliability testing was designed to measure how accurate and consistent the grading of eight predictive tools, conducted by two independent researchers, compared to each other, and compared to the grading of the same tools by the authors.

Results

The literature search generated a list of 1,186 relevant publications. A total of 882 unique emails were identified and extracted from the publications. In six weeks, from the 4th of October to the 15th of November 2018, a total of eighty-one valid responses were received from international experts, with a response rate of 9.2%. Valid responses were those completing all survey sections, responding to all statements, and answering all questions in each section. We also received seven invalid responses, which started the survey but did not complete all sections, missed some statements, or missed some questions on one or more sections.

Experts Agreement on GRASP Criteria
On average, the eighty-one respondents 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. Respondents strongly agreed with six of the eight closed-ended agreement statements. They somewhat agreed with one, and were neutral about another one, of the eight closed-ended agreement statements. Table 2 shows the average agreement score, standard deviation, and the 95% confidence interval of the respondents on each of the eight closed-ended statements. Figure 3 shows the averages and distributions of respondents’ agreements on each statement. The country distributions of the respondents are shown in Table 7 in the Appendix.

Table 2: Average Scores, Standard Deviations, and Confidence Intervals of Expert Respondents Agreement with the GRASP Framework Evaluation Criteria

<table>
<thead>
<tr>
<th>SN</th>
<th>Question</th>
<th>Mean Score</th>
<th>Meaning</th>
<th>SD</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Predictive Performance: We should consider the evidence on validating the tool’s predictive performance.</td>
<td>4.88</td>
<td>Strongly Agree</td>
<td>0.43</td>
<td>[4.87, 4.88]</td>
</tr>
<tr>
<td>2</td>
<td>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).</td>
<td>4.44</td>
<td>Strongly Agree</td>
<td>0.87</td>
<td>[4.44, 4.45]</td>
</tr>
<tr>
<td>3</td>
<td>Usability: We should consider the evidence on the tool’s usability.</td>
<td>4.68</td>
<td>Strongly Agree</td>
<td>0.70</td>
<td>[4.67, 4.68]</td>
</tr>
<tr>
<td>4</td>
<td>Potential Effect: We should consider the evidence on the tool’s potential effect.</td>
<td>4.62</td>
<td>Strongly Agree</td>
<td>0.68</td>
<td>[4.61, 4.62]</td>
</tr>
<tr>
<td>5</td>
<td>Usability is Higher: The evidence level on tools’ usability should be considered higher than the evidence level on tools’ potential effect.</td>
<td>2.96</td>
<td>Neither Agree nor Disagree</td>
<td>1.23</td>
<td>[2.95, 2.97]</td>
</tr>
<tr>
<td>6</td>
<td>Impact: We should consider the evidence on the tool’s impact on healthcare effectiveness, efficiency or safety.</td>
<td>4.78</td>
<td>Strongly Agree</td>
<td>0.57</td>
<td>[4.78, 4.79]</td>
</tr>
<tr>
<td>7</td>
<td>Evidence Levels of Post-Implementation Impact: The evidence level could be High (based on experimental studies), Medium (observational studies), or Low (subjective studies).</td>
<td>4.18</td>
<td>Somewhat Agree</td>
<td>1.14</td>
<td>[4.17, 4.19]</td>
</tr>
<tr>
<td>8</td>
<td>Evidence Direction: Based on the conclusions of published studies, the overall evidence direction could be Positive, Negative or Mixed.</td>
<td>4.25</td>
<td>Strongly Agree</td>
<td>0.78</td>
<td>[4.25, 4.26]</td>
</tr>
<tr>
<td></td>
<td>Overall Average</td>
<td>4.35</td>
<td>Strongly Agree</td>
<td>1.01</td>
<td>[4.349, 4.354]</td>
</tr>
</tbody>
</table>

Experts Comments, Suggestions and Recommendations

While the total valid responses were eighty-one; almost 80%; 64 respondents, provided their suggestions or discussed some recommendations for two or more of the six open-ended free text questions. These questions asked experts for their feedback regarding adding, removing or changing any of the GRASP
framework evaluation criteria, their feedback regarding defining and capturing successful tools’ predictive performance, when different clinical predictive tasks have different predictive requirements, and their feedback regarding managing conflicting evidence of studies while there is variability in the quality and specifications of published evidence. Using the coding strategy to quantify the feedback of experts, their responses were classified in relation to each of the GRASP framework evaluation criteria. Their responses were also highlighted as positive when they suggested to add to the existing criteria or positively change the wording of some of them. On the other hand, their responses were highlighted as negative, when they suggested to remove the existing criteria or negatively change the wording of some of them. The strength of their feedback, in terms of being important or less important was coded a score from 1 to 10. In addition to quantifying the qualitative feedback of experts, the authors worked together on deciding which suggestions were more valuable, feasible, and essential to be included in the GRASP framework evaluation criteria.

