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: March 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-15929/v1

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. To overcome this challenge, 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: An online survey was developed to collect responses of a wide international group of experts, who published studies on developing, implementing or evaluating clinical decision support tools. The interrater reliability of the framework, to assign grades to eight predictive tools by two independent users, was evaluated.

Results: Among 882 invited experts, 81 provided valid responses. On a five-points Likert scale, experts overall strongly agreed to GRASP evaluation criteria (4.35). Experts strongly agreed to six criteria: predictive performance (4.87) and predictive performance levels (4.44), usability (4.68) and potential effect (4.61), post-implementation impact (4.78) and evidence direction (4.26). Experts somewhat agreed to one criterion: post-implementation impact levels (4.16). Experts were neutral about one criterion; usability is higher than potential effect (2.97). Sixty-four respondents provided recommendations to open-ended questions regarding adding, removing or changing evaluation criteria. Forty-three respondents suggested the potential effect should be higher than the usability. Experts highlighted the importance of reporting quality of studies and strength of evidence supporting grades assigned to predictive tools. Accordingly, GRASP concept and its detailed report were updated. The updated framework's interrater reliability, to produce accurate and consistent results by two independent users, was tested and found to be initially reliable.

Conclusion: The GRASP framework grades predictive tools based on critical appraisal of published evidence across three dimensions: phase of evaluation, level of evidence, and direction of evidence. The final grade of a tool is based on the highest phase of evaluation, supported by the highest level of positive evidence, or mixed evidence that supports positive conclusion. GRASP aims to provide clinicians with a high-level, evidence-based, and comprehensive, yet simple and feasible, approach to evaluate predictive tools, considering their predictive performance before implementation, usability and potential effect during planning for implementation, and post-implementation impact on healthcare outcomes.

Background

Clinical decision support (CDS) systems have been proved to enhance evidence-based clinical practice and support healthcare cost-effectiveness [1–6]. According to the definition developed by Edward Shortliffe, there are three levels of CDS functions, these include; 1) managing health information through providing tools for search and retrieval, 2) focusing users' attention through flagging abnormal values or possible drug-to-drug interactions, and 3) providing patient specific recommendations based on the

clinical scenario, which usually follow rules and algorithms, cost-benefit analysis or clinical pathways [7, 8]. Clinical predictive tools, here referred to simply as predictive tools, belong to the third level of CDS and include various developed applications; ranging from the simplest manual clinical prediction rules to the most sophisticated machine learning algorithms [9, 10]. These research-based applications provide clinicians and other healthcare professionals with diagnostic, prognostic, and therapeutic decision support through predicting clinical and other related healthcare outcomes. 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 [11, 12].

Many studies discuss that there is an inappropriate but common practice of developing new predictive tools instead of validating or updating existing ones [12–16]. 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 [17–21]. Currently, clinicians rely on their previous experience, subjective evaluation or recent exposure to predictive tools in making selection decisions. Objective methods and evidence based approached are rarely used in such decisions [22, 23]. Some clinicians, especially those developing clinical 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 [24–26].

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) [27]. 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. If a tool’s usability and/or potential effect have been tested, it is assigned phase B. Finally, if a tool has been implemented in clinical practice, and there is published evidence evaluating its impact, it is assigned phase A.

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 conditions and the original predictive tool specifications, 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 Fig. 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. The GRASP framework concept is shown in Fig. 1 and the GRASP framework detailed report is presented in Table 3 in the Appendix [27].

Study Objectives

Updating new clinical instruments, healthcare models and evaluation frameworks through the feedback of experts is a well-established approach, especially in the area of CDS [12, 28, 29]. Using a mixed approach of qualitative and quantitative methods in research proved to be useful in healthcare, because of the complexity of the studied topics [30]. Using open-ended questions in quantitative surveys adds significant value and depth to both the results and conclusions of studies conducted [31]. The aim of this study is to update the GRASP framework and evaluate its initial reliability. The primary objective is to update the criteria used by the GRASP framework, for grading and assessment of predictive tools,

through the feedback of a wide international group of healthcare experts in the areas of developing, implementing and evaluating clinical decision support systems and predictive tools. The secondary objective is to evaluate the initial reliability of this 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 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 of the study is to measure the experts' agreement and update the design and content of the GRASP framework. The analysis includes evaluating the degree of agreement of experts on how essential the different criteria used to grade predictive tools are, 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' feedback on adding, removing or updating any of the criteria, used to grade predictive tools, and their further suggestions and recommendations will also be analysed and considered. Figure 2 shows a flow diagram of the study design.

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 [32–35]. Experts were identified as researchers who have published at least one paper on developing, implementing or evaluating predictive tools and clinical decision support systems. Expert 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 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 survey completion time, which was estimated at 20 minutes, and a participation consent was submitted to the identified experts with the link to the online survey. A reminder email, in two weeks, was sent to the experts who have not responded or completed the survey.

