# NERDCAT-CPR

## A clinician's guide to appraising clinical prediction rules

The latest version of this tool and other NERDCATs can be found at **nerdcat.org**

Full references to supporting literature and articles used as examples in this document can be found at: **http://nerdcat.org/useful-references/**

### 1. What are clinical prediction rules?

Clinical prediction rules (CPRs), which go by *many* different names, (e.g. risk assessment models, prognostic tools), are mnemonics, clinical calculators or algorithms that use information about a patient to predict their risk of experiencing an outcome of interest. Two of the most widely recognized CPRs in medicine are the Framingham Risk Score and the CHADS2 score.

### 2. What… I need to appraise these too?

As with drug trials, there is significant variation in the quality and applicability of the evidence for CPRs. Using a CPR that is not right for your patient will mislead you and may lead to harmful decisions. Therefore, as with drug trials, we need to assess studies of clinical prediction rules for the evidentiary trifecta: Generalizability, risk of bias, and clinical and statistical significance of the results.

### 3. Do I need to learn a bunch of new statistics?

No! There are some concepts unique to diagnostics (gasp, but I'm a pharmacist!) and prognostics that you will need to become familiar with, which we will discuss below.

## A. Levels of Evidence

Just as drugs go through different phases of trials prior to regulatory approval, clinical prediction rules must go through a rigorous process before they are ready for clinical use. Table 1 shows a generally accepted overview of the different steps involved in the development of a clinical prediction rule. Table 2 shows a hierarchy of evidence for clinical prediction rules that is analogous to the 5S hierarchy of evidence for therapeutic interventions.

**Table 1. Development of a Clinical Prediction Rule**

| Step | Level of evidence |
|---|---|
| Derivation | 3 (lowest) |
| Internal validation | 2b |
| External validation | 2a |
| Impact analysis | 1 (highest) |

**Table 2. Hierarchy of Evidence for Clinical Prediction Rules**

| | |
|---|---|
| **Level 1** | Rule that can be used in a wide variety of settings with confidence that they can change clinician behavior and improve patient outcomes. Requires:<br>• ≥1 prospective validation in a different population<br>• ≥1 impact analysis, demonstrating change in clinician behavior with beneficial consequences |
| **Level 2** | Rule that can be used in various settings with confidence in their accuracy. Requires:<br>• ≥1 large prospective study including a broad spectrum of patients & clinicians<br>or<br>• Several smaller settings that differ from one another |
| **Level 3** | Rule that clinicians may consider using with caution and only if patients in the study are similar to those in the clinician's setting<br>• Validated in only 1 narrow prospective sample |
| **Level 4** | Rule that needs further evaluation before it can be applied clinically<br>• Derived but not validated<br>• Validated only in split samples, large retrospective databases, or by statistical techniques |

### *Derivation*

The first step is the creation of the tool (*derivation*), which optimally starts with prospective collection of data on known (or commonly believed) risk factors and other available information. These characteristics are eventually entered into a multivariable analysis, which assesses each variable's independent association with the outcome of interest.

The variables that are found to be independently associated with the outcome of interest (the independent risk factors) can then be given a score, programmed into an online calculator, or turned into an algorithm for use. Even if such a fancy tool now exists, it doesn't mean that it's yet appropriate for clinical use!

> Note 1: In some cases, risk factors aren't derived from a rigorous statistical process, and are instead simply chosen by consensus. This approach is inherently flawed because the relative importance of the individual risk factors and their independence from one-another cannot be assessed without multivariable analysis.

> Note 2: A variation of the above involves finding that certain variables are not independently associated with the outcome of interest, but including them into the clinical prediction rule anyway (e.g. 'HAS-BLED' should really just be 'ABE'). This (predictably!) weakens the predictive ability of the clinical prediction rule.

### *Internal Validation*

Most clinical prediction rules are tested in the cohort of patients from which they were derived (known as *internal validation*). This is an important step in the development of a clinical prediction rule, but the estimated predictive power of the tool is typically overestimated when assessed in the cohort from which it originated.

A tool that has been internally validated can be applied VERY cautiously in a population and setting that is very similar to those of the derivation cohort (e.g. new patients treated in the same ward from the same hospital with the same prevalence of medical conditions).

