Cowley et al. Diagnostic and Prognostic Research (2019) 3:16
https://doi.org/10.1186/s41512-019-0060-y

Diagnostic and Prognostic Research

REVIEW Open Access

Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature

Laura E. Cowley*, Daniel M. Farewell, Sabine Maguire and Alison M. Kemp

Abstract

Clinical prediction rules (CPRs) that predict the absolute risk of a clinical condition or future outcome for individual patients are abundant in the medical literature; however, systematic reviews have demonstrated shortcomings in the methodological quality and reporting of prediction studies. To maximise the potential and clinical usefulness of CPRs, they must be rigorously developed and validated, and their impact on clinical practice and patient outcomes must be evaluated. This review aims to present a comprehensive overview of the stages involved in the development, validation and evaluation of CPRs, and to describe in detail the methodological standards required at each stage, illustrated with examples where appropriate. Important features of the study design, statistical analysis, modelling strategy, data collection, performance assessment, CPR presentation and reporting are discussed, in addition to other, often overlooked aspects such as the acceptability, cost-effectiveness and longer-term implementation of CPRs, and their comparison with clinical judgement. Although the development and evaluation of a robust, clinically useful CPR is anything but straightforward, adherence to the plethora of methodological standards, recommendations and frameworks at each stage will assist in the development of a rigorous CPR that has the potential to contribute usefully to clinical practice and decision-making and have a positive impact on patient care.

Keywords: Clinical prediction rule, Prediction model, Risk model, Model development, Model validation, Impact studies, Model reporting, Implementation, Diagnosis, Prognosis, Study design

Background

The aim of a clinical prediction rule (CPR) is to estimate the probability of a clinical condition or a future outcome by considering a small number of highly valid indicators [1, 2]. CPRs include three or more predictors, from patients' clinical findings, history or investigation results [3]. Their purpose is to assist clinicians in making decisions under conditions of uncertainty and enhance diagnostic, prognostic or therapeutic accuracy and decision-making, with the ultimate aim of improving the quality of patient care [1, 2, 4]. The predicted probabilities from a CPR allow clinicians to stratify patients into risk groups and help them to decide whether further assessment or treatment is necessary [5]. Some CPRs can help to 'rule in' a condition by identifying patients who are very likely to have a condition and who thus require additional diagnostic testing or treatment, whilst others aim to 'rule out' a condition by identifying patients who are very unlikely to have a condition, thus reducing unnecessary testing without compromising patient care [2, 4]. CPRs that aim to predict the probability of a condition being present are termed *diagnostic* or *screening* rules; those that aim to predict the probability of a future outcome are termed *prognostic* rules; and those that aim to predict the probability that a specific treatment or intervention will be effective are termed *prescriptive* rules [2].

To maximise the predictive accuracy and clinical utility of CPRs, it is vital that they are rigorously developed, validated and evaluated. However, numerous systematic reviews have demonstrated shortcomings in the methodological quality and reporting of prediction studies, which restricts the CPR's usefulness in practice [6–15]. Methodological standards for the development of CPRs were originally outlined by Wasson and colleagues [16]. With the increase in popularity of CPRs inspired by the evidence-based medicine movement, these standards

* Correspondence: CowleyLE@cardiff.ac.uk
Division of Population Medicine, School of Medicine, Neuadd Meirionnydd, Heath Park, Cardiff University, Wales CF14 4YS, UK

have since been modified and updated by a number of authors over the years [3, 4, 17–19]. Experts have provided thorough and accessible overviews of the principles and methods involved in conducting diagnostic and prognostic research [20–32] and devised frameworks to enhance the conduct and interpretation of prediction studies [33–35]. They have also provided guidance and recommendations for researchers to consider when developing and evaluating CPRs, without aiming to dictate how analyses should be conducted. These recognise that there is no clear consensus on many aspects of model development, that the field is continually evolving and that methodological standards will therefore require updating accordingly [36]. Guidelines for the *reporting* of clinical prediction research have also been developed, namely the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [36].

This review aims to outline the stages and methodological standards involved in the development and evaluation of CPRs, illustrated with examples where appropriate.

Terminology used in this review

In the literature, the term 'clinical prediction rule' is used interchangeably with the terms clinical prediction tool [37], clinical decision rule [17], clinical decision tool [38], clinical prediction algorithm [39], prognostic score [40], prognostic model [21], risk prediction model [23], risk model [30], risk score [41], scoring tool [42], scoring system [43] or risk index [44]. Reilly and Evans [32] distinguish between *assistive prediction* rules that simply provide clinicians with diagnostic or prognostic predicted probabilities without recommending a specific clinical course of action, and *directive decision* rules that explicitly suggest additional diagnostic tests or treatment in line with the obtained score. Decision rules intend to directly influence clinician behaviour, while prediction rules intend to help clinicians predict risk without providing recommendations, with the assumption that accurate predictions will lead to better decisions [32]. Some researchers also distinguish between prediction *models* that provide predicted probabilities along the continuum between certified impossibility ($Pi = 0$) and absolute certainty ($Pi = 1$) [45], and prediction *rules* that classify patients into risk groups, by applying a clinically relevant cut-off that balances the likelihood of benefit with the likelihood of harm [19, 46]. Such cut-offs are known as 'decision thresholds'; a threshold must be applied if a prediction model aims to influence decision-making [19]. In this review, the term 'clinical prediction rule' is used to refer to diagnostic, prognostic or prescriptive rules/models derived from multivariable statistical analyses, which predict the probability of a condition or outcome, *with or without* the use of a clinical cut-off or recommendation for further action.

Stages in the development of clinical prediction rules

It is widely acknowledged in the literature that there are three *main* stages in the development of CPRs (Fig. 1); derivation; external validation; and impact analysis to determine their impact on patient care [4, 20, 22–25, 32, 33]. Stiell and Wells [17] identified a further three important stages, namely identifying the need for a CPR, determining the cost-effectiveness of a CPR and long-term dissemination and implementation of a CPR. Therefore all six stages are summarised in Table 1 and discussed in detail below.

Detailed methodological and practical recommendations pertaining to the three main stages of development have been published, as each requires a different methodological approach [3, 4, 16–36]. These three stages also correspond to increasing hierarchies of evidence, as outlined in Table 2 [4, 32, 33]. A CPR that has been *derived*, but not externally validated, corresponds to the lowest level of evidence and is not recommended for use in clinical practice, except arguably in rare instances when a CPR is developed for use in only one setting. It has been suggested that a CPR that has been successfully externally *validated* in a setting, or population, similar to the one from which it was derived ('narrow' validation), can be used cautiously in similar future patients [32]. Similarly, it is proposed that a CPR should be consistently successfully externally validated in multiple settings or populations ('broad' validation), before clinicians can use its predictions confidently in future patients [32]. Finally, it is recommended that an *impact analysis* is conducted and that the CPR demonstrates improvements to patient care, before it can be used as a decision rule for the management and treatment of patients [32]. Ideally, the impact of a CPR should also be tested in multiple settings. Impact analysis studies correspond to the highest level of evidence [32].

Stage 1: identifying the need for a clinical prediction rule

Before developing a CPR, researchers need to ensure that there is a clinical need for the rule. CPRs are most valuable when decision-making is challenging, when there is evidence that clinicians are failing to accurately diagnose a condition, and when there are serious consequences associated with an incorrect diagnosis [2, 4]. CPRs are also valuable when there is a need to simplify or speed up the diagnostic or triage process, for example in patients presenting to the emergency department with chest pain and suspected acute cardiac ischaemia [49]. CPRs are most likely to be adopted into clinical practice, and to demonstrate improvements in patient care and reductions in health care costs, when they improve the overall efficiency of clinical practice [17]. For example, ankle injuries are frequently seen in the emergency department. Prior to the implementation of the Ottawa

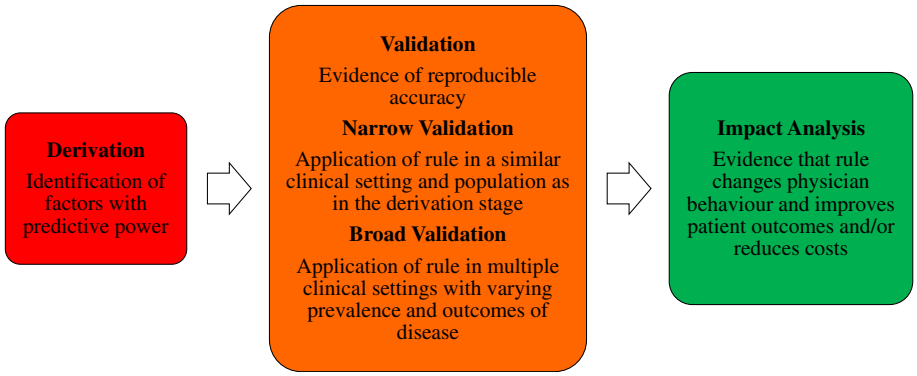

Fig. 1 The three main stages in the development and evaluation of clinical prediction rules. Adapted from McGinn, 2016 [47]

Ankle Rule, clinicians ordered a high proportion of radiographs that were negative for fracture, when the majority of them believed that a fracture was highly unlikely [50]. The rule was found to lead to a reduction in both radiography [51] and health care costs [52], and in one survey 70% of Canadian and UK emergency department clinicians reported frequent use of the rule [53].

Before developing a CPR, researchers should consider whether a new CPR is needed, as many are developed for the same target population or to predict the same outcome [8, 10, 11, 54–57]. The characteristics, performance and level of evidence of existing CPRs should be systematically reviewed using validated search filters for locating prediction studies, and the Critical Appraisal and Data Extraction for Systematic Reviews of prediction modelling studies (CHARMS) checklist [58, 59]. The recently published Prediction model Risk Of Bias ASsessment Tool (PROBAST) can be used to assess the risk of bias and applicability of CPRs [60]. Researchers can also assess the performance of existing CPRs on their own collected data [61]. Existing CPRs with potential should be updated, validated or tested in an impact study before a new CPR is developed [54, 62, 63]. If a new CPR is derived, researchers should clearly justify why it is required, with reference to existing CPRs, to avoid research waste and duplication of efforts [64]. Qualitative research with clinicians can be useful in determining whether a proposed CPR is clinically relevant, and to assess the credibility of the proposed predictor variables [65, 66].

Stage 2: derivation of a clinical prediction rule according to methodological standards

Once a need for a new CPR is established, and a researcher has an appropriate clinical question, a CPR must be derived according to strict methodological standards [23]. There are various elements to consider, pertaining to the study design, statistical techniques employed and the assessment, presentation and reporting of the CPR. Researchers should consider writing and publishing a study protocol and registering the study prior to the derivation of a new CPR, in the interests of transparency [67, 68].

Study design for the derivation of a clinical prediction rule

The first stage in the development of a CPR is the derivation of the rule. This involves an examination of the ability of multiple potential variables from the clinical findings, history or investigation results to predict the target outcome of interest. Predicted probabilities are derived from the statistical analysis of patients with known outcomes, and the outcome of interest serves as the reference standard by which the performance of the CPR is assessed. The performance of a CPR is dependent upon the quality of the underlying data, and the dataset used to derive the CPR should be representative of the target population it is intended for [17, 30, 69, 70].

The optimal study design for the derivation of a diagnostic CPR is a cross-sectional cohort study, while for prognostic CPRs, the preferred design is a longitudinal cohort study [30]. In general, case-control studies are inappropriate, as they do not allow for the estimation of absolute outcome risk [21, 23, 71]; however, nested case-control or case-cohort studies can be used [71, 72]. Prospective cohort studies are preferred to retrospective cohort studies, to optimise measurement and documentation of predictive and outcome variables [21, 23]. For prescriptive CPRs, study designs that include a control group, such as randomised controlled trials (RCTs), are essential to ensure that treatment effect modifiers and non-specific prognostic predictors are distinguishable from one another [73, 74]. The study design should be adequately detailed and include the study setting, inclusion and exclusion criteria and patient demographics and characteristics [17]. To enhance generalisability, multicentre studies are recommended [30].