The Study Survey

The online survey was developed using Qualtrics experience management platform [36]. The online survey, as illustrated through the screenshots in the Appendix, included eight five-points Likert scale closed-ended agreement questions and six open-ended suggestions and recommendations questions distributed over seven sections. The introduction informed the participants about the aim of developing the GRASP framework, its design, and the task they are requested to complete. In addition to informing them that they can request feedback and acknowledgement as well as providing them with contacts to ask for further information or to make complaints. The second section asked experts about their level of agreement with the evaluation of the published evidence on the tools' predictive performance before implementation, such as internal and external validation. The third section asked experts about their level of agreement with the evaluation of the published evidence on the tools' usability and/or potential effect on healthcare during implementation.

The fourth section asked experts about their level of agreement with the evaluation of the published evidence on the tools' post-implementation impact on clinical effectiveness, patient safety or healthcare efficiency. The fifth section asked experts about their level of agreement with the evaluation of the direction of the published evidence on the tools, being positive, negative or mixed. 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 asked experts to provide their free-text feedback on the best methods suggested to define and capture successful tools' predictive performance, when different clinical prediction tasks have different predictive performance requirements. Experts were also asked about managing conflicting evidence of studies when there is variability in the quality and/or sub-populations of the published evidence.

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 long experience in evidence-based methods and systematic reviews. 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 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 [37]. 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 [38]. 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 [39]. 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 questions 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 [40, 41]. 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. The qualitative analysis should help categorise such 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 [42]. 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%.

Experts Agreement on GRASP Criteria

The overall average agreement of the eighty-one respondents to the eight closed-ended questions was 4.35, which means the respondents overall strongly agreed to the criteria of the GRASP framework. Respondents strongly agreed to six of the eight closed-ended agreement questions, regarding the criteria used by the GRASP framework for evaluating predictive tools. They somewhat agreed to one, and were neutral about another one, of the eight closed-ended agreement questions. Table 1 shows the average agreements of the respondents on each of the eight closed-ended questions and Figure 3 shows the averages and distributions of respondents' agreements on each question. The country distributions of the respondents are shown in Table 6 in the Appendix.

Table 1: Average Agreements of Expert Respondents on Framework Criteria

SN	Question	Score	Meaning
1	Predictive Performance: We should consider the evidence on validating the tool's predictive performance.	4.87	Strongly Agree
2	Predictive Performance Level: The evidence level could be High (internal + multiple external validation), Medium (internal + external validation once), or Low (internal validation only).	4.44	Strongly Agree
3	Usability: We should consider the evidence on the tool's usability.	4.68	Strongly Agree
4	Potential Effect: We should consider the evidence on the tool's potential effect.	4.61	Strongly Agree
5	Usability is Higher: The evidence level on tools' usability should be considered higher than the evidence level on tools' potential effect.	2.97	Neither Agree nor Disagree
6	Impact: We should consider the evidence on the tool's impact on healthcare effectiveness, efficiency or safety.	4.78	Strongly Agree
7	Impact Level: The evidence level could be High (based on experimental studies), Medium (observational studies), or Low (subjective studies).	4.16	Somewhat Agree
8	Evidence Direction: Based on the conclusions of published studies, the overall evidence direction could be Positive, Negative or Mixed.	4.26	Strongly Agree
Overall Average		**4.35**	**Strongly Agree**

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.

Predictive Performance and Performance Levels

The respondents 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. The respondents 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

The respondents 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 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 said that the potential effect is more important than the usability and should have a higher evidence level. 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 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

The respondents discussed that the method and quality of the post-implementation impact study should be reported in the GRASP framework detailed report. Again, respondents 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

Respondents 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. Respondents 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 largely on the quality of the evidence, in case there are conflicting conclusions from multiple studies.

Defining and Capturing Predictive Performance

Respondents 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. Respondents 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

Respondents 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 questions 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 4 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 5 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 7. 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.

Table 2: Grades Assigned by the Two Independent Researchers and the Paper Authors

Tools	Grading by Researcher 1	Grading by Researcher 2	Grading by Paper Authors
Centor Score [43]	**B2**	**B3**	**B3**
CHALICE Rule [44]	**B2**	**B2**	**B2**
Dietrich Rule [45]	**C0**	**C0**	**C0**
LACE Index [46]	**C1**	**C1**	**C1**
Manuck Scoring System [47]	**C2**	**C2**	**C2**
Ottawa Knee Rule [48]	**A1**	**A2**	**A1**
PECARN Rule [49]	**A2**	**A2**	**A2**
Taylor Mortality Model [50]	**C3**	**C3**	**C3**

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 [25, 51]. The predictive performance is evaluated using measures of discrimination and calibration [52]. 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 [53]. Discrimination is usually measured through sensitivity, specificity, and the area under the curve (AUC) [54]. On the other hand, calibration could be summarised using the Hosmer-Lemeshow test or the Brier score [55]. The external validation of predictive tools is essential to reflect the reliability and generalisability of the tools [56]. 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 [57]. 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 [25].

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 [2, 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 [59, 60]. 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 [61]. 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 [62, 63]. CDS tools with poor usability will eventually fail, even if they provide the best performance or potential effect on healthcare [7, 64]. 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 [65, 66]. 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 [67, 68], in addition to learnability, memorability and freedom of errors [69, 70].