Predictive Performance and Performance Levels

Most of the respondents, 59/64, discussed that the method, type, and quality of internal and external validation studies should be reported in the GRASP framework detailed report. When external validation studies are conducted multiple times using different patient populations, in different healthcare settings, at different institutions, in different countries, over different times, or by different researchers then the tool is said to have a broad validation range, which means it is more reliable to be used across these different variations of healthcare settings. Some of the respondents, 21/64, said that the tool’s predictive performance is considered stable and reliable, when multiple external validation studies produce homogeneous predictive performances, e.g. similar sensitivities and specificities. They also discussed adding the concept of “Strength of Evidence”; which should be mainly based on the quality of the reported study and how much the conditions of the study are close to the original specifications of the predictive tool, in terms of clinical area, population, and target outcomes. It should be part of the components of deciding the direction of evidence (positive, negative, or mixed). It should also be reported in the detailed GRASP framework report, so that users can consider when selecting among two or more tools of the same assigned grade. For example, two predictive tools are assigned grade C1 (each was externally validated multiple times) but one of them shows a 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.

Usability and Potential Effect

Most of the respondents, 49/64, discussed that the methods and quality of the usability studies and the potential effect studies should be reported in the GRASP framework detailed report. Some of the respondents, 19/64, discussed that the potential effect and usability are not measured during implementation, rather they are measured during the planning for implementation, which is before wide-
scale implementation. They also suggested that the details on the potential effect should report the focus on clinical patient outcomes, healthcare outcomes, or provider behaviour. Most of the respondents, 51/64, said that the potential effect is more important than the usability and should have a higher evidence level. This feedback came basically from respondents who gave a low rate, “strongly disagree” and “disagree”, to statement number Q6: “The evidence level on tools' usability should be considered higher than the evidence level on tools' potential effect”. They argued that a highly usable tool that has no potential effect on healthcare is useless, while a less usable tool that has a promising potential effect is surely better. Some respondents, 11/64, discussed that evaluating both the potential effect and the usability should be considered together as a higher evidence than any of them alone.

Post-Implementation Impact and Impact Levels

Most of the respondents, 57/64, discussed that the method and quality of the post-implementation impact study should be reported in the GRASP framework detailed report. Again, some respondents, 13/64, discussed adding the concept of “Strength of Evidence”. Within each evidence level of the post-implementation impact we could have several sub-levels, or at least a classification of the quality of studies. For example, not all observational studies are equal in quality; a case series would be very different to a case control or large-scale prospective cohort study. Within the experimental studies we could also have different sub-levels of evidence, quasi-experimental vs. randomised controlled trial for example. These sub-levels should be included in the GRASP framework detailed report, when reporting the individual studies, this will provide the reader with more details on the strength and quality of the evidence on the tools.

Direction of Evidence

Most of the respondents, 47/64, discussed that the direction of evidence should consider the quality and strength of evidence. Most respondents here used the terms; “quality of evidence” and “strength of evidence”, synonymously. Many respondents, 38/64, discussed that quality of evidence or the strength of evidence should consider many elements of the published study, such as the methods used, the appropriate population, appropriate settings, the clinical practice, the sample size, the type of data collection; retrospective vs prospective, the outcomes, the institute of study and any other quality measures. The direction of evidence depends mainly on the quality and strength of the evidence in case there are conflicting conclusions from multiple studies.

Defining and Capturing Predictive Performance

Some of the respondents, 24/64, discussed that the predictive performance evaluation depends basically on the intended prediction task, so this is different from one tool to another, based on the task that each tool does. The clinical condition under prediction and the cost-effectiveness of treatment would highly
influence the predictive performance evaluation. Predictive performance evaluation depends also on the actions recommended based on the tool. For example, screening tools should perform with high sensitivity, high negative predictive value, and low likelihood ratio, since there is a following level of checking by clinicians or other tests, while diagnostic tools should always perform with high specificity, high positive predictive value, and high likelihood ratio, since the decisions are based here directly on the outcomes of the tool, and some of these decisions might be risky to the patient or expensive to the healthcare organisation. A few of the respondents, 7/64, discussed that for diagnostic tools, predictive performance is more likely to be expressed through sensitivity and specificity, while for prognostic tools, it is better to express predictive performance through probability/risk estimation. Predictive tools must always be adjusted to the settings, populations, and the intended tasks before their adoption and implementation in the clinical practice.