Statistical analysis

Commonly used statistical methods for the derivation of CPRs include multivariable regression techniques, and recursive partitioning techniques, such as classification

Table 1 Stages in the development and evaluation of clinical prediction rules

Stage of development	Methodological standards
Stage 1. Identifying the need for a CPR	• Consider conducting qualitative research with clinicians to determine clinical relevance and credibility of CPR
	• Conduct a systematic review of the literature to identify and evaluate existing CPRs developed for the same purpose
	• Consider updating, validating or testing the impact of existing CPRs
Stage 2. Derivation of a CPR according to methodological standards	Study design for the derivation of a CPR
	• Consider registering the study and publishing a protocol
	• Ensure the dataset is representative of the population for whom the CPR is intended
	• Conduct a prospective multicentre cohort study
	Statistical analysis
	• Conduct multivariable regression analysis (logistic for binary outcomes, Cox for long-term prognostic outcomes) • Identify the model to be used, plus rationale if other methods used
	Missing data
	• Use multiple imputation
	Selection of candidate predictors for inclusion in a multivariable model
	• Only include relevant predictors based on evidence in the literature/clinical experience
	• Aim for a sample size with a minimum of ten events per predictor, preferably more
	• Avoid selection based on univariable significance testing
	• Avoid categorising continuous predictors
	Selection of predictors during multivariable modelling
	• Backward elimination of predictors is preferred
	• Avoid data-driven selection and incorporate subject-matter knowledge into the selection process
	Definition and assessment of predictor and outcome variables
	• Define predictor and outcome variables clearly
	• Consider inter-rater reliability of predictor measurement and potential measurement error
	• Aim for blind assessment of predictor and outcome variables
	Internal validation
	• Use cross-validation or bootstrapping and adjust for optimism
	• Ensure to repeat each step of model development if using bootstrapping
	CPR performance measures
	• Assess and report both calibration and discrimination
	• Consider decision curve analysis to estimate the clinical utility of the CPR
	Presentation of a CPR
	• Report the regression coefficients of the final model, including the intercept or baseline hazard
	• Consider a clinical calculator if the CPR is complex
	Reporting the derivation of a CPR
	• Adhere to the TRIPOD guidelines [36]
Stage 3. External validation and refinement of a CPR	Study design for the external validation of a CPR
	• Conduct a prospective multicentre cohort study
	• Aim for a sample size with a minimum of 100 outcome events, preferably 200
	• Consider using a framework of generalisability to enhance the interpretation of the findings [34]
	Types of external validation
	• Conduct temporal, geographical and domain validation studies to ensure maximum generalisability
	• If multiple validations have been performed, conduct a meta-analysis to summarise the overall performance of the CPR, using a published framework [35]
	Refinement of a CPR: model updating or adjustment
	• Consider updating, adjusting or recalibrating the CPR if poor performance is found in an external validation study
	• Consider further external validation of updated CPRs
	Comparing the performance of CPRs
	• Compare the CPR with other existing CPRs for the same condition
	• Ensure the statistical procedures used for comparison are appropriate; consider a decision-analytic approach
	Reporting the external validation of a CPR
	• Adhere to the TRIPOD guidelines [36]
Stage 4. Impact of a CPR on clinical practice	Study design for an impact analysis
	• Consider whether the CPR is ready for implementation
	• Conduct a cluster randomised trial with centres as clusters, or a before–after study
	• Perform appropriate sample size calculations
	• Consider decision-analytic modelling as an intermediate step prior to a formal impact study
	Measures of impact of a CPR
	• Report the safety and efficacy of the CPR
	• Report the impact of the CPR on clinician behaviour if assessed

Table 1 Stages in the development and evaluation of clinical prediction rules *(Continued)*

Stage of development	Methodological standards
	Acceptability of a CPR
	• Evaluate the acceptability of the CPR using the validated OADRI [48], or using qualitative or vignette methods
	Comparison of a CPR with unstructured clinical judgement
	• Compare the sensitivity and specificity of the CPR with clinicians own predictions/ decisions
	The four phases of impact analysis for CPRs
	• Follow the framework for the impact analysis of CPRs [33]
	• Ensure extensive preparatory and feasibility work is conducted prior to a formal impact study
	Reporting the impact analysis of a CPR
	• There are currently no published reporting guidelines for impact studies of CPRs; this is an area for future research
Stage 5. Cost-effectiveness	• Conduct a formal economic evaluation, with sensitivity analyses to examine the uncertainty of the model projections
Stage 6. Long-term implementation and dissemination	• Devise and evaluate targeted implementation strategies to ensure maximum uptake
	Barriers and facilitators to the use of CPRs
	• Assess barriers to the use of the CPR and devise strategies to overcome these

CPR clinical prediction rule, *TRIPOD* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis, *OADRI* Ottawa Acceptability of Decision Rules Instrument

and regression tree analysis [75]. Methods based on univariable analysis, where individual risk factors are simply totalled and assigned arbitrary weightings, should be avoided, as they are much less accurate than methods based on multivariable analysis [76]. This is because the final model may include predictors that are potentially related to each other and not independently associated with the outcome of interest [76]. Multivariable methods overcome the limitations of univariable analysis by enabling improved assessment of the association of the predictors with the target outcome [76].

In the case of multivariable regression, logistic regression models are required to predict binary events such as the presence or absence of a condition, while Cox regression models are suitable for time-to-event outcomes. Such models estimate regression coefficients (e.g. log odds or hazard ratios) of each predictor. Regression coefficients are mutually adjusted for the other predictors, and thus represent the contribution of each predictor to the probability of the outcome [23]. The probability of an outcome can be computed for a patient by combining the observed values of the predictors and their corresponding regression coefficients with the model intercept, or estimated baseline hazard [23]. For logistic models, the model intercept and the weighted values applicable to each patient are summed [16]. Specific values are assigned to each predictor, which are multiplied by the corresponding coefficients. In the case of a model with only binary categorical predictors, the predictors are multiplied by 0 or 1, depending on whether they are absent (0) or present (1), as per the model in Table 3 [77]. Exponentiating the final risk score gives the odds, and the probability (absolute risk) is calculated by use of the inverse logistic link function [78]. In this way, the probability of an outcome can be estimated from any combination of the predictor values [36]. The estimated probability for an individual without any of the predictors depends only on the intercept [23]. In this case, the value for each of the predictors will be 0; when each of these is multiplied by its relevant coefficient the value of 0 is retained [78]. For Cox regression models, the baseline hazard is estimated separately [26, 29].

Recursive partitioning involves repeatedly splitting patients into subpopulations including only individuals with a specific outcome [79], and was the method used to derive the Ottawa Ankle Rule [80]. CPRs can also be derived using discriminant function analysis [3], and machine learning algorithms based on artificial neural networks [1]. Artificial intelligence and

Table 2 Hierarchies of evidence in the development and evaluation of clinical prediction rules

Level of evidence	Definitions and standards of evaluation	Implications for clinicians
Level 1: Derivation of CPR	Identification of predictors using multivariable model; blinded assessment of outcomes.	Needs validation and further evaluation before it is used clinically in actual patient care.
Level 2: Narrow validation of CPR	Validation of CPR when tested prospectively in one setting; blinded assessment of outcomes.	Needs validation in varied settings; may use CPR cautiously in patients similar to derivation sample.
Level 3: Broad validation of CPR	Validation of CPR in varied settings with wide spectrum of patients and clinicians.	Needs impact analysis; may use CPR predictions with confidence in their accuracy.
Level 4: Narrow impact analysis of CPR used for decision-making	Prospective demonstration in one setting that use of CPR improves clinicians' decisions (quality or cost-effectiveness of patient care).	May use cautiously to inform decisions in settings similar to that studied.
Level 5: Broad impact analysis of CPR used for decision-making	Prospective demonstration in varied settings that use of CPR improves clinicians' decisions for wide spectrum of patients.	May use in varied settings with confidence that its use will benefit patient care quality or effectiveness.

Adapted from Reilly and Evans 2016 [32]. *CPR* clinical prediction rule

Table 3 Clinical prediction rule for postoperative nausea and vomiting (PONV) [77]

Risk of PONV = 1/(1 + exp. – [2.28 + 1.27 × female sex + 0.65 × history of PONV or motion sickness + 0.72 × non-smoking + 0.78 × postoperative opioid use])

machine learning approaches are becoming increasingly more common [81, 82].

Missing data

In clinical research, investigators almost always encounter missing observations involving predictor or outcome variables, even in carefully designed studies and in spite of their best efforts to maximise data quality [83]. There are three types of missing data mechanisms: (1) missing completely at random (MCAR), (2) missing at random (MAR) and (3) missing not at random (MNAR) [84]. When data are MCAR, this means that there are no systematic differences between the missing and observed values; for example, laboratory tests may be missing because of a dropped test tube or broken equipment. When data are MAR, this means that the probability of a missing value depends on the observed values of other variables (but not the unobserved values); for example, missing blood pressure measurements may be lower than observed measurements because younger people may be more likely to have missing measurements; in this case, data can be said to be MAR given age [85]. When data are MNAR, this means that the probability of a missing value depends on the unobserved values or other unobserved predictors, conditional on the observed data; for example, people with high blood pressure may be more likely to miss a doctor's appointment due to headaches [85]. Missing values are rarely MCAR, that is, their 'missingness' is usually directly or indirectly related to other subject or disease characteristics, including the outcome [23, 25]. Missing data is frequently addressed with case-wise deletion, which excludes all participants with missing values from the analysis [85]. However, when data are plausibly MAR, this reduces sample size and statistical power and biases the results [85], leading to inaccurate estimates of predictor-outcome relationships and the predictive performance of the model, since the participants with complete data are not a random subsample of the original sample [84, 86, 87].

Multiple imputation is a popular approach to the problem of missing data [83, 85, 86, 88–91], as it quantifies the uncertainty in the imputed values, by generating multiple different plausible imputed datasets, and pooling the results obtained from each of them [85, 91]. Multiple imputation involves three stages [85, 89, 91–93]. First, as the name suggests, multiple imputed datasets are created, based on the distribution of the observed data. This first stage accounts for uncertainty in estimating the missing values by adding variability into the values across the imputed datasets. In the second stage, standard statistical techniques are used to fit the models that are of interest in the substantive analysis to each of the imputed datasets. Estimated associations in each of the imputed datasets will be different, due to the variability introduced in stage one. In the third and final stage, the multiple results are averaged together, and standard errors are calculated using Rubin's combination rules [91], which account for both within-and between-imputation variability and the number of imputed datasets, and therefore the uncertainty of the imputed values. Multiple imputation typically assumes that data are MAR [93]. Importantly, the MAR assumption is just that; an assumption, rather than a property of the data [85]. The MCAR assumption can be tested, but it is not possible to differentiate between MAR and MNAR from the observed data [26, 85]. Most missing data are expected to be at least partly MNAR [85, 94, 95]. Sensitivity analyses can help to determine the effect of different assumptions about the missing data mechanism; work in this area is ongoing [96–100]. Other statistically principled approaches to dealing with missing data have been developed, based on random effects models [101, 102], Bayesian methods or maximum likelihood estimation [103] or, where data are longitudinal, joint models [104, 105]. Guidelines for reporting on the treatment of missing data in clinical and epidemiological research studies have been suggested by Sterne and colleagues [85]. Guidance also exists for handling missing data when deriving and validating CPRs [83, 106, 107]. It has been demonstrated that the outcome should be used for imputation of missing predictor values [87]. It is also becoming increasingly apparent that a real-time strategy to impute missing values is desirable when applying a CPR in clinical practice [108–110]. This is because one or more predictor variables may be unobserved for a particular patient, and thus the CPRs risk prediction cannot be estimated at the time of decision-making [108]. Real-time multiple imputation is not typically straightforward, as it requires access to the derivation dataset via, for example, a website [108, 110]. Of note, although multiple imputation is a widely advocated approach for handling missing data in CPR studies, a recent study showed that implementing simpler imputation methods resulted in similar predictive utility of a CPR to predict undiagnosed diabetes, when compared to multiple imputation [111].

Selection of candidate predictors for inclusion in a multivariable model

Candidate predictors are variables that are preselected for consideration in a multivariable model, and differ from those that are subsequently selected for inclusion in the final model [23]. Candidate predictors should be selected without studying the predictor-outcome relationship in the data; in other words, predictors should not be excluded as candidates solely because they are

not statistically significant in univariable analysis [25, 26, 29, 112–114]. Predictor variables do not have to be causally related to the outcome of interest [21, 115]. Effects modelled in studies examining *causality* are expressed with relative risk estimates such as odds ratios, while risk *predictions* are presented as probabilities on an absolute scale between 0 and 1. Relative risk estimates are used in prediction research to calculate an absolute probability of an outcome for a patient, as described above, and can also be reported alongside risk predictions. All variables thought to be related to the target outcome can be selected as candidate predictors for inclusion in a multivariable model; however, when the number of outcome events in the dataset is small, there is a risk of overfitting the data when a large number of predictor variables are included. Thus the CPR will perform well on the derivation data, but poorly on new data [29, 69, 113, 116]. CPRs with a smaller number of predictors are also easier to use in practice. To overcome this problem, only the most clinically relevant candidate predictors should be chosen from the larger pool of potential predictor variables, without looking into the data [5, 117]. In addition, sample size recommendations for studies deriving CPRs are often based on the concept of events-per-variable (EVP), whereby the researcher controls the ratio of the number of outcome events to the number of coefficients estimated prior to any data-driven variable selection [31]. A rule-of-thumb of ten EPV has been suggested [29, 31, 114, 118]. Simulation studies examining the effect of this rule-of-thumb have yielded conflicting results [119–123]. One study found that when the EPV was less than ten, there were a range of circumstances in which coverage and bias were within acceptable levels [119]. Another found that 20 EPV or more are required when low-prevalence predictors are included in a model [123], while another suggested that problems may arise even when the EPV exceeds ten, as CPR performance may depend on many other factors [120]. Research in this area continues to evolve, as new guidance is clearly needed to support sample size considerations for the derivation of CPRs [121]. Recently, van Smeden and colleagues have suggested that sample size should be guided by three influential parameters: the number of predictors, total sample size and the events fraction [122].

