BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

# Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support

Mohamed Khalifa[1], Farah Magrabi[1] and Blanca Gallego[1,2*]

## Abstract

**Background:** Clinical predictive tools quantify contributions of relevant patient characteristics to derive likelihood of diseases or predict clinical outcomes. When selecting predictive tools for implementation at clinical practice or for recommendation in clinical guidelines, clinicians are challenged with an overwhelming and ever-growing number of tools, most of which have never been implemented or assessed for comparative effectiveness. To overcome this challenge, we have developed a conceptual framework to Grade and Assess Predictive tools (GRASP) that can provide clinicians with a standardised, evidence-based system to support their search for and selection of efficient tools.

**Methods:** A focused review of the literature was conducted to extract criteria along which tools should be evaluated. An initial framework was designed and applied to assess and grade five tools: LACE Index, Centor Score, Well's Criteria, Modified Early Warning Score, and Ottawa knee rule. After peer review, by six expert clinicians and healthcare researchers, the framework and the grading of the tools were updated.

**Results:** GRASP framework grades predictive tools based on 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 a positive conclusion. Ottawa knee rule had the highest grade since it has demonstrated positive post-implementation impact on healthcare. LACE Index had the lowest grade, having demonstrated only pre-implementation positive predictive performance.

**Conclusion:** GRASP framework builds on widely accepted concepts to provide standardised assessment and evidence-based grading of predictive tools. Unlike other methods, GRASP is based on the critical appraisal of published evidence reporting the tools' predictive performance before implementation, potential effect and usability during implementation, and their post-implementation impact. Implementing the GRASP framework as an online platform can enable clinicians and guideline developers to access standardised and structured reported evidence of existing predictive tools. However, keeping GRASP reports up-to-date would require updating tools' assessments and grades when new evidence becomes available, which can only be done efficiently by employing semi-automated methods for searching and processing the incoming information.

**Keywords:** Predictive analytics, Clinical prediction, Clinical decision support, Evidence-based medicine

* Correspondence: b.gallego@unsw.edu.au
[1]Australian Institute of Health Innovation, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, Australia
[2]Centre for Big Data Research in Health, Faculty of Medicine, Univerisity of New South Wales, Sydney, Australia

## Background

Modern healthcare is building upon information technology to improve the effectiveness, efficiency, and safety of healthcare and clinical processes [1–7]. In particular, clinical decision support (CDS) systems operate in three levels [8, 9]: 1) Managing information, through facilitating the reach to clinical knowledge, the search for, and the retrieval of relevant information needed to make clinical decisions, 2) Focusing users' attention, through flagging abnormal results, providing lists of possible explanations for such results, or generating clinical and drug interaction alerts, and 3) Recommending specific actions and decisions, tailored for the clinical condition of the patient, which is often supported by clinical models and smart clinical guidelines.

Clinical predictive tools (here referred to simply as predictive tools) belong to the third level of CDS and include various applications ranging from the simplest manually applied clinical prediction rules to the most sophisticated machine learning algorithms [10, 11]. Through the processing of relevant clinical variables, clinical predictive tools derive the likelihood of diseases and predict their possible outcomes in order to provide patient specific diagnostic, prognostic, or therapeutic decision support [12, 13].

### Why do we need grading and assessment of predictive tools?

Traditionally, the selection of predictive tools for implementation in clinical practice has been conducted based on subjective evaluation of and exposure to particular tools [14, 15]. This is not optimal since many clinicians lack the required time and knowledge to evaluate predictive tools, especially as their number and complexity have increased tremendously in recent years.

More recently, there has been an increase in the mention and recommendation of selected predictive algorithms in clinical guidelines. However, this represents only a small proportion of the amount and variety of proposed clinical predictive tools, which have been designed for various clinical contexts, target many different patient populations and comprise a wide range of clinical inputs and techniques [16–18]. Unlike in the case of treatments, clinicians generating guildeines have no available methods to objectively summarise or interpret the evidence behind clinical predictive tools. This is made worse by the complex nature of the evaluation process itself and the variability in the quality of the published evidence [19–22].

Although most reported tools have been internally validated, only some have been externally validated and very few have been implemented and studied for their post-implementation impact on healthcare [23, 24]. Various studies and reports indicate that there is an unfortunate practice of developing new tools instead of externally validating or updating existing ones [13, 25–28]. More importantly, while a few pre-implementation studies compare similar predictive tools along some predictive performance measures, comparative studies for post-implementation impact or cost-effectiveness are very rare [29–38]. As a result, there is lack of a reference against which predictive tools can be compared or benchmarked [12, 39, 40].

In addition, decision makers need to consider the usability of a tool, which depends on the specific healthcare and IT settings in which it is embedded, and on users' priorities and perspectives [41]. The usability of tools is consistently improved when the outputs are actionable or directive [42–45]. Moreover, clinicians are keen to know if a tool has been endorsed by certain professional organisations they follow, or recommended by specific clinical guidelines they know [26].

### Current methods for appraising predictive tools

Several methods have been proposed to evaluate predictive tools [13, 24, 46–60]. However, most of these methods are not based on the critical appraisal of the existing evidence. Two exceptions are the TRIPOD statement [61, 62], which provides a set of recommendations for the reporting of studies developing, validating, or updating predictive tools; and the CHARMS checklist [63], which provides guidance on critical appraisal and data extraction for systematic reviews of predictive tools. Both of these methods examine only the pre-implementation predictive performance of the tools, ignoring their usability and post-implementation impact. In addition, none of the currently available methods provide a grading system to allow for benchmarking and comparative effectiveness of tools.

On the other hand, looking beyond predictive tools, the GRADE framework, grades the quality of published scientific evidence and strength of clinical recommendations, in terms of their post-implementation impact. GRADE has gained a growing consensus, as an objective and consistent method to support the development and evaluation of clinical guidelines, and has increasingly been adopted worldwide [64, 65]. According to GRADE, information based on randomised controlled trials (RCTs) is considered the highest level of evidence. However, the level of evidence could be downgraded due to study limitations, inconsistency of results, indirectness of evidence, imprecision, or reporting bias [65–67]. The strength of a recommendation indicates the extent to which one can be confident that adherence to the recommendation will do more good than harm, it also requires a balance between simplicity and clarity [64, 68].

The aim of this study is to develop a conceptual framework for evidence-based grading and assessment of predictive tools. This framework is based on the critical appraisal of information provided in the published evidence reporting the evaluation of predictive tools. The framework should provide clinicians with standardised objective information on predictive tools to support their search for and selection of effective tools for their intended tasks. It should support clinicians' informed decision making, whether they are implementing predictive tools at their clinical practices or recommending such tools in clinical practice guidelines to be used by other clinicians.

## Methods

Guided by the work of Friedman and Wyatt, and their suggested three phases approach, which became an internationally acknowledged standard for evaluating health informatics technologies [20, 21, 41, 69], we aimed to extract the main criteria along which predictive tools can be similarly evaluated before, during and after their implementation.

We started with a focused review of the literature in order to examine and collect the evaluation criteria, and measures proposed for the appraisal of predictive tools along these three phases. The concepts used in the search included "clinical prediction", "tools", "rules", "models", "algorithms", "evaluation", and "methods". The search was conducted using four databases; MEDLINE, EMBASE, CINAHL and Google Scholar, with no specific timeframe. This literature review was then extended to include studies describing methods evaluating CDS systems and more generally, health information systems and technology. Following the general concepts of the PRISMA guidelines [70], the duplicates of the retrieved studies, from the four databases, were first removed. Studies were then screened, based on their titles and abstracts, for relevance, then the full text articles were assessed for eligibility and only the eligible studies were included in the review. We included three types of studies evaluating predictive tools and other CDS systems; 1) studies describing the methods or processes of the evaluation, 2) studies describing the phases of the evaluation, and 3) studies describing the criteria and measures used in the evaluation. The first author manually extracted the methods of evaluation described in each type of study, and this was then revised and confirmed by the second and last authors. Additional file 1: Figure S1 shows the process of study selection for inclusion in the focused review of the literature.

Using the extracted information, we designed an initial version of the framework and applied it to asses and grade five predictive tools. We reviewed the complete list of 426 tools published by the MDCalc medical reference website for decision support tools and applications and calculators (https://www.mdcalc.com) [71]. We excluded tools which are not clinical - their output is not related to providing individual patient care, such as scores of ED crowding and calculators of waiting times. We also excluded tools which are not predictive - their output is not the result of statistically generating new information but rather the result of a deterministic equation, such as calculators deriving a number from laboratory results. The five example tools were then randomly selected, using a random number generator [72], from a shorter list of 107 eligible predictive tools, after being alphabetically sorted and numbered.

A comprehensive and systematic search for the published evidence, on each of the five predictive tools, was conducted, using MEDLINE, EMBASE, CINAHL and Google Scholar, and refined in four steps. 1) The primary studies, describing the development of the tools, were first identified and retrieved. 2) All secondary studies that cited the primary studies or that referred to the tools' names or to any of their authors, anywhere in the text, were retrieved. 3) All tertiary studies that cited the secondary studies or that were used as references by the secondary studies were retrieved. 4) Secondary and tertiary studies were examined to exclude non-relevant studies or those not reporting the validation, implementation or evaluation of the tools. After the four steps, eligible evidence was examined and grades were assigned to the predictive tools. Basic information about the tool, such as year of publication, intended use, target population, target outcome, and source and type of input data were extracted, from the primary studies, to inform the first "Tool Information" section of the framework. Eligible studies were then examined in detail for the reported evaluations of the predictive tools. Additional file 1: Figure S2 shows the process of searching the literature for the published evidence on the predictive tools.

The framework and its application to the selected five predictive tools were then peer reviewed by six expert healthcare professionals. Three of these professionals are clinicians, who work in hospitals and have over 10 years of experience using CDS systems, while the other three are healthcare researchers, who work in research organisations and have over 20 years of experience in developing, implementing or evaluating CDS systems. The reviewers were provided with the framework's concept design and its detailed report template. They were also provided with the summarised and detailed grading of the five predictive tools, as well as the justification and published evidence underpinning the grade assignment. After a brief orientation session, reviewers were asked to feedback on how much they agreed with each of the framework's dimensions and corresponding evaluation criteria, the 'Tool information' section as well as the grading of the five exemplar predictive tools. The framework was then refined and the grading of the five predictive tools was updated based on the reviewers' feedback. Figure 1 shows the flowchart of the GRASP framework overall development process.

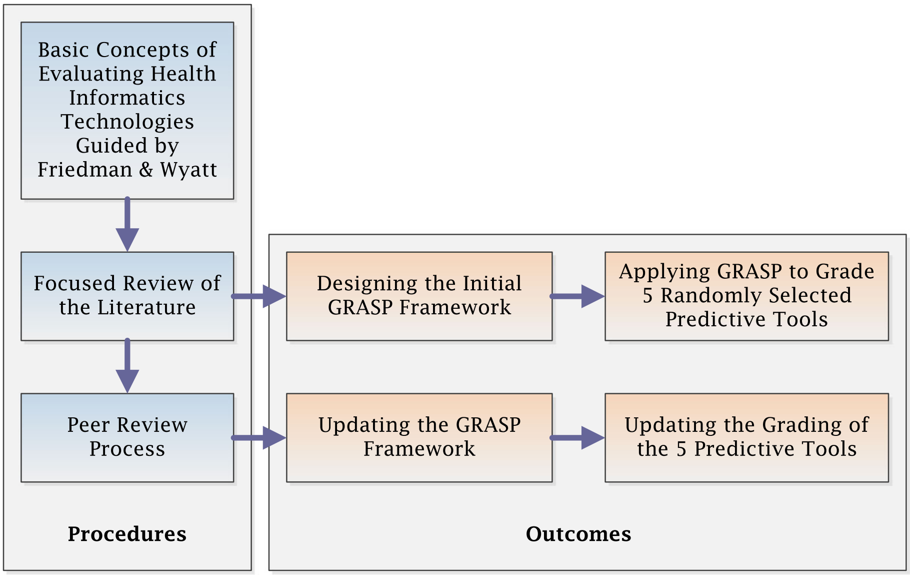

Basic Concepts of Evaluating Health Informatics Technologies Guided by Friedman & Wyatt
Focused Review of the Literature
Peer Review Process
Procedures
Designing the Initial GRASP Framework
Applying GRASP to Grade 5 Randomly Selected Predictive Tools
Updating the GRASP Framework
Updating the Grading of the 5 Predictive Tools
Outcomes

**Fig. 1** The GRASP Framework Development Flowchart

## Results

### The focused review of literature

The search in the four databases, after removing the duplicates, identified a total of 831 studies. After screening the titles and abstracts, 647 studies were found not relevant to the topic. The full text of the remaining 184 studies were then examined to exclude non-eligible studies, which were 134 studies, based on the inclusion criteria. Only 50 studies were identified as eligible. Twenty three of the 50 studies described methods for the evaluation of predictive tools [13, 17, 24, 40, 44, 46–61, 63, 73], ten studies described the evaluation of CDS systems [2, 74–82], and 11 studies described the evaluation of hospital information systems and technology [83–93]. One study described the NASSS framework, a guideline to help predict and evaluate the success of healthcare technologies [94]; and five studies described the GRADE framework for evaluating clinical guidelines and protocols [64–68]. The following three subsections describe the methods used to evaluate predictive tools as described in the focussed literature review. A summary of the evaluation criteria and examples of corresponding measures for each phase can be found in Additional file 1: Table S1.

### Before implementation – predictive performance

During the development phase, the internal validation of the predictive performance of a tool is the first step to make sure that the tool is doing what it is intended to do [54, 55]. Predictive performance is defined as the ability of the tool to utilise clinical and other relevant patient variables to produce an outcome that can be used to supports diagnostic, prognostic or therapeutic decisions made by clinicians and other healthcare professionals [12, 13]. The predictive performance of a tool is evaluated using measures of discrimination and calibration [53]. Discrimination refers to the ability of the tool to distinguish between patients with and without the outcome under consideration. This can be quantified with measures such as sensitivity, specificity, and the area under the receiver operating characteristic curve – AUC (or concordance statistic, c). The D-statistic is a measure of discrimination for time-to-event outcomes, which is commonly used in validating the predictive performance of prognostic models using survival data [95]. The log-rank test, or sometimes referred to as the Mantel-Cox test, is used to establish if the survival distributions of two samples of patients are statistically different. They are commonly used to validate the discrimination power of clinical prognostic models [96]. On the other hand, calibration refers to the accuracy of prediction, and indicates the extent to which expected and observed outcomes agree [48, 56]. Calibration is measured by plotting the observed outcome rates against their corresponding predicted probabilities. This is usually presented graphically with a calibration plot that shows a calibration line, which can be described with a slope and an intercept [97]. It is sometimes summarised using the Hosmer-Lemeshow test or the Brier score [98]. To avoid over-fitting, tools' predictive performance must always be assessed out-of-sample, either via cross-validation or bootstrapping [56]. Of more interest than the internal validity is the external validity (reliability or generalisability), where the predictive performance of a tool is estimated in independent validation samples of patients from different populations [52].