Managing Conflicting Evidence

Some of the respondents, 27/64, discussed that deciding on the conflicting evidence should consider the quality of each study or the strength of evidence, to decide on the overall direction of evidence. Measures include the proper methods used in the study, if the population is appropriate, if the settings are appropriate, if the study is conducted at the clinical practice, if the sample size is large, if the data collection was prospective not retrospective, if the outcomes are clearly reported, if the institute of the study is credible, if the study involved multiple sites or hospitals, and any other quality measures related to the methods or the data. We should rely primarily on conclusions from high-quality low risk of bias studies, as recommended in other fields, e.g. systematic reviews. A well designed and conducted study should have more credibility than a poorly designed and conducted study. If different results are obtained for sub-populations, this should be further investigated and explained. The predictive tool may only perform well in certain sub-populations, based on the intended tasks. If we have evidence from settings outside the target population of the tool, then these shouldn't have much weight, or less weight, on the evidence to support the tool, such as non-equivalent studies; which are conducted to validate a tool for a different population, predictive task, or clinical settings. Much of the important information is in the details of the evidence variability. So, it is important to report this in the framework detailed report, to provide as much details as possible for each reported study to help end users make more accurate decisions based on their own settings, intended tasks, target populations, practice priorities, and improvement objectives.

Updating the GRASP Framework

Based on the respondents’ feedback, on both the closed-ended evaluation criteria agreement statements and the open-ended suggestions and recommendations questions, the GRASP framework concept was updated, as shown in Figure 4. Regarding Phase C; the pre-implementation phase including the evidence on predictive performance evaluation, the three levels of internal validation, external validation once, and
external validation multiple times, were additionally assigned “Low Evidence”, “Medium Evidence”, and “High Evidence” labels respectively. Phase B: During Implementation, has been renamed to “Planning for Implementation”. The Potential Effect is now made of higher evidence level than Usability and the evidence of both potential effect and usability together is higher than any one of them alone. Now we have three levels of evidence; B1 = both potential effect and usability are reported, B2 = Potential effect evaluation is reported, and B3 = Usability testing is reported. Figure 5 in the Appendix shows a clean copy of the updated GRASP framework concept.

The GRASP framework detailed report was also updated, as shown in Table 5 in the Appendix. More details were added to the predictive tools information section, such as the internal validation method, dedicated support of research networks, programs, or professional groups, the total citations of the tool, number of studies discussing the tool, the number of authors, sample size used to develop the tool, the name of the journal which published the tool and its impact factor. Table 6 in the Appendix shows the Evidence Summary. This summary table provides users with more information in a structured format on each study discussing the tools, whether these were studies of predictive performance, usability, potential effect, or post-implementation impact. Information includes study name, country, year of development, and phase of evaluation. The evidence summary provides more quality related information, such as the study methods, the population and sample size, settings, practice, data collection method, and study outcomes. Furthermore, the evidence summary provides information on the strength of evidence and a label, to highlight the most prominent or important predictive functions, potential effects or post-implementation impacts of the tools.

We developed a new protocol to decide on the strength of evidence. The strength of evidence protocol considers two main criteria of the published studies. Firstly, it considers the degree of matching between the evaluation study conditions 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. Figure 8 in the Appendix shows the strength of evidence protocol.