> Note 2: A variation of internal validation is *temporal validation*, in which investigators calculate the predictive power of the clinical prediction rule in the same clinical setting, but with a population collected at a different time point (e.g. a clinical prediction rule derived from VGH patients treated in the ICU from 2000-2005 is validated in VGH ICU patients treated from 2006-2010).

### *External Validation*

Before the use of a clinical prediction rule can be widely disseminated into practice, it must first show reproducible predictive power in a broad patient population (*external validation*). This is optimally done in a prospective cohort of consecutively enrolled patients that are representative of the target population.

> Note 3: A common trend is to validate a clinical prediction rule in an RCT population. This is inadequate, as most RCTs selectively exclude patients with the very risk factors that we are attempting to validate (e.g. one of the strongest predictors of death or MI in the GRACE prognostic score is declining creatinine clearance. Previous clinical prediction rules derived and validated in RCTs (e.g. TIMI) failed to find as a predictor variable because these trials excluded patients with significant renal impairment).

Ideally, clinicians should be collecting the data using the tool as intended for practice (pocket nomograph, smartphone app, mental arithmetic, etc.)

### *Impact Analysis*

The final step in the development of a clinical prediction rule is assessing its ability to improve patient outcomes. The optimal study design to assess this with a cluster RCT, wherein entire institutions are randomized to use of standard of care or use of the clinical prediction rule to guide therapy. Another alternative includes a before-and-after design, wherein clinicians in the "before" group treat patients according to the historic standard of care, and clinicians in the "after" group use the clinical prediction rule to guide their practice.

Though they represent the best available evidence in terms of clinical prediction rules, impact analysis trials have rarely been conducted, likely due to the low number of clinical prediction rules that have made it through external validation and pose a significant improvement over the standard of care.

## CRITICALLY APPRAISING A DERIVATION OR VALIDATION STUDY

| Study Population | |
|---|---|
| **Source of data** – The study type dictates the reliability of data collection and generalizability of the tool.<br>• The optimal study for derivation of a clinical prediction rule is a prospective cohort designed specifically to collect data for its development.<br>• Retrospective cohorts may be used, but data collection may be inaccurate or incomplete.<br>• Randomized-controlled trial (RCT) data may be used, but tools derived from RCTs tend to perform less well in later validation studies (be less generalizable) due to the strict eligibility criteria employed in these studies. | |
| **Description of study setting** – To apply the results of the trial, you must be confident that the study setting was similar enough to where you're planning to apply this tool, e.g.:<br>• International vs individual country<br>• Inpatient vs outpatient<br>• Tertiary site versus community hospital<br>• ER vs medical ward vs surgical ward vs ICU | |
| **Description of patient population** – Similar to the above, the study population should be representative of the population seen in this setting.<br>• Clear description of inclusion and exclusion criteria.<br>• Consecutive rather than selective enrolment is desirable (look for figure describing patient inclusion/exclusion, those eligible but not included).<br>• SCRAPP mnemonic from baseline characteristics table. | |
| **Adequate proportion of patients with predictor variables** – If a predictor variable is present in too few patients in the cohort, it may not be sufficiently powered to be statistically significant in the statistical model. | |

| Predictor variables | |
|---|---|
| **Description of all predictor variables considered, including those that did and did not make it in the final tool** – All outcomes that were evaluated for inclusion in the tool should be mentioned to assure you that all potentially important variables were considered. | |
| **Clear and reproducible definition** – The predictor variables, whether from patient history, physical exam, laboratory tests or investigations should be described with sufficient detail to evaluate and apply it in your practice. | |
| **Reliability** – The results of the predictor variable should be consistent and reproducible when assessed by the same (intraobserver reliability) and another (interobserver reliability) clinician. If a tool relies on predictor variables that require clinician interpretation, it should provide a test of agreement in the assessment by different evaluators.<br>• The kappa test for agreement should be used. Predictor variables with $\kappa$ <0.6 are considered unreliable and should not be included in the model. | |
| **Availability at point of decision** – Predictor variables should be available to clinicians at the point in time when they are intended to make a decision. For example, 'length of stay ≥6 days' was a significant predictor of in-hospital VTE in a clinical prediction rule derived from a retrospective cohort. This variable has no value to a clinician trying to decide whether to order thromboprophylaxis on admission. | |
| **Blind assessment** – The investigator collecting the predictor variables should be unaware of the patient's outcome. In a prospective study, where predictor variables were collected prior to the outcome, is blind by definition. | |