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 [71–73]. 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 [74, 75]. 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 [76–78]. 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 [79, 80].

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 [81, 82]. 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 [83]. 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 impossible 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 [84, 85].

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.

Challenges, Limitations, and Future Work

It might be easy to analyse the feedback of experts using closed-ended questions. However, analysing the feedback of experts using open-ended questions is rather difficult [86]. 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 [87]. 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 [88–90]. 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 [91]. 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 an online 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.

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 an online 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.

Conclusion

The GRASP framework grades predictive tools based on the critical appraisal of the published evidence across three dimensions: 1) Phase of evaluation; 2) Level of evidence; and 3) Direction of evidence. The final grade of a tool is based on the highest phase of evaluation, supported by the highest level of positive evidence, or mixed evidence that supports positive 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 an online 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. 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

CDS: Clinical Decision Support. CHALICE: Children's Head Injury Algorithm for the Prediction of Important Clinical Events. E.g.: For example. GRASP: Grading and Assessment of Predictive tools. LACE: Length of Stay, Admission Acuity, Comorbidity and Emergency Department Visits. PECARN: Paediatric Emergency Care Applied Research Network.

Declarations

Acknowledgments

We would like to 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, NadÃ¨ge 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.

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, Univeristy of New South Wales, Lowy Cancer Research Centre, Cnr High &, Botany St, Kensington NSW 2052, Australia, Email: b.gallego@unsw.edu.au

References

1. Chaudhry, B., et al., *Systematic review: impact of health information technology on quality, efficiency, and costs of medical care.* Ann Intern Med, 2006. **144**(10): p. 742-52.
2. Garg, A.X., et al., *Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review.* Jama, 2005. **293**(10): p. 1223-1238.
3. Kawamoto, K., et al., *Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success.* Bmj, 2005. **330**(7494): p. 765.
4. Oman, K.S., *Evidence-based practice: An implementation guide for healthcare organizations*. 2010: Jones & Bartlett Publishers.
5. Osheroff, J.A. *Improving outcomes with clinical decision support: an implementer's guide*. 2012. Himss.

6. Osheroff, J.A., et al., *A roadmap for national action on clinical decision support.* Journal of the American medical informatics association, 2007. **14**(2): p. 141-145.
7. Musen, M.A., B. Middleton, and R.A. Greenes, *Clinical decision-support systems*, in *Biomedical informatics*. 2014, Springer. p. 643-674.
8. Shortliffe, E.H. and J.J. Cimino, *Biomedical informatics: computer applications in health care and biomedicine*. 2013: Springer Science & Business Media.
9. Adams, S.T. and S.H. Leveson, *Clinical prediction rules.* Bmj, 2012. **344**: p. d8312.
10. Wasson, J.H., et al., *Clinical prediction rules: applications and methodological standards.* New England Journal of Medicine, 1985. **313**(13): p. 793-799.
11. Beattie, P. and R. Nelson, *Clinical prediction rules: what are they and what do they tell us?* Australian Journal of Physiotherapy, 2006. **52**(3): p. 157-163.
12. Steyerberg, E.W., *Clinical prediction models: a practical approach to development, validation, and updating*. 2008: Springer Science & Business Media.
13. Altman, D.G., et al., *Prognosis and prognostic research: validating a prognostic model.* Bmj, 2009. **338**: p. b605.
14. Bouwmeester, W., et al., *Reporting and methods in clinical prediction research: a systematic review.* PLoS medicine, 2012. **9**(5): p. e1001221.
15. Hendriksen, J., et al., *Diagnostic and prognostic prediction models.* Journal of Thrombosis and Haemostasis, 2013. **11**(s1): p. 129-141.
16. Moons, K.G., et al., *Prognosis and prognostic research: what, why, and how?* Bmj, 2009. **338**: p. b375.
17. Ebell, M.H., *Evidence-based diagnosis: a handbook of clinical prediction rules*. Vol. 1. 2001: Springer Science & Business Media.
18. Babl, F.E., et al., *Accuracy of PECARN, CATCH, and CHALICE head injury decision rules in children: a prospective cohort study.* The Lancet, 2017. **389**(10087): p. 2393-2402.
19. Easter, J.S., et al., *Comparison of PECARN, CATCH, and CHALICE rules for children with minor head injury: a prospective cohort study.* Annals of emergency medicine, 2014. **64**(2): p. 145-152. e5.
20. Kappen, T., et al., *General Discussion I: Evaluating the Impact of the Use of Prediction Models in Clinical Practice: Challenges and Recommendations.* Prediction Models and Decision Support, 2015: p. 89.
21. Taljaard, M., et al., *Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol.* BMJ open, 2014. **4**(10): p. e006701.
22. Ansari, S. and A. Rashidian, *Guidelines for guidelines: are they up to the task? A comparative assessment of clinical practice guideline development handbooks.* PloS one, 2012. **7**(11): p. e49864.
23. Kish, M.A., *Guide to development of practice guidelines.* Clinical Infectious Diseases, 2001. **32**(6): p. 851-854.
24. Shekelle, P.G., et al., *Developing clinical guidelines.* Western Journal of Medicine, 1999. **170**(6): p. 348.

25. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* European heart journal, 2014. **35**(29): p. 1925-1931.
26. Tranfield, D., D. Denyer, and P. Smart, *Towards a methodology for developing evidence-informed management knowledge by means of systematic review.* British journal of management, 2003. **14**(3): p. 207-222.
27. Khalifa, M., F. Magrabi, and B. Gallego, *Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support.* BMC medical informatics and decision making, 2019. **19**(1): p. 207.
28. McCoy, A.B., et al., *A framework for evaluating the appropriateness of clinical decision support alerts and responses.* Journal of the American Medical Informatics Association, 2011. **19**(3): p. 346-352.
29. Kung, J., et al., *From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (R-AMSTAR) for grading of clinical relevance.* The open dentistry journal, 2010. **4**: p. 84.
30. Östlund, U., et al., *Combining qualitative and quantitative research within mixed method research designs: a methodological review.* International journal of nursing studies, 2011. **48**(3): p. 369-383.
31. Harland, N. and E. Holey, *Including open-ended questions in quantitative questionnaires—theory and practice.* International Journal of Therapy and Rehabilitation, 2011. **18**(9): p. 482-486.
32. Bond, R.R., et al., *A usability evaluation of medical software at an expert conference setting.* Computer methods and programs in biomedicine, 2014. **113**(1): p. 383-395.
33. Hamborg, K.-C., B. Vehse, and H.-B. Bludau, *Questionnaire based usability evaluation of hospital information systems.* Electronic journal of information systems evaluation, 2004. **7**(1): p. 21-30.
34. Lehrer, D. and J. Vasudev, *Visualizing Information to Improve Building Performance: A study of expert users.* 2010.
35. Santiago-Delefosse, M., et al., *Quality of qualitative research in the health sciences: Analysis of the common criteria present in 58 assessment guidelines by expert users.* Social Science & Medicine, 2016. **148**: p. 142-151.
36. Qualtrics Experience Management Solutions, Q. *Qualtrics Experience Management Solutions, Qualtrics*. 2018 [cited 2018 1 January]; Available from: https://www.qualtrics.com/.
37. Tinsley, H.E. and D.J. Weiss, *Interrater reliability and agreement*, in *Handbook of applied multivariate statistics and mathematical modeling*. 2000, Elsevier. p. 95-124.
38. Croux, C. and C. Dehon, *Influence functions of the Spearman and Kendall correlation measures.* Statistical methods & applications, 2010. **19**(4): p. 497-515.
39. Pardo, L., *Statistical inference based on divergence measures*. 2018: Chapman and Hall/CRC.
40. Allen, I.E. and C.A. Seaman, *Likert scales and data analyses.* Quality progress, 2007. **40**(7): p. 64-65.
41. Boone, H.N. and D.A. Boone, *Analyzing likert data.* Journal of extension, 2012. **50**(2): p. 1-5.
42. Bazeley, P. and K. Jackson, *Qualitative data analysis with NVivo*. 2013: Sage Publications Limited.

43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. Steyerberg, E.W. and F.E. Harrell Jr, *Prediction models need appropriate internal, internal-external, and external validation.* Journal of clinical epidemiology, 2016. **69**: p. 245.
52. Steyerberg, E.W., et al., *Internal validation of predictive models: efficiency of some procedures for logistic regression analysis.* Journal of clinical epidemiology, 2001. **54**(8): p. 774-781.
53. Alba, A.C., et al., *Discrimination and calibration of clinical prediction models: users' guides to the medical literature.* Jama, 2017. **318**(14): p. 1377-1384.
54. Hajian-Tilaki, K., *Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation.* Caspian journal of internal medicine, 2013. **4**(2): p. 627.
55. Schmid, C.H. and J.L. Griffith, *Multivariate classification rules: calibration and discrimination.* Encyclopedia of biostatistics, 2005. **5**.
56. Steyerberg, E.W., et al., *Internal and external validation of predictive models: a simulation study of bias and precision in small samples.* Journal of clinical epidemiology, 2003. **56**(5): p. 441-447.
57. Collins, G.S., et al., *External validation of multivariable prediction models: a systematic review of methodological conduct and reporting.* BMC medical research methodology, 2014. **14**(1): p. 40.
58. Bates, D.W., et al., *Big data in health care: using analytics to identify and manage high-risk and high-cost patients.* Health Affairs, 2014. **33**(7): p. 1123-1131.

