Methodologic Standards for Interpreting Clinical Decision Rules in Emergency Medicine: 2014 Update

Steven M. Green, MD; David L. Schriger, MD, MPH; Donald M. Yealy, MD

Clinical decision rules are increasingly prominent in medicine, particularly in emergency care. The quality, use, and impact of current published decision rules widely vary, requiring clinicians to be critical consumers. We present an approach to assist in the appraisal of clinical decision rules and in judging when to use such rules. [Ann Emerg Med. 2014;■:1-6.]

0196-0644/$-see front matter

http://dx.doi.org/10.1016/j.annemergmed.2014.01.016

INTRODUCTION

Clinical decision rules are widespread, particularly in emergency care, given the need for prompt but accurate decisionmaking. These algorithms or scoring systems seek to either improve diagnosis or decrease unneeded testing.[1-3] One definition of a clinical decision rule is "a clinical tool that quantifies the individual contributions that various components of the history, physical examination, and basic laboratory results make towards the diagnosis, prognosis, or likely response to treatment in an individual patient."[2,3]

Decision "rule" is a misnomer in that these are not inflexible or absolute, but generally positioned to advise rather than supersede clinical judgment. Accordingly, some prefer to call them clinical decision instruments or decision aids. Although such phrasing is technically more accurate, we have retained the term "decision rule" in this article, given the pervasive appearance in scientific venues for more than 2 decades and no signs of replacement on the horizon.

In 1999, *Annals of Emergency Medicine* published "Methodologic Standards for the Development of Clinical Decision Rules in Emergency Medicine."[1] In this article, Stiell and Wells[1] set forth a structure to guide development of these tools. We seek to update the publication and focus on rule assessment.

WHY AN UPDATE?

This article reflects the evolution of decision rule methodology and philosophy in the past 15 years as investigators have subsequently deployed novel approaches and interpretation during rule creation and dissemination, creating new challenges and controversies. Similarly, some issues briefly noted or implicit in the 1999 guideline now warrant emphasis.

Given the large and increasing pool of clinical decision rules, clinicians can easily feel confused and overwhelmed. Our goal is to ease sorting through the choices, helping readers decide whether a given rule warrants further development or, ultimately, incorporation into clinical practice. Conversely, when is a rule challenged to the point at which it should be reworked or abandoned?

WHEN DO WE NEED A CLINICAL DECISION RULE?

When rules are created or advocated, our first question should be whether they are likely to be helpful in practice; simply put, is it worth the trouble? Specific questions should be posed.

First, does the rule have reasonable potential to improve clinical care? Is there convincing evidence that emergency physicians are inaccurate in diagnosing a specific disease entity or that for the specified condition they are inefficient in their diagnostic testing? If we are already doing a reasonable job, then the potential benefit of adding a rule is slim.

Next, the feasibility of use must be considered; is the rule likely to be used or ignored? The willingness of a clinician to adopt a rule is multifactorial and not always objective and calculated. Some physicians, while acknowledging the limits of their decisionmaking, may be reluctant to use a rule with imperfect sensitivity, particularly when the consequences of a missed diagnosis are high. Other physicians believe that they, treating 1 patient at a time, will perform better than any "generic" rule. Any instrument that has an ungainly number of elements or a complex algorithm is unlikely to be adopted. Whatever the reason, there seems little point in developing a rule that is unlikely to be implemented.

When investigators believe that a decision rule might improve care over current practice, they perform a derivation study in which potential rule components are systematically identified and the proposed tool is developed. If successful, the same investigators or, ideally, others should attempt to validate the rule in a new sample. Below we present guidelines for derivation and validation of rules, and the reporting of the methods and results of such efforts (see Figure).

IS THE METHODOLOGY OF THE RULE SOUND?

Reports of the development and evaluation of a clinical decision rule should identify and explain the following:

What Is the Population and Setting?

A decision rule evaluated in one practice setting may not extrapolate to a different patient population. For example, a

For a Clinical Decision Rule to be considered successful, all of the following should apply:

General
❒ **Need**: Is there reason to believe that the rule could improve patient care? Is it likely to be used?