### During implementation – potential effect & usability

Before wide implementation, it is important to learn about the estimated potential effect of a predictive tool, when used in the clinical practice, on three main categories of measures: 1) Clinical effectiveness, such as improving patient outcomes, estimated through clinical effectiveness studies, 2) healthcare efficiency, including saving costs and resources, estimated through feasibility and cost-effectiveness studies, and 3) patient safety, including minimising complications, side effects, and medical errors. These categories are defined by the Institute of Medicine as objectives for improving healthcare performance and outcomes, and are differently prioritised by clinicians, healthcare professionals and health administrators [99, 100]. 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 [41, 101]. A few predictive tools have been studied for their potential to enhance clinical effectiveness and improve patient outcomes. For example, the spinal manipulation clinical prediction rule was tested, before implementation, on a small sample of patients to identify those with low back pain most likely to benefit from spinal manipulation [102]. Other tools have been studied for their potential to improve healthcare efficiency and save costs. For example, using a decision analysis model, and assuming all eligible children with minor blunt head trauma were managed using the CHALICE rule (Children's Head Injury Algorithm for the Prediction of Important Clinical Events), it was estimated that CHALICE would reduce unnecessary expensive head computed tomography (CT) scans, by 20%, without risking patients' health [103–105]. Similarly, the use of the PECARN (Paediatric Emergency Care Applied Research Network) head injury rule was estimated to potentially improve patient safety through minimising the exposure of children to ionising radiation resulting in fewer radiation-induced cancers and lower net quality adjusted life years loss [106, 107].

In addition, it is important to learn about the usability of predictive tools. Usability is defined as the extent to which a system can be used by the specified users to achieve specified and quantifiable objectives in a specified context of use [108, 109]. There are several methods to make a system more usable and many definitions have been developed, based on the perspective of what usability is and how it can be evaluated, such as the mental effort needed and the user attitude or the user interaction, represented in the easiness of use and acceptability of systems [110, 111]. Usability can be evaluated through measuring the effectiveness of task management with accuracy and completeness, measuring efficiency of utilising resources in completing tasks and measuring users' satisfaction, comfort with, and positive attitudes towards, the use of the tools [112, 113]. More advanced techniques, such as think aloud protocols and near live simulations, are recently used to evaluate usability [114]. Think aloud protocols are a major method in usability testing, since they produce a larger set of information and a richer content. They are conducted either retrospectively or concurrently, where each method has its own way of detecting usability problems [115]. The near live simulations provide users, during testing, with an opportunity to go through different clinical scenarios while the system captures interaction challenges and usability problems [116, 117]. Some researchers add learnability, memorability and freedom of errors to the measures of usability. Learnability is an important aspect of usability and a major concern in the design of complex systems. It is the capability of a system to enable the users to learn how to use it. Memorability, on the other hand, is the capability of a system to enable the users to remember how to use it, when they return back. Learnability and memorability are measured through subjective survey methods, asking users about their experience after using systems, and can also be measured by monitoring users' competence and learning curves over successive sessions of system usage [118, 119].

### After implementation – post-implementation impact

Some predictive tools have been implemented and used in the clinical practice for years, such as the PECARN head injury rule or the Ottawa knee and ankle rules [120–122]. In such cases, clinicians might be interested to learn about their post-implementation impact. The post-implementation impact of predictive tools is defined as the achieved change or influence, of a predictive tool, on different healthcare aspects, processes or outcomes, after the tool has been successfully implemented and used in the clinical practice, as designed by its developers [2, 42]. Similar to the measures of potential effect, post-implementation impact is reported along three main categories of measures: 1) Clinical effectiveness, such as improving patient outcomes, 2) Healthcare efficiency, such as saving costs and resources, and 3) Patient safety, such as minimising complications, side effects, and medical errors. These three categories of post-implementation impact measures are differently prioritised by clinicians, healthcare professionals and health administrators. In this phase of evaluation, we follow the main concepts of the GRADE framework, where the level of evidence for a given outcome is firstly determined by the study design [64, 65, 68]. High quality experimental studies, such as randomised and nonrandomised controlled trials, and the systematic reviews of their findings, come on top of the evidence levels followed by observational well-designed cohort or case-control studies and lastly subjective studies, opinions of respected authorities, and reported of expert committees or panels [65–67]. For simplicity, we did not include GRADE's detailed criteria for higher and lower quality of studies. However, effect sizes and potential biases are reported as part of the framework, so that consistency of findings, trade-offs between benefits and harms, and other considerations can also be assessed.

## Developing the GRASP framework

Our suggested GRASP framework (abbreviated from Grading and Assessment of Predictive Tools) is illustrated in Table 1. Based on published evidence, the GRASP framework uses three dimensions to grade predictive tools: ***Phase of Evaluation, Level of Evidence* and *Direction of Evidence*.**

### *Phase of evaluation*

Assigns a letter A, B and/or C based on the highest phase of evaluation reported in the published literature. A tool is assigned the lowest phase, C, if its predictive performance has been tested and reported for validity; phase B if its usability and/or potential impact have been validated; and the highest phase, A, if it has been implemented in clinical practice and its post-implementation impact has been reported.

### *Level of evidence*

Assigns a numerical score within each phase of evaluation based on the level of evidence associated with the evaluation process. Tools in phase C of evaluation can be assigned three levels. A tool is assigned the lowest level, C3, if it has been tested for internal validity; C2 if it has been tested for external validity once; and C1 if it has been tested for external validity multiple times. Similarly, tools in phase A of evaluation can be assigned the lowest level of evidence, A3, if their post-implementation impact has been evaluated only through subjective or descriptive studies; A2 if it has been evaluated via observational studies; and A1 if post-implementation impact has been measured using experimental evidence. Tools in phase B of evaluation are assigned grade B2 if they have been tested for potential impact and B1 if they have been tested for usability. Effect sizes for each outcome of interest together with study type, clinical settings and patient populations are also reported.

### *Direction of evidence*

Due to the large heterogeneity in study design, outcome measures and patient subpopulations contained in the studies, synthesising measures of predictive performance, usability, potential effect or post-implementation impact into one quantitative value is not possible. Furthermore, acceptable values of predictive performance or post-implementation impact measures depend on the clinical context and the task at hand. For example, tools like Ottawa knee rule [122] and Wells' criteria [123, 124] are considered effective only when their sensitivity is very close to 100%, since their task is to identify patients with fractures or pulmonary embolism before sending them home. On the other hand, tools like LACE Index [125] and Centor score [126] are accepted to show sensitivities of around 70%, since their tasks, to predict 30 days readmission risk or identify that pharyngitis is bacterial, aim to screen patients who may benefit from further interventions. Therefore, for each phase and level of evidence, we assign a direction of evidence, based on the conclusions reported in the studies, and provide the user with the option to look at the summary of the findings for further information.

Positive evidence is assigned when all studies reported positive conclusions while negative evidence is assigned when all studies reported negative or equivocal conclusions. In the presence of mixed evidence, studies are farther ranked according to their quality as well as their degree of matching with the orginal tool specifications (namely target population, target outcome and settings). Mixed evidence is then classified considering this ranking as supporting an overall positive conclusion or supporting an overall negative conclusion. Details and illustration of this protocol can be found in Additional file 1: Table S2 and Figure S3.

The final grade of a predictive tool is based on the highest phase of evaluation, supported by the highest level of positive evidence, or mixed evidence that supports an overall positive conclusion. Figure 2 shows the GRASP framework concept; a visual presentation of the framework three dimensions, phase of evaluation, level of evidence, and direction of evidence, explaining how each tool is assigned the final grade. Table 1 shows the GRASP framework detailed report.

## Applying the GRASP framework to grade five predictive tools

In order to show how GRASP works, we applied it to grade five randomly selected predictive tools; LACE Index for Readmission [125], Centor Score for Streptococcal Pharyngitis [126], Wells' Criteria for Pulmonary Embolism [123, 124, 127], The Modified Early Warning Score (MEWS) for Clinical Deterioration [128] and Ottawa Knee Rule [122]. In addition to these seven primary studies, describing the development of the five predictive tools, our systematic search for the published evidence revealed a total of 56 studies; validating, implementing, and evaluating the five predictive tools. The LACE Index was evaluated and reported in six studies, the Centor Score in 14 studies, the Wells' Criteria in ten studies, the MEWS in 12 studies, and the Ottawa Knee Rule in 14 studies. To apply the GRASP framework and

**Table 1** The GRASP Framework Detailed Report

| | | | |
|---|---|---|---|
| **Name** | Name of predictive tool (report tool's creators and year in the absence of a given name) | | |
| **Authors/Year** | Name of developer, country and year of publication | | |
| **Intended use** | Specific aim/intended use of the predictive tool | | |
| **Intended user** | Type of practitioner intended to use the tool | | |
| **Category** | Diagnostic/Therapeutic/Prognostic/Preventive | | |
| **Clinical area** | Clinical specialty | | |
| **Target Population** | Target patient population and health care settings in which the tool is applied | | |
| **Target Outcome** | Event to be predicted (including prediction lead time if needed) | | |
| **Action** | Recommended action based on tool's output | | |
| **Input source** | • Clinical (including Diagnostic, Genetic, Vital signs, Pathology)<br>• Non-Clinical (including Healthcare Utilisation) | | |
| **Input type** | • Objective (Measured input; from electronic systems or clinical examination)<br>• Subjective (Patient reported; history, checklist ...etc.) | | |
| **Local context** | Is the tool developed using location-specific data? (e.g. life expectancy tables) | | |
| **Methodology** | Type of algorithm (e.g. parametric/non-parametric) | | |
| **Endorsement** | Organisations endorsing the tool and/or guidelines recommending its utilisation | | |
| **Automation Flag** | Automation status (manual/automated) | | |
| **Tool Citations** | Total citations of the tool | | Number of studies reporting the tool |
| **Phase of Evaluation** | **Level of Evidence** | **Grade** | **Evaluation Studies** |
| **Phase C:**<br>**Before implementation**<br>**Is it possible?** | Insufficient internal validation | **C0** | Tested for internally validity but was either insufficiently internally validated or validation was insufficiently reported. |
| | Internal validation | **C3** | Tested for internally validity (reported calibration & discrimination; sensitivity, specificity, positive and negative predictive values & other predictive performance measures). |
| | External validation | **C2** | Tested for external validity, using one external dataset. |
| | External validation multiple times | **C1** | Tested multiple times for external validity, using more than one external dataset. |
| **Phase B:**<br>**During implementation**<br>**Is it practicable?** | Potential effect | **B2** | Reported estimated potential effect on clinical effectiveness, patient safety or healthcare efficiency. |
| | Usability | **B1** | Reported usability testing (effectiveness, efficiency, satisfaction, learnability, memorability, and minimizing errors). |
| **Phase A:**<br>**After implementation:**<br>**Is it desirable?** | Evaluation of Post-Implementation Impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency | **A3** | Based on subjective studies; e.g. the opinion of a respected authority, clinical experience, a descriptive study, or a report of an expert committee or panel. |
| | | **A2** | Based on observational studies; e.g. a well-designed cohort or case-control study. |
| | | **A1** | Based on experimental studies; e.g. a well-designed, widely applied randomised/nonrandomised controlled trial. |
| **Final Grade** | **Grade ABC,123** | | **A1** **A2** **A3** **B1** **B2** **C1** **C2** **C3** |
| **Direction of Evidence** | ● Positive Evidence | | ◕ Mixed Evidence Supporting Positive Conclusion |
| | ○ Negative Evidence | | ◔ Mixed Evidence Supporting Negative Conclusion |
| **Justification** | Explains how the final grade is assigned based on evidence; which conclusions were taken into consideration, as positive evidence, and which were considered negative. | | |
| **References** | Details of studies that support the justification: phase of evaluation, level of evidence, direction of evidence, study type, study settings, methodology, results, findings and conclusions (highlighted according to the colour code). | | These two sections are included in the full GRASP report on each tool. |
| **Label/Colour Code** | • Positive Findings<br>• Negative Findings | | • Important Findings<br>• Less Relevant Findings |

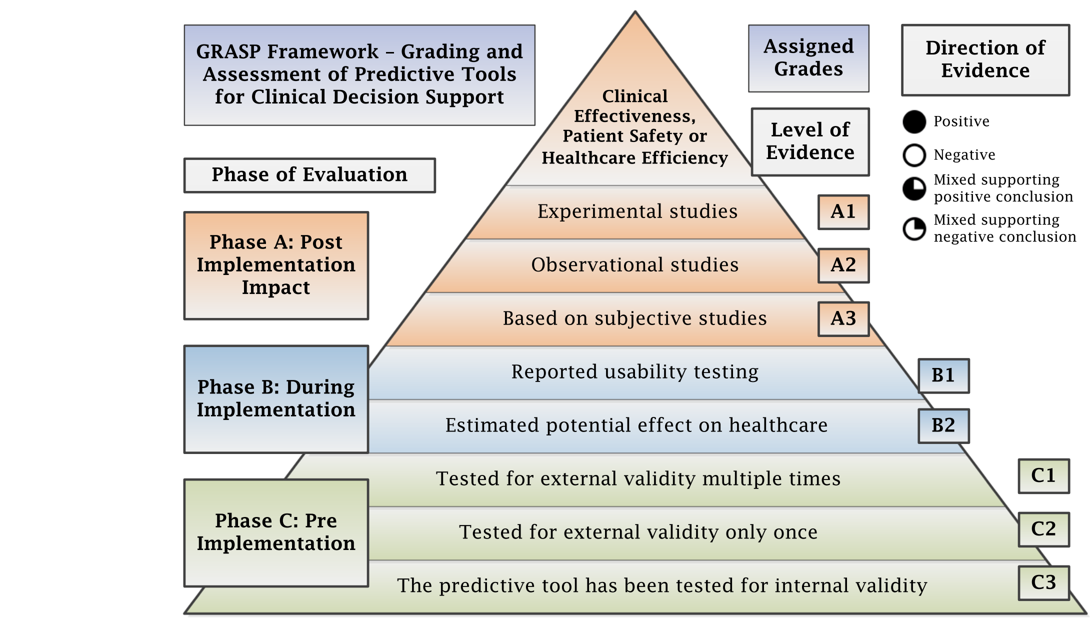

GRASP Framework – Grading and Assessment of Predictive Tools for Clinical Decision Support
Phase of Evaluation
Phase A: Post Implementation Impact
Phase B: During Implementation
Phase C: Pre Implementation
Clinical Effectiveness, Patient Safety or Healthcare Efficiency
Experimental studies
Observational studies
Based on subjective studies
Reported usability testing
Estimated potential effect on healthcare
Tested for external validity multiple times
Tested for external validity only once
The predictive tool has been tested for internal validity
Assigned Grades
Level of Evidence
A1
A2
A3
B1
B2
C1
C2
C3
Direction of Evidence
Positive
Negative
Mixed supporting positive conclusion
Mixed supporting negative conclusion

**Fig. 2** The GRASP Framework Concept

assign a grade to each predictive tool, the following steps were conducted; 1) The primary study or studies were first examined for the basic information about the tool and the reported details of development and validation. 2) Other studies were examined for their phases of evaluation, levels of evidence and direction of evidence. 3) Mixed evidence was sorted into positive or negative. 4) The final grade was assigned and supported by the detailed justification. A summary of grading the five tools is shown in Table 2 and a detailed GRASP report

**Table 2** Summary of Grading the Five Predictive Tools

| Tool Name | Tool Information | | | | Tool Grade | Impact After Implementation | | | During Implementation | | Predictive performance Before Implementation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Country | Year | Citations | Studies | | Experimental Studies | Observational Studies | Subjective Studies | Usability | Potential Effect | External Validation Multiple Times | External Validation Only Once | Internal Validation |
| | | | | | | A1 | A2 | A3 | B1 | B2 | C1 | C2 | C3 |
| LACE Index [125] | Canada | 2010 | 455 | 7 | C1 | | | | | | ◕ | | ● |
| Centor Score [126] | USA | 1981 | 715 | 15 | B1 | ◔ | | | ● | | ● | | ● |
| Wells' Criteria [123, 124, 127] | Canada | 1998 | 1,260 | 13 | A2 | | ● | | ● | | ● | | ● |
| Modified Early Warning Score [128] | UK | 2001 | 1,176 | 13 | A2 | | ◕ | | | | ◕ | | ● |
| Ottawa Knee Rule [122] | Canada | 1995 | 227 | 15 | A1 | ● | | | | | ● | | ● |
| Evidence Direction | ● Positive Evidence | | | | ◕ Mixed Evidence Supporting Positive Conclusion | | | | | | | | |
| | ○ Negative Evidence | | | | ◔ Mixed Evidence Supporting Negative Conclusion | | | | | | | | |

on each tool is provided in the Additional file 1: Tables S3-S7.