59. Friedman, C.P. and J.C. Wyatt, *Evaluation of Biomedical and Health Information Resources*, in *Biomedical Informatics*. 2014, Springer. p. 355-387.
60. Friedman, C.P. and J.C. Wyatt, *Evaluation methods in medical informatics*. 2013: Springer Science & Business Media.
61. Bright, T.J., et al., *Effect of clinical decision-support systemsa systematic review.* Annals of internal medicine, 2012. **157**(1): p. 29-43.
62. Bevan, N., *Measuring usability as quality of use.* Software Quality Journal, 1995. **4**(2): p. 115-130.
63. Bevan, N. and M. Macleod, *Usability measurement in context.* Behaviour & information technology, 1994. **13**(1-2): p. 132-145.
64. Li, A.C., et al., *Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support.* International journal of medical informatics, 2012. **81**(11): p. 761-772.
65. Bevan, N., *Usability*, in *Encyclopedia of Database Systems*. 2009, Springer. p. 3247-3251.
66. Dix, A., *Human-computer interaction*, in *Encyclopedia of database systems*. 2009, Springer. p. 1327-1331.
67. Frøkjær, E., M. Hertzum, and K. Hornbæk. *Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?* in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2000. ACM.
68. Khajouei, R., et al., *Clinicians satisfaction with CPOE ease of use and effect on clinicians' workflow, efficiency and medication safety.* international journal of medical informatics, 2011. **80**(5): p. 297-309.
69. Jeng, J., *Usability assessment of academic digital libraries: effectiveness, efficiency, satisfaction, and learnability.* Libri, 2005. **55**(2-3): p. 96-121.
70. Nielsen, J., *Usability metrics: Tracking interface improvements.* Ieee Software, 1996. **13**(6): p. 12.
71. Kaplan, B., *Evaluating informatics applications—clinical decision support systems literature review.* International journal of medical informatics, 2001. **64**(1): p. 15-37.
72. Bates, D.W., et al., *Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality.* Journal of the American Medical Informatics Association, 2003. **10**(6): p. 523-530.
73. Lorenzi, N.M., et al., *How to successfully select and implement electronic health records (EHR) in small ambulatory practice settings.* BMC medical informatics and decision making, 2009. **9**(1): p. 15.
74. Blum, D., et al., *Computer-based clinical decision support systems and patient-reported outcomes: a systematic review.* The Patient-Patient-Centered Outcomes Research, 2015. **8**(5): p. 397-409.
75. Hunt, D.L., et al., *Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review.* Jama, 1998. **280**(15): p. 1339-1346.
76. Guyatt, G., et al., *GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables.* Journal of clinical epidemiology, 2011. **64**(4): p. 383-394.

77. Guyatt, G.H., et al., *GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology.* Journal of clinical epidemiology, 2011. **64**(4): p. 380-382.
78. Guyatt, G.H., et al., *Rating quality of evidence and strength of recommendations: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations.* BMJ: British Medical Journal, 2008. **336**(7650): p. 924.
79. Black, N., *Why we need observational studies to evaluate the effectiveness of health care.* Bmj, 1996. **312**(7040): p. 1215-1218.
80. Dreyer, N.A., et al., *Why observational studies should be among the tools used in comparative effectiveness research.* Health Affairs, 2010. **29**(10): p. 1818-1825.
81. Debray, T.P., et al., *A new framework to enhance the interpretation of external validation studies of clinical prediction models.* Journal of clinical epidemiology, 2015. **68**(3): p. 279-289.
82. Toll, D., et al., *Validation, updating and impact of clinical prediction rules: a review.* Journal of clinical epidemiology, 2008. **61**(11): p. 1085-1094.
83. Lalkhen, A.G. and A. McCluskey, *Clinical tests: sensitivity and specificity.* Continuing Education in Anaesthesia Critical Care & Pain, 2008. **8**(6): p. 221-223.
84. Bowling, A., *Research methods in health: investigating health and health services*. 2014: McGraw-hill education (UK).
85. Petitti, D.B., *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine*. 2000: OUP USA.
86. Nagy, S., et al., *Using research in healthcare practice*. 2010: Lippincott, Williams and Wilkins.
87. Vaismoradi, M., H. Turunen, and T. Bondas, *Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study.* Nursing & health sciences, 2013. **15**(3): p. 398-405.
88. Boulkedid, R., et al., *Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review.* PloS one, 2011. **6**(6): p. e20476.
89. Falzarano, M. and G.P. Zipp, *Seeking consensus through the use of the Delphi technique in health sciences research.* Journal of allied health, 2013. **42**(2): p. 99-105.
90. Nair, R., R. Aggarwal, and D. Khanna. *Methods of formal consensus in classification/diagnostic criteria and guideline development*. in *Seminars in arthritis and rheumatism*. 2011. Elsevier.
91. Iqbal, S. and L. Pipon-Young, *Methods-The Delphi method--A guide from Susanne Iqbal and Laura Pipon-Young.* Psychologist, 2009. **22**(7): p. 598.