The GRASP Framework Reliability

The two independent researchers assigned grades to the eight predictive tools and produced a detailed report on each one of them. The summary of the two independent researchers assigned grades, compared to the grades assigned by the authors of this study, are shown in Table 2. A more detailed information on the justification of the assigned grades is shown in the Appendix in Table 8. The Spearman's rank correlation coefficient was 0.994 (p<0.001) comparing the first researcher to the authors, 0.994 (p<0.001) comparing the second researcher to the authors, and 0.988 (p<0.001) comparing the two
researchers to each other. This shows a statistically significant and strong correlation, indicating a strong interrater reliability of the GRASP framework. Accordingly, the GRASP framework produced reliable and consistent grades when it was used by independent users. Providing their feedback to the five open-ended questions, after assigning the grades to the eight tools, both two independent researchers found GRASP framework design logical, easy to understand, and well organized. They both found GRASP useful, considering the variability of tools’ quality and levels of evidence. They both found it easy to use. They both thought the criteria used for grading were logical, clear, and well structured. They did not wish to add, remove, or change any of the criteria. However, they asked for adding some definitions and clarifications to the evaluation criteria of the GRASP framework, which was included in the update of the framework. The screenshot of the post-task questionnaire given to the two independent reviewers is shown in the Appendix.

Table 3: Grades Assigned by the Two Independent Researchers and the Paper Authors

<table>
<thead>
<tr>
<th>Tools</th>
<th>Grading by Researcher 1</th>
<th>Grading by Researcher 2</th>
<th>Grading by Paper Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centor Score [42]</td>
<td>B2</td>
<td>B3</td>
<td>B3</td>
</tr>
<tr>
<td>Dietrich Rule [44]</td>
<td>C0</td>
<td>C0</td>
<td>C0</td>
</tr>
<tr>
<td>LACE Index [45]</td>
<td>C1</td>
<td>C1</td>
<td>C1</td>
</tr>
<tr>
<td>Ottawa Knee Rule [47]</td>
<td>A1</td>
<td>A2</td>
<td>A1</td>
</tr>
<tr>
<td>PECARN Rule [48]</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
</tr>
<tr>
<td>Taylor Mortality Model [49]</td>
<td>C3</td>
<td>C3</td>
<td>C3</td>
</tr>
</tbody>
</table>

Discussion

Brief Summary

It is a challenging task for most clinicians to critically evaluate a growing number of predictive tools, proposed in the literature, in order to select effective tools for implementation at their clinical practice or for recommendation in clinical practice guidelines, to be used by other clinicians. Although most of these predictive tools have been assessed for predictive performance, only a few have been implemented and evaluated for comparative effectiveness or post-implementation impact. Clinicians need an evidence-based approach to provide them with standardised objective information on predictive tools to support their search for and selection of effective tools for their clinical tasks. To achieve this objective and overcome this major challenge, the GRASP framework was developed. The GRASP framework was first conceptualised by the authors in May 2017 and by March 2018 the development process was completed. Updating the GRASP framework through the feedback of a wide international group of experts was conducted in October 2018 and completed in March 2019. The GRASP framework concept, evaluation criteria, and the detailed report have been updated based on the feedback of experts. Using the updated version of the framework, the interrater reliability testing was conducted in May 2019. Positive results
showed that the GRASP framework can be used reliably and consistently by independent users to grade predictive tools.

**Predictive Performance**

The internal validation of the predictive performance of a tool is essential to make sure the tool is doing the prediction task as designed [17, 50]. The predictive performance is evaluated using measures of discrimination and calibration [51]. While discrimination refers to the ability of the tool to distinguish between patients with and without the outcome under consideration, calibration refers to the accuracy of the prediction, and show how much the predicted and the observed outcomes agree [52]. Discrimination is usually measured through sensitivity, specificity, and the area under the curve (AUC) [53]. On the other hand, calibration could be summarised using the Hosmer-Lemeshow test or the Brier score [54]. The external validation of predictive tools is essential to reflect the reliability and generalisability of the tools [55]. Predictive tools are more reliable and trustworthy, not only when their predictive performance is better, but more importantly when they undergo high quality, multiple, and wide range external validation [56]. The high quality is usually reflected in the type and size of data samples used in the validation, while repeating the external validations on different patient populations, at different institutions, in different healthcare settings, and by different researchers shows higher reliability of the predictive tools [17].

**Usability and Potential Effect**

In addition to the predictive performance, clinicians are usually interested to learn about the potential effects of the tools on improving patient outcomes, saving time, costs and resources, or supporting patient safety [57, 58]. They need to know more about the expected impact of using the tool on different healthcare aspects, processes or outcomes, assuming the tool has been successfully implemented in the clinical practice [22, 23]. If a CDS tool has less potential to improve healthcare processes or clinical outcomes it will not be easily adopted or successfully implemented in the clinical practice [59]. Some clinicians might also be interested to learn about the usability of predictive tools; whether these tools can be used by the specified users to achieve specified and quantifiable objectives in the specified context of use [20, 21]. CDS tools with poor usability will eventually fail, even if they provide the best performance or potential effect on healthcare [60, 61]. Usability includes several measurable criteria, based on the perspectives of the stakeholders, such the mental effort needed, the user attitude, interaction, easiness of use, and acceptability of systems [62, 63]. Usability can also be evaluated through measuring the effectiveness of task management with accuracy and completeness, the efficiency of utilising resources, and the users’ satisfaction, comfort with, and positive attitudes towards, the use of the tools [64, 65], in addition to learnability, memorability and freedom of errors [66, 67].