Rule Derivation Methodology
❒ **Setting and Population**: Was the rule derived in a setting and patient population representative of where it would be applied?
❒ **Outcome**: Is the rule's predicted outcome measure clinically important and consistently assessed?
❒ **Predictor Variables**: Were all reasonable predictor variables assessed for inclusion? Were these variables prospectively recorded by observers who were blinded to the outcome, and shown to be reliable?
❒ **Rule Creation**: Was there appropriate variable coding, sample size, and analytic technique? Do the authors state and justify their desired test characteristics in advance?

Rule Derivation Reporting
❒ **Performance**: Are both the sensitivity and specificity of the rule clearly and prominently reported, together with their confidence intervals?
❒ **Success**: Does the rule appear successful? Was it compared to decisions made by unaided physicians? Does it improve on unaided clinical performance? Is its accuracy sufficient for the clinical question at hand, taking into consideration the range of confidence intervals associated with the rule's performance? Does it provide a clear two-way course of action, or is there compelling evidence to verify that it can be successfully applied in one-way fashion?
❒ **Practicality**: Is the rule sensible and easy to apply? Can it be readily memorized or otherwise implemented?
❒ **Provisional Nature**: Does the study acknowledge the need for validation prior to clinical use?

Rule Validation Methodology
❒ **Setting and Population**: Was the rule validated in a setting and patient population representative of where it would be applied?
❒ **Replication**: Does the validation methodology closely replicate the definitions and variable coding used in the derivation study?

Rule Validation Reporting
❒ **Performance**: Are both the sensitivity and specificity of the rule clearly and prominently reported, together with their confidence intervals?
❒ **Success**: Is the rule successful? Does it improve on baseline clinical performance? Is its accuracy sufficient for the clinical question at hand, taking into consideration the range of confidence intervals associated with the rule's performance? Does it provide a clear two-way course of action, or is there compelling evidence to verify that it can be successfully applied in one-way fashion?
❒ **Reliability**: Was the global rule found to be reliable?

Figure. Success Criteria for Clinical Decision Rules.

syncope rule derived in a relatively young, inner-city emergency department (ED) population may not remain predictive in a suburban ED in which elderly patients predominate. Ultimately, a rule must be validated in a setting comparable to the one in which it will be used.

How Were Patients Selected?

Verify the specific inclusion and exclusion criteria used. Many clinical decision rules do not study all patients considered for a specific diagnosis, but instead enroll a limited sample at higher or lower risk of the disease, or sample only those for whom a certain diagnostic test has been ordered. It is important to know whether the population used to derive or test the rule had the same approximate pretest probability of disease as those to whom you hope to apply the rule. Inclusion of an undue number of obviously positive or clearly negative cases may create a rule whose performance deteriorates in the subset of patients for whom the rule might truly be helpful, ie, clinical situations that are not straightforward.

The sampling of above patients meeting inclusion and exclusion criteria should ideally be consecutive because any method that falls short of this can introduce selection biases that cannot necessarily be adjusted for later.

Does the Outcome Matter?

The target outcome for the clinical decision rule should be clear, reproducible, clinically important, and ideally patient-oriented. For example, we do not just want to predict abnormal computed tomography (CT) findings with blunt head trauma; we want to predict abnormalities that dictate intervention. Meaningful outcomes might include death, operative intervention, hospitalization, and need for mechanical

ventilation or vasopressor support. Some investigators choose composite outcomes; unfortunately, these mask the differential effect on the component suboutcomes. When using a composite outcome, investigators should report each component individually so readers can assess its relative importance.

To avoid the circular logic known as incorporation bias, the rule's outcome should be classified in a manner independent of its predictor variables. For example, a decision rule to identify acute myocardial infarction will naturally identify troponin elevation as a predictor if troponin is one means of assigning the myocardial infarction diagnosis.

Was the Outcome Routinely Assessed?

Was the outcome measured for everyone enrolled? In some studies, it is not possible to routinely assess the reference criterion (eg, always perform appendectomy to obtain a pathologic diagnosis.) In these cases, was the proxy measure (eg, clinical follow-up) sufficient in quality and timing to closely approximate the outcome? Were patients unable to be assessed different from those who were, perhaps altering rule performance?

Were All Potentially Important Predictor Variables Included?