**LACE Index** is a prognostic tool designed to predict 30 days readmission or death of patients after discharge from hospitals. It uses multivariable logistic regression analysis of four administrative data elements; length of stay, admission acuity, comorbidity (Charlson Comorbidity Index) and emergency department (ED) visits in the last 6 months, to produce a risk score [125]. The tool has been tested for external validity twice; using a sample of 26,045 patients from six hospitals in Toronto and a sample of 59,652 patients from all hospitals in Alberta, Canada. In both studies, the LACE Index showed positive external validity and superior predictive performance to the previous similar tools endorsed by the Centres for Medicare and Medicaid Services in the United States [129, 130].

Two studies examined the predictive performance of LACE Index on small sub-population samples; 507 geriatric patients in the United Kingdom and 253 congestive heart failure patients in the United States, and found that the index performed poorly [131, 132]. Two more studies reported that the LACE Index performed well but not better that their own developed tools [133, 134]. Using the mixed evidence protocol, the mixed evidence here supports external validity, since the two negative conclusion studies have been conducted on very small samples of patients and on different subpopulations than the one the LACE Index was developed for. There was no published evidence on the usability, potential effect or post-implementation impact of the LACE Index. Accordingly, the LACE Index has been assigned Grade C1.

**Centor Score** is a diagnostic tool that uses a rule-based algorithm on clinical data to estimate the probability that pharyngitis is streptococcal in adults who present to the ED complaining of sore throat [126]. The score has been tested for external validity multiple times and all the studies reported positive conclusions [135–142]. This qualifies Centor score for Grade C1. One study conducted a multicentre cluster RCT usability testing of the integration of Centor score into electronic health records. The study used “Think Aloud” testing with ten primary care providers, post interaction surveys in addition to screen captures and audio recordings to evaluate usability. Within the same study, another “Near Live” testing, with eight primary care providers, was conducted. Conclusions reported positive usability of the tool and positive feedback of users on the easiness of use and usefulness [143]. This qualifies Centor score for Grade B1.

Evidence of the post-implementation impact of Centor score is mixed. One RCT conducted in Canada reported a clinically important 22% reduction in overall antibiotic prescribing [144]. Four other studies, three of which were RCTs, reported that implementing Centor score did not reduce antibiotic prescribing in clinical practice [145–148]. Using the mixed evidence protocol, we found that the mixed evidence does not support positive post-implementation impact of Centor score. Therefore, Centor score has been assigned Grade of B1.

**Wells’ Criteria** is a diagnostic tool used in the ED to estimate pre-test probability of pulmonary embolism [123, 124]. Using a rule-based algorithm on clinical data, the tool calculates a score that excludes pulmonary embolism without diagnostic imaging [127]. The tool was tested for external validity multiple times [149–153] and its predictive performance has been also compared to other predictive tools [154–156]. In all studies, Wells’ criteria was reported externally valid, which qualifies it for Grade C1. One study conducted usability testing for the integration of the tool into the electronic health record system of a tertiary care centre’s ED. The study identified a strong desire for the tool and received positive feedback on the usefulness of the tool itself. Subjects responded that they felt the tool was helpful, organized, and did not compromise clinical judgment [157]. This qualifies Wells’ criteria for Grade B1. The post-implementation impact of Well’s Criteria on efficiency of computed tomography pulmonary angiography (CTPA) utilisation has been evaluated through an observational before-and-after intervention study. It was found that the Well’s Criteria significantly increased the efficiency of CTPA utilisation and decreased the proportion of inappropriate scans [158]. Therefore, Well’s Criteria has been assigned Grade A2.

**The Modified Early Warning Score (MEWS)** is a prognostic tool for early detection of inpatients’ clinical deterioration and potential need for higher levels of care. The tool uses a rule-based algorithm on clinical data to calculate a risk score [128]. The MEWS has been tested for external validity multiple times in different clinical areas, settings and populations [159–165]. All studies reported that the tool is externally valid. However, one study reported MEWS poorly predicted the in-hospital mortality risk of patients with sepsis [166]. Using the mixed evidence protocol, the mixed evidence supports external validity, qualifying MEWS for Grade C1. No literature has been found regarding its usability or potential effect.

The MEWS has been implemented in different healthcare settings. One observational before-and-after intervention study failed to prove positive post-implementation impact of the MEWS on patient safety in acute medical admissions [167]. However, three more recent observational before-and-after intervention studies reported positive post-implementation impact of the MEWS on patient safety. One study reported significant increase in frequency of patient observation and decrease in serious adverse events after intensive care unit (ICU) discharge [168]. The second reported significant increase in frequency of vital signs recording, 24 h post-ICU discharge and 24 h preceding unplanned ICU admission [169]. The third, an 8 years study, reported that the post-implementation 4 years showed significant reductions in the incidence of cardiac arrests, the proportion of patients admitted to ICU and their in-hospital mortality [170]. Using the mixed evidence protocol, the mixed evidence supports positive post-implementation impact. The MEWS has been assigned Grade A2.

**Ottawa Knee Rule** is a diagnostic tool used to exclude the need for an X-ray for possible bone fracture in patients presenting to the ED, using a simple five items manual check list [122]. It is one of the oldest, most accepted and successfully used rules in CDS. The tool has been tested for external validity multiple times. One systematic review identified 11 studies, 6 of them involved 4249 adult patients and were appropriate for pooled analysis, showing high sensitivity and specificity predictive performance [171]. Furthermore, two studies discussed the post-implementation impact of Ottawa knee rule on healthcare efficiency. One nonrandomised controlled trial with before-after and concurrent controls included a total of 3907 patients seen during two 12-month periods before and after the intervention. The study reported that the rule decreased the use of knee radiography without patient dissatisfaction or missed fractures and was associated with reduced waiting times and costs per patient [172]. Another nonrandomised controlled trial reported that the proportion of ED patients referred for knee radiography was reduced. The study also reported that the practice based on the rule was associated with significant cost savings [173]. Accordingly, the Ottawa knee rule has been assigned Grade A1.

In the Additional file 1, a summary of the predictive performance of the five tools is shown in Additional file 1: Table S8. The c-statistics of LACE Index, Centor Score, Wells' Criteria and MEWS are reported in Additional file 1: Figure S4. The usability of Centor Score and Wells Criteria are reported in Additional file 1: Table S9 and post-implementation impact of Wells Criteria, MEWS and Ottawa knee rule is reported in Additional file 1: Table S10.

### Peer review of the GRASP framework

On peer-review, experts found the GRASP framework logical, helpful and easy to use. The reviewers strongly agreed to all criteria used for evaluation. The reviewers suggested adding more specific information about each tool, such as the author's name, the intended user of the tool and the recommended action based on the tool's findings. The reviewers showed a clear demand for knowledge regarding the applicability of tools to their local context. Two aspects were identified and discussed with the reviewers. Firstly, the operational aspect of how easy it would be to implement a particular tool and if the data required to use the tool is readily available in their clinical setting. Secondly, the validation aspect of adopting a tool developed using local predictors, such as life expectancy (which is location specific) or information based on billing codes (which is hospital specific). Following this discussion, elements related to data sources and context were added to the information section of the framework. One reviewer suggested assigning grade C0 to the reported predictive tools that did not meet C3 criteria, i.e. those tools which were tested for internally validity but were either insufficiently internally validated or the internal validation was insufficiently reported in the study, in order to differentiate them from those tools for which neither predictive performance nor post-implementation impact have been reported in the literature.

## Discussion

### Brief summary

It is challenging for clinicians to critically evaluate the growing number of predictive tools proposed to them by colleagues, administrators and commercial entities. Although most of these tools have been assessed for predictive performance, only a few have been implemented or evaluated for comparative predictive performance or post-implementation impact. In this manuscript, we present GRASP, a framework that provides clinicians with an objective, evidence-based, standardised method to use in their search for, and selection of tools. GRASP builds on widely accepted concepts, such as Friedman and Wyatt's evaluation approach [20, 21, 41, 69] and the GRADE system [64–68].

The GRASP framework is composed of two parts; 1) the GRASP framework concept, which shows the grades assigned to predictive tools based on the three dimensions: Phase of Evaluation, Level of Evidence and Direction of Evidence, and 2) the GRASP framework detailed report, which shows the detailed quantitative information on each predictive tool and justifies how the grade was assigned. The GRASP framework is designed for two levels of users: 1) Expert users, who will use the framework to assign grades to predictive tools and report their details, through the critical appraisal of published evidence about these tools. This step is essential to provide decision making clinicians with grades and detailed reports on predictive tools. Expert users include healthcare researchers who specialise in evidence-based methods and have experience in developing, implementing or evaluating predictive tools. 2) End users, who will use the GRASP framework detailed report of tools and their final grades, produced by expert users, to compare existing predictive tools and select the most suitable tool(s) for their predictive tasks, target objectives, and desired outcomes. End users include clinicians and other healthcare professionals involved in the decision making and selection of predictive tools for implementation at their clinical practice or for recommendation in clinical practice guidelines to be used by other clinicians and healthcare professionals. The two described processes; the grading of predictive tools by expert users and the selection decisions made by end users, should occur before the recommended tools are made available for use in the hands of the practicing clinicians.

### Comparison with previous literature

Previous approaches to the appraisal of predictive tools from the published literature, namely the TRIPOD statement [61, 62] and the CHARMS checklist [63], examine only their predictive performance, ignoring their usability and post-implementation impact. On the other hand, the GRADE framework appraises the published literature in order to evaluate clinical recommendations based on their post-implementation impact [64–68]. More broadly, methods for the evaluation of health information systems and technology focus on the integration of systems into tasks, workflows and organisations [84, 91]. The GRASP framework takes into account all phases of the development and translation of a predictive algorithm: predictive performance before implementation, using similar concepts as those utilised in TRIPOD and CHARMS; usability and potential effect during implementation, and post-implementation impact on patient outcomes and processes of care after implementation, using similar concepts as those utilised in the GRADE system. The GRASP grade is not the result of combining and synthesising selected measures of predictive performance (e.g. AUC), potential effect (e.g. potential saved money), usability (e.g. user satisfaction) or post-implementation impact (e.g. increased efficacy) from the existing literature, like in a meta-analysis; but rather the result of combining and synthesising the reported qualitative conclusions.

Walker and Habboushe, at the MDCalc website; classified and reported the most commonly used medical calculators and other clinical decision support applications. However, the website does not provide users with a structured grading system or an evidence-based method for the assessment of the presented tools. Therefore, we believe that our proposed framework can be adopted and used by MDCalc, and similar clinical decision support resources, to grade their tools.

### Quality of evidence and conflicting conclusions

One of the main challenges in assessing and comparing clinical predictive tools is dealing with the large variability in the quality and type of studies in the published literature. This heterogeneity makes it impractical to quantitatively synthese measures of predictive performance, usability, potential effect or post-implementation impact into single numbers. Furthermore, as discussed earlier, a particular value of a predictive performance metric that is considered good for some tasks and clinical settings may be considered insufficient for others. In order to avoid complex decisions regarding the quality and strength of reported measures, we chose to assign a direction of evidence, based on positive or negative conclusions as reported in the studies under consideration, since synthesizing qualitative conclusions is the only available option which adds some value. We then provide the end user with the option to look at a summary of the reported measures, in the GRASP detailed report, for further details.

It is not uncommon to encounter conflicting conclusions when a tool has been validated in different patient subpopulations. For example, LACE Index for readmission showed positive external validity when tested in adult medical inpatients [129, 130], but showed poor predictive performance when tested in a geriatric subpopulation [131]. Similarly, the MEWS for clinical deterioration demonstrated positive external validity when tested in emergency patients [159, 161, 165], medical inpatients [160, 164], surgical inpatients [162], and trauma patients [163], but not when tested in a subpopulation of patients with acute sepsis [166]. Part of these disagreements could be explained by changes in the distributions of important predictors, which affect the discriminatory power of the algorithms. For example, sepsis patients have similarly disturbed physiological measures such as those used to generate MEWS. In addition, conflicting conclusions may be encountered when a study examines the validity of a proposed tool in a healthcare setting or outcome different from those the tool was primarily developed for.

### Integration and socio-technical context

As is the case with other healthcare interventions, examining the post-implementation impact of predictive tools is challenging, since it is confounded by co-occurrent socio-technical factors [174–176]. This is complicated further by the fact that predictive tools are often integrated into electronic health record systems, since this facilitates their use, and are influenced by their usability [42, 177]. The usability, therefore, is an essential and major contributing factor in the wide acceptance and successful implementation of predictive tools and other CDS systems [42, 178]. It is clearly essential to involve user clinicians in the design and usability evaluations of predictive tools before their implementation. This should eliminate their concerns that integrating predictive tools into their workflow would increase their workload, consultation times, or decrease their efficiency and productivity [179].

Likewise, well designed post-implementation impact evaluation studies are required in order to explore the influence of organisational factors and local differences on the success or failure of predictive tools [180, 181].

Data availability, IT systems capabilities, and other human knowledge and organisational regulatory factors are crucial for the adoption, acceptance, and successful implementation of predictive tools. These factors and differences need to be included in the tools assessments, as they are important when making decisions about selecting predictive tools, in order to estimate the feasibility and resources needed to implement the tools. We have to acknowledge that it is not possible to include such wide range of variables in deciding or presenting the grades assigned by the framework to the predictive tools, which remain simply at a high-level. However, all the necessary information, technical specifications, and requirements of the tools, as reported in the published evidence, should be fully accessible to the users, through the framework's detailed reports on the predictive tools. Users can compare such information, of one or more tools, to what they have at their healthcare settings, then make selection and implementation decisions.

### Local data

There is a challenging trade-off between the generalisability and the customisation of a given predictive tool. Some algorithms are developed using local data. For example, Bouvy's prognostic model, for mortality risk in patients with heart failure, uses life quality and expectancy scores from the Netherlands [182]. Similarly, Fine's prediction rule identifies low-risk patients with community-acquired pneumonia based on national rates of acquired infections in the United States [183]. This necessitates adjustment of the algorithm to the local context, therefore producing a new version of the tool, which requires re-evaluation.

### Other considerations

GRASP evaluates predictive tools based on the critical appraisal of the existing published evidence. Therefore, it is subject to publication bias, since statistically positive results are more likely to be published than negative or null results [184, 185]. The usability and potential effect of predictive tools are less studied and hence the published evidence needed to support level B of the grading system is often lacking. We have nevertheless chosen to keep this element in GRASP since it is an essential part of the safety evaluation of any healthcare technology. It also allows for early redesign and better workflow integration, which leads to higher utilisation rates [114, 157, 186]. By keeping it, we hope to encourage tool developers and evaluators to increase their execution and reporting of these type of studies.

The grade assigned to a tool provides relevant evidence-based information to guide the selection of predictive tools for clinical decision support, but it is not prescriptive. An A1 tool is not always better than an A2 tool. A user may prefer an A2 tool showing improved patient safety in two observational studies rather than an A1 tool showing reduced cost in one experimental study. The grade is a code (not an ordinal quantity) that provides information on three relevant dimensions: phase of evaluation, level of evidence, and direction of evidence as reported in the literature.