Relevant predictors may be chosen based on a combination of clinical experience, expert opinion surveys, qualitative studies and formal systematic reviews and meta-analyses of the literature [26, 33, 36, 65, 124]. Strategies for reducing the number of candidate predictors include removing those that are highly correlated with others, and combining similar predictors [29]. Other considerations include selecting predictors that will be readily available for clinicians to observe or measure in the target setting, and selecting predictors that are relatively easy to measure and demonstrate high inter-rater reliability between clinicians [17, 21]. In terms of handling continuous predictors, researchers strongly advise against converting continuous variables into categorical variables, due to information loss and reduced predictive accuracy [125–128]. Similarly, it should not be assumed that continuous variables have a linear relationship [127]. Instead, methods that permit more flexibility in the functional form of the association between the predictors and outcome should be considered [127, 129]; two common approaches are fractional polynomials and restricted cubic splines [130, 131]. However, if sample size is limited, assuming a linear relationship between continuous variables may make a model less sensitive to extreme observations.

Penalised regression can be used to alleviate the problem of overfitting [116]. This approach involves placing a constraint on the values of the estimated regression coefficients in order to shrink them towards zero [116]. This has the effect of yielding less extreme risk predictions, and thus may improve the accuracy of predictions when the CPR is applied in new patients [113, 132]. The two most popular penalised methods are ridge regression [133] and lasso regression [134]. Unlike ridge regression, lasso regression also selects predictors as a consequence of its penalisation [116]. Ridge regression is usually preferred when a set of pre-specified predictors is available, while lasso regression may be preferred if a simpler model with fewer predictors is required [116, 132].

Selection of predictors during multivariable modelling

There is no consensus regarding how predictors should be selected while developing the final model [25]. Two common strategies include the 'full model approach' and the 'predictor selection approach' [23]. An alternative approach, known as 'all possible subsets regression', is less commonly used [28]. In the full model approach, all previously identified candidate predictors are included, and no further analysis is performed. Although this approach precludes selection bias and overfitting, it requires in-depth knowledge about the most relevant candidate predictors [26, 29]. In the predictor selection approach, predictors are chosen either by 'backward elimination' or 'forward selection', based on pre-defined criteria. Backward elimination begins with all predictors in the model and removes predictors, while forward selection begins with an empty model, and predictors are added successively. All possible subsets regression can build models with combinations of predictors not generated by the standard forward or backward procedures, because every conceivable combination of predictors is assessed to find the best fitting model [135]. With all methods, a series of statistical tests are performed to

assess the 'goodness of fit' between the different models. Models can be compared by setting a pre-defined significance level and using the log likelihood ratio test, or using other model selection criterion such as the Akaike information criterion, or the Bayesian information criterion [23, 25]. Backward elimination is favoured, as it allows for the assessment of the effects of all predictors concurrently, and can take into account all correlations between predictors [136, 137]. Multiple testing in all possible subsets regression can easily lead to overfitting. However, with all methods, the choice of significance level impacts upon the number of final predictors; the use of smaller significance levels (e.g. $p < 0.05$) produces models with fewer predictors at the risk of excluding potentially important predictors, while the use of larger significance levels (e.g. $p < 0.25$) may result in the inclusion of less important predictors [25].

Predictor selection by so-called automated, data-dependent significance testing may generate overfitted, 'optimistic' models, particularly when the derivation dataset is small [23, 28, 128, 138, 139]. Thus, the Akaike information criterion is preferred, as it discourages overfitting by comparing models based on their fit to the data and penalising for the complexity of the model [25]. In addition, it may be acceptable to retain a non-significant predictor in a model, if there is substantial evidence of its predictive ability in the literature [26].

Definition and assessment of predictor and outcome variables

To ensure that the CPR can be accurately applied in practice, predictor and outcome variables should be clearly defined, and outcome variables should be clinically important [17]. Predictor variables must be reliable to enable their assessment in clinical practice; reliability refers to the reproducibility of the findings by the same clinician (intra-rater reliability) or between different clinicians (inter-rater reliability). Some researchers recommend that the reliability of predictor variables be explicitly evaluated, and that only those demonstrating good agreement beyond that expected by chance alone should be considered for inclusion [17]. A recent study found that measurement error of predictor variables is poorly reported, and that researchers seldom state explicitly when the predictors should be measured, and the CPR applied [140]. Another study demonstrated that predictor measurement heterogeneity across settings can have a detrimental impact on the performance of a CPR at external validation [141]. Ideally, the outcome variable should be assessed independently of the predictor variables to avoid circular reasoning or 'incorporation bias', when the results of the CPR or its predictor variables are used in the determination of the outcome [142]. However, it is acknowledged that this is not always possible, particularly for conditions that require a consensus diagnosis based on all available patient information [143]. It is well known that misclassification in the outcome variable may cause serious problems with prediction accuracy [144, 145].

Internal validation

Prediction models are known to perform better in the dataset from which they are derived, in comparison to applying them in new but plausibly related patients [146, 147]. 'Plausibly related patients' may be defined as those who are suspected of having the same condition or who are at risk of the same outcome examined in the derivation study [148]. This enhanced performance occurs simply because a model is designed to optimally fit the available data [23]. The performance of a model is most likely to be overestimated when the derivation dataset is small, and uses a large number of candidate predictors. Therefore, regardless of the approaches used in the derivation stage of development, internal validation is required to examine and correct the amount of overfitting or 'optimism' in the model, and thus the stability of the model [23].

Internal validation does not validate a model itself, but the process used to fit the model [26, 29]. Optimism is estimated using the original derivation dataset only. A number of methods are available for this purpose, including split-sampling, cross-validation and bootstrapping. Split-sampling is the simplest method, and is performed by dividing the derivation dataset into a 'training' sample and a 'test' sample prior to modelling. The CPR is then derived using the training sample, and its performance is assessed using the test sample [20]. However, the test sample usually comprises one-third of the original derivation dataset and is likely to be relatively small, resulting in imprecise performance estimates [149, 150]. This approach also squanders the test data that could have been used in the derivation of the CPR [23, 150]. In cross-validation, the CPR is derived using the whole derivation dataset, and the whole dataset is then reused to assess performance [20]. It is randomly split into equal samples: five or ten samples are commonly used. In the case of five samples, the model is refitted using four of the five samples and its performance tested using the fifth; this process is repeated five times until each of the five samples has been used as the test data, and an average of the estimated performance is taken. To improve stability, the overall procedure can be replicated several times, using different random subsamples [149]. The preferred internal validation method is bootstrapping, particularly when the derivation dataset is small or a large number of candidate predictors are assessed [23, 29]. The idea is to mimic random sampling from the target population by repeatedly drawing samples of the same size with replacement from the derivation dataset [151]. Sampling with replacement renders

bootstrap samples similar, but not identical, to the original derivation sample [23]. Each step of model development is repeated in each bootstrap sample (typically 500), most likely yielding different models with varying performance. Each bootstrap model is then applied to the original derivation sample, yielding a difference in model performance. The average of these differences indicates the optimism in the performance metrics of the model that was initially derived in the derivation dataset [23, 26, 29, 151], and enabling adjustment of the overall performance to better approximate the expected model performance in novel samples [23]. Bootstrapping also estimates a uniform shrinkage factor to enable adjustment of the estimated regression coefficients for over-fitting [26, 29, 151]. However, no internal validation procedures can be a substitute for external validation; internal validation only addresses sampling variability, while external validation considers variation in the patient population [147].

Clinical prediction rule performance measures

CPR predictive performance can be assessed in terms of overall performance, calibration and discrimination [26]. 'Overall performance' can be quantified by calculating the distance between observed and predicted outcomes, using measures such as R^2 or the Brier score [152]. 'Calibration' reflects the agreement between the predicted probabilities produced by the model and the observed outcome frequencies [23]. For example, if a model predicts a 20% probability of residual tumour for a testicular cancer patient, residual tumour should be observed in about 20 out of 100 of these patients [46]. 'Internal calibration' refers to agreement between predicted probabilities and observed outcome frequencies in the derivation dataset, where poor calibration may indicate lack of model fit or model misspecification [153]. 'External calibration' refers to agreement between predicted probabilities and observed outcome frequencies in novel datasets external to the one from which the model was derived, where poor calibration may indicate an overfitted model [153]. Calibration can be visualised by categorising individuals into quantiles based on their predicted probabilities, and plotting the observed outcome frequencies against the mean predicted probabilities [25]. Such a plot is the graphical equivalent of the Hosmer and Lemeshow goodness-of-fit test [154], which, although frequently used, may lack statistical power to identify overfitting [25, 26]. Alternatively, binary outcomes can be regressed on the predicted probabilities of the fitted model to estimate the observed outcome probabilities using smoothing techniques such as the loess algorithm [29, 153]. A comprehensive overview of calibration is given in Van Calster et al. [155].

Discrimination reflects the ability of a CPR to discriminate between patients with, and without, the outcome of interest. The predicted probabilities for patients *with* the outcome should be higher than the predicted probabilities for those who do not have the outcome [46]. The easiest way to assess discrimination is by calculation of the discrimination slope, which is simply the absolute difference in the average predicted probabilities for patients with and without the outcome [26]. Discrimination can also be visualised with a simple box plot. The most widely used measure to assess discrimination is the concordance index (c-index) [156], or, for logistic models its equivalent, the area under the receiver operating characteristic curve (AUROC) [157]. These measures represent the chance that, given one patient with the outcome and one without, the CPR will assign a higher predictive probability to the patient with the outcome compared to the one without. A c-index or AUROC of 0.5 indicates predictions that are no better than random predictions, and a value of 1 represents perfect discrimination between patients with and without the outcome [29]. In theory, a CPR may demonstrate good *discrimination* (classifying patients into the correct risk categories), but poor *calibration* (inaccurately estimating the absolute probability of an outcome), and vice versa [158]. A model that cannot discriminate between patients with and without the outcome has little use as a CPR; however, poor calibration can be corrected without compromising discriminatory performance [19, 114]. Van Calster and Vickers [159] found that poorly calibrated models diminish the clinical usefulness of a CPR, and can be harmful for clinical decision-making under certain circumstances, emphasising the importance of developing well-calibrated CPR's. On the other hand, a CPR with poor calibration but good discrimination at a particular risk threshold may be appropriate if the aim is to prioritise patients for assessment or treatment, by identifying those with a very low risk of the target outcome relative to the rest of the population [160].

Performance measures such as sensitivity, specificity, positive and negative predictive values and positive and negative likelihood ratios are used to assess performance following the application of a risk threshold. Choosing a risk threshold can often be arbitrary, and it can therefore be useful to consider a range of thresholds when assessing performance [19]. Ideally, a CPR will have both a high sensitivity and a high specificity, and therefore correctly identify the majority of patients who truly have the condition, as well as correctly exclude the majority of patients who do not actually have the condition. However, this scenario rarely occurs in clinical practice. More often than not, the definition of a threshold is based on clinical considerations about the relative consequences of false positive and false negative classifications. Sensitivity and specificity are inversely proportional, so that as sensitivity increases, specificity decreases and vice versa [161]. Defining a high cut-off

point will result in good specificity and few false positives, but poor sensitivity and many false negatives. A test with a high specificity is useful for ruling in a disease if a person tests positive. This is because it rarely misdiagnoses those who do not have the condition of interest. Defining a low cut-off point will result in good sensitivity and few false negatives, but poor specificity and many false positives. A test with a high sensitivity is useful for ruling out disease if a person tests negative. This is because it rarely misdiagnoses those who have the condition of interest [161]. Receiver operating characteristic (ROC) curves display the sensitivity and specificity of a CPR across the full range of cut-off values, and can be used to choose an optimal cut-off threshold [162]. Other approaches to determining clinical cut-offs have also been proposed [163].