In a decision rule derivation study, the investigators must create a candidate list of clinical items (eg, signs, symptoms, diagnostic results) to be tested for their ability to predict the outcome. The list should include all factors that clinicians are currently using to assess the predictor and any others with a biologically plausible association.

Clinicians should scrutinize the predictor variable list to ensure that it is complete. A development effort that fails to consider well-established predictors, such as anticoagulant use in head trauma, is unlikely to produce an optimal rule. Similarly, the predictor variable list should include results of rapidly available tests. For example, a rule to predict when to obtain a CT for blunt abdominal trauma that omitted focused abdominal sonography for trauma is less relevant to clinicians who routinely use this latter tool. Furthermore, each element must be easily obtained in the ED. A rule based on culture results, endoscopy findings, or tests that take hours or days to complete is not feasible.

Are the Predictor Variables Objective or Collected Prospectively?

Except in rare circumstances of objective, contemporaneously recorded variables that cannot be confused (eg, admission versus discharge, laboratory values, electronic time stamps), we recommend prospective collection of all predictor data to avoid assessment biases. Many clinical findings cannot be reliably determined with chart review.[4] Large administrative databases created with chart review have the same limitation.[5]

Were the Predictor Variables Recorded Before Knowledge of the Outcome?

To prevent conscious or subconscious bias on the part of investigators, predictor variables should be recorded without knowledge of the outcome. For a decision rule designed to predict meningitis, a clinician aware of positive lumbar puncture results might be more prone to recording neck stiffness as present.

Are the Potential Predictor Variables Reliable?

Even with prospective collection, many history and physical examination findings that might predict an outcome are subjective enough that they are not reliably assessed between clinicians. Nonreproducible variables should not be inserted into clinical decision rules because unreliable input often impairs rule performance in practice. Two physicians often do not reliably record the same Glasgow Coma Scale score,[6] and insertion of this scale may create error in illness or risk assessment.

Derivation studies should report or reference the reliability of their predictor variables, usually expressed as raw agreement and at times also chance-corrected agreement. The latter is commonly quantified with Cohen's κ statistic for binary categories and the weighted κ for multiple categories. When to use κ and what specific κ threshold might constitute acceptable reliability are complex questions well beyond the scope of this article. However, each derivation article should report a predetermined plan for measuring and verifying reliability.[7]

Additional guidelines specific to rule derivation studies are as follows.

How Were the Predictor Variables Coded?

Although many clinical variables are studied as simply present versus absent, others are graded (eg, scores) or continuous (eg, WBC count). Investigators can code these latter variables in ways that may or may not reflect how a given clinician might apply them. If the WBC is coded as high versus normal, with a threshold of 10,000/mm^3, this may not be meaningful to a clinician who prefers to consider 15,000/mm^3 or greater as abnormal. How were missing data points coded? Were they simply assumed negative?

Was the Derivation Sample Large Enough?

A decision rule model can easily give the illusion of precision—or "overfit the data"—when the derivation sample is too small. One widely used rule of thumb when using typical multivariable logistic regression for rule derivation is that there be *a minimum* of 10 subjects with the *least* common outcome for each predictor variable evaluated.[8] Regardless of the development technique, readers should be highly skeptical when this rule of thumb is violated. They should also recognize that although overfitting is likely when the 10:1 rule is not met, it can still occur even when this rule is met.

Is the Analytic Technique Appropriate?

The most common techniques used to develop rules are multivariable logistic regression and binary recursive partitioning in its various forms (eg, classification and regression trees [CART]). Each technique has its advocates, and some techniques may be preferred in certain situations. Authors should describe what they did in sufficient detail that an interested reader could replicate the technique.

Was the Goal of the Rule Explicitly Specified A Priori?

No rule is perfect and all rules require some tradeoff between sensitivity and specificity. The optimal compromise depends on the relative harm of a false positive versus a false negative. Investigators need to specify the minimal combination of sensitivity and specificity they require at validation for the rule to be considered clinically useful.

Is the Need for External Validation Before Clinical Application Acknowledged?