| *FOR VALIDATION ONLY* | |
|---|---|
| **Comparability of predictor variable definition & ascertainment in validation study to those in derivation study** – Differences in how the predictor variable is defined, collected and measured in a validation study compared to the original derivation will affect the reproducibility of the clinical prediction rule. | |

| **Outcome** | |
|---|---|
| **Clinical importance** – The outcome should be of direct clinical importance (e.g. death, stroke) or strongly associated with a clinically important outcome (e.g. diabetes mellitus). | |
| **Clear and reproducible definition** – The outcome should be described with sufficient detail to evaluate it in your practice. | |
| **Blind assessment –** The investigator evaluating the presence/absence of the outcome should be unaware of the status of predictors (i.e. risk factors) in the patients. Not necessary if the outcome is all-cause mortality. | |
| *FOR VALIDATION ONLY* | |
| **Comparability of outcome definition & ascertainment in validation study to outcome in derivation study** – Differences in how and when the outcome is defined and ascertained in a validation study compared to the original derivation will affect the reproducibility of the clinical prediction rule. | |

| Accuracy | |
|---|---|
| **Description of multivariable analysis** – In order to explore the association between the predictor variables with each other and with the outcome, sophisticated multivariable analysis is required, such as *logistic regression* or *recursive partitioning*. Recursive partitioning is used when a highly sensitive tool is desired because missing the outcome would be catastrophic. | |
| **Handling of missing data** – Missing data should be reported and appropriately managed in order to avoid biasing the results.<br>• *Multiple imputation* should be used rather than excluding patients with incomplete data. This is analogous to using intention-to-treat analysis in an intervention RCT rather than per-protocol analysis.<br>• Also consider: Predictor variables that are missing in a substantial proportion of study patients may be difficult to acquire in real practice, and should be omitted from the clinical prediction rule. | |
| **Overfitting/Optimism** – Predictions provided by clinical prediction rules are expected to be "overly optimistic" in internal validation compared with subsequent external validation. This is simply because the model was designed to optimally fit the derivation cohort but becomes less accurate when tested in new but similar individuals (*overfitting*).<br>• A study should have 10 or more outcomes per predictor variable being considered for the clinical prediction rule to avoid overfitting.<br>• The *bootstrapping* method is preferred over other techniques such as randomly splitting the original cohort into a derivation and validation cohort, which is statistically inefficient. Bootstrapping allows for adjustment of the c-index to better approximate the expected accuracy in a future study. | |

| | |
|---|---|
| **Discrimination** – The model's ability to distinguish between patients who do and do not experience the outcome.<br>• When the clinical prediction tool has a range of scores, the *c-statistic* can be used. This test gives the chance when given 2 patients – one who has that outcome and one who does not – that the clinical prediction tool will assign a higher probability of outcome to the one who has the outcome. The score ranges from 0.5 to 1.0; a value of 0.5 means that the model is no better than chance at picking out the higher risk patient, and a value of 1.0 means that the model has perfect discrimination.<br>• With a binary score cutoff, a 2x2 table is used to derive values such as *sensitivity*, *specificity* and *likelihood ratios* (see gray box). | |
| **Calibration (a.k.a. goodness of fit)** – The model's ability to correctly estimate the probability of experiencing the outcome.<br>• Evaluated by plotting the observed proportions of events against the predicted probabilities for groups defined by ranges of predicted risk (e.g. low, moderate, high risk). The observed probability of events is the actual % of outcome events occurring in a given subgroup, whereas the predicted probability is the % that's spit out of the statistical model.<br>• Can be formally assessed with statistical analysis, such as the *Hosmer-Lemeshow goodness-of-fit test*, though these tests are insensitive to poor calibration (i.e. conclude that the model is well-calibrated when it is not). | |

**Stats Corner**

To assess the discrimination of a clinical prediction tool at specific cutoffs, we can use a 2x2 table. The terms to be familiar with are:

**Sensitivity (Sn):** The proportion of people who eventually develop the outcome who "test positive".

- **SnNout:** When a clinical prediction tool has a very high sensitivity, a negative result tends to rule out the risk of an outcome.