**Post-Implementation Impact**
Clinicians are interested to learn about the post-implementation impact of CDS tools, on different healthcare aspects, processes, and outcomes, before they consider their implementation in the clinical practice [24, 68, 69]. The most interesting part of the impact studies for clinicians is the effect size of the CDS tools and their direct impact on physicians’ performance and patients’ outcomes [70, 71]. Clinicians consider that high quality experimental studies, such as randomised controlled trials, are the highest level of evidence, followed by observational well-designed cohort or case-control studies and lastly subjective studies, opinions of respected authorities, and reports of expert committees or panels [72-74]. For many years, experimental methods have been viewed as the gold standard for evaluation, while observational methods were considered to have little or no value. However, this ignores the limitations of randomised controlled trials, which may prove unnecessary, inappropriate, inadequate, or sometimes impossible. Furthermore, high-quality observational studies have an important role in comparative effectiveness research because they can address issues that are otherwise difficult or impossible to study. Therefore, we need to understand the complementary roles of the two approaches and appreciate the scientific rigour in evaluation, regardless of the method used [75, 76].

**Direction of Evidence and Conflicting Conclusions**

It is not uncommon to encounter conflicting conclusions when a predictive tool is validated or implemented and evaluated in different patient subpopulations or for different prediction tasks or outcomes [77, 78]. The cut-off value that determines what a good predictive performance is, for example, depends not only on the clinical condition under consideration but largely on the requirements, conditions, and consequences of the decisions made accordingly [79]. When conducting systematic reviews and/or meta-analysis on a group of homogeneous predictive tools, predicting the same type of outcomes for the same subpopulation, then the task is feasible. On the other hand, when trying to develop a standard to define good performance, potential effect, or post-implementation impact, for different types of predictive tools, predicting different types of outcomes for different subpopulation, then the challenges are much greater. One of the main challenges here is dealing with the huge variability in the quality, types, and conditions of studies published in the literature. This variability makes it difficult to synthesise different measures of predictive performance, usability, potential effect or post-implementation impact into simple quantitative values, like in meta-analysis or systematic reviews [80, 81].

**The GRASP Framework Overall**

The grades assigned to predictive tools, using the GRASP framework, provide relevant evidence-based information to guide the selection of predictive tools for clinical decision support. However, the framework is not meant to be precisely prescriptive. An A1 tool is not always and absolutely better than an A2 tool. A clinician may prefer an A2 tool showing improved patient safety in two observational studies rather than an A1 tool showing reduced healthcare costs in three experimental studies. It all
depends on the objectives and priorities the users are trying to achieve, through implementing and using predictive tools in their clinical practice. More than one predictive tool could be endorsed, in clinical practice guidelines, each supported by its requirements and conditions of use and recommended for its most prominent outcome of predictive performance, potential effect, or post-implementation impact on healthcare and clinical outcomes.

The GRASP framework remains a high-level approach to provide clinicians with an evidence-based and comprehensive, yet simple and feasible, method to evaluate and select predictive tools. However, when clinicians need further information, the framework detailed report provides them with the required details to support their decision making. The GRASP framework is designed for two levels of users: 1) Expert users, such as healthcare researchers experienced in evidence-based evaluation methods. These will use the framework to critically evaluate published evidence, assign grades to predictive tools, and report their details. 2) End users, such as clinicians and healthcare professionals responsible for selecting tools for implementation at their clinical practice or for recommendation in clinical practice guidelines. These will use the GRASP framework detailed reports on tools and their assigned grades, produced by expert users, to compare existing predictive tools and select the most suitable tools for their intended tasks.