### Study limitations and future work

One of the limitations of the GRASP framework is that the Direction of Evidence dimension is based on the conclusions of the considered studies on each predictive tool, which confers some subjectivity to this dimension. However, the end-user clinicians are provided with the full details of all available studies on each tool, through the GRASP detailed report, where they can access the required objective information to support their decisions. In addition, we applied the GRASP framework to only five predictive tools and consulted a small number of healthcare experts for their feedback. This could have limited the conclusions about the framework's coverage and/or validity. Although GRASP framework is not a predictive tool, it could be thought of as a technology of Grade C3, since it has only been internally validated after development. However, conducting a large-scale validation study of the framework, extending the application of the framework to a larger number of predictive tools, and studying its effect on end-users' decisions is out of the scope of this study and is left for future work.

To validate, update, and evaluate the GRASP framework, the authors are currently working on three more studies. The first study should validate the design and content of the framework, through seeking the feedback of a wider international group of healthcare experts, who have published work on developing, implementing or evaluating predictive tools. This study should help to update the criteria used, by the framework, to grade predictive tools and improve the details provided, by the framework, to the end users. The second study should evaluate the impact of using the framework on improving the decisions made by clinicians, regarding evaluating and selecting predictive tools. The experiment should compare the performance and outcomes of clinicians' decisions with and without using the framework. Finally, the third study aims to apply the framework to a larger consistent group of predictive tools, used for the same clinical task. This study should show how the framework provides clinicians with an evidence-based method to compare, evaluate and select predictive tools, through reporting and grading tools based on the critical appraisal of published evidence.

## Conclusion

The GRASP framework builds on widely accepted concepts to provide standardised assessment and evidence-based grading of predictive tools. Unlike other methods, GRASP is based on the critical appraisal of published evidence reporting the tools' predictive performance before implementation, potential effect and usability during implementation, and their post-implementation impact. Implementing the GRASP framework as an online platform can enable clinicians and guideline developers to access standardised and structured reported evidence of existing predictive tools. However, keeping the GRASP framework reports up-to-date would require updating tools' assessments and grades when new evidence becomes available. Therefore, a sustainable GRASP system would require employing automated or semi-automated methods for searching and processing the incoming published evidence. Finally, we recommend that GRASP framework be applied to predictive tools by working groups of professional organisations, in order to provide consistent results and increase reliability and credibility for end users. These professional organisations should also be responsible for making their associates aware of the availability of such evidence-based information on predictive tools, in a similar way of announcing and disseminating clinical practice guidelines.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12911-019-0940-7.

**Additional file 1: Table S1.** Phases, Criteria, and Measures of Evaluating Predictive Tools. **Table S2.** Evaluating Evidence Direction Based on the Conclusions of Studies. **Table S3.** LACE Index for Readmission – Grade C1. **Table S4.** Centor Score for Streptococcal Pharyngitis – Grade B1. **Table S5.** Wells' Criteria for Pulmonary Embolism – Grade A2. **Table S6.** Modified Early Warning Score (MEWS) – Grade A2. **Table S7.** Ottawa Knee Rule – Grade A1. **Table S8.** Predictive Performance of the Five Tools – Before Implementation. **Table S9.** Usability of Two Predictive Tools – During Implementation. **Table S10.** Post-Implementation Impact of Three Predictive Tools. **Figure S1.** Study Selection for the Focused Review of Literature. **Figure S2.** Searching the Literature for Published Evidence on Predictive Tools. **Figure S3.** The Mixed Evidence Protocol. **Figure S4.** Reported C-Statistic of LACE Index, Centor Score, Wells Criteria and MEWS [187–192].

### Abbreviations

AUC: Area under the curve; CDS: Clinical decision support; CHALICE: Children's Head Injury Algorithm for the Prediction of Important Clinical Events; CHARMS: Critical appraisal and data extraction for systematic reviews of prediction modelling studies; CT: Computed tomography; CTPA: Computed tomography pulmonary angiography; E.g.: For example; ED: Emergency department; GRADE: The Grading of Recommendations Assessment, Development and Evaluation; GRASP: Grading and assessment of predictive tools; I.e.: In other words or more precisely; ICU: Intensive care unit; IT: Information technology; LACE: Length of Stay, Admission Acuity, Comorbidity and Emergency Department Visits; MDCalc: Medical calculators; MEWS: Modified Early Warning Score; NASSS: Nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability; PECARN: Paediatric Emergency Care Applied Research Network; RCTs: Randomised controlled trials; ROC: Receiver operating characteristic curve; TRIPOD: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; UK: United Kingdom; US: United States

### Acknowledgments

We would like to thank Professor William Runciman, Professor of Patient Safety, Dr. Megan Sands, Senior Staff Specialist in Palliative Care, Dr. Thilo Schuler, Radiation Oncologist, Dr. Louise Wiles, Patient Safety Research Project Manager, and research fellows Dr. Virginia Mumford and Dr. Liliana Laranjo, for their contribution in testing the GRASP framework and for their valuable feedback and suggestions.

### Authors' contributions

MK contributed to the detailed design of the framework and led its execution and implementation. BG contributed to the conception of the framework and was responsible for the overall supervision of this work. FM provided advice on the enhancement and utility of the methodology. 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.

### 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

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

### Ethics approval and consent to participate

No ethics approval was required for any element of this study.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 2 October 2018 Accepted: 16 October 2019
Published online: 29 October 2019

### References

1. Middleton B, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc. 2013;20(e1):e2–8.
2. 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):765.
3. Osheroff JA. Improving outcomes with clinical decision support: an implementer's guide. New York: Imprint HIMSS Publishing; 2012.
4. Osheroff JA, et al. A roadmap for national action on clinical decision support. J Am Med Inform Assoc. 2007;14(2):141–5.
5. Øvretveit J, et al. Improving quality through effective implementation of information technology in healthcare. Int J Qual Health Care. 2007;19(5):259–66.
6. Castaneda C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma. 2015;5(1):4.
7. Capobianco E. Data-driven clinical decision processes: it's time: BioMed Central; 2019.
8. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Biomedical informatics: Springer; 2014. p. 643–74.
9. Shortliffe EH, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine: Springer Science & Business Media; 2013.
10. Adams ST, Leveson SH. Clinical prediction rules. BMJ. 2012;344:d8312.
11. Wasson JH, et al. Clinical prediction rules: applications and methodological standards. N Engl J Med. 1985;313(13):793–9.
12. Beattie P, Nelson R. Clinical prediction rules: what are they and what do they tell us? Aust J Physiother. 2006;52(3):157–63.
13. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating: Springer Science & Business Media; 2008.

14. Ansari S, Rashidian A. Guidelines for guidelines: are they up to the task? A comparative assessment of clinical practice guideline development handbooks. PLoS One. 2012;7(11):e49864.
15. Kish MA. Guide to development of practice guidelines. Clin Infect Dis. 2001; 32(6):851–4.
16. Ebell MH. Evidence-based diagnosis: a handbook of clinical prediction rules, vol. 1: Springer Science & Business Media; 2001.
17. Kappen T, et al. General discussion I: evaluating the impact of the use of prediction models in clinical practice: challenges and recommendations. In: Prediction models and decision support; 2015. p. 89.
18. 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):e006701.
19. Berner ES. Clinical decision support systems, vol. 233: Springer; 2007.
20. Friedman CP, Wyatt J. Evaluation methods in biomedical informatics: Springer Science & Business Media; 2005.
21. Friedman CP, Wyatt JC. Challenges of evaluation in biomedical informatics. In: Evaluation methods in biomedical informatics; 2006. p. 1–20.
22. Lobach DF. Evaluation of clinical decision support. In: Clinical decision support systems: Springer; 2016. p. 147–61.
23. Plüddemann A, et al. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. Br J Gen Pract. 2014;64(621):e233–42.
24. Wallace E, et al. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. BMJ Open. 2016;6(3):e009957.
25. Altman DG, et al. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
26. Bouwmeester W, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9(5):e1001221.
27. Hendriksen J, et al. Diagnostic and prognostic prediction models. J Thromb Haemost. 2013;11(s1):129–41.
28. Moons KG, et al. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
29. Christensen S, et al. Comparison of Charlson comorbidity index with SAPS and APACHE scores for prediction of mortality following intensive care. Clin Epidemiol. 2011;3:203.
30. Das K, et al. Comparison of APACHE II, P-POSSUM and SAPS II scoring systems in patients underwent planned laparotomies due to secondary peritonitis. Ann Ital Chir. 2014;85(1):16–21.
31. Desautels T, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform. 2016;4(3):e28.
32. Faruq MO, et al. A comparison of severity systems APACHE II and SAPS II in critically ill patients. Bangladesh Crit Care J. 2013;1(1):27–32.
33. Hosein FS, et al. A systematic review of tools for predicting severe adverse events following patient discharge from intensive care units. Crit Care. 2013; 17(3):R102.
34. Kim YH, et al. Performance assessment of the SOFA, APACHE II scoring system, and SAPS II in intensive care unit organophosphate poisoned patients. J Korean Med Sci. 2013;28(12):1822–6.
35. Köksal Ö, et al. The comparison of modified early warning score and Glasgow coma scale-age-systolic blood pressure scores in the assessment of nontraumatic critical patients in emergency department. Niger J Clin Pract. 2016;19(6):761–5.
36. Moseson EM, et al. Intensive care unit scoring systems outperform emergency department scoring systems for mortality prediction in critically ill patients: a prospective cohort study. J Intensive Care. 2014;2(1):40.
37. Reini K, Fredrikson M, Oscarsson A. The prognostic value of the modified early warning score in critically ill patients: a prospective, observational study. Eur J Anaesthesiol. 2012;29(3):152–7.
38. Yu S, et al. Comparison of risk prediction scoring systems for ward patients: a retrospective nested case-control study. Crit Care. 2014;18(3):R132.
39. Laupacis A, Sekar N. Clinical prediction rules: a review and suggested modifications of methodological standards. JAMA. 1997;277(6):488–94.
40. Moons KG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012. https://doi.org/10.1136/heartjnl-2011-301247.
41. Friedman CP, Wyatt JC. Evaluation of biomedical and health information resources. In: Biomedical informatics: Springer; 2014. p. 355–87.
42. Bates DW, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003;10(6):523–30.
43. Gong Y, Kang H. Usability and clinical decision support. In: Clinical decision support systems: Springer; 2016. p. 69–86.
44. Kappen TH, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. J Clin Epidemiol. 2016;70:136–45.
45. Sittig DF, et al. A survey of factors affecting clinician acceptance of clinical decision support. BMC Med Inform Decis Mak. 2006;6(1):6.
46. Collins GS, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1):40.
47. Debray T, et al. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32(18):3158–80.
48. Debray TP, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460.
49. Debray TP, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279–89.
50. Harris AH. Path from predictive analytics to improved patient outcomes: a framework to guide use, implementation, and evaluation of accurate surgical predictive models. Ann Surg. 2017;265(3):461–3.
51. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969–75.
52. Steyerberg EW, et al. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56(5):441–7.
53. Steyerberg EW, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54(8):774–81.
54. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245.
55. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014; 35(29):1925–31.
56. Steyerberg EW, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.). 2010;21(1):128.
57. Toll D, et al. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61(11):1085–94.
58. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. In: Seminars in oncology: Elsevier; 2010.
59. Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Urology. 2010; 76(6):1298–301.
60. Wallace E, et al. Framework for the impact analysis and implementation of clinical prediction rules (CPRs). BMC Med Inform Decis Mak. 2011;11(1):62.
61. Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med. 2015;13(1):1.
62. Moons KG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–W73.
63. Moons KG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014; 11(10):e1001744.
64. Atkins D, et al. Grading quality of evidence and strength of recommendations. BMJ (Clinical research ed). 2004;328(7454):1490.
65. Guyatt GH, et al. Rating quality of evidence and strength of recommendations: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924.
66. Guyatt G, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94.
67. Guyatt GH, et al. GRADE guidelines: a new series of articles in the journal of clinical epidemiology. J Clin Epidemiol. 2011;64(4):380–2.
68. Atkins D, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches the GRADE working group. BMC Health Serv Res. 2004;4(1):38.
69. Friedman CP, Wyatt JC. Challenges of evaluation in medical informatics. In: Evaluation methods in medical informatics: Springer; 1997. p. 1–15.

70. Moher D, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264–9.
71. Walker, G. and J. Habboushe. MD+Calc (Medical reference for clinical decision tools and content). 2018. Available from: https://www.mdcalc.com/. Cited 15 Sept 2018.
72. Ltd, R.a.I.S. Random.org. 2019. Available from: https://www.random.org/. Cited 1 Jan 2018.
73. Moons KG, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
74. Bright TJ, et al. Effect of clinical decision-support systemsa systematic review. Ann Intern Med. 2012;157(1):29–43.
75. Garg AX, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293(10):1223–38.
76. Hunt DL, et al. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA. 1998;280(15):1339–46.
77. Johnston ME, et al. Effects of computer-based clinical decision support systems on clinician performance and patient outcome: a critical appraisal of research. Ann Intern Med. 1994;120(2):135–42.
78. Kaplan B. Evaluating informatics applications—clinical decision support systems literature review. Int J Med Inform. 2001;64(1):15–37.
79. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med. 2003;163(12):1409–16.
80. McCoy AB, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc. 2011;19(3):346–52.
81. Pearson S-A, et al. Do computerised clinical decision support systems for prescribing change practice? A systematic review of the literature (1990-2007). BMC Health Serv Res. 2009;9(1):154.
82. Wright A, Sittig DF. A framework and model for evaluating clinical decision support architectures. J Biomed Inform. 2008;41(6):982–90.
83. Ammenwerth E, et al. Evaluation of health information systems—problems and challenges. Int J Med Inform. 2003;71(2):125–35.
84. Ammenwerth E, Iller C, Mahler C. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Med Inform Decis Mak. 2006;6(1):3.
85. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy Plan. 2009;24(3):217–28.
86. 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):742–52.
87. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems?: a framework for investigation and systematic review. JAMA. 1998;280(15):1347–52.
88. Kaufman D, et al. Applying an evaluation framework for health information system design, development, and implementation. Nurs Res. 2006;55(2):S37–42.
89. Kazanjian A, Green CJ. Beyond effectiveness: the evaluation of information systems using a comprehensive health technology assessment framework. Comput Biol Med. 2002;32(3):165–77.
90. Lau F, Hagens S, Muttitt S. A proposed benefits evaluation framework for health information systems in Canada. Health Q (Toronto, Ont.). 2007; 10(1):112–6, 118.
91. Yusof MM, et al. An evaluation framework for health information systems: human, organization and technology-fit factors (HOT-fit). Int J Med Inform. 2008;77(6):386–98.
92. Yusof MM, et al. Investigating evaluation frameworks for health information systems. Int J Med Inform. 2008;77(6):377–85.
93. Yusof MM, Paul RJ, Stergioulas LK. Towards a framework for health information systems evaluation. In: System sciences, 2006. HICSS'06. Proceedings of the 39th annual Hawaii international conference on: IEEE; 2006.
94. Greenhalgh T, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res. 2017;19(11):e367.
95. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med. 2004;23(5):723–48.
96. Kleinbaum DG, Klein M. Kaplan-Meier survival curves and the log-rank test. In: Survival analysis: Springer; 2012. p. 55–96.
97. Janssen K, et al. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61(1):76–86.
98. Schmid CH, Griffith JL. Multivariate classification rules: calibration and discrimination. In: Encyclopedia of biostatistics, vol. 5; 2005.
99. Berwick DM. A user's manual for the IOM's 'Quality Chasm'report. Health Aff. 2002;21(3):80–90.
100. Porter ME. What is value in health care? N Engl J Med. 2010;363(26):2477–81.
101. Friedman CP, Wyatt JC. Evaluation methods in medical informatics: Springer Science & Business Media; 2013.
102. Childs JD, et al. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. Ann Intern Med. 2004;141(12):920–8.
103. Alali AS, et al. Economic evaluations in the diagnosis and management of traumatic brain injury: a systematic review and analysis of quality. Value Health. 2015;18(5):721–34.
104. Barrett J. The use of clinical decision rules to reduce unnecessary head CT scans in pediatric populations: The University of Arizona; 2016.
105. Holmes M, et al. The cost-effectiveness of diagnostic management strategies for children with minor head injury. Arch Dis Child. 2013; 98(12):939–44.
106. Gökharman FD, et al. Pediatric emergency care applied research network head injuryprediction rules: on the basis of cost and effectiveness. Turk J Med Sci. 2017;47(6):1770–7.
107. Nishijima DK, et al. Cost-effectiveness of the PECARN rules in children with minor head trauma. Ann Emerg Med. 2015;65(1):72–80.e6.
108. Bevan N. Measuring usability as quality of use. Softw Qual J. 1995;4(2):115–30.
109. Bevan N, Macleod M. Usability measurement in context. Behav Inform Technol. 1994;13(1–2):132–45.
110. Bevan N. Usability. In: Encyclopedia of database systems: Springer; 2009. p. 3247–51.
111. Dix A. Human-computer interaction. In: Encyclopedia of database systems: Springer; 2009. p. 1327–31.
112. Frøkjær E, Hertzum M, Hornbæk K. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: Proceedings of the SIGCHI conference on human factors in computing systems: ACM; 2000.
113. Khajouei R, et al. Clinicians satisfaction with CPOE ease of use and effect on clinicians' workflow, efficiency and medication safety. Int J Med Inform. 2011;80(5):297–309.
114. Li AC, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. Int J Med Inform. 2012;81(11):761–72.
115. Van Den Haak M, De Jong M, Jan Schellens P. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav Inform Technol. 2003;22(5):339–51.
116. Borycki E, et al. Usability methods for ensuring health information technology safety: evidence-based approaches contribution of the IMIA working group health informatics for patient safety. Yearb Med Inform. 2013;22(01):20–7.
117. Richardson S, et al. "Think aloud" and "near live" usability testing of two complex clinical decision support tools. Int J Med Inform. 2017;106:1–8.
118. Jeng J. Usability assessment of academic digital libraries: effectiveness, efficiency, satisfaction, and learnability. Libri. 2005;55(2–3):96–121.
119. Nielsen J. Usability metrics: tracking interface improvements. IEEE Softw. 1996;13(6):12.
120. Kuppermann N, et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet. 2009;374(9696):1160–70.
121. Stiell IG, et al. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med. 1992;21(4):384–90.
122. Stiell IG, et al. Derivation of a decision rule for the use of radiography in acute knee injuries. Ann Emerg Med. 1995;26(4):405–13.
123. Wells PS, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism-increasing the models utility with the SimpliRED D-dimer. Thromb Haemost. 2000;83(3):416–20.
124. Wells PS, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. Ann Intern Med. 1998;129(12):997–1005.
125. 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. Can Med Assoc J. 2010;182(6):551–7.