Figures

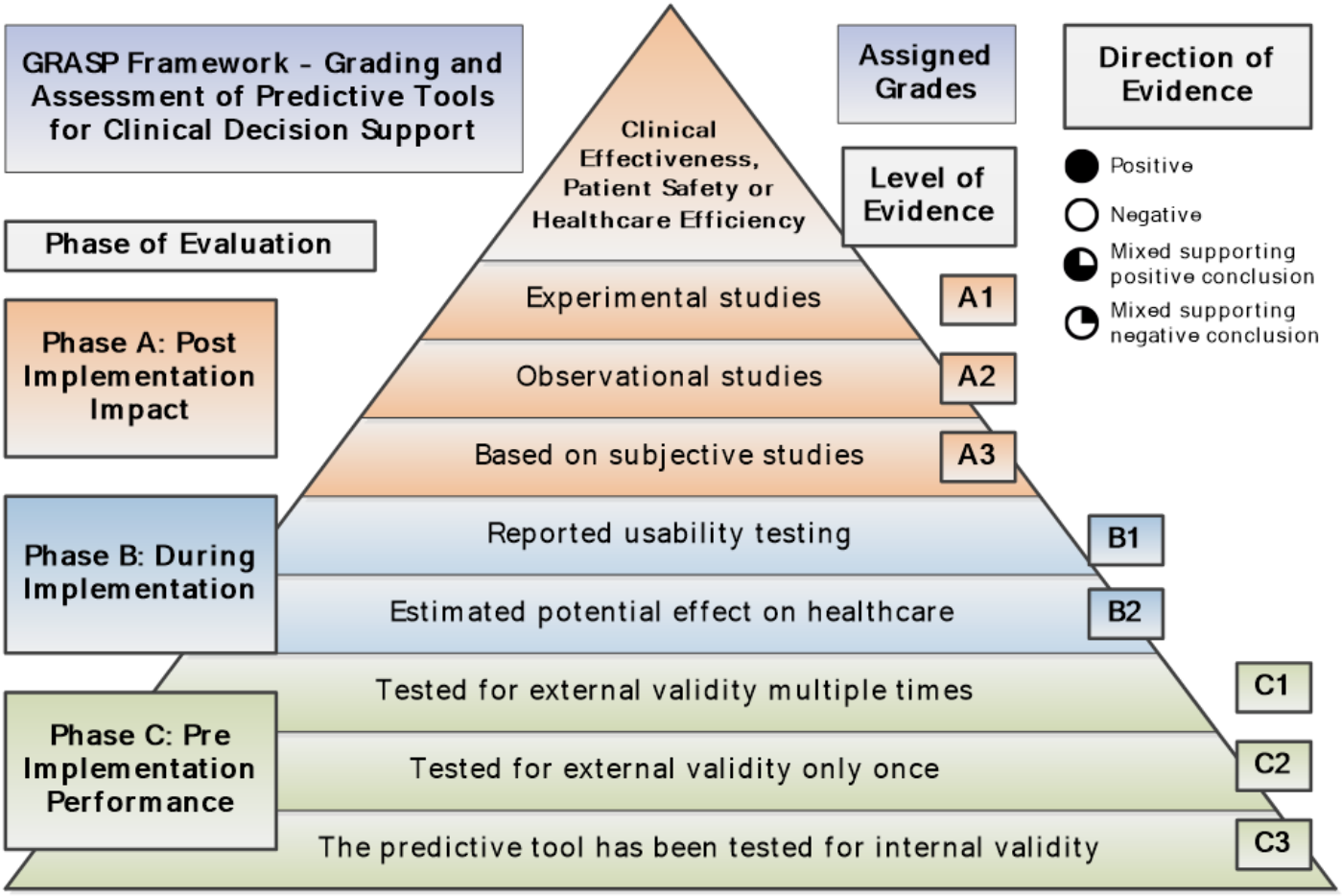

Figure 1

The GRASP Framework Concept [27]