**Other Methods, Approaches, and Frameworks**

Some suggested approaches and frameworks are focusing on general methods of developing, validating, and implementing clinical predictive tools [4, 17, 82, 83]. Other approaches and frameworks are specifically focused on evaluating performance and external validation of predictive tools. For example, Steyerberg developed a framework that combine predictive tools performance evaluation measures and discussed methods of internal and external validation [50, 51, 55, 84]. Likewise, Debray developed a framework to improve the interpretation of external validation studies [77, 85]. Similarly, Collins developed a method to report external validation of predictive tools [56]. Moreover, the TRIPOD statement [86, 87] and the CHARMS checklist [88] both critically appraise the existing evidence on the performance of predictive tools. The first suggests a set of recommendations to report studies developing, validating, or updating predictive tools and the second provides guidance on critical appraisal and data extraction for systematic reviews of predictive tools.

Furthermore, other approaches and frameworks focused on describing methods of to evaluate the post-implementation impact of predictive tools. A framework for predictive tools implementation impact analysis was developed by Wallace [89, 90]. A similar framework to evaluate the impact of predictive tools on clinical outcomes was developed by Harris [91]. Toll also suggested an approach to report the impact of predictive tools [78]. However, none of the available methods, designed to evaluate the performance or impact of predictive tools, provide a grading system to allow comparing the tools. On a wider level and through evaluating the published evidence, the GRADE framework grades the quality of published evidence and strength of clinical recommendations, in terms of their post-implementation impact [72-74, 92, 93].
Challenges, Limitations, and Future Work

It might be easy to analyse the feedback of experts using closed-ended statements and questions. However, analysing the feedback of experts using open-ended questions is rather difficult [94]. Qualitative content and thematic analysis of free text feedback is challenging, since the extraction of significance becomes more difficult with diverse opinions, different experiences, and variable perspectives [40]. As another different challenge, it is advised by many healthcare researchers to use Delphi techniques to reach to consensus among experts, through successive rounds of feedback, when developing clinical guidelines or selecting evaluation criteria and indicators [95-97]. However, many researchers recommend that Delphi method panels should include a number between ten and fifty members, as this count is more appropriate and realistic, considering the large amount of the collected data and the subsequent multiple rounds of analysis each panellist generates [98]. Based on our expectation, that the number of respondents would be almost a hundred, and due to time and resources limitations, we preferred using a single round of a web-based survey of experts’ feedback.

Even though we have contacted a large number of 882 experts, in the area of developing, implementing and evaluating predictive tools and CDS systems, we got a very low response rate of 9.2%, and received only 81 valid responses. This low response rate could have been improved if participants were motivated by some incentives, more than just acknowledging their participation in the study, or if more support was provided through the organisations these participants belong to, which needs much more resources to synchronise these efforts. For the sake of keeping the survey feasible for most busy experts, the number of the closed ended as well as the open-ended questions were kept limited and the required time to complete the whole survey was kept in the range of 20 minutes. However, some of the participants could have been willing to provide more detailed feedback, through interviews for example, which was out of the scope of this study and was not initially possible to conduct with all the invited experts, otherwise we would have received a much lower response rate.

One of the challenges facing GRASP framework is the proper evaluation of the quality of the included evidence, which represents the core foundation of the grades assigned to the predictive tools. The mixed evidence protocol, the evidence summary tables, and the strength of evidence protocol are GRASP tools used to evaluate the quality of studies before considering their evidence in grading predictive tools. However, it is wise to use all available methods, frameworks, approaches, guidelines, and tools to support the GRASP framework through ensuring the quality, robustness, and comprehensiveness of the included evidence. These include Debray framework [77, 85], Steyerberg framework [50, 51, 55, 84], Collins approach [56], the TRIPOD statement [86, 87] and the CHARMS checklist [88], Wallace framework [89, 90], Harris framework [91], Toll approach [78], and the GRADE guidelines [72-74, 92, 93].