In recent years, some novel model performance measures have been proposed that quantify the clinical usefulness of a CPR, by taking into account the costs and benefits of clinical decisions. These measures include relative utility curves and decision curves [164, 165]. Decision curves in particular are becoming a popular method of evaluating whether clinical decisions based on CPRs would do more good than harm [166]. Decision curve analysis assumes that a given probability threshold is directly related to the cost to benefit ratio, and uses this threshold to weight false positive and false negative predictions. The cost to benefit ratio thus defines the relative weight of false-positive decisions to true-positive decisions [164]. Model performance can subsequently be summarised as a net benefit, by subtracting the proportion of false-positive patients from the proportion of true-positive patients, weighting by the relative costs of a false-positive and a false-negative result. The net benefit of a CPR can be derived across and plotted against the whole range of threshold probabilities, yielding a decision curve, similar to ROC curves that plot the full range of cut-offs for a sensitivity/specificity pair [164].

Presentation of a clinical prediction rule

The final step in the derivation of a CPR is to consider the format in which it should be presented. It is imperative that the regression coefficients and intercept of a final model are presented, and confidence intervals around predicted probabilities can also be provided [23, 26]. If the final regression formula (as in Table 3) is not provided, a CPR could not be applied by future users [36]. A model can be developed into a simple web-based calculator or application to enhance the usability of a CPR. This may be beneficial for complex CPRs, and would facilitate their integration into the electronic health record, allowing them to be used at the point of clinical care [167]. Nomograms, graphical decision trees and other novel visualisation techniques could also be used [26, 168], which may aid in the interpretation and understanding of a CPR [168]; however, these must be presented alongside the full model formula. Scoring systems are often used to simplify CPRs and facilitate use, where regression coefficients are converted to integer point values that can be easily totalled and related back to the predicted probabilities [169]. However, this transformation leads to a loss of information and therefore reduced predictive accuracy [170].

Reporting the derivation of a clinical prediction rule

Numerous systematic reviews have shown that reporting of the derivation of CPRs is deficient [6–8]. As a result, the TRIPOD guidelines were produced [36], and should be followed by all researchers working in this field.

Stage 3: external validation and refinement of a clinical prediction rule

As previously noted, CPRs perform better in the dataset from which they are derived compared to their application in plausibly related or 'similar but different' individuals, even after internal validation and adjustment [24]. Diminished performance can be due to overfitting, unsatisfactory model derivation, the absence of important predictors, differences in how the predictor variables are interpreted and measured, differences in the patient samples ('case mix') and differences in the prevalence of the disease [26, 148]. There is no guarantee that even well-developed CPRs will be generalisable to new individuals. In one external validation study, a CPR to detect serious bacterial infections in children with fever of unknown source demonstrated considerably worse predictive performance, such that it was rendered useless for clinical care [146]. It is therefore essential to assess the performance of a CPR in individuals outside the derivation dataset; this process is known as external validation [28].

External validation is not simply repeating the steps involved at the derivation stage in a new sample to examine whether the same predictors and regression coefficients are obtained; neither is it refitting the model in a new sample and comparing the performance to that observed in the derivation sample [24, 31]. External validation involves taking the original fully specified model, with its predictors and regression coefficients as estimated from the derivation study; measuring and documenting the predictor and outcome variables in a new patient sample; applying the original model to these data to predict the outcome of interest; and quantifying the predictive performance of the model by comparing the predictions with the observed outcomes [20]. Performance should be assessed using calibration, discrimination and measures to quantify clinical usefulness such as decision curve analysis [164]. A CPR can also be refined if it demonstrates poor performance in an external validation study. Regrettably, few CPRs are externally validated

[27, 171, 172]. A systematic review of CPRs for children identified 101 CPRs addressing 36 conditions; of these, only 17% had narrow validation and only 8% had broad validation [171].

Study design for the external validation of a clinical prediction rule

Ideally, a validation study should be conducted prospectively, by enrolling new individuals in a specifically predesigned study, and the CPR should be applied to all patients meeting the study inclusion criteria [17, 23]. However, validation studies can be conducted retrospectively, using existing datasets. If adequate data on the predictor and outcome variables is available [23]. Investigators conducting a validation study should receive brief training on the accurate application of the CPR. If possible, all patients should be subjected to the reference standard, to establish their true outcome and enable comparison with the CPR prediction. However, in some cases, this may not be feasible or practical, and an appropriate and sensible proxy outcome may be used instead [173]. Stiell and Wells [17] recommend that the inter-rater reliability of the interpretation of the CPR result is assessed, to determine if the CPR is being applied accurately and consistently. In terms of sample size, for a logistic regression model with six predictors, a minimum of 100 patients with the outcome of interest and 100 patients without the outcome of interest has been suggested [174]. Other authors propose that external validation studies require a minimum of 100 events, but ideally 200 events [175]. A minimum of 200 events and 200 non-events has been suggested in order to reliably assess moderate calibration and produce useful calibration plots [155]. The characteristics of patients included in a validation study should be described in detail, and compared with those included in the derivation study. To enhance the interpretation of external validation studies, it is possible to quantify the degree of relatedness between derivation and validation datasets, to determine the extent to which the CPR can be generalised to different populations [34]. Authors have also proposed benchmark values to distinguish between a case-mix effect and incorrect regression coefficients in external validation studies, and therefore assist in the interpretation of a CPR's performance in validation samples [176]. Similarly, a model-based concordance measure has recently been derived that enables quantification of the expected change in a CPR's discriminative ability owing to case-mix heterogeneity [177].

Types of external validation

Many types of external validation are recognised in the literature, but all types consider patients that differ in some respect from the patients included in the derivation study [26]. The greater the differences between the patients in the derivation and validation samples, the stronger the test of generalisability of the CPR [24]. Three types of external validation have received the most attention, namely *temporal* validation, *geographical* validation and *domain* validation [148].

In *temporal* validation studies, the CPR is tested on patients in the same centre(s) but over a different time period [147]. *Geographical* validation studies examine the generalisability of the CPR to other centres, institutes, hospitals or countries [147]. Patient characteristics are likely to vary between locations, and predictor and outcome variables are likely to be interpreted and measured differently in different places, leading to greater differences between the derivation and validation populations than in a temporal validation study [24, 148]. In *domain* validation, the CPR is tested in very different patients than those from whom it was derived, for example in patients from a different setting (e.g. primary or secondary care), or in patients of different ages (e.g. adults vs. children). The case mix of patients included in a domain validation study will clearly differ from the derivation population [148]. Differences between the derivation and validation populations are generally smallest in a temporal validation study, and greatest in a domain validation study; therefore, good performance of a CPR in a temporal validation study may only provide weak evidence that the CPR can be generalised to new patients, while good performance in a domain validation study can be considered as the strongest evidence of generalisability [148]. Other types of external validation studies include *methodologic* validation which refers to testing using data collected via different methods, *spectrum* validation which refers to testing in patients with different disease severity or prevalence of the outcome of interest and fully *independent* validation which refers to testing by independent investigators at different sites [26, 147]. A recent study of cardiovascular risk CPRs found that very few were externally validated by independent researchers; to increase the chance of fully independent validation, researchers should report all the information required for risk calculation, to ensure replicability [178]. Some authors have found that CPRs demonstrate worse performance in fully independent external validation studies compared to temporal or geographical external validation studies [26, 28], while others have found no difference [179]. When multiple external validations of a CPR have been performed, it is useful to conduct a formal meta-analysis to summarise its overall performance across different settings and to assess the circumstances under which the CPR may need adjusting; a recently published framework provides guidance on how to do this [35].

Refinement of a clinical prediction rule: model updating or adjustment

When researchers encounter an inferior performance of a CPR in an external validation study compared with that found in the derivation study, there is a temptation to reject the CPR and derive an entirely new one in the often considerably smaller validation dataset [148, 180]. This approach leads to a loss of scientific information captured in the derivation study and an abundance of CPRs developed for the same clinical situation, leaving clinicians in a quandary over which one to use [24, 148]. However, a reduction in performance is to be expected in an external validation study [24, 26, 148]. The recommended alternative is to update, adjust or recalibrate the CPR using the validation data, thereby combining information captured in the original CPR with information from new patients and improving generalisability [22, 181, 182]. Several methods for updating CPRs are available. When the outcome prevalence in the validation study is different to that in the derivation study, calibration in the validation sample will be affected, but can be improved by adjusting the baseline risk (intercept) of the original model to the patients in the validation sample [180]. If the CPR is overfitted or underfitted, calibration can be improved by simultaneously adjusting all of the regression coefficients [24]. To improve discrimination, individual regression coefficients can be re-estimated, or additional predictors can be added [24, 180]. Ideally, updated CPRs that are adjusted to validation samples should themselves be externally validated, just like newly derived CPRs [148].

Comparing the performance of clinical prediction rules

Once a CPR has been externally validated, it is useful to compare its performance with the performance of other existing CPRs for the same condition [61]. Improvements in discrimination can be assessed by quantifying the difference in the AUROC or equivalent c-index between two CPRs [183]; however, this approach is inappropriate in the case of nested models that are fitted in the same data set [184]. Novel metrics have been proposed that quantify the extent to which a new CPR improves the classification of individuals with and without the outcome of interest into predefined risk groups [46]. These include the net reclassification improvement (NRI), and the integrated discrimination improvement (IDI) [185]. Various decision-analytic approaches to model comparison have also been proposed [186]. All of these measures can be used for comparing both nested and non-nested models. However, both the NRI and IDI statistics have come under intense scrutiny in the literature and many researchers caution against their use, as positive values may arise simply due to poorly fitted models [30, 187–191]. Therefore, the NRI and IDI statistics cannot be recommended [192]. Decision-analytic methods are increasingly recommended as they incorporate misclassification costs and therefore indicate the clinical usefulness of CPRs [186]. A systematic review of comparisons of prediction models for cardiovascular disease found that formal and consistent statistical testing of the differences between models was lacking and that appropriate risk reclassification measures were rarely reported [193]. A recent commentary provides a useful and comprehensive overview of the advantages and disadvantages of the various methods available for quantifying the added value of new biomarkers [194].

Reporting the external validation of a clinical prediction rule

External validation studies of CPRs are often poorly reported [9]; researchers should adhere to the TRIPOD checklist and accompanying guidelines [36].

Stage 4: impact of a clinical prediction rule on clinical practice

Since the ultimate aim of a CPR is to improve the quality of patient care, the effect of a validated CPR on clinician behaviour and patient outcomes should be examined in what are known as impact analysis studies [22, 24]. It is increasingly recognised that CPR's should be regarded as complex interventions, as the introduction of a CPR into clinical practice with subsequent management decisions consists of multiple interacting components [108, 195–201]. The impact of a CPR on clinical practice will depend on several interacting factors, including the accuracy and applicability of the CPR, clinicians' interpretation of probabilities and clinicians' adherence to and acceptance of the CPR [196]. Evaluating the impact of a CPR has been described as 'the next painful step' in the development process [202]. Impact analysis studies clearly differ from validation studies as they must be comparative, typically requiring a control group of clinicians providing usual care [22, 24, 32]. It is possible to assess the impact of both assistive CPRs that simply provide predicted probabilities, and directive *decision* rules that suggest a specific course of action based on probability categories [32]. Assistive CPRs respect clinicians' individual judgement and leave room for intuition, whereas directive rules may be more likely to influence clinician behaviour [32, 203, 204]. However, it is not guaranteed that clinicians will follow CPR, or the recommendations provided by directive rules [32]. Therefore, an impact study must demonstrate that clinical behaviour can be altered and patient care improved by the CPR, prior to widespread dissemination and implementation [17].

Unfortunately, even fewer CPRs undergo an impact assessment than undergo external validation. In the

systematic review of 101 CPRs for children, none had impact analysis performed [171]. An evaluation of 434 primary care CPRs found that only 12 had undergone impact analysis [172]. A subsequent systematic review of the impact of primary care CPRs found 18 studies relating to 14 CPRs, with 10/18 studies demonstrating an improvement in primary outcome when the CPR was used compared to usual care [205]. This review cautioned that the small number of impact analysis studies found precluded the possibility of drawing firm conclusions about the overall effectiveness of CPRs in primary care, with the authors pointing out that the methodological quality of the included studies was unclear due to incomplete reporting [205]. Another recent systematic review of the impact of CPRs found that the intermediate consequences of a CPR such as clinical management decisions were the primary outcome in the majority of studies, while few studies aimed to establish the effect of a CPR on patient outcomes [206]. In addition, in many of the included studies, the risk of bias was either high or unclear [206]. Finally, a study describing the distribution of derivation, validation and impact studies in four reviews of leading medical journals since 1981 demonstrated that a minority of studies concerned CPR impact (10/201), with the pattern remaining stable over time [27].