126. Centor RM, et al. The diagnosis of strep throat in adults in the emergency room. Med Decis Mak. 1981;1(3):239–46.
127. Wells PS, et al. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer. Ann Intern Med. 2001;135(2):98–107.
128. Subbe C, et al. Validation of a modified early warning score in medical admissions. QJM. 2001;94(10):521–6.
129. Au AG, et al. Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. Am Heart J. 2012; 164(3):365–72.
130. Gruneir A, et al. Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. Open Med. 2011;5(2):e104.
131. Cotter PE, et al. Predicting readmissions: poor performance of the LACE index in an older UK population. Age Ageing. 2012;41(6):784–9.
132. Wang H, et al. Using the LACE index to predict hospital readmissions in congestive heart failure patients. BMC Cardiovasc Disord. 2014;14(1):97.
133. Low LL, et al. Predicting 30-day readmissions: performance of the LACE index compared with a regression model among general medicine patients in Singapore. Biomed Res Int. 2015;2015:169870.
134. Yu S, et al. Predicting readmission risk with institution-specific prediction models. Artif Intell Med. 2015;65(2):89–96.
135. Aalbers J, et al. Predicting streptococcal pharyngitis in adults in primary care: a systematic review of the diagnostic accuracy of symptoms and signs and validation of the Centor score. BMC Med. 2011;9(1):67.
136. Alper Z, et al. Diagnosis of acute tonsillopharyngitis in primary care: a new approach for low-resource settings. J Chemother. 2013;25(3):148–55.
137. Ebell MH, et al. Does this patient have strep throat? JAMA. 2000;284(22):2912–8.
138. Fine AM, Nizet V, Mandl KD. Large-scale validation of the Centor and McIsaac scores to predict group A streptococcal pharyngitis. Arch Intern Med. 2012;172(11):847–52.
139. McIsaac WJ, et al. Empirical validation of guidelines for the management of pharyngitis in children and adults. JAMA. 2004;291(13):1587–95.
140. Meland E, Digranes A, Skjærven R. Assessment of clinical features predicting streptococcal pharyngitis. Scand J Infect Dis. 1993;25(2):177–83.
141. Poses RM, et al. The importance of disease prevalence in transporting clinical prediction rules: the case of streptococcal pharyngitis. Ann Intern Med. 1986;105(4):586–91.
142. Wigton RS, Connor JL, Centor RM. Transportability of a decision rule for the diagnosis of streptococcal pharyngitis. Arch Intern Med. 1986;146(1):81–3.
143. Feldstein DA, et al. Design and implementation of electronic health record integrated clinical prediction rules (iCPR): a randomized trial in diverse primary care settings. Implement Sci. 2017;12(1):37.
144. McIsaac WJ, Goel V. Effect of an explicit decision-support tool on decisions to prescribe antibiotics for sore throat. Med Decis Mak. 1998;18(2):220–8.
145. Little, P., et al., Randomised controlled trial of a clinical score and rapid antigen detection test for sore throats. 2014.
146. McIsaac WJ, et al. A clinical score to reduce unnecessary antibiotic use in patients with sore throat. Can Med Assoc J. 1998;158(1):75–83.
147. Poses RM, Cebul RD, Wigton RS. You can lead a horse to water-improving physicians' knowledge of probabilities may not affect their decisions. Med Decis Mak. 1995;15(1):65–75.
148. Worrall G, et al. Diagnosing streptococcal sore throat in adults. Can Fam Physician. 2007;53(4):666–71.
149. Geersing G-J, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. BMJ. 2012;345:e6564.
150. Gibson NS, et al. Further validation and simplification of the Wells clinical decision rule in pulmonary embolism. Thromb Haemost. 2008;99(1):229.
151. Page P. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. JAMA. 2006;295(2):172–9.
152. Posadas-Martínez ML, et al. Performance of the Wells score in patients with suspected pulmonary embolism during hospitalization: a delayed-type cross sectional study in a community hospital. Thromb Res. 2014;133(2):177–81.
153. Söderberg M, et al. The use of d-dimer testing and Wells score in patients with high probability for acute pulmonary embolism. J Eval Clin Pract. 2009; 15(1):129–33.
154. Arslan ED, et al. Prediction of pretest probability scoring systems in pulmonary embolism: wells, Kline and Geneva. Int J Clin Med. 2013;3(07):731.
155. Klok F, et al. Comparison of the revised Geneva score with the Wells rule for assessing clinical probability of pulmonary embolism. J Thromb Haemost. 2008;6(1):40–4.
156. Turan O, et al. The contribution of clinical assessments to the diagnostic algorithm of pulmonary embolism. Adv Clin Exp Med. 2017;26(2):303.
157. Press A, et al. Usability testing of a complex clinical decision support tool in the emergency department: lessons learned. JMIR Hum Factors. 2015;2(2):e14.
158. Murthy C, et al. The impact of an electronic clinical decision support for pulmonary embolism imaging on the efficiency of computed tomography pulmonary angiography utilisation in a resource-limited setting. S Afr Med J. 2016;106(1):62–4.
159. Armagan E, et al. Predictive value of the modified early warning score in a Turkish emergency department. Eur J Emerg Med. 2008;15(6):338–40.
160. Burch V, Tarr G, Morroni C. Modified early warning score predicts the need for hospital admission and inhospital mortality. Emerg Med J. 2008;25(10):674–8.
161. Dundar ZD, et al. Modified early warning score and VitalPac early warning score in geriatric patients admitted to emergency department. Eur J Emerg Med. 2016;23(6):406–12.
162. Gardner-Thorpe J, et al. The value of modified early warning score (MEWS) in surgical in-patients: a prospective observational study. Ann R Coll Surg Engl. 2006;88(6):571–5.
163. Salottolo K, et al. A retrospective cohort study of the utility of the modified early warning score for interfacility transfer of patients with traumatic injury. BMJ Open. 2017;7(5):e016143.
164. Tanriöver MD, et al. Daily surveillance with early warning scores help predict hospital mortality in medical wards. Turk J Med Sci. 2016;46(6):1786–91.
165. Wang A-Y, et al. Periarrest modified early warning score (MEWS) predicts the outcome of in-hospital cardiac arrest. J Formos Med Assoc. 2016; 115(2):76–82.
166. Tirotta D, et al. Evaluation of the threshold value for the modified early warning score (MEWS) in medical septic patients: a secondary analysis of an Italian multicentric prospective cohort (SNOOPII study). QJM. 2017;110(6):369–73.
167. Subbe C, et al. Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions. Anaesthesia. 2003;58(8):797–802.
168. De Meester K, et al. Impact of a standardized nurse observation protocol including MEWS after intensive care unit discharge. Resuscitation. 2013; 84(2):184–8.
169. Hammond NE, et al. The effect of implementing a modified early warning scoring (MEWS) system on the adequacy of vital sign documentation. Aust Crit Care. 2013;26(1):18–22.
170. Moon A, et al. An eight year audit before and after the introduction of modified early warning score (MEWS) charts, of patients admitted to a tertiary referral intensive care unit after CPR. Resuscitation. 2011;82(2):150–4.
171. Bachmann LM, et al. The accuracy of the Ottawa knee rule to rule out knee fractures A systematic review. Ann Intern Med. 2004;140(2):121–4.
172. Stiell IG, et al. Implementation of the Ottawa knee rule for the use of radiography in acute knee injuries. JAMA. 1997;278(23):2075–9.
173. Nichol G, et al. An economic analysis of the Ottawa knee rule. Ann Emerg Med. 1999;34(4):438–47.
174. Khong PCB, Holroyd E, Wang W. A critical review of the theoretical frameworks and the conceptual factors in the adoption of clinical decision support systems. Comput Inform Nurs. 2015;33(12):555–70.
175. Meeks DW, et al. Exploring the sociotechnical intersection of patient safety and electronic health record implementation. J Am Med Inform Assoc. 2013;21(e1):e28–34.
176. Sheehan B, et al. Informing the design of clinical decision support services for evaluation of children with minor blunt head trauma in the emergency department: a sociotechnical analysis. J Biomed Inform. 2013;46(5):905–13.
177. Karsh B-T. Clinical practice improvement and redesign: how change in workflow can be supported by clinical decision support. Rockville: Agency for Healthcare Research and Quality; 2009. p. 200943.
178. Cresswell KM, Bates DW, Sheikh A. Ten key considerations for the successful implementation and adoption of large-scale health information technology. J Am Med Inform Assoc. 2013;20(e1):e9–e13.
179. Carroll C, et al. Involving users in the design and usability evaluation of a clinical decision support system. Comput Methods Prog Biomed. 2002; 69(2):123–35.

180. Li J. A sociotechnical approach to evaluating the impact of ICT on clinical care environments. Open Med Inform J. 2010;4:202.
181. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. 2010;19(Suppl 3):i68–74.
182. Bouvy ML, et al. Predicting mortality in patients with heart failure: a pragmatic approach. Heart. 2003;89(6):605–9.
183. Fine MJ, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med. 1997;336(4):243–50.
184. Dickersin K, et al. Publication bias and clinical trials. Control Clin Trials. 1987; 8(4):343–53.
185. Song F, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess. 2010;14(8):1–193.
186. Mann DM, et al. Rationale, design, and implementation protocol of an electronic health record integrated clinical prediction rule (iCPR) randomized trial in primary care. Implement Sci. 2011;6(1):109.
187. Aubert CE, et al. Prospective validation and adaptation of the HOSPITAL score to predict high risk of unplanned readmission of medical patients. Swiss Med Wkly. 2016;146:w14335.
188. Hung S-K, et al. Comparison of the mortality in emergency department Sepsis score, modified early warning score, rapid emergency medicine score and rapid acute physiology score for predicting the outcomes of adult splenic abscess patients in the emergency department. PLoS One. 2017; 12(11):e0187495.
189. Keene CM, et al. The effect of the quality of vital sign recording on clinical decision making in a regional acute care trauma ward. Chin J Traumatol. 2017;20(5):283–7.
190. Heitz CR, et al. Performance of the maximum modified early warning score to predict the need for higher care utilization among admitted emergency department patients. J Hosp Med. 2010;5(1):E46–52.
191. Bulloch B, et al. Validation of the Ottawa knee rule in children: a multicenter study. Ann Emerg Med. 2003;42(1):48–55.
192. Jalili M, Gharebaghi H. Validation of the Ottawa knee rule in Iran: a prospective study. Emerg Med J. 2010. https://doi.org/10.1136/emj.2009.080267.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Additional file

### *Phases, Criteria, and Measures of Evaluation*

Table S1. Phases, Criteria, and Measures of Evaluating Predictive Tools

| **Phase of Evaluation** | **Criteria of Evaluation** | **Definitions** | **Example Measures*** |
|---|---|---|---|
| Before Implementation | Predictive Performance | The ability of the predictive tool to utilise clinical variables and quantify relevant patient characteristics to produce an outcome that can be used to supports diagnostic, prognostic or therapeutic decisions made by clinicians and other healthcare professionals [12, 13]. | Discrimination:<br>• Sensitivity<br>• Specificity<br>• AUC, ROC, and C-Statistic<br>• D-Statistic<br>• Log-Rank Test.<br>Calibration:<br>• Calibration Plots & Curves<br>• Hosmer-Lemeshow test<br>• The Brier score. |
| During Implementation | Usability | The degree to which the predictive tool can be used by the specified users to achieve specified and quantifiable objectives in a specified context of use [108, 109]. | • Effectiveness of task management (accuracy and completeness).<br>• Efficiency of utilising resources.<br>• Users' satisfaction, comfort with, and positive attitudes towards, the use of the tools.<br>• Learnability<br>• Memorability<br>• Freedom of Errors. |
| | Potential Effect | 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 used in the clinical practice, as designed by its developers [41, 101]. | • Clinical Effectiveness (Clinical Patient Outcomes).<br>• Patient Safety (Complications, Side Effects, or Medical Errors).<br>• Healthcare Efficiency (Utilisation of Resources, Such as Time and Money). |
| After Implementation | Post-Implementation Impact | The achieved change or influence of a predictive tool on different healthcare aspects, processes or outcomes, after the tool has been successfully implemented and used in the clinical practice, as designed by its developers [2, 42]. | • Clinical Effectiveness (Clinical Patient Outcomes).<br>• Patient Safety (Complications, Side Effects, or Medical Errors).<br>• Healthcare Efficiency (Utilisation of Resources, Such as Time and Money). |

* These measures of evaluation are examples, the list is not meant to be exhaustive; literature on predictive tools may evaluate them along other measures.