**Specificity (Sp):** The proportion of people who do not develop the outcome who "test negative".

- **SpPin:** When a clinical prediction tool has a very high specificity, a positive result rends to rule in the risk of an outcome.

**Positive likelihood ratio (LR+):** The relative likelihood of this test result coming from a person who eventually has the outcome versus someone who does not.

- Values ≥5 are thought to be clinically useful

**Negative likelihood ratio (LR-):** The relative likelihood of this test result coming from an individual who does NOT have the outcome versus someone with it.

- Values ≤0.1 are thought to be clinically useful

Next, let's make a trusty 2x2 table using the first validation of the CHADS2 score (I chose the CHADS2 ≥1 cutoff because current clinical practice guidelines recommend oral anticoagulant therapy based on this cutoff):

| | Stroke | No stroke |
|---|---|---|
| **CHADS ≥ 1 ("+ test")** | 92 (true+) | 1521 (false+) |
| **CHADS = 0 ("- test")** | 2 (false-) | 118 (true-) |

***Sn*** *= true+/(true+ plus false-) = 92/(92+2) =* ***0.98***

***Sp*** *= true-/(false+ plus true-) = 118/(1521+118) =* ***0.07***

***LR+*** *= Sn/(1-Sp) = 0.98/(1-0.07) =* ***1.05 (95% CI 1.02-1.09)***

***LR-*** *= (1-Sn)/Sp = (1-0.98)/0.07 =* ***0.29 (95% CI 0.07-1.18)***

To use LRs, we first need the incidence of the outcome in the overall population (called the pre-test probability). In the cohort used to first validate the CHADS2 score, the incidence of stroke was 5.4%. Using a smartphone app like MedCalc, we can take this pre-test probability and our LR to calculate our refined probability of the outcome based on our patient's individualized characteristics (the post-test probability). In this case, the probability of stroke in a patient with CHADS2 = 0 (using LR- since the test was "negative") is 1.6% (95% CI 0.4% to 6.3%).

As with relative risks and odds ratios, a LR confidence interval that includes 1 is not statistically significant. Based on the calculated LRs, we can make a few conclusions about the CHADS2 score's performance in this cohort:

1) The LR+ has a narrow confidence interval, but given that it's so small (1.05), simply classifying a patient as CHADS2 ≥1 does not provide any clinically important discrimination.
2) The LR-'s confidence interval crosses 1 and is therefore not statistically significant. Therefore – at least in this cohort of patients, a CHADS2 score of 0 does not identify patients who won't benefit from oral anticoagulation. We would thus be doing a disservice to our patients by withholding warfarin in these patients based on this score alone.

| **Transportability (aka Generalizability):** The clinical prediction rule is accurate in patients drawn from a population different than the derivation cohort. *The more diverse the previous settings in which the clinical prediction rule has been tested and found to be accurate, the more likely it is that it will generalize to an untested setting.* | |
|---|---|
| **Historical transportability** – Accuracy is maintained when the rule is applied to patients from a different time point. | |
| **Geographic transportability** – Accuracy is maintained when the rule is applied to patients from a different geographical location. | |
| **Domain transportability** – Accuracy is maintained when the rule is applied to a different setting (e.g. applying a rule that was developed in inpatients to outpatients). | |
| **Methodologic transportability** – Accuracy is maintained when the predictor variables are collected using different methods (e.g. resident interprets CXR instead of attending, different cutoffs for predictor variables). | |
| **Spectrum transportability** – Accuracy is maintained in a patient population that is, on average, more or less advanced in disease process or has a somewhat different disease process or trajectory (e.g. applying Framingham risk score in CKD patients).<br>• Calibration of clinical prediction rules is often compromised when tested in a sample of patients with very different levels of disease severity.<br>• Discrimination of clinical prediction rules is altered in a population with a narrower spectrum of disease from both sides (intermediate-risk patients). | |
| **Follow-up interval transportability –** Accuracy is maintained when the outcome is predicted over a longer or shorter period. | |