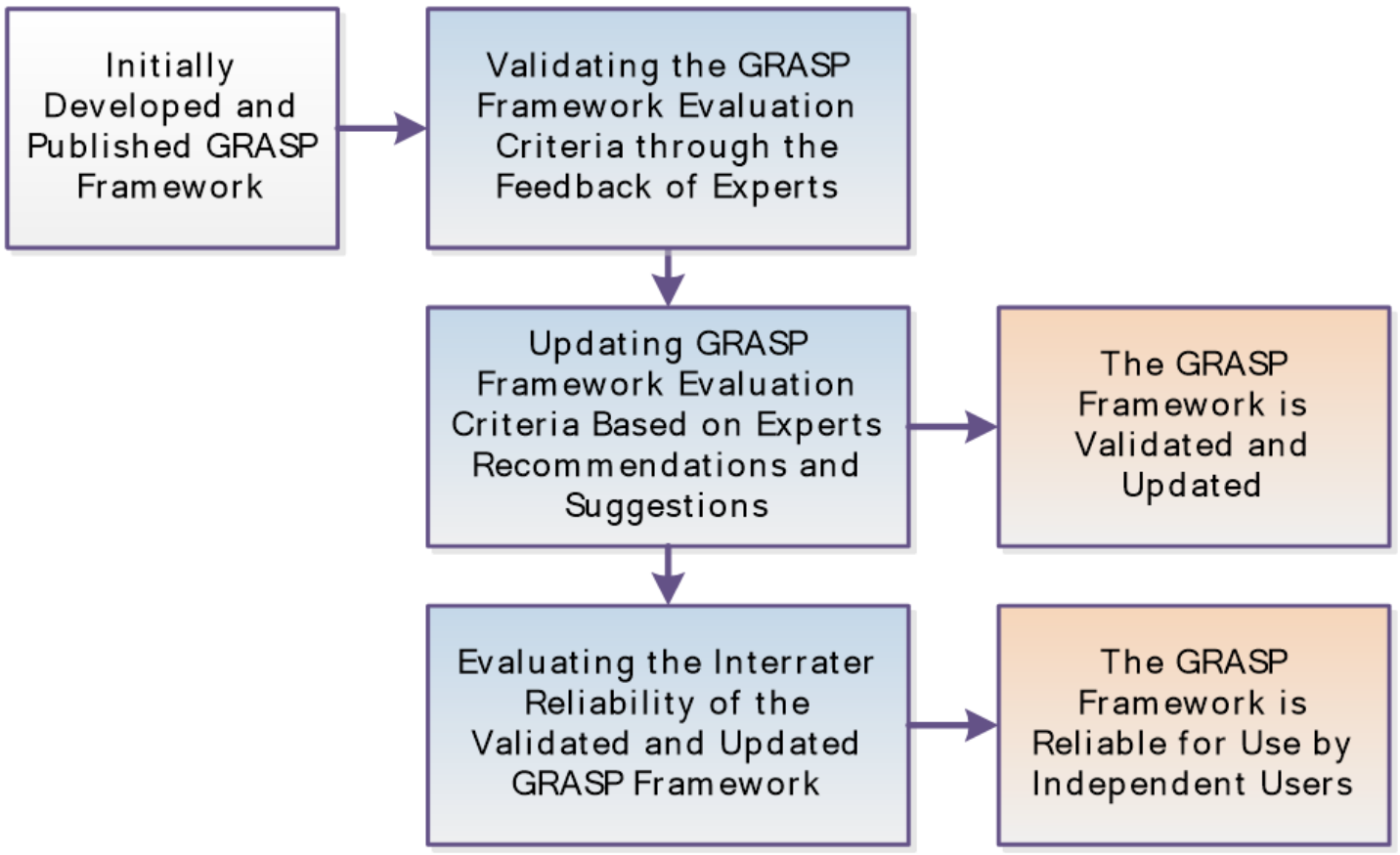

Figure 2

Flow Diagram of the Study Design

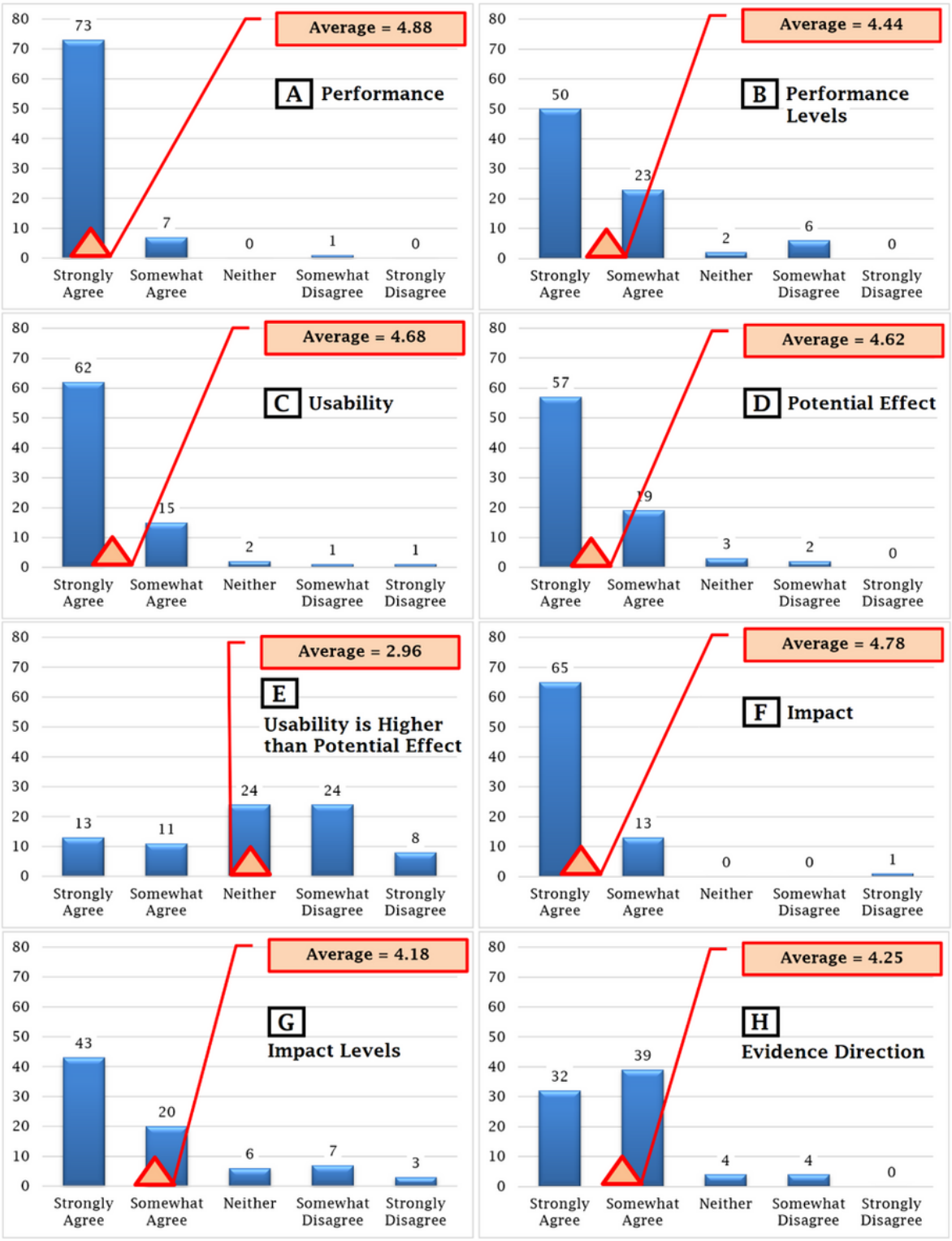

Figure 3

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