To evaluate the impact of the GRASP framework on clinicians’ decisions and examine the application of the framework to grade predictive tools, the authors are currently working on two more studies. The first study should evaluate the impact of using the framework on improving the decisions made by end user clinicians and healthcare professionals, regarding selecting predictive tools for their clinical tasks.
Through a web-based survey of a wide international group of clinicians and healthcare professionals, the study should compare the performance and outcomes of making the selection decisions with and without using the framework. The first study should also evaluate the usability and usefulness of the GRASP framework on a wider scale, seeking the feedback of over a hundred clinicians and healthcare professionals, which should give a more accurate and reliable measure of usability and usefulness, than the feedback provided currently by only two researchers. The second study aims to apply the framework to a large consistent group of predictive tools, used for the same clinical prediction task. This study should show how the framework provides clinicians with an evidence-based method to compare, evaluate, and select predictive tools, through grading and reporting tools based on the critical appraisal of their published evidence. Furthermore, the relative complexity of the GRASP framework makes the simple graphical presentation of the GRASP concept, and the detailed report, not sufficient by themselves to provide adequate guidance for individuals seeking to use the framework independently and reliably. This motivates the authors to develop and publish a detailed instruction manual to provide step-by-step guidance on how to use the GRASP framework and how to grade and assess predictive tools using evidence accurately and consistently.

Conclusion

The GRASP framework aims to provide clinicians with a high-level, evidence-based, and comprehensive, yet simple and feasible, approach to evaluate and compare clinical predictive tools, considering their predictive performance before implementation, potential effect and usability during planning for implementation, and post-implementation impact on healthcare processes and clinical outcomes. To enable end user clinicians and clinical practice guideline developers to access detailed information, reported evidence and assigned grades of predictive tools, it is essential to discuss implementing the GRASP framework into a web-based platform. However, maintaining such grading system up to date is a challenging task, as this requires the continuous updating of the predictive tools grading and assessments, when new evidence becomes published and available. It is important to discuss using automated or semi-automated methods for searching and processing new information to keep the GRASP framework updated. Furthermore, we recommend that the GRASP framework be utilised by working groups of experts of professional organisations to grade predictive tools, in order to provide consistent results and increase reliability and credibility for end users. These professional organisations should also support disseminating such evidence-based information on predictive tools, in a similar way of announcing and disseminating new updates of clinical practice guidelines.

Finally, the GRASP framework is still in the development phase. This is a first pass experiment, to update the framework and its evaluation criteria, and further work will confirm what else needs to be added, removed, or changed. Wider scale validation studies are needed to fine-tune and improve the framework. Accordingly, we can only claim its validity and reliability after it has been used, over a period of time, by the real expert users and end users; using the framework to grade predictive tools and then using the assigned grades and detailed reports to select the most effective tools.
Abbreviations


Declarations

Acknowledgments


Funding

This work was financially supported by the Commonwealth Government Funded Research Training Program, Australia, to help the authors to carry out the study. The funding body played no role in the design of the study and collection, analysis, and interpretation of the data and in writing the manuscript.

Availability of data and materials

Not applicable.

Authors’ contributions

MK mainly contributed to the conception, detailed design, and conduction of the study. BG and FM supervised the study from the scientific perspective. BG was responsible for the overall supervision of the work done, while FM was responsible for providing advice on the enhancement of the methodology used and data analysis conducted. All the authors have been involved in drafting the manuscript and revising
it. Finally, all the authors gave approval of the manuscript to be published and agreed to be accountable for all aspects of the work.

**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.

**Consent to publish**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

Mohamed Khalifa, MD, MSc, MRCSEd, CPHIMS

Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine and Health Sciences, Macquarie University, 75 Talavera Rd, North Ryde

Sydney, NSW 2113, Australia, Email: dr.m.khalifa@gmail.com

Farah Magrabi, PhD

Associate Professor, Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine and Health Sciences, Macquarie University, 75 Talavera Rd, North Ryde, Sydney, NSW 2113, Australia, Email: farah.magrabi@mq.edu.au

Blanca Gallego Luxan, PhD

Associate Professor, Head of Clinical Machine Learning Unit, Centre for Big Data Research in Health, Faculty of Medicine, University of New South Wales, Lowy Cancer Research Centre, Cnr High & Botany St, Kensington NSW 2052, Australia, Email: b.gallego@unsw.edu.au

**References**


34. Qualtrics Experience Management Solutions, Q. *Qualtrics Experience Management Solutions, Qualtrics*. 2018 [cited 2018 1 January]; Available from: [https://www.qualtrics.com/](https://www.qualtrics.com/).


**Figures**
Figure 1

The GRASP Framework Concept
Figure 2

Flow Diagram of the Study Design
Figure 3

Averages & distributions of respondents’ agreements: A - Predictive Performance, B - Performance Levels, C - Usability, D - Potential Effect, E - Usability is Higher than Potential Effect, F - Impact, G - Impact Levels, and H - Evidence Direction
Figure 4
The Updated GRASP Framework Concept

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.pdf