### *Evaluating Evidence Direction*

Table S2. Evaluating Evidence Direction Based on the Conclusions of Studies

| **Conclusions of Studies** | | | **Overall Direction of Evidence** |
|---|---|---|---|
| **Positive *** | **Equivocal **** | **Negative ***** | |
| ✓ | | | Positive |
| ✓ | | ✓ | Mixed |
| ✓ | ✓ | ✓ | |
| ✓ | ✓ | | |
| | | ✓ | Negative |
| | ✓ | ✓ | |
| | ✓ | | |
| * Positive Conclusion | • The tool shows positive valid predictive performance, usability, potential effect, or post-implementation impact, which are desirable and/or superior to other methods/tools, if the study includes a comparison. | | |
| ** Equivocal Conclusion | • The tool shows positive valid predictive performance or usability, which are acceptable, but not superior to other methods/tools, if the study includes a comparison.<br>• The tool does not show positive potential effect or post-implementation impact. These are inferior to other methods/tools, if the study includes a comparison. | | |
| *** Negative Conclusion | • The tool shows that predictive performance or usability is poor, not acceptable, or inferior to other methods/tools, if the study includes a comparison.<br>• The tool shows negative potential effect or post-implementation impact (leads to deterioration instead of improvement), whether in comparison or not. | | |

### *GRASP Detailed Reports on Predictive Tools*

Table S3. LACE Index for Readmission – Grade C1

| | |
|---|---|
| **Name** | LACE Index for Readmission |
| **Authors/Year** | Dr. Carl van Walraven, Canada, 2010 |
| **Intended use** | Predicts 30 days readmission or death risk of medical and surgical inpatients after discharge |
| **Intended user** | Used by nurses at patient discharge |
| **Category** | Prognostic |
| **Clinical area** | All medical/surgical areas |
| **Target Population** | Hospitalised patients |
| **Target Outcome** | 30 days readmission or death |
| **Action** | Inform the clinical team about patients at high risk for readmission |
| **Input source** | Objective data (Data is available in the EHR – electronic health record, or manually obtained from the patient medical record). |
| **Input type** | Administrative data: Length of stay (days), Admission acuity (yes/no), Comorbidity (Charlson Index), Number of ED visits within 6 months. |
| **Local context** | Input does not depend on local context of data |
| **Methodology** | Multivariable logistic regression analysis |

| | | | |
|---|---|---|---|
| **Endorsement** | Recommended by:<br>• Texas Healthcare Association, USA.<br>• American Heart Association, USA.<br>• Michigan Care Management Resource Center, USA | | |
| **Automation Flag** | Manual | | |
| **Tool Citations** | 455 | Reported in 7 studies | |
| **Phase of Evaluation** | **Level of Evidence** | **Grade** | **Evaluation Studies** |
| **Phase C: Before implementation Does the tool work? Is it possible?** | Internal validation | **C3** | Developed and tested for internal validity:<br>• van Walraven et al, 2010 [125] |
| | External validation | **C2** | Tested for externally validity:<br>• Gruneir et al, 2011 [130] |
| | External validation multiple times | **C1** | Tested for external validity again:<br>• Au et al, 2012 [129]<br>Negative conclusion validation/performance studies:<br>• Cotter et al, 2012 [131]<br>• Wang et al, 2014 [132]<br>• Low et al, 2015 [133]<br>• Yu et al, 2015 [134] |
| **Phase B: During implementation: Is the tool practicable?** | Potential effect | **B2** | Not reported |
| | Usability | **B1** | Not reported |
| **Phase A: After implementation: Is the tool desirable?** | Evaluation of Post-Implementation Impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency | **A3** | No subjective studies reported |
| | | **A2** | No observational studies reported |
| | | **A1** | No experimental studies reported |
| **Final Grade** | **Grade C1** | | A1 \| A2 \| A3 \| B1 \| B2 \| ◕ \| C2 \| ● |
| **Direction of Evidence** | ● Positive Evidence | | ◕ Mixed Evidence Supporting Positive Conclusion |
| | ○ Negative Evidence | | ◔ Mixed Evidence Supporting Negative Conclusion |
| **Justification** | LACE Index is a prognostic tool designed to predict 30 days readmission or death after discharge from hospital. It uses multivariable logistic regression analysis of administrative data: length of stay, admission acuity, comorbidity (Charlson Comorbidity Index) and emergency department (ED) visits in the last six months, to produce a risk score [125]. The tool has been tested for external validity twice: using a sample of 26,045 patients from six hospitals in Toronto and a sample of 59,652 patients from all hospitals in Alberta. The LACE Index showed external validity and superior predictive performance to previous tools endorsed by the Centres for Medicare and Medicaid Services [129, 130]. Two studies examined LACE Index predictive performance on small sub-population samples: 507 geriatric patients in the UK, and 253 congestive heart failure patients in the USA, and found that the index performed poorly [131, 132]. Two more studies reported that LACE Index works well but their own developed tools performed better [133, 134]. Using the mixed evidence protocol, the mixed evidence supports external validity, since the two negative conclusion studies have been conducted on very small samples of patients and a different subpopulation than the one LACE was developed for. There was no published evidence on the usability, potential effect or post-implementation impact of LACE Index. Accordingly, LACE Index has been assigned Grade C1. | | |

Table S4. Centor Score for Streptococcal Pharyngitis – Grade B1

| | |
|---|---|
| **Name** | Centor Score for Streptococcal Pharyngitis |
| **Authors/Year** | Dr. Robert M. Centor, USA, 1981. Modified later by Dr. Warren McIsaac, Canada, 1998. |
| **Intended use** | Estimate the probability that pharyngitis is streptococcal in adult patients presenting to the emergency department with sore throat |
| **Intended user** | Used by physicians at ED as part of the clinical examination |
| **Category** | Diagnostic |
| **Clinical area** | Infectious diseases |
| **Target** | Patients visiting the emergency department |

|  |  |  |  |
|---|---|---|---|
| **Population** |  |  |  |
| **Target Outcome** | Streptococcal pharyngitis |  |  |
| **Action** | Consider rapid strep testing and/or culture |  |  |
| **Input source** | Objective data (clinical examination) + Subjective data (symptoms described by patient) |  |  |
| **Input type** | Clinical data: Age (3-14, 15-44 & >45 years), Exudate or swelling on tonsils (yes/no), Tender/swollen anterior cervical lymph nodes (yes/no), Temp >38°C (100.4°F) (yes/no), Cough (present/absent). Data is obtained from the patient. |  |  |
| **Local context** | Input does not depend on local context of data |  |  |
| **Methodology** | Rule-based algorithm |  |  |
| **Endorsement** | Recommended by:<br>• Department of Health, New South Wales, Australia<br>• American Academy of Family Physicians, United States<br>• The National Institute for Health and Care Excellence, United Kingdom |  |  |
| **Automation Flag** | Manual |  |  |
| **Tool Citations** | 715 | Reported in 15 studies |  |
| **Phase of Evaluation** | **Level of Evidence** | **Grade** | **Evaluation Studies** |
| **Phase C: Before implementation Does the tool work? Is it possible?** | Internal validation | **C3** | Developed and tested for internal validity:<br>• Centor et al, 1981 [126] |
|  | External validation | **C2** | Tested for external validity:<br>• Wigton, Connor & Centor, 1986 [142] |
|  | External validation multiple times | **C1** | Tested for external validity multiple times:<br>• Poses et al, 1986 [141]<br>• Meland, Digranes & Skjærven, 1993 [140]<br>• Ebell et al, 2000 [137]<br>• McIsaac et al, 2004 [139]<br>• Aalbers et al, 2011 [135]<br>• Fine, Nizet & Mandl, 2012 [138]<br>• Alper et al, 2013 [136] |
| **Phase B: During implementation: Is the tool practicable?** | Potential effect | **B2** | Not reported |
|  | Usability | **B1** | Reported usability testing is positive:<br>• Feldstein et al, 2017 [143] |
| **Phase A: After implementation: Is the tool desirable?** | Evaluation of Post-Implementation Impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency | **A3** | No subjective studies reported |
|  |  | **A2** | No observational studies reported |
|  |  | **A1** | One RCT show positive post-implementation impact of Centor score on reducing unnecessary antibiotics prescribing:<br>• McIsaac et al, 1998 [144]<br>One observational study + 3 RCTs show negative conclusions (No impact of Centor score on antibiotics prescribing):<br>• McIsaac et al, 1998 [146]<br>• Poses, Cebul & Wigton, 1995 [147]<br>• Worrall et al, 2007 [148]<br>• Little et al, 2014 [145] |
| **Final Grade** | **Grade B1** |  | ◔ A2 A3 ● B2 ● C2 ● |
| **Direction of Evidence** | ● Positive Evidence |  | ◕ Mixed Evidence Supporting Positive Conclusion |
|  | ○ Negative Evidence |  | ◔ Mixed Evidence Supporting Negative Conclusion |
| **Justification** | Centor score is a diagnostic tool that uses a rule-based algorithm on clinical data to estimate the probability that pharyngitis is streptococcal in adults who present to ED complaining of sore throat [126]. The score has been tested for external validity multiple times and all studies reported positive conclusions [135-142]. This qualifies Centor score for Grade C1. One study conducted a multicentre cluster RCT usability testing of the integration of Centor score into electronic health records. The study used "Think Aloud" testing with ten primary care providers, post interaction surveys in addition to screen captures and audio recordings to evaluate usability. Within the same study, another "Near Live" testing, with eight primary care providers, was conducted. Conclusions reported positive usability of the tool and positive feedback of users on the easiness of use and usefulness [143]. This qualifies Centor score for Grade B1. Evidence of the post-implementation impact of Centor score post-implementation is mixed. One RCT conducted in Canada reported a clinically important 22% reduction in overall antibiotic prescribing [144]. Four other studies, three of which were RCTs, reported that implementing Centor score did not reduce antibiotic prescribing in clinical practice [145-148]. Using the mixed evidence protocol, the mixed evidence does not |  |  |

| | |
|---|---|
| | support positive post-implementation impact of Centor score. Therefore, Centor score has been assigned Grade B1. |

Table S5. Wells' Criteria for Pulmonary Embolism – Grade A2

| | | | |
|---|---|---|---|
| **Name** | Wells' Criteria for Pulmonary Embolism | | |
| **Authors/Year** | Dr. Phil Wells, Canada, 1998. | | |
| **Intended use** | Calculates the pre-test probability (risk)v of pulmonary embolism at the bedside without imaging | | |
| **Intended user** | Used by physicians at ED as part of the clinical examination | | |
| **Category** | Diagnostic | | |
| **Clinical area** | Cardiovascular diseases | | |
| **Target Population** | Patients visiting the emergency department | | |
| **Target Outcome** | Pulmonary embolism | | |
| **Action** | Rule out high risk patients with computed tomography angiography | | |
| **Input source** | Objective data (clinical examination) + Subjective data (symptoms described by patient). | | |
| **Input type** | Clinical data: Clinical signs and symptoms of DVT (yes/no), PE is #1 diagnosis OR equally likely (yes/no), Heart rate > 100 (yes/no), Immobilization at least 3 days OR surgery in the previous 4 weeks (yes/no), Previous, objectively diagnosed PE or DVT (yes/no), Haemoptysis (yes/no), Malignancy w/ treatment within 6 months or palliative (yes/no). | | |
| **Local context** | Input does not depend on local context of data | | |
| **Methodology** | Rule-based algorithm | | |
| **Endorsement** | Recommended by:<br>• New South Wales Agency for Clinical Innovation, Australia<br>• The Royal Australian College of General Practitioners, Australia | | |
| **Automation Flag** | Manual | | |
| **Tool Citations** | 1,260 | Reported in 13 studies | |
| **Phase of Evaluation** | **Level of Evidence** | **Grade** | **Evaluation Studies** |
| **Phase C: Before implementation Does the tool work? Is it possible?** | Internal validation | **C3** | Developed and tested for internal validity:<br>• Wells et al, 1998 [124]<br>• Wells et al, 2000 [123]<br>• Wells et al, 2001 [127] |
| | External validation | **C2** | Tested for external validity:<br>• Page, 2006 [151] |
| | External validation multiple times | **C1** | Tested for external validity multiple times:<br>• Gibson et al, 2008 [150]<br>• Klok et al, 2008 [155]<br>• Söderberg et al, 2009 [153]<br>• Geersing et al, 2012 [149]<br>• Arslan et al, 2013 [154]<br>• Posadas-Martínez et al, 2014 [152]<br>• Turan et al, 2017 [156] |
| **Phase B: During implementation: Is the tool practicable?** | Potential effect | **B2** | Not reported |
| | Usability | **B1** | Reported usability testing is positive:<br>• Press et al, 2015 [157] |
| **Phase A: After implementation: Is the tool desirable?** | Evaluation of Post-Implementation Impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency | **A3** | No subjective studies reported |
| | | **A2** | Observational before-and-after intervention study showing positive post-implementation impact of Wells' Criteria on healthcare efficiency:<br>• Murthy et al, 2016 [158] |
| | | **A1** | No experimental studies reported |
| **Final Grade** | **Grade A2** | | A1 ● A3 ● B2 ● C2 ● |
| **Direction of** | ● Positive Evidence | | ◕ Mixed Evidence Supporting Positive Conclusion |

| | | |
|---|---|---|
| **Evidence** | ○ Negative Evidence | ◔ Mixed Evidence Supporting Negative Conclusion |
| **Justification** | Wells' criteria is a diagnostic tool used in ED to estimate pre-test probability of pulmonary embolism [123, 124]. Using a rule-based algorithm on clinical data, the tool calculates a score that excludes pulmonary embolism without diagnostic imaging [127]. The tool was tested for external validity multiple times [149-153] and its predictive performance has been also compared to other predictive tools [154-156]. In all studies, Wells' criteria was reported valid, which qualifies it for Grade C1. One study conducted usability testing for the integration of the tool into the electronic health record system of a tertiary care centre's ED. The study identified a strong desire for the tool and received positive feedback on the usefulness of the tool itself. Subjects responded that they felt the tool was helpful, organized, and did not compromise clinical judgment [157]. This qualifies Wells' criteria for Grade B1. The post-implementation impact of Well's Criteria on efficiency of computed tomography pulmonary angiography (CTPA) utilisation has been evaluated through an observational before-and-after intervention study. It was found that the Well's Criteria significantly increased the efficiency of CTPA utilisation and decreased the proportion of inappropriate scans [158]. Therefore, Well's Criteria has been assigned Grade A2. | |

Table S6. Modified Early Warning Score (MEWS) – Grade A2

| | | | |
|---|---|---|---|
| **Name** | Modified Early Warning Score (MEWS) for Clinical Deterioration | | |
| **Authors/Year** | Dr. Christian Peter Subbe, UK, 2001 | | |
| **Intended use** | Early detection of inpatients' clinical deterioration, calculate chance of ICU admission or death within 60 days and potential need for higher levels of care. | | |
| **Intended user** | Used by nurses at bedside | | |
| **Category** | Prognostic | | |
| **Clinical area** | General Medicine | | |
| **Target Population** | Hospitalised patients | | |
| **Target Outcome** | Clinical deterioration/death | | |
| **Action** | Consider higher level of care for patient (e.g. transfer to ICU) | | |
| **Input source** | Objective (Data from EHR – electronic health record) | | |
| **Input type** | Clinical data: Systolic BP, Heart rate, Respiratory rate, Temperature, AVPU Score. | | |
| **Local context** | Input does not depend on local context of data | | |
| **Methodology** | Rule-based algorithm | | |
| **Endorsement** | Recommended by:<br>• Australian Commission on Safety and Quality in Health Care, Australia<br>• National Health Services, United Kingdom | | |
| **Automation Flag** | Automated (However, in some hospitals a manual version is still used by nurses) | | |
| **Tool Citations** | 1,176 | Reported in 13 studies | |
| **Phase of Evaluation** | **Level of Evidence** | **Grade** | **Evaluation Studies** |
| **Phase C: Before implementation Does the tool work? Is it possible?** | Internal validation | **C3** | Developed and tested for internal validity:<br>• Subbe et al, 2001 [128] |
| | External validation | **C2** | Tested for External validity:<br>• Armagan et al, 2008 [159] |
| | External validation multiple times | **C1** | Tested for external validity multiple times:<br>• Burch, Tarr & Morroni, 2008 [160]<br>• Dundar et al, 2016 [161]<br>• Gardner-Thorpe et al, 2006 [162]<br>• TANRIÖVER et al, 2016 [164]<br>• Wang et al, 2016 [165]<br>• Salottolo et al, 2017 [163]<br>One negative conclusion validation/performance study:<br>• Tirotta et al, 2017 [166] |
| **Phase B: During implementation: Is the tool practicable?** | Potential effect | **B2** | Not reported |
| | Usability | **B1** | Not reported |