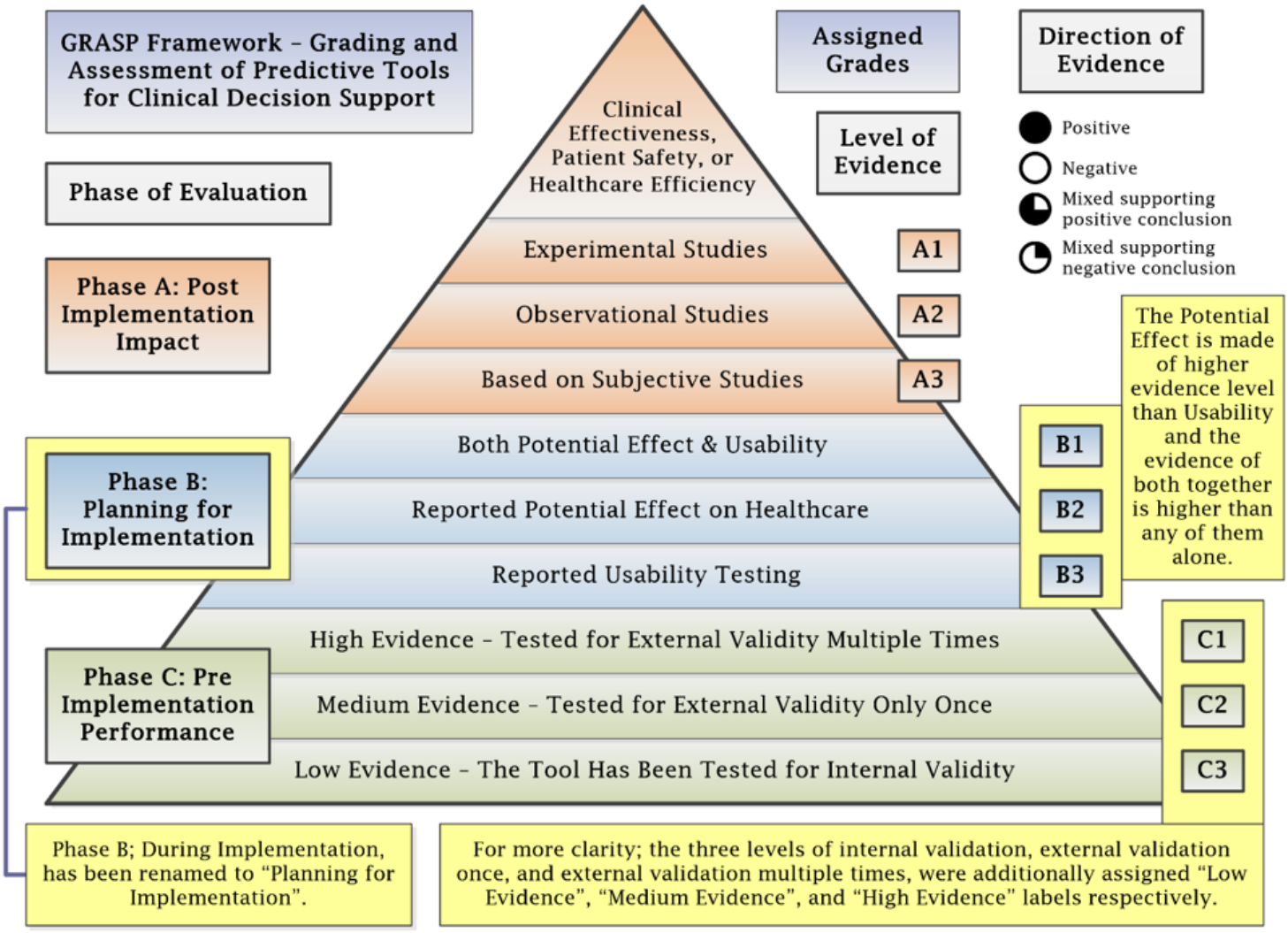

Figure 4

The Updated GRASP Framework Concept

Supplementary Files

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

- Appendix.pdf