Study design for an impact analysis

Before carrying out a formal impact study, researchers must consider whether the CPR is ready for implementation [108, 207]. If possible, the predictive performance of the CPR should be verified in the new setting, and the CPR tailored to the new setting to enhance performance [108]. The optimal study design for an impact analysis is a cluster randomised trial with centres as clusters [22]. Randomising individual patients is not recommended as clinicians may learn the rule and apply it to patients randomised to the control group [22]. Randomising clinicians is preferable but requires more patients, and may lead to contamination of experience between clinicians in the same centre [24, 208]. An attractive variant of a cluster randomised trial is the stepped-wedge cluster randomised trial. In a stepped-wedge design, all centres apply care-as-usual, and then use the CPR at different, randomly allocated time periods [209]. This design allows for the comparison of outcomes both within and between hospitals, generates a wealth of data regarding potential barriers to implementation and is particularly beneficial if the CPR turns out to have a promising effect [210]. When the outcome of interest in an impact study is clinician behaviour or decision-making, a cross-sectional randomised study without patient follow-up is sufficient, with randomisation at either the patient or clinician level. However, to determine the impact of a CPR on patient outcomes or cost-effectiveness, follow-up of patients is essential [22].

Given the significant practical, logistic and economic challenges associated with cluster randomised trials, non-randomised approaches are possible and are often used. Cluster randomised trials can be expensive and time-consuming and it may be difficult to recruit an adequate number of clusters [24, 108]. A suggested rule-of-thumb is to regard four clusters per arm as the absolute minimum number required [211]; however, methods for determining sample size in cluster randomised trials have been proposed by a number of authors [212–214]. A popular design is a before–after study, in which outcomes are assessed in a time period before a CPR is available and compared with outcomes measured in a time period after it is introduced; this design is susceptible to temporal confounding [24]. Finally, a relatively low-cost and simple design is a before–after study within the same clinicians. In this design, clinicians are asked to indicate their treatment or management decision or perceived risk of disease for the same patient both before, and after, receiving the CPR prediction [24]. Single centre impact studies are recommended to inform the planning of multicentre randomised trials [32]. As with derivation and validation studies, a sample size calculation should be performed, with consideration of all relevant impact measures, and where possible assessment of outcome measures should be blinded to the CPR predictions and recommendations [32, 33]. Clinicians must undergo training in order to correctly interpret and use the CPR [17].

The impact of CPRs can also be estimated indirectly using decision analytic modelling, which integrates information on CPR predictions and information about the effectiveness of treatments from therapeutic intervention studies [215, 216]. Such studies cost less, and take less time, than RCTs; however, they are limited by the quality of available evidence, and only provide theoretical indications of the impact CPRs may have on patient outcomes. Thus it has been suggested that they should not replace RCTs but rather be performed as an intermediate step prior to an RCT [217].

Measures of impact of a clinical prediction rule

During an impact analysis study, the sensitivity and specificity of the CPR should be recalculated to determine its accuracy in the new study population [17]. However, measures of CPR *accuracy* are not synonymous with measures of *impact*, and only represent the *potential* impact of the CPR [32]. This is because clinicians are unlikely to follow the logic of the CPR or its recommendations in every case; they may not use the CPR at all, they may not use it correctly, they may deliberately disregard its predictions or suggestions or they may be unable to use it for other reasons [32]. Measures that are assessed in traditional RCTs include safety, which refers to any adverse events resulting

from the implementation of an intervention, and efficacy, which relates to the extent that an intervention helps to improve patient outcomes, for example by reducing mortality rates [218]. In addition, Reilly and Evans [32] propose that the impact of a CPR is assessed in terms of its 'safety' and 'efficiency', where safety is defined as the proportion of patients found to have the outcome of interest and who received the appropriate intervention, and efficiency is defined as the proportion of patients *without* the outcome of interest and who *did not* receive the intervention. The sensitivity and specificity of a CPR will only be the same as its safety and efficiency if clinicians follow the logic and recommendations of the CPR exactly [32]. Therefore, in an impact analysis study, a CPR may demonstrate more, or less, actual impact than its potential impact. The effect of clinicians' incorrect use of the CPR, or their deviations from its logic or suggestions can provide important insights into its impact under specific circumstances, and may reveal complex interactions between clinicians and the CPR [32]. For example, Reilly and colleagues [219] found that when clinicians did not consult a CPR for suspected acute cardiac ischemia at all, or overruled its recommendations, their decisions were less efficient than if they had followed the CPR in every case.

Acceptability of a clinical prediction rule

If the use of a CPR is warranted but it is not used, the considerable time, money and effort that goes into its development and evaluation is wasted. Assessing the acceptability of a CPR is therefore crucial for successful implementation. Even valid and reliable CPRs may not be accepted or used by clinicians [17]. Impact studies allow researchers to evaluate the acceptability of a CPR to clinicians, patients or others who may use it, as well as its ease of use and barriers to its uptake [22]. If a CPR proves to be acceptable, its long-term and widespread dissemination and implementation would be justified; if not, the CPR could undergo modification and further evaluation [48]. Acceptability of a CPR and attitudes towards it can be determined via survey, qualitative, simulation or clinical vignette studies [33, 48, 220–222]. The validated Ottawa Acceptability of Decision Rules survey instrument can be used both to measure the overall acceptability of a CPR, and to assess specific barriers to its use, which can inform potential improvements to the CPR as well as the design of dedicated implementation strategies [48]. Qualitative studies can be invaluable for determining the acceptability of a CPR but are relatively rare [200, 220, 222–225].

Comparison of a clinical prediction rule with unstructured clinical judgement

For a CPR to improve the diagnostic accuracy of clinicians, its performance in distinguishing between patients with and without the condition of interest should be superior to that of unstructured clinical judgement alone. Therefore, a vital metric is the comparison of the accuracy of the CPR-predicted probabilities of disease, or recommended decisions, with the accuracy of clinicians own estimated disease probabilities or management decisions [18]. The sensitivity and specificity of clinicians' predictions or decisions are generally measured under usual practice, and compared to the sensitivity and specificity of the CPR predictions or decisions when applied to the same patients [226, 227]. Some studies have used clinical vignettes [228] while others have used multivariable logistic models to assess the added value of a CPR over and above clinical judgement alone [229]. If it can be demonstrated that the performance of a CPR is superior to unaided clinician judgement, this may aid clinicians' acceptance and use of the CPR [32]. Although comparison of a CPR to clinician suspicion regularly takes place at the impact analysis stage, some researchers have recommended that this is carried out during the derivation or validation stages, arguing that if the CPR does not add anything beyond clinical judgement, then the use of the CPR and an impact study would not be warranted [230]. In addition, Finnerty and colleagues [231] recommend that comparison is undertaken in multiple settings, as the performance of a CPR may be superior to clinical judgement in certain settings, but inferior or no different in other settings. A recent systematic review comparing CPRs with clinical judgement concluded that the differences between the two methods of judgement are likely due to different diagnostic thresholds, and that the preferred judgement method in a given situation would therefore depend on the relative benefits and harms resulting from true positive and false positive diagnoses [232]. Brown and colleagues' [200] found that the use and potential advantages of a CPR may be much more complex than originally thought, and that CPRs may be useful for purposes not previously reported, such as enhancing communication with colleagues and patients, and medico-legal purposes. Recent studies in the child protection field have demonstrated that CPRs may provide clinicians with additional confidence in their decision-making, even if they do not alter their management actions based on the CPRs risk prediction [220, 233].

The four phases of impact analysis for clinical prediction rules

Despite the abundance of methodological guidelines for the derivation and validation of CPRs [26], there is a lack of clear guidance for the design, conduct and reporting of impact analysis studies of CPRs. To this end, Wallace and colleagues [33] formulated an iterative four-phased framework for the impact analysis of CPRs, specifying the importance of substantial preparatory and feasibility

work prior to the conduct of a full-scale formal experimental study (Fig. 2). Phase 1 involves determining whether the CPR is ready for impact analysis, i.e. whether it has been rigorously derived and broadly validated according to pre-defined methodological standards. Phase 2 includes assessing the acceptability of the CPR and identifying potential barriers to its uptake and implementation, as well as assessing the feasibility of conducting an impact study. Evaluating the feasibility of carrying out an impact study involves consideration of multiple factors including the most appropriate study design for measuring relevant outcomes, and how the CPR will be delivered at the point of care or integrated into the clinical workflow. Phase 3 involves formally testing the impact of the CPR using a comparative study design. Phase 4 involves long-term dissemination and implementation of the CPR, which corresponds to stage 6 in the development of CPRs, discussed below.

Reporting the impact analysis of a clinical prediction rule

There are currently no published reporting guidelines for studies analysing the impact of CPRs. This is a gap in the literature, and a priority for future research. However, researchers assessing the impact of CPRs in an RCT may refer to guidelines on the reporting of clinical trials, such as the Consolidated Standards of Reporting Trials (CONSORT) statement [218].

Stage 5: cost-effectiveness of the clinical prediction rule

If an impact analysis study shows that a CPR demonstrates safety and efficiency, alters clinician behaviour and improves clinical care, a formal economic evaluation can be carried out to determine the cost-effectiveness of the CPR. The aim is to establish the health care savings associated with routine use of the CPR in clinical practice [17]. Economic evaluation is usually based on decision analytic models [234]. Any economic evaluation must make reasonable assumptions about the accuracy and effectiveness of the CPR and the costs involved [17]. Sensitivity analyses should be performed by re-running models with alternative assumptions, to examine the uncertainty of the model projections [234]. In reality, many economic evaluations are conducted prior to an impact analysis study or even an external validation study, perhaps because they are relatively quick and low cost to perform, and provide a significant part of the justification for the development and implementation of a CPR.

Stage 6: long-term implementation and dissemination of the clinical prediction rule

The gap between evidence and practice has been consistently demonstrated in health services research [235], and there is no guarantee that a CPR will be widely disseminated or used, even if it is shown to have a positive impact on clinical care and cost benefits. Therefore, in order to maximise the uptake of a CPR, an active

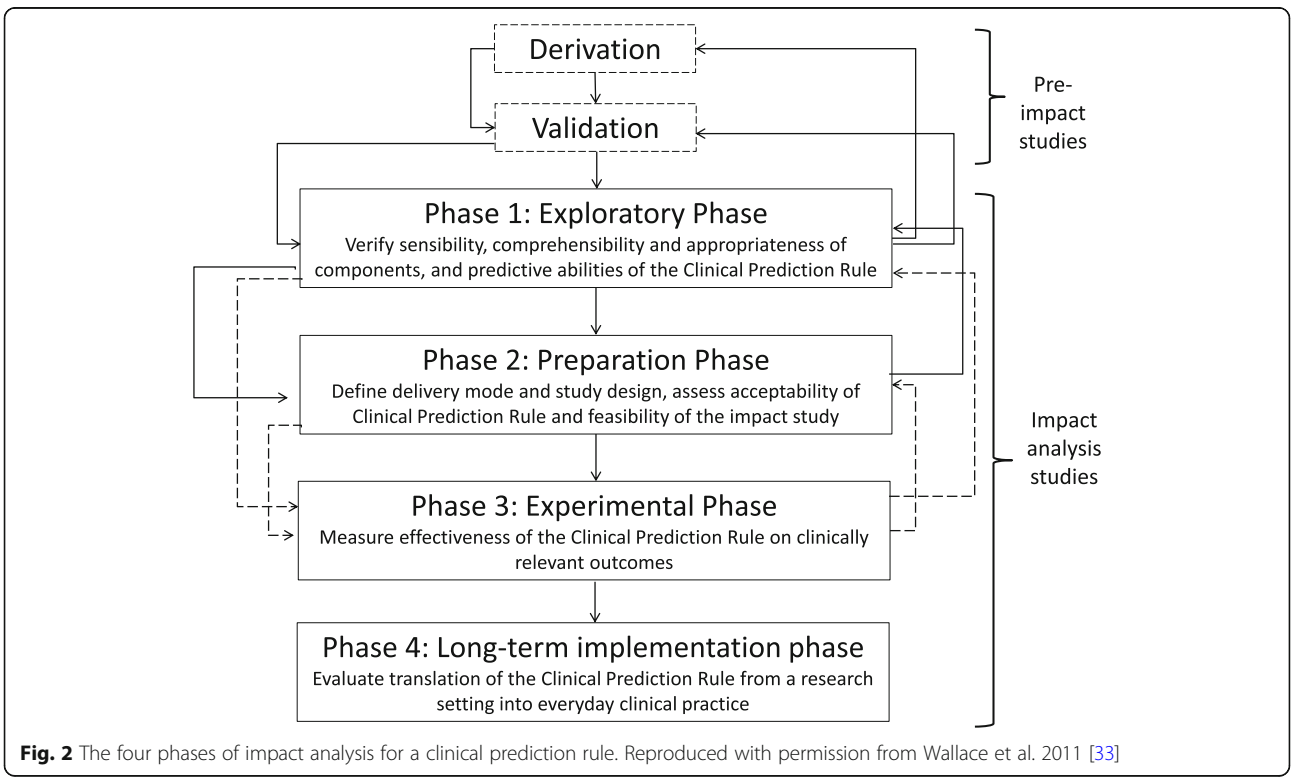

Fig. 2 The four phases of impact analysis for a clinical prediction rule. Reproduced with permission from Wallace et al. 2011 [33]