<table>
<tr><td rowspan="3"><b>Phase A:<br>After implementation: Is the tool desirable?</b></td><td rowspan="3">Evaluation of Post-Implementation Impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency</td><td><b>A3</b></td><td colspan="8">No subjective studies reported</td></tr>
<tr><td><b>A2</b></td><td colspan="8">One observational before-and-after intervention study failed to prove positive post-implementation impact of the MEWS on patient safety:<br>• Subbe et al, 2003 [167]<br>Three observational before-and-after intervention studies showed positive post-implementation impact of the MEWS on patient safety:<br>• Moon et al, 2011 [170]<br>• De Meester et al, 2013 [168]<br>• Hammond et al, 2013 [169]</td></tr>
<tr><td><b>A1</b></td><td colspan="8">No experimental studies reported</td></tr>
<tr><td><b>Tool Grade</b></td><td colspan="2"><b>Grade A2</b></td><td>A1</td><td>◕</td><td>A3</td><td>B1</td><td>B2</td><td>◕</td><td>C2</td><td>●</td></tr>
<tr><td rowspan="2"><b>Direction of Evidence</b></td><td colspan="2">● Positive Evidence</td><td colspan="8">◕ Mixed Evidence Supporting Positive Conclusion</td></tr>
<tr><td colspan="2">○ Negative Evidence</td><td colspan="8">◔ Mixed Evidence Supporting Negative Conclusion</td></tr>
<tr><td><b>Justification</b></td><td colspan="10">The MEWS is a prognostic tool for early detection of inpatients' clinical deterioration and potential need for higher levels of care. The tool uses a rule-based algorithm on clinical data to calculate a risk score [128]. The MEWS has been tested for external validity multiple times in different clinical areas, settings and populations [159-165]. All studies reported the tool is externally valid. However, one study reported MEWS poorly predicted the in-hospital mortality risk of patients with sepsis [166]. Using the mixed evidence protocol, the mixed evidence supports external validity, qualifying MEWS for Grade C1. No literature has been found regarding its usability or potential effect. The MEWS has been implemented in different healthcare settings. One observational before-and-after intervention study failed to prove positive post-implementation impact of the MEWS on patient safety in acute medical admissions [167]. However, three more recent observational before-and-after intervention studies reported positive post-implementation impact of the MEWS on patient safety. One study reported significant increase in frequency of patient observation and decrease in serious adverse events after intensive care unit (ICU) discharge [168]. The second reported significant increase in frequency of vital signs recording, 24h post-ICU discharge and 24h preceding unplanned ICU admission [169]. The third, an eight years study, reported that the post-implementation four years showed significant reductions in the incidence of cardiac arrests, the proportion of patients admitted to ICU and their in-hospital mortality [170]. Using the mixed evidence protocol, the mixed evidence supports positive post-implementation impact. The MEWS has been assigned Grade A2.</td></tr>
</table>

Table S7. Ottawa Knee Rule – Grade A1

| | |
|---|---|
| **Name** | Ottawa Knee Rule |
| **Authors/Year** | Dr. Ian Stiell, Canada, 1995 |
| **Intended use** | Exclude the need for an X-ray for possible bone fracture in adult patients |
| **Intended user** | Used by emergency physicians as part of the clinical examination |
| **Category** | Diagnostic |
| **Clinical area** | Orthopaedics |
| **Target Population** | Patients visiting the emergency department |
| **Target Outcome** | Bone fracture |
| **Action** | Refer patient to knee imaging |
| **Input source** | Objective data (clinical examination) + Subjective data (symptoms described by patient) |
| **Input type** | Clinical data: Age ≥55 (yes/no), Isolated tenderness of the patella (no other bony tenderness) (yes/no), Tenderness at the fibular head (yes/no), Unable to flex knee to 90° (yes/no), Unable to bear weight both immediately and in ED (4 steps, limping is okay) (yes/no). Data is obtained from the patient. |
| **Local context** | Input does not depend on local context of data |
| **Methodology** | Set of rules |
| **Endorsement** | Recommended by:<br>• Department of Emergency Medicine, Faculty of medicine, Ottawa University, Canada<br>• The Royal College of Radiologists, United Kingdom<br>• The National Institute for Health and Care Excellence, United Kingdom |

| | | | |
|---|---|---|---|
| **Automation Flag** | Manual | | |
| **Tool Citations** | 227 | Reported in 15 studies | |
| **Phase of Evaluation** | **Level of Evidence** | **Grade** | **Evaluation Studies** |
| **Phase C: Before implementation Does the tool work? Is it possible?** | Internal validation | **C3** | Developed and tested for internal validity:<br>- Stiell et al, 1995 [122] |
| | External validation | **C2** | Tested for externally validity |
| | External validation multiple times | **C1** | Externally tested for externally validity (One systematic review reported 11 validation studies):<br>- Bachmann et al, 2004 [171] |
| **Phase B: During implementation: Is the tool practicable?** | Potential effect | **B2** | Not reported |
| | Usability | **B1** | Not reported |
| **Phase A: After implementation: Is the tool desirable?** | Evaluation of Post-Implementation Impact on Clinical Effectiveness, Patient Safety or Healthcare Efficiency | **A3** | No subjective studies reported |
| | | **A2** | No observational studies reported |
| | | **A1** | Two nonrandomised controlled studies reported positive post-implementation impact of Ottawa knee rule on healthcare efficiency:<br>- Stiell et al, 1997 [172]<br>- Nichol et al, 1999 [173] |
| **Final Grade** | **Grade A1** | | ● A2 A3 B1 B2 ● C2 ● |
| **Direction of Evidence** | ● Positive Evidence | | ◕ Mixed Evidence Supporting Positive Conclusion |
| | ○ Negative Evidence | | ◔ Mixed Evidence Supporting Negative Conclusion |
| **Justification** | Ottawa knee rule is a diagnostic tool used to exclude the need for an X-ray for possible bone fracture in patients presenting to the ED, using a simple five items manual check list [122]. It is one of the oldest, most accepted and successfully used rules in CDS. The tool has been tested for external validity multiple times. One systematic review identified 11 studies, 6 of them involved 4,249 adult patients and were appropriate for pooled analysis, showing high sensitivity and specificity [171]. Furthermore, two studies discussed the impact of implementing Ottawa knee rule on healthcare efficiency. One nonrandomised controlled trial with before-after and concurrent controls included a total of 3,907 patients seen during two 12-month periods before and after the intervention. The study reported that the rule decreased the use of knee radiography without patient dissatisfaction or missed fractures and was associated with reduced waiting times and costs per patient [172]. Another nonrandomised controlled trial reported that the proportion of ED patients referred for knee radiography was reduced. The study also reported that the practice based on the rule was associated with significant cost savings [173]. The Ottawa knee rule has been assigned Grade A1. | | |

## *Predictive Performance, Usability and Post-implementation Impact Tables of Predictive Tools*

Table S8: Predictive Performance of the Five Tools – Before Implementation

| Tool | Discrimination | | Calibration |
|---|---|---|---|
| | **AUC/C-Statistic** | **Sensitivity, Specificity, Cut-Off** | **Hosmer–Lemeshow goodness-of-fit** |
| LACE Index | • 0.68 (95% CI, 0.68–0.69) [125]<br>• 0.68 [129]<br>• 0.56 (95% CI, 0.46–0.66) [187] | • 66.3%, 53.3%, 50% [133] | • 14.1 (P=0.59) [125] |
| Centor Score | • 0.78 [126]<br>• 0.72 [138]<br>• 0.84 [136] | • 90%, 92%, 50% [126]<br>• 49%, 82%, 50% [135]<br>• 92%, 73%, 50% [139]<br>• 92%, 63%, 50% [136] | • Not reported |
| Wells' Criteria | • 0.71 [153]<br>• 0.75 [154]<br>• 0.79 (95% CI, 0.75-0.82) [152]<br>• 0.79 (95% CI, 0.72–0.87) [155]<br>• 0.76 [156]<br>• 0.74 (95% CI,0.72-0.76) [150] | • 83%, 48%, 50% [153]<br>• 65%, 81%, 50% [152]<br>• 100%, 56%, 50% [156]<br>• 95%, 51%, 50% [149] | • Not reported |
| MEWS | • 0.73 (95% CI, 0.69–0.77) – Hospitalisation [161]<br>• 0.89 (95% CI 0.84–0.94) – In-hospital mortality [161]<br>• 0.79 (95% CI, 0.74-0.83) – Mortality [163]<br>• 0.56 (95% CI, 0.51 to 0.62) – ICU Admission [163]<br>• 0.85 (95% CI, 0.77-0.91) [164]<br>• 0.80 (95% CI, 0.72–0.88) [37]<br>• 0.76 – Mortality [188] | • 88%, 68%, 50% (MEWS≥3) [162]<br>• 53%, 91%, 50% (MEWS≥4) Mortality [163]<br>• 17%, 94%, 50% (MEWS≥4) ICU Admission [163]<br>• 86%, 94%, 50% (MEWS≥4) [164]<br>• 75%, 83%, 50% (MEWS≥4) [189]<br>• 57%, 86%, 50% (MEWS≥4) [188] | • P=0.06 [190] |
| Ottawa Knee Rule | • Not reported | • 98.5%, 48.6%, 50% [171]*<br>• 100% [172]<br>• 100%, 42.8%, 50% [191]<br>• 95%, 44%, 50% [192] | • Not reported |

* A systematic review study.

Table S9. Usability of Two Predictive Tools – During Implementation

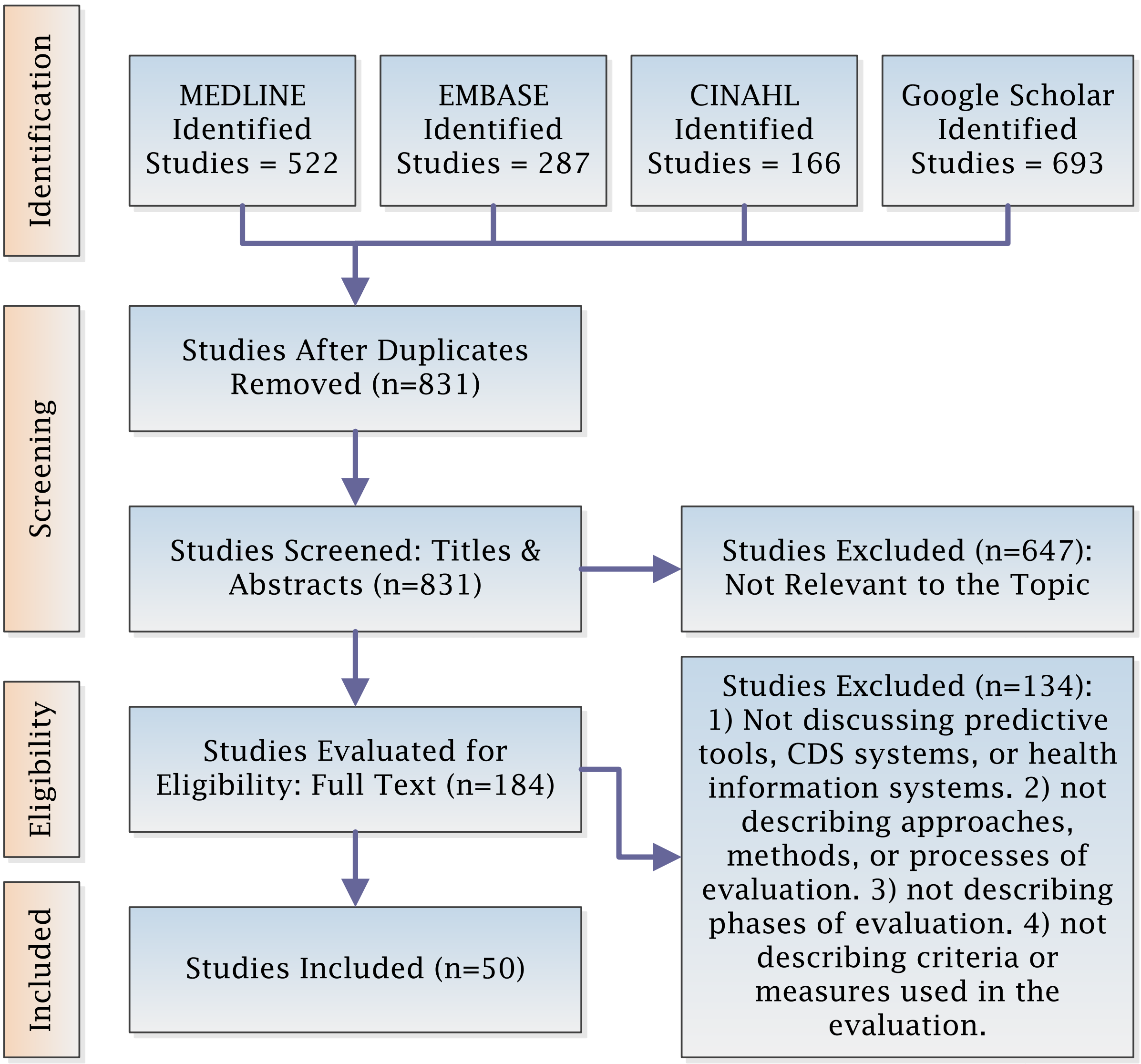
| Tool | Study Type | Method | Outcomes |
|---|---|---|---|
| Centor Score | Usability testing [143] | Think Aloud + Near Live | • Positive usability & feedback of users<br>○ Easiness of use<br>○ Usefulness |
| Wells' Criteria | Usability testing [157] | Think Aloud + Near Live | • Positive usability & feedback of users<br>○ Tool is helpful<br>○ Organized<br>○ Did not compromise clinical judgment |

Table S10. Post-Implementation Impact of Three Predictive Tools

| Tool | Study Type | Study Settings | Outcome | Effect Size |
|---|---|---|---|---|
| Wells' Criteria | Prospective before-and-after intervention study [158] | Public-sector tertiary-level and referral teaching hospital in South Africa | Efficiency of CTPA utilisation | 17.4% vs 30.7% (p=0.036) |
| | | | Inappropriate CTPA scans | 82.6% vs 69.3% (p=0.015) |
| MEWS | Prospective before-and-after intervention study [168] | A University Hospital, in Belgium | Frequency of patient observation | 0.99 vs 1.07 (p=0.005) |
| | | | Serious adverse events after ICU discharge | 5.7% vs 3.5% |
| | Prospective before-and-after intervention study [169] | The department of intensive care medicine, at a tertiary referral hospital in Brisbane, Australia | Vital signs documentation after ICU discharge | 210% (95% CI 148, 288%, p <0.001). |
| | | | Vital signs documentation before unplanned ICU admissions | 44% (95% CI, 3, 102%, p = 0.035). |
| | Retrospective analysis of prospectively collected data before-and-after intervention study [170] | The department of perioperative and critical care at a university teaching hospital in the United Kingdom | Cardiac arrest calls | 0.2% vs 0.4% (p<0.0001) |
| | | | Patients admitted to ICU | 2% vs 3% (p=0.004) |
| | | | In-hospital mortality of cardiac arrest patients | 42% vs 52% (p=0.05) |
| Ottawa Knee Rule | Nonrandomised controlled trial with before-after & concurrent controls [172] | The Emergency departments of two teaching and two community hospitals in Canada | Reduced time spent by patient | 85.7 minutes vs 118.8 minutes |
| | | | Cost savings per patient | US $80 vs US $183 |
| | Nonrandomised controlled trial with before-after & concurrent controls [173] | The Emergency departments of an academic and a community hospital in Canada. | Reduced proportion of knee injury patients referred to radiology | 77.6% vs 57.1% |
| | | | Cost savings per patient | $31 (95% CI, 22 to 44) to $34 (95% CI, 24 to 47). |