Table 4 Barriers to the use of clinical prediction rules in practice identified in the literature

Theme	Subtheme	Barrier
Knowledge	Awareness	Unaware:
		• That CPR exists
		• Of clinical problem or burden of clinical problem to which CPR applies
		Unable to choose from multiple CPRs
	Familiarity	Unfamiliar with CPR
	Understanding	Lack of knowledge and understanding of the purpose, development and application of CPRs in general
	Forgetting	Clinician forgets to use CPR despite best intentions
Attitudes	Negative beliefs about CPRs	Belief that:
		• CPRs threaten autonomy
		• CPRs are too 'cook-book', and oversimplify the clinical assessment process
		• Clinical judgement is superior to CPRs
		• Clinical judgement is not error prone
		• Use of CPRs causes intellectual laziness
		• The development of the CPR was biased
		• Patients will deem clinicians less capable if using a CPR
		• CPRs only apply to the less experienced
		• Probabilities are not helpful for decision-making
		Dislike of the term 'rule'
		Clinician had a false negative result when using a CPR in the past
		Existing CPRs are not ready for clinical application
	Outcome expectancy	Belief that:
		• CPRs will not lead to improved patient or process outcomes
		• The information provided by the CPR is not sufficient to alter clinical decisions
		Clinician:
		• Fears unintended consequences of use
		• Is uncertain about using the CPR in patients with an atypical presentation
		• Worries that improving efficiency threatens patient safety
	Self-efficacy	Belief that the CPR is too difficult to use
		Clinician uncertain how to interpret or use CPR output
	Motivation	Clinician lacks motivation to use the CPR
Behaviour	Patient factors	Patients expectations are not consistent with the CPR
	Features of the CPR	Clinician:
		• Finds CPR too complicated
		• Finds CPR 'too much trouble' to apply
		Perception that:
		• The CPR is not an efficient use of time • The CPR does not have face validity or that important predictors are missing
		• The CPR does not fit in with usual work flow or approach to decision-making
		• The CPR is not generalisable to the clinician's patient
		• The CPR is static and does not consider the dynamic nature of clinical practice
		• Overruling the CPR is often justified
		Data required for the CPR is difficult to obtain
	Environmental factors	Lack of:
		• Time
		• Organisational support
		• Peer support for use
		Perceived increased risk of litigation
		Insufficient incentives or reimbursement for use of the CPR

Adapted from Sanders 2015 [253]. *CPR* clinical prediction rule

dissemination and implementation plan must be in place. Simple passive diffusion of study results via publication in journals or presentations at conferences is unlikely to significantly change clinical practice [236]. Examples of dissemination include actively targeting specific audiences via direct mail or the press, while implementation involves the use of local administrative, educational, organisational and behavioural strategies to put the CPR into effect in clinical practice [236]. Active broad dissemination of the widely accepted Ottawa ankle rule via an educational intervention found no impact of the rule on clinicians' use of ankle radiography [237], leading the authors to recommend implementation strategies at the local level instead. Some implementation strategies have been found to be more effective than others in changing clinician behaviour. A systematic review found the most effective approaches to be reminders in the form of posters, pocket cards, sheets or computer-embedded prompts, face-to-face local clinician education and the use of multiple interventions simultaneously [238]. Incorporation of CPRs into clinical guidelines may also be of benefit; a recent study found that clinical guidelines and local policies that mandated the use of CPRs were effective in increasing their adoption in clinical practice [200]. In addition, the integration of CPRs into the clinical workflow via electronic health records may promote their use [239]. Since impact in a research study does not ensure impact in real-world clinical practice, follow-up of clinicians can be conducted to assess the long-term use and effect of the CPR [17, 33].

Barriers and facilitators to the use of clinical prediction rules

Clearly, identifying the barriers and facilitators to the implementation of CPRs is crucial for the development of targeted implementation strategies that may encourage

clinicians to use the CPR. The adoption of CPRs into clinical practice is influenced by various factors including clinician characteristics, patient factors, features of the CPR itself and environmental factors [32, 66, 221, 224, 225, 240–252].

Table 4 provides an overview of the barriers to the adoption of CPRs identified in the literature [253], grouped according to their effect on clinician knowledge, attitudes or behaviours [254]. Barriers relating to knowledge include lack of awareness of the CPR or the burden of the clinical problem it applies to, unfamiliarity with the CPR and a lack of understanding of the purpose of CPRs in general [225, 240–242]. Clinicians may also be unaware of a CPR due to the increasing volume of CPRs, particularly when they are developed for the same condition [61, 243]. Common barriers relating to clinician attitude include a conviction that clinical judgement is superior to the CPR, and distrust of the accuracy of the CPR [32, 224, 240, 241, 244, 245]. Barriers relating to behaviour include organisational factors [251], the complexity of the CPR and the time it takes to apply; survey studies suggest that clinicians much prefer a CPR that is simple to use, memorable and saves time [221, 246, 247]. Complex models such as those based on machine and artificial learning algorithms may introduce additional barriers relating to applicability and usability, due to their potential lack of reproducibility and transparency [60, 82]. Other studies have demonstrated that clinicians will be unlikely to use a CPR if there are predictors missing which are deemed to be important, or if the predictor variables are not logically related to the outcome variable [32, 225]. Reilly and Evans [32] offer a number of strategies for overcoming barriers to the use of CPRs. These include emphasising the discretionary use of the CPR, comparing clinical judgement with the CPR, checking whether any excluded factors affect the CPR predictions, performing a simulated impact analysis and soliciting clinicians input regarding the logic and format of the CPR, among others [32].

Summary

For CPRs to be useful in clinical practice, they must be properly planned [67], derived using appropriate statistical techniques [23] and externally validated in multiple settings and by independent investigators to determine their predictive accuracy [148]. In addition, CPRs must undergo impact analysis to determine their effect on clinician behaviour and relevant patient outcomes [22]. There are numerous factors to consider when deriving, validating and assessing the impact of a CPR including the study design, preparatory work, statistical analysis, modelling strategy, performance/impact measures, the presentation of the CPR and the reporting of the study methodology. New CPRs should only be derived when there is a clear clinical need for them [17]. There is an urgent need to change the focus from the derivation of CPRs, to the validation and impact analysis of existing ones [33]. The CPR must be presented in full, and the study methods reported adequately, to ensure its quality, risk of bias and clinical utility can be evaluated; the TRIPOD guidelines should be followed to ensure completeness of reporting requirements [36]. Feasibility and preparatory work is essential to determine whether a formal impact study of the CPR is warranted [33, 108], and survey and qualitative work should be undertaken to verify whether the CPR is acceptable and relevant to clinicians [48, 65, 220, 222]. If a CPR is found to have a positive impact on patient outcomes, its cost-effectiveness should be evaluated, and a targeted implementation and dissemination strategy devised, with consideration of possible barriers to implementation, to maximise uptake [17].

In summary, the development and evaluation of a robust, clinically useful CPR with high predictive accuracy is challenging, and research in the field concerning derivation, validation and impact evaluation continues to evolve. However, adhering to the existing methodological standards and recommendations in the literature at every step will help to ensure a rigorous CPR that has the potential to contribute usefully to clinical practice and decision-making.

Abbreviations

AUROC: Area under the receiver operating characteristic curve; CPR: Clinical prediction rule; EPV: Events per variable; IDI: Integrated discrimination improvement; MAR: Missing at random; MCAR: Missing completely at random; MNAR: Missing not at random; NRI: Net reclassification improvement; RCT: Randomised controlled trial; ROC: Receiver operating characteristic curve; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

Acknowledgements

We would like to thank Health and Care Research Wales for funding this work.

Funding

This work was supported by Health and Care Research Wales (grant number HS-14-24). The funders had no involvement in the study design, the collection, analysis or interpretation of the data, the writing of the manuscript or the decision to submit the manuscript for publication.

Availability of data and materials

Not applicable.

Authors' contributions

LEC conducted the literature search, drafted the manuscript, produced the tables, boxes and figures and edited the manuscript. DMF, SM and AMK critically revised the manuscript for important intellectual content. All authors approved the final version submitted for publication.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 13 August 2018 Accepted: 12 May 2019
Published online: 22 August 2019

References

1. Adams ST, Leveson SH. Clinical prediction rules. BMJ. 2012;344:d8312.
2. Beattie P, Nelson R. Clinical prediction rules: what are they and what do they tell us? Aust J Physiother. 2006;52(3):157–63.
3. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277(6):488–94.
4. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-based medicine working group. JAMA. 2000;284(1):79–84.
5. Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. J Thromb Haemost. 2013;11(Suppl 1):129–41.
6. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9(5):1–12.
7. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010;8:20.
8. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9:103.
9. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.
10. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol. 2016;214(1):79–90 e36.
11. Ettema RG, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KG. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. Circulation. 2010;122(7):682–9.
12. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol. 2013;66(3):268–77.
13. Nayak S, Edwards DL, Saleh AA, Greenspan SL. Performance of risk assessment instruments for predicting osteoporotic fracture risk: a systematic review. Osteoporos Int. 2014;25(1):23–49.
14. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Investig. 2009;27(3):235–43.
15. Collins GS, Michaelsson K. Fracture risk assessment: state of the art, methodologically unsound, or poorly reported? Curr Osteoporos Rep. 2012;10(3):199–207.
16. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985;313(13):793–8.
17. Stiell I, Wells G. Methodologic standards for the development of clinical decision rules in emergency medicine. Ann Emerg Med. 1999;33(4):437–47.
18. Green SM, Schriger DL, Yealy DM. Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. Ann Emerg Med. 2014;64(3):286–91.
19. 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.
20. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
21. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
22. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
23. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683–90.
24. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691–8.
25. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604.
26. Steyerberg E. Clinical prediction models: a practical approach to development, validation and updating. New York: Springer-Verlag; 2009.
27. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
28. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453–73.
29. Harrell F. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
30. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. BJOG. 2017;124(3):423–32.
31. Collins GS, Ma J, Gerry S, Ohuma E, Odondi LO, Trivella M, et al. Risk prediction models in perioperative medicine: methodological considerations. Curr Anesthesiol Rep. 2016;6(3):267–75.
32. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144(3):201–9.
33. Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al. Framework for the impact analysis and implementation of clinical prediction rules (CPRs). BMC Med Inform Decis Mak. 2011;11:62.
34. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279–89.
35. Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res. 2018. https://doi.org/10.1177/0962280218785504.
36. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, 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.
37. Lo BWY, Fukuda H, Nishimura Y, Farrokhyar F, Thabane L, Levine MAH. Systematic review of clinical prediction tools and prognostic factors in aneurysmal subarachnoid hemorrhage. Surg Neurol Int. 2015;6:135.
38. Hopper AD, Cross SS, Hurlstone DP, McAlindon ME, Lobo AJ, Hadjivassiliou M, et al. Pre-endoscopy serological testing for coeliac disease: evaluation of a clinical decision tool. BMJ. 2007;334:729.
39. LaValley MP, Lo GH, Price LL, Driban JB, Eaton CB, McAlindon TE. Development of a clinical prediction algorithm for knee osteoarthritis structural progression in a cohort study: value of adding measurement of subchondral bone density. Arthritis Res Ther. 2017;19:95.
40. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008;5(8):e165.
41. Ferro JM, Bacelar-Nicolau H, Rodrigues T, Bacelar-Nicolau L, Canhão P, Crassard I, et al. Risk score to predict the outcome of patients with cerebral vein and dural sinus thrombosis. Cerebrovasc Dis. 2009;28(1):39–44.
42. Woo J, Leung J, Wong S, Kwok T, Lee J, Lynn H. Development of a simple scoring tool in the primary care setting for prediction of recurrent falls in men and women aged 65 years and over living in the community. J Clin Nurs. 2009;18(7):1038–48.
43. Scholz NN, Bäsler KK, Saur PP, Burchardi HH, Felder SS. Outcome prediction in critical care: physicians' prognoses vs. scoring systems. Eur J Anaesthesiol. 2004;21(8):606–11.
44. Kheterpal S, Tremper KK, Heung M, Rosenberg AL, Englesbe M, Shanks AM, Campbell DA. Development and validation of an acute kidney injury risk index for patients undergoing general surgery results from a national data set. Anesthesiology. 2009;110(3):505–15.
45. Pace N, Eberhart L, Kranke P. Quantifying prognosis with risk predictions. Eur J Anaesthesiol. 2012;29(1):7–16.
46. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21(1):128–38.
47. McGinn T. Putting meaning into meaningful use: a roadmap to successful integration of evidence at the point of care. JMIR Med Inform. 2016;4(2):e16.
48. Brehaut JC, Graham ID, Wood TJ, Taljaard M, Eagles D, Lott A, et al. Measuring acceptability of clinical decision rules: validation of the Ottawa acceptability of decision rules instrument (OADRI) in four countries. Med Decis Mak. 2010;30(3):398–408.
49. Sarasin FP, Reymond JM, Griffith JL, Beshansky JR, Schifferli JA, Unger PF, et al. Impact of the acute cardiac ischemia time-insensitive predictive