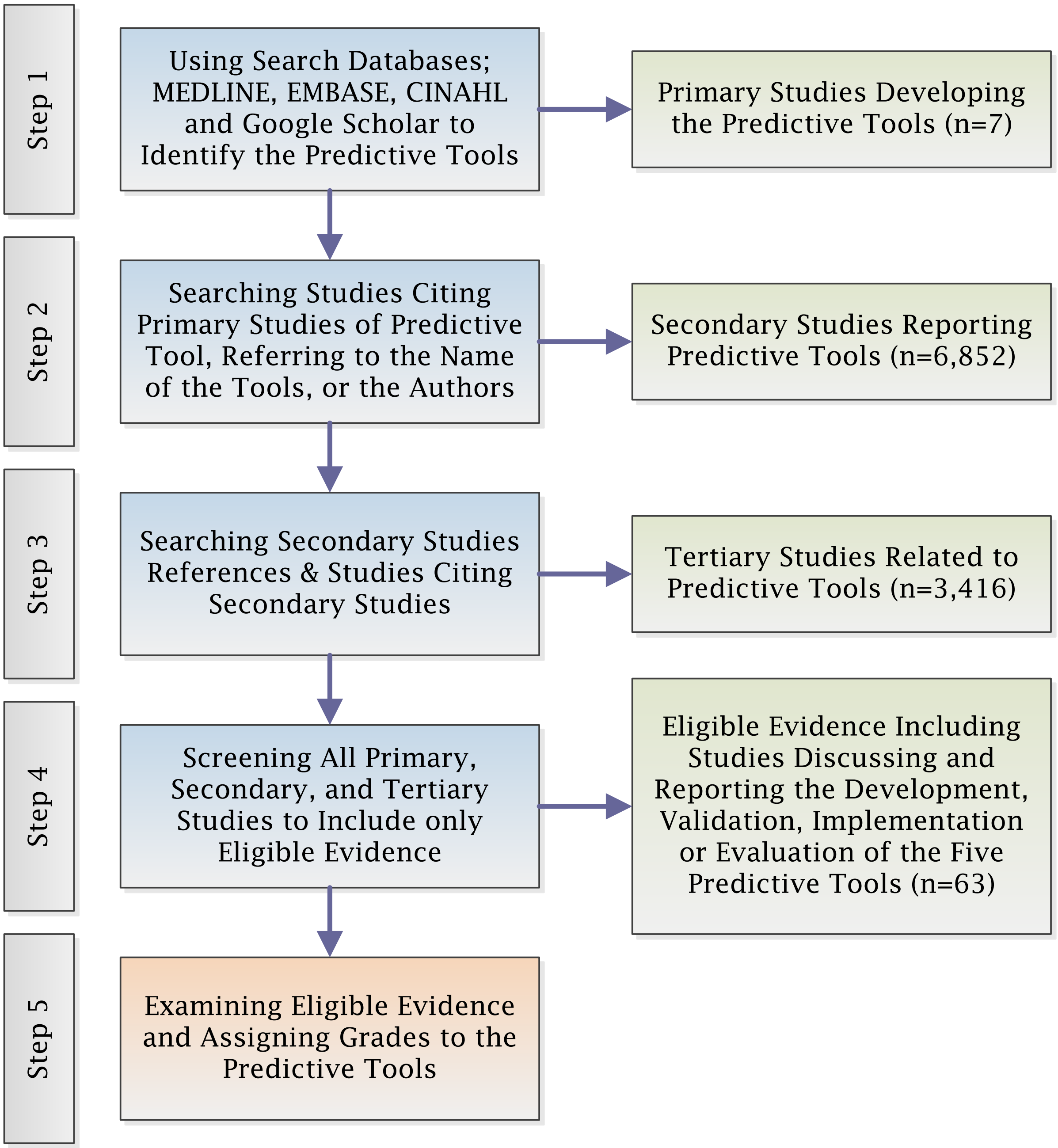
***Study Selection Process***

**Figure S1.** Study Selection for the Focused Review of Literature


Identification
MEDLINE Identified Studies = 522
EMBASE Identified Studies = 287
CINAHL Identified Studies = 166
Google Scholar Identified Studies = 693
Screening
Studies After Duplicates Removed (n=831)
Studies Screened: Titles & Abstracts (n=831)
Studies Excluded (n=647): Not Relevant to the Topic
Eligibility
Studies Evaluated for Eligibility: Full Text (n=184)
Studies Excluded (n=134): 1) Not discussing predictive tools, CDS systems, or health information systems. 2) not describing approaches, methods, or processes of evaluation. 3) not describing phases of evaluation. 4) not describing criteria or measures used in the evaluation.
Included
Studies Included (n=50)

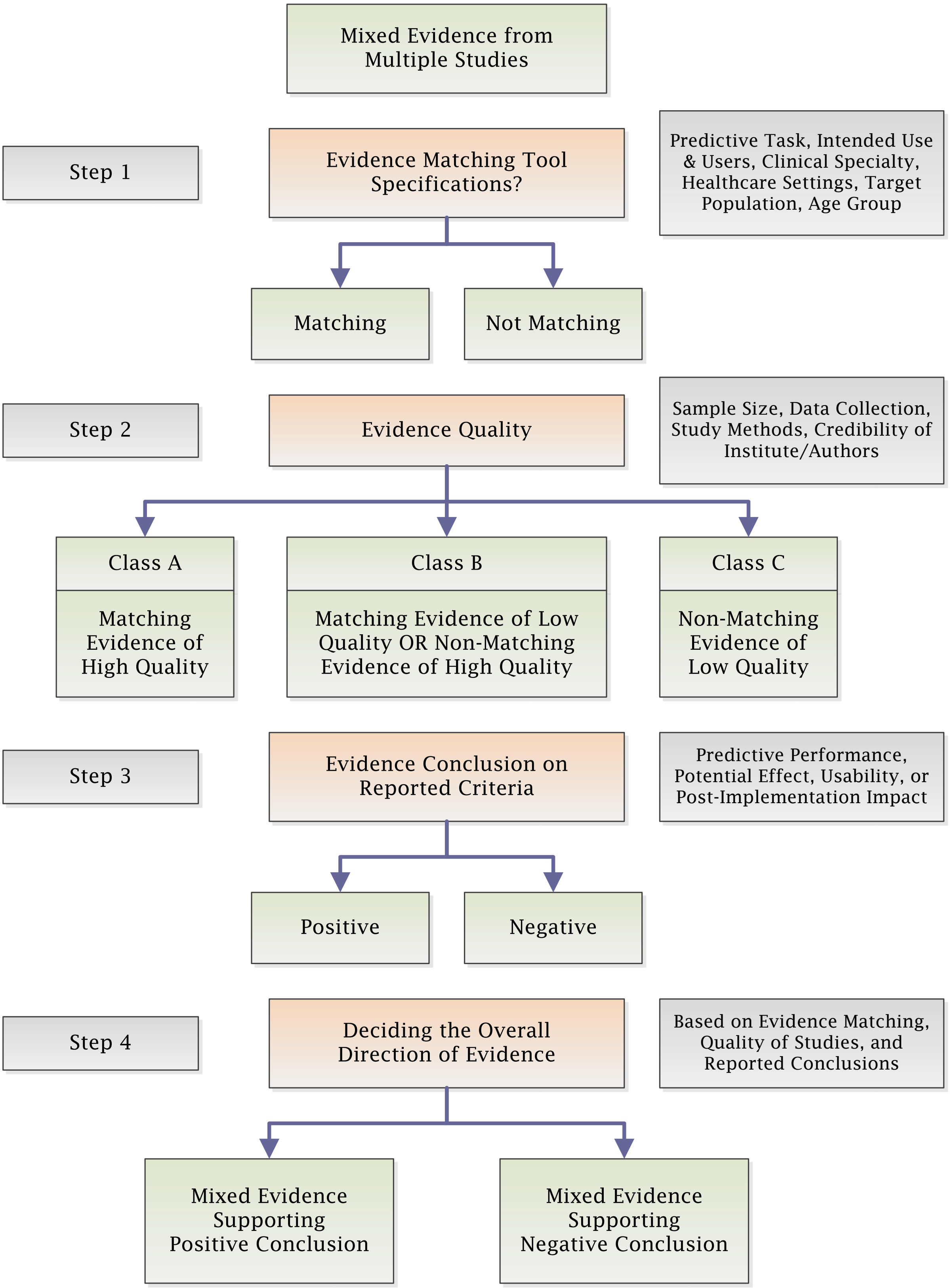
## *Searching the Literature for Published Evidence on Predictive Tools*

**Figure S2.** Searching the Literature for Published Evidence on Predictive Tools


Step 1
Using Search Databases; MEDLINE, EMBASE, CINAHL and Google Scholar to Identify the Predictive Tools
Primary Studies Developing the Predictive Tools (n=7)
Step 2
Searching Studies Citing Primary Studies of Predictive Tool, Referring to the Name of the Tools, or the Authors
Secondary Studies Reporting Predictive Tools (n=6,852)
Step 3
Searching Secondary Studies References & Studies Citing Secondary Studies
Tertiary Studies Related to Predictive Tools (n=3,416)
Step 4
Screening All Primary, Secondary, and Tertiary Studies to Include only Eligible Evidence
Eligible Evidence Including Studies Discussing and Reporting the Development, Validation, Implementation or Evaluation of the Five Predictive Tools (n=63)
Step 5
Examining Eligible Evidence and Assigning Grades to the Predictive Tools

***The Mixed Evidence Protocol***

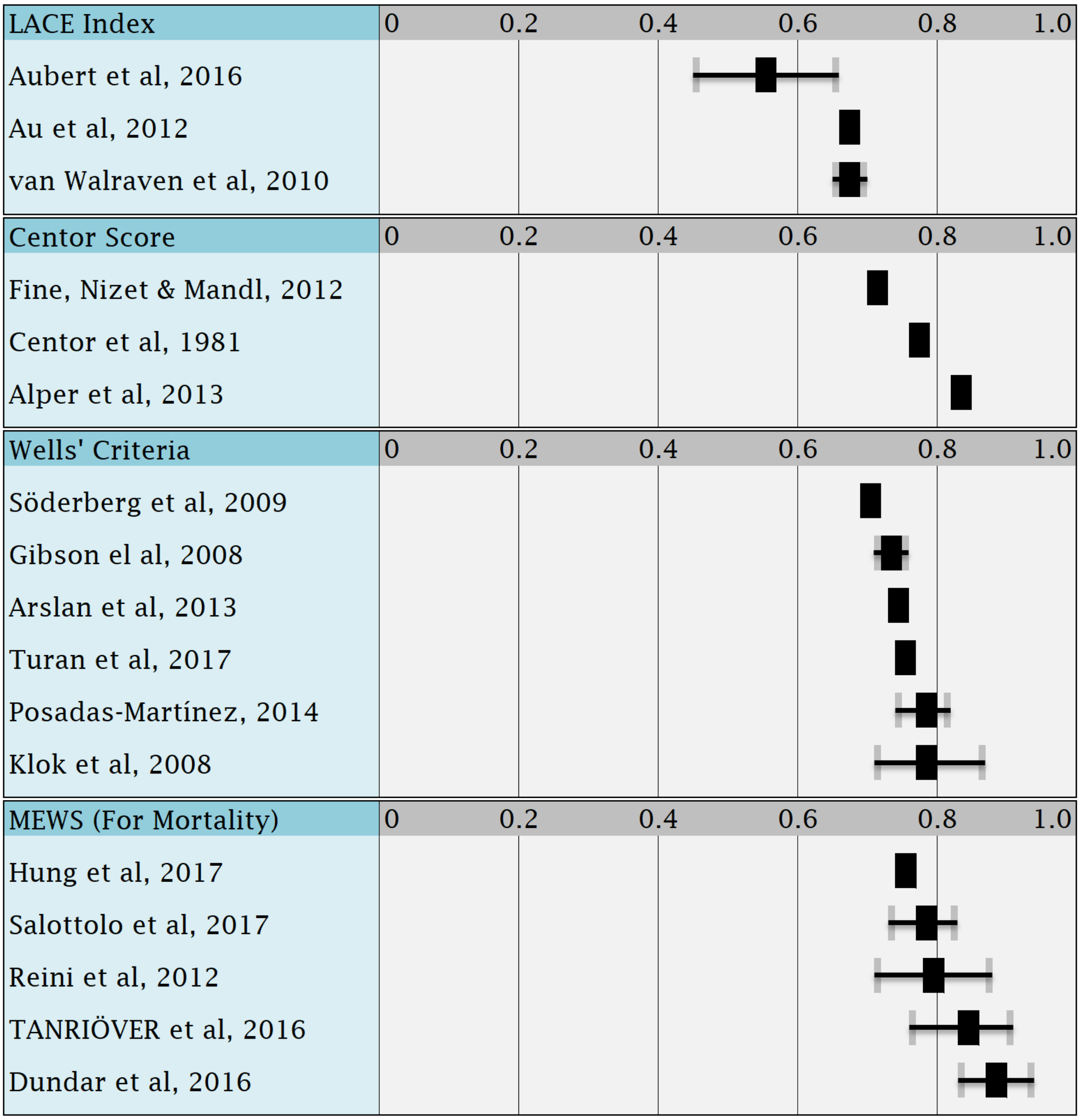
**Figure S3.** The Mixed Evidence Protocol


Mixed Evidence from Multiple Studies
Step 1
Evidence Matching Tool Specifications?
Predictive Task, Intended Use & Users, Clinical Specialty, Healthcare Settings, Target Population, Age Group
Matching
Not Matching
Step 2
Evidence Quality
Sample Size, Data Collection, Study Methods, Credibility of Institute/Authors
Class A
Matching Evidence of High Quality
Class B
Matching Evidence of Low Quality OR Non-Matching Evidence of High Quality
Class C
Non-Matching Evidence of Low Quality
Step 3
Evidence Conclusion on Reported Criteria
Predictive Performance, Potential Effect, Usability, or Post-Implementation Impact
Positive
Negative
Step 4
Deciding the Overall Direction of Evidence
Based on Evidence Matching, Quality of Studies, and Reported Conclusions
Mixed Evidence Supporting Positive Conclusion
Mixed Evidence Supporting Negative Conclusion

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

### *Performance Figures of Predictive Tools*

**Figure S4.** Reported C-Statistic of LACE Index, Centor Score, Wells Criteria and MEWS


LACE Index
Aubert et al, 2016
Au et al, 2012
van Walraven et al, 2010
Centor Score
Fine, Nizet & Mandl, 2012
Centor et al, 1981
Alper et al, 2013
Wells' Criteria
Söderberg et al, 2009
Gibson el al, 2008
Arslan et al, 2013
Turan et al, 2017
Posadas-Martínez, 2014
Klok et al, 2008
MEWS (For Mortality)
Hung et al, 2017
Salottolo et al, 2017
Reini et al, 2012
TANRIÖVER et al, 2016
Dundar et al, 2016