instrument (ACI-TIPI) on the speed of triage decision making for emergency department patients presenting with chest pain: a controlled clinical trial. J Gen Intern Med. 1994;9(4):187–94.
50. Stiell IG, McDowell I, Nair RC, Aeta H, Greenberg G, McKnight RD, Ahuja J. Use of radiography in acute ankle injuries: physicians' attitudes and practice. CMAJ. 1992;147(11):1671–8.
51. Stiell IG, McKnight R, Greenberg GH, McDowell I, Nair RC, Wells GA, et al. Implementation of the Ottawa ankle rules. JAMA. 1994;271(11):827–32.
52. Anis AH, Stiell IG, Stewart DG, Laupacis A. Cost-effectiveness analysis of the Ottawa ankle rules. Ann Emerg Med. 1995;26(4):422–8.
53. Graham ID, Stiell IG, Laupacis A, McAuley L, Howell M, Clancy M, et al. Awareness and use of the Ottawa ankle and knee rules in 5 countries: can publication alone be enough to change practice? Ann Emerg Med. 2001;37(3):259–66.
54. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
55. Shariat SF, Karakiewicz PI, Margulis V, Kattan MW. Inventory of prostate cancer predictive tools. Curr Opin Urol. 2008;18(3):279–96.
56. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak. 2006;6:38.
57. Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, Kent DM. Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. Circ Cardiovasc Qual Outcomes. 2015;8(4):368–75.
58. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7(2):e32844.
59. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744.
60. Moons KM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1–W33.
61. Collins GS, Moons KG. Comparing risk prediction models. BMJ. 2012;344:e3186.
62. Dekker FW, Ramspek CL, van Diepen M. Con: most clinical risk scores are useless. Nephrol Dial Transplant. 2017;32(5):752–5.
63. Masconi K, Matsha T, Erasmus R, Kengne A. Recalibration in validation studies of diabetes risk prediction models: a systematic review. Int J Stat Med Res. 2015;4(4):347–69.
64. Ban JW, Wallace E, Stevens R, Perera R. Why do authors derive new cardiovascular clinical prediction rules in the presence of existing rules? A mixed methods study. PLoS One. 2017;12(6):e0179102.
65. de Salis I, Whiting P, Sterne JA, Hay AD. Using qualitative research to inform development of a diagnostic algorithm for UTI in children. Fam Pract. 2013;30(3):325–31.
66. Haskins R, Osmotherly PG, Southgate E, Rivett DA. Australian physiotherapists' priorities for the development of clinical prediction rules for low back pain: a qualitative study. Physiotherapy. 2015;101(1):44–9.
67. Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med. 2014;11(7):e1001671.
68. Altman DG. The time has come to register diagnostic and prognostic research. Clin Chem. 2014;60(4):580–2.
69. Han K, Song K, Choi BW. How to develop, validate, and compare clinical prediction models involving radiological parameters: study design and statistical methods. Korean J Radiol. 2016;17(3):339–50.
70. Lee Y-h, Bang H, Kim DJ. How to establish clinical prediction models. Endocrinol Metab (Seoul). 2016;31(1):38–44.
71. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol. 2008;8:48.
72. Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using case-cohort data. BMC Med Res Methodol. 2013;13:113.
73. Nee RJ, Coppieters MW. Interpreting research on clinical prediction rules for physiotherapy treatments. Man Ther. 2011;16(2):105–8.
74. Hancock M, Herbert RD, Maher CG. A guide to interpretation of studies investigating subgroups of responders to physical therapy interventions. Phys Ther. 2009;89(7):698–704.
75. Labarère J, Renaud B, Fine MJ. How to derive and validate clinical prediction models for use in intensive care medicine. Intensive Care Med. 2014;40(4):513–27.
76. Grobman WA, Stamilio DM. Methods of clinical prediction. Am J Obstet Gynecol. 2006;194(3):888–94.
77. van den Bosch JE, Kalkman CJ, Vergouwe Y, Van Klei WA, Bonsel GJ, Grobbee DE, Moons KG. Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting. Anaesthesia. 2005;60(4):323–31.
78. Hilbe J. Logistic regression models. Boca Raton: Chapman & Hall/CRC; 2009.
79. Marshall RJ. The use of classification and regression trees in clinical epidemiology. J Clin Epidemiol. 2001;54(6):603–9.
80. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med. 1992;21(4):384–90.
81. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
82. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. CoRR. 2018; abs/1812.10404.
83. Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63(2):205–14.
84. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 2002.
85. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
86. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087–91.
87. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–101.
88. Janssen KJM, Donders ART, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, Moons KGM. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol. 2010;63(7):721–7.
89. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–66.
90. van der Heijden GJMG, Donders AR, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol. 2006;59(10):1102–9.
91. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
92. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67.
93. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30(4):377–99.
94. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330–51.
95. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549–76.
96. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res. 2007;16(3):259–75.
97. Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. Stat Med. 2003;22(16):2553–75.
98. Leurent B, Gomes M, Faria R, Morris S, Grieve R, Carpenter JR. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. Pharmacoeconomics. 2018;36(8):889–901.
99. Leacy FP, Floyd S, Yates TA, White IR. Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. Am J Epidemiol. 2017;185(4):304–15.
100. Héraud-Bousquet V, Larsen C, Carpenter J, Desenclos J-C, Le Strat Y. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. BMC Med Res Methodol. 2012;12:73.
101. Carpenter JR, Kenward MG. MAR methods for quantitative data. In: missing data in randomised controlled trials— a practical guide. Birmingham: National Institute for Health Research; 2008.
102. Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. Stat Model. 2009;9(3):173–97.

103. Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.
104. Dobson A, Diggle P, Henderson R. Joint modelling of longitudinal measurements and event time data. Biostatistics. 2000;1(4):465–80.
105. Rizopoulos D. Joint models for longitudinal and time-to-event data with applications in R. New York: Chapman and Hall/CRC; 2012.
106. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57.
107. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010;10:112.
108. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. BMC Diagn Progn Res. 2018;2:11.
109. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KGM. Adaptation of clinical prediction models for application in local settings. Med Decis Mak. 2012;32(3):E1–E10.
110. Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, Moons KGM. Dealing with missing predictor values when applying clinical prediction models. Clin Chem. 2009;55(5):994–1001.
111. Masconi KL, Matsha TE, Erasmus RT, Kengne AP. Effects of different missing data imputation techniques on the performance of undiagnosed diabetes risk prediction models in a mixed-ancestry population of South Africa. PLoS One. 2015;10(9):e0139210.
112. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907–16.
113. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Mak. 2001;21(1):45–56.
114. Harrell FEJ, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–87.
115. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289–310.
116. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, Omar RZ. How to develop a more accurate risk prediction model when there are few events. BMJ. 2015;351:h3868.
117. Heinze G, Dunkler D. Five myths about variable selection. Transpl Int. 2017;30(1):6–10.
118. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.
119. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and cox regression. Am J Epidemiol. 2007;165(6):710–8.
120. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993–1000.
121. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, Reitsma JB. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163.
122. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, Reitsma JB. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res. 2018. https://doi.org/10.1177/0962280218784726.
123. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016;76:175–82.
124. Battle CE, Hutchings H, Evans PA. Expert opinion of the risk factors for morbidity and mortality in blunt chest wall trauma: results of a national postal questionnaire survey of emergency departments in the United Kingdom. Injury. 2013;44(1):56–9.
125. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26(30):5512–28.
126. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127–41.
127. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35(23):4124–35.
128. Steyerberg EW, Uno H, Ioannidis JPA, van Calster B, Ukaegbu C, Dhingra T, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol. 2018;98:133–43.
129. Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley; 2009.
130. Harrell FEJ, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst. 1988;80(15):1198–202.
131. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. J R Stat Soc Ser C Appl Stat. 1994;43(3):429–67.
132. Ambler G, Seaman S, Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. Stat Med. 2012;31(11–12):1150–61.
133. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Ser C Appl Stat. 1992;41(1):191–201.
134. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267–88.
135. Hosmer DW, Jovanovic B, Lemeshow S. Best subsets logistic regression. Biometrics. 1989;45(4):1265–70.
136. Mantel N. Why stepdown procedures in variable selection. Technometrics. 1970;12(3):621–5.
137. Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. Clin Chem. 2004;50(3):473–6.
138. Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52(10):935–42.
139. Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: selection bias dominates. J Clin Epidemiol. 2011;64(12):1464–5.
140. Whittle R, Peat G, Belcher J, Collins GS, Riley RD. Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. J Clin Epidemiol. 2018. https://doi.org/10.1016/j.jclinepi.2018.05.008.
141. Luijken K, Groenwold RHH, van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on performance of prediction models: a measurement error perspective. arXiv: 180610495 [statME]. 2018:arXiv:1806.10495.
142. Worster A, Carpenter C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. CJEM. 2008;10(2):174–5.
143. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? J Clin Epidemiol. 2002;55(7):633–6.
144. Wang LE, Shaw PA, Mathelier HM, Kimmel SE, French B. Evaluating risk-prediction models using data from electronic health records. Ann Appl Stat. 2016;10(1):286–304.
145. van Doorn S, Brakenhoff TB, Moons KGM, Rutten FH, Hoes AW, Groenwold RHH, Geersing GJ. The effects of misclassification in routine healthcare databases on the accuracy of prognostic prediction models: a case study of the CHA2DS2-VASc score in atrial fibrillation. BMC Diagn Progn Res. 2017;1:18.
146. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, Moons KG. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56(9):826–32.
147. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130(6):515–24.
148. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61(11):1085–94.
149. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774–81.
150. Steyerberg EW. Validation in prediction research: the waste by data-splitting. J Clin Epidemiol. 2018. https://doi.org/10.1016/j.jclinepi.2018.07.010.
151. Efron B, Tibshirani R. An introduction to the bootstrap. Boca Raton: Chapman & Hall/CRC; 1993.
152. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J. 2008;50(4):457–79.
153. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33(3):517–35.
154. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley; 2000.

155. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76.
156. Pencina MJ, D'Agostino RBS. Evaluating discrimination of risk prediction models: the C statistic. JAMA. 2015;314(10):1063–4.
157. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
158. Baron JA, Sorensen HT. Clinical epidemiology. In: Olsen J, Saracci R, Trichopoulos D, editors. Teaching epidemiology: a guide for teachers in epidemiology, public health and clinical medicine. New York: Oxford University Press; 2010. p. 411–28.
159. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Mak. 2014;35(2):162–9.
160. Meurer WJ, Tolles J. Logistic regression diagnostics: understanding how well a model predicts outcomes. JAMA. 2017;317(10):1068–9.
161. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008;56(1):45–50.
162. Søreide K. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. J Clin Pathol. 2009;62(1):1–5.
163. Ebell MH, Locatelli I, Senn N. A novel approach to the determination of clinical decision thresholds. BMJ Evid Based Med. 2015;20(2):41–7.
164. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74.
165. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc. 2009;172(4):729–48.
166. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6.
167. Feldstein DA, Hess R, McGinn T, Mishuris RG, McCullagh L, Smith PD, 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.
168. Van Belle V, Van Calster B. Visualizing risk prediction models. PLoS One. 2015;10(7):e0132614.
169. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: the Framingham study risk score functions. Stat Med. 2004;23(10):1631–60.
170. Cole TJ. Algorithm AS 281: scaling and rounding regression coefficients to integers. J R Stat Soc Ser C Appl Stat. 1993;42(1):261–8.
171. Maguire JL, Kulik DM, Laupacis A, Kuppermann N, Uleryk EM, Parkin PC. Clinical prediction rules for children: a systematic review. Pediatrics. 2011;128(3):e666–e77.
172. Keogh C, Wallace E, O'Brien KK, Galvin R, Smith SM, Lewis C, et al. Developing an international register of clinical prediction rules for use in primary care: a descriptive analysis. Ann Fam Med. 2014;12(4):359–66.
173. Stiell IG, Greenberg GH, Wells GA, McDowell I, Cwinn AA, Smith NA, et al. Prospective validation of a decision rule for the use of radiography in acute knee injuries. JAMA. 1996;275(8):611–5.
174. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5):475–83.
175. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35(2):214–26.
176. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172(8):971–80.
177. van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. Stat Med. 2016;35(23):4136–52.
178. Ban J-W, Stevens R, Perera R. Predictors for independent external validation of cardiovascular risk clinical prediction rules: cox proportional hazards regression analyses. BMC Diagn Progn Res. 2018;2:3.
179. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25–34.
180. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61(1):76–86.
181. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation. 1999;99(16):2098–104.
182. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004;23(16):2567–86.
183. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
184. Demler OV, Pencina MJ, D'Agostino RBS. Misuse of DeLong test to compare AUCs for nested models. Stat Med. 2012;31(23):2577–87.
185. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157–72.
186. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. Med Decis Mak. 2013;33(4):490–501.
187. Leening MJ, Steyerberg EW, Van Calster B, D'Agostino RB Sr, Pencina MJ. Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. Stat Med. 2014;33(19):3415–8.
188. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med. 2014;160(2):122–31.
189. Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. The net reclassification index (NRI): a misleading measure of prediction improvement even with independent test data sets. Stat Biosci. 2015;7(2):282–95.
190. Burch PM, Glaab WE, Holder DJ, Phillips JA, Sauer JM, Walker EG. Net reclassification index and integrated discrimination index are not appropriate for testing whether a biomarker improves predictive performance. Toxicol Sci. 2017;156(1):11–3.
191. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2014;33(19):3405–14.
192. Antolini L, Tassistro E, Valsecchi MG, Bernasconi DP. Graphical representations and summary indicators to assess the performance of risk predictors. Biom J. 2018. https://doi.org/10.1002/bimj.201700186.
193. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318.
194. Cook NR. Quantifying the added value of new biomarkers: how and how not. BMC Diagn Progn Res. 2018;2:14.
195. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ. 2012;344:e686.
196. White H. Theory-based impact evaluation: principles and practice. J Dev Effect. 2009;1(3):271–84.
197. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ. 2015;350:h1258.
198. Dowding D, Lichtner V, Closs SJ. Using the MRC framework for complex interventions to develop clinical decision support: a case study. Stud Health Technol Inform. 2017;235:544-8.
199. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ. 2011;343:d7163.
200. Brown B, Cheraghi-Sohi S, Jaki T, Su T-L, Buchan I, Sperrin M. Understanding clinical prediction models as 'innovations': a mixed methods study in UK family practice. BMC Med Inform Decis Mak. 2016;16:106.
201. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ. 2008;337:a1655.
202. Lee TH. Evaluating decision aids. J Gen Intern Med. 1990;5(6):528–9.
203. Kappen TH, Vergouwe Y, van Wolfswinkel L, Kalkman CJ, Moons KG, van Klei WA. Impact of adding therapeutic recommendations to risk assessments from a prediction model for postoperative nausea and vomiting. Br J Anaesth. 2015;114(2):252–60.
204. Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. BMJ. 2004;328(7435):343–5.

205. Wallace E, Uijen MJM, Clyne B, Zarabzadeh A, Keogh C, Galvin R, et al. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. BMJ Open. 2016;6(3):e009957.
206. Sanders SL, Rathbone J, Bell KJL, Glasziou PP, Doust JA. Systematic review of the effects of care provided with and without diagnostic clinical prediction rules. BMC Diagn Progn Res. 2017;1:13.
207. Kappen T, Peelen LM. Prediction models: the right tool for the right problem. Curr Opin Anesthesiol. 2016;29(6):717–26.
208. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. BMJ. 2004;328(7441):702–8.
209. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ. 2015;350:h391.
210. Poldervaart JM, Reitsma JB, Koffijberg H, Backus BE, Six AJ, Doevendans PA, Hoes AW. The impact of the HEART risk score in the early assessment of patients with acute chest pain: design of a stepped wedge, cluster randomised trial. BMC Cardiovasc Disord. 2013;13:77.
211. Hayes RJ, Moulton LH. Cluster randomised trials. Boca Raton: CRC Press; 2017.
212. Campbell MK, Elbourne DR, Altman DG. CONSORT group. CONSORT statement: extension to cluster randomised trials. BMJ. 2004;328(7441):702–8.
213. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. Int J Epidemiol. 2015;44(3):1051–67.
214. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. BMJ. 2017;358:j3064.
215. Schaafsma JD, van der Graaf Y, Rinkel GJ, Buskens E. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. J Clin Epidemiol. 2009;62(12):1248–52.
216. Koffijberg H, van Zaane B, Moons KG. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. BMC Med Res Methodol. 2013;13:12.
217. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG, Ioannidis JP. Reply to letter by Ferrante di Ruffano et al.: patient outcomes in randomized comparisons of diagnostic tests are still the ultimate judge. J Clin Epidemiol. 2016;69:267–8.
218. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.
219. Reilly BM, Evans AT, Schaider JJ, Das K, Calvin JE, Moran LA, et al. Impact of a clinical decision rule on hospital triage of patients with suspected acute cardiac ischemia in the emergency department. JAMA. 2002;288(3):342–50.
220. Cowley LE, Maguire S, Farewell DM, Quinn-Scoggins HD, Flynn MO, Kemp AM. Acceptability of the predicting abusive head trauma (PredAHT) clinical prediction tool: a qualitative study with child protection professionals. Child Abuse Negl. 2018;81:192–205.
221. Ballard DW, Rauchwerger AS, Reed ME, Vinson DR, Mark DG, Offerman SR, et al. Emergency physicians' knowledge and attitudes of clinical decision support in the electronic health record: a survey-based study. Acad Emerg Med. 2013;20(4):352–60.
222. Johnson EL, Hollen LI, Kemp AM, Maguire S. Exploring the acceptability of a clinical decision rule to identify paediatric burns due to child abuse or neglect. Emerg Med J. 2016;33(7):465–70.
223. Mullen S, Quinn-Scoggins HD, Nuttall D, Kemp AM. Qualitative analysis of clinician experience in utilising the BuRN tool (burns risk assessment for neglect or abuse tool) in clinical practice. Burns. 2018;44(7):1759–66.
224. Haskins R, Osmotherly PG, Southgate E, Rivett DA. Physiotherapists' knowledge, attitudes and practices regarding clinical prediction rules for low back pain. Man Ther. 2014;19(2):142–51.
225. Kelly J, Sterling M, Rebbeck T, Bandong AN, Leaver A, Mackey M, Ritchie C. Health practitioners' perceptions of adopting clinical prediction rules in the management of musculoskeletal pain: a qualitative study in Australia. BMJ Open. 2017;7(8):e015916.
226. Atabaki SM, Hoyle JDJ, Schunk JE, Monroe DJ, Alpern ER, Quayle KS, et al. Comparison of prediction rules and clinician suspicion for identifying children with clinically important brain injuries after blunt head trauma. Acad Emerg Med. 2016;23(5):566–75.
227. Mahajan P, Kuppermann N, Tunik M, Yen K, Atabaki SM, Lee LK, et al. Comparison of clinician suspicion versus a clinical prediction rule in identifying children at risk for intra-abdominal injuries after blunt torso trauma. Acad Emerg Med. 2015;22(9):1034–41.
228. Reilly BM, Evans AT, Schaider JJ, Wang Y. Triage of patients with chest pain in the emergency department: a comparative study of physicians' decisions. Am J Med. 2002;112(2):95–103.
229. Broekhuizen BD, Sachs A, Janssen K, Geersing GJ, Moons K, Hoes A, Verheij T. Does a decision aid help physicians to detect chronic obstructive pulmonary disease? Br J Gen Pract. 2011;61(591):e674–e79.
230. Schriger DL, Newman DH. Medical decisionmaking: let's not forget the physician. Ann Emerg Med. 2012;59(3):219–20.
231. Finnerty N, Rodriguez R, Carpenter C, Sun B, Theyyunni N, Ohle R, et al. Clinical decision rules for diagnostic imaging in the emergency department: a research agenda. Acad Emerg Med. 2015;22(12):1406–16.
232. Sanders S, Doust J, Glasziou P. A systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment. PLoS One. 2015;10(6):e0128233.
233. Cowley LE, Farewell DM, Kemp AM. Potential impact of the validated predicting abusive head trauma (PredAHT) clinical prediction tool: a clinical vignette study. Child Abuse Negl. 2018;86:184–96.
234. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. BMJ. 2011;342:d1766.
235. Grimshaw J, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L, et al. Changing provider behavior: an overview of systematic reviews of interventions. Med Care. 2001;39(8 Suppl 2):II2–II45.
236. Stiell IG, Bennett C. Implementation of clinical decision rules in the emergency department. Acad Emerg Med. 2007;14(11):955–9.
237. Cameron C, Naylor CD. No impact from active dissemination of the Ottawa ankle rules: further evidence of the need for local implementation of practice guidelines. CMAJ. 1999;160(8):1165–8.
238. Davis DA, Taylor-Vaisey A. Translating guidelines into practice. A systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines. CMAJ. 1997;157(4):408–16.
239. Katz MH. Integrating prediction rules into clinical work flow. JAMA Intern Med 2013;173(17):1591–91.
240. Boutis K, Constantine E, Schuh S, Pecaric M, Stephens D, Narayanan UG. Pediatric emergency physician opinions on ankle radiograph clinical decision rules. Acad Emerg Med. 2010;17(7):709–17.
241. Pluddemann A, Wallace E, Bankhead C, Keogh C, Van der Windt D, Lasserson D, et al. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. Br J Gen Pract. 2014;64(621):e233–e42.
242. Kappen TH, van Loon K, Kappen MA, van Wolfswinkel L, Vergouwe Y, van Klei WA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. J Clin Epidemiol. 2016;70:136–45.
243. Keogh C, Fahey T. Clinical prediction rules in primary care: what can be done to maximise their implementation? Clin Evid. 2010. https://core.ac.uk/download/pdf/60774649.pdf. Accessed 12 June 2018.
244. Runyon MS, Richman PB, Kline JA. Emergency medicine practitioner knowledge and use of decision rules for the evaluation of patients with suspected pulmonary embolism: variations by practice setting and training level. Acad Emerg Med. 2007;14(1):53–7.
245. Pearson SD, Goldman L, Garcia TB, Cook EF, Lee TH. Physician response to a prediction rule for the triage of emergency department patients with chest pain. J Gen Intern Med. 1994;9(5):241–7.
246. Brehaut JC, Stiell IG, Visentin L, Graham ID. Clinical decision rules "in the real world": how a widely disseminated rule is used in everyday practice. Acad Emerg Med. 2005;12(10):948–56.
247. Brehaut JC, Stiell IG, Graham ID. Will a new clinical decision rule be widely used? The case of the Canadian C-spine rule. Acad Emerg Med. 2006;13(4):413–20.
248. Graham ID, Stiell IG, Laupacis A, O'Connor AM, Wells GA. Emergency physicians' attitudes toward and use of clinical decision rules for radiography. Acad Emerg Med. 1998;5(2):134–40.
249. Eichler K, Zoller M, Tschudi P, Steurer J. Barriers to apply cardiovascular prediction rules in primary care: a postal survey. BMC Fam Pract. 2007;8:1.
250. Beutel BG, Trehan SK, Shalvoy RM, Mello MJ. The Ottawa knee rule: examining use in an academic emergency department. West J Emerg Med. 2012;13(4):366–72.
251. Sheehan B, Nigrovic LE, Dayan PS, Kuppermann N, Ballard DW, Alessandrini E, 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.

252. van der Steen JT, Albers G, Licht-Strunk E, Muller MT, Ribbe MW. A validated risk score to estimate mortality risk in patients with dementia and pneumonia: barriers to clinical impact. Int Psychogeriatr. 2011;23(1):31–43.
253. Sanders S. Clinical prediction rules for assisting diagnosis (doctoral thesis). Australia: Faculty of Heath Sciences & Medicine, Bond University; 2015.
254. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, Rubin HR. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999;282(15):1458–65.