Even when developed with the best derivation methodology, clinical decision rules should not be applied clinically until they are successfully validated in a different sample, ideally by different investigators. Internal validation within the original data set, although useful if conducted with appropriate methods, is not external validation. As already noted, clinical decision rules typically perform better in the derivation sample because they are statistically modeled to depict that specific data set. When tested in a new patient sample, decision rules typically perform less well or at times fail altogether. One example is the San Francisco Syncope Rule, in which performance decreased when tested in multiple new samples.[9] Ideally, validity is established in varied settings by different investigators.

Can the Final Rule Itself Be Reliably Assessed?

We have already discussed the importance of verifying the interrater reliability of clinical variables used to develop a decision rule. However, it is also of importance whether the final global rule itself can be reliably calculated. As part of each validation study, a subset of enrolled patients should have the rule independently assessed and calculated by 2 clinicians, and the degree of agreement should be reported.

Can the Rule Be Refined?

At times, validation or impact research suggests ways in which a rule could be simplified or refined. If attempted, a new validation study is then warranted to verify the accuracy of the modified rule.

ARE RESULTS OF THE RULE CLEARLY REPORTED?

Are Sensitivity and Specificity Both Emphasized?

A highly sensitive rule may lack sufficient specificity to be practical, or vice versa. Neither value should be interpreted in the absence of the other; articles that stress a singular "more-impressive" value are misleading. We prefer seeing the complete 2×2 rule performance table. Report positive and negative predictive values in a subordinate manner to sensitivity and specificity (if at all) because they vary with disease prevalence and may not be stable.

Does the Rule Improve On Baseline Clinical Practice?

Does the study clearly contrast the result (or theoretical result if a derivation study) of applying the rule with what occurs in its absence? For example, does applying the rule improve diagnosis or decrease test use compared with unstructured clinical judgment (ie, gestalt)? The fundamental purpose of a decision rule is to improve clinical care, not just to predict what we are already doing. The Ottawa Ankle Rules improve on clinical judgment because they are just as sensitive while being more specific. A rule that replicates but does not improve on bedside judgment has little added value.[10,11] Accordingly, the contrast of the rule to unstructured clinical judgment is a vital metric mandatory in any studies creating or validating a clinical decision rule.

Some authors might argue that even if their rule did not improve on their local clinical judgment, it might aid in other ED settings in which practitioners have less training or experience with the disease entity in question. Such a presumption of differential skills should be supported with data, particularly because many studies are conducted in academic centers in which the clinical gestalt of a trainee may not be the same as that of an experienced clinician.

Is the Rule's Performance Sufficiently Precise?

A sensitivity of 100% is ideally sought for a rule wishing to exclude a disease or outcome. However, if investigators report 100% sensitivity for a new rule that identified only 10 subjects with the condition, then the lower limit of the 95% confidence interval (CI) for this value is 70%, meaning that repetitions of this experiment could reasonably produce values that low. The larger the sampling of diseased patients included in this calculation, the smaller the variation around the point estimate. The lower CI for 100 of 100 is 96.3%, for 250 of 250 it is 98.5%, and for 500 of 500 it is 99.2%. If clinicians will only be comfortable implementing a rule that guarantees a sensitivity above 99%, then the validation study must enroll a large sample such that the lower bound of the CI exceeds this threshold.

HOW IS THE RULE INTERPRETED?

Is the Rule Successful?

Does the rule improve on baseline clinical performance? Is its accuracy sufficient for the clinical question at hand, taking into consideration the range of CIs associated with the rule's performance?

Is the Rule Sensible?

Do the components of the rule make sense? Is there face validity?

Is the Rule Easy to Apply?

Is the rule simple and straightforward? Most decision rules are ignored, and the more complicated they are, the more likely they are to be ignored. Is the rule such that clinicians can reliably list the components by memory? If beyond typical recall capacity, and if there is not an easy method to apply it (eg, a pop-up screen when ordering a test or choosing a diagnosis), then the rule is unlikely to be consistently applied. If any support tool requires extra data entry to input the required variables, it may not be adopted.

Does the Rule Provide a 2-Way Course of Action?

The outcome of applying the rule should be a specific course of action if the rule is positive and another course of action if it is negative. For example, if the rule is positive, order the radiograph, and if negative, do not.

The output of some rules is instead just a probability of a given diagnosis or of mortality, leaving clinicians to decide how to act according to this knowledge. Although this may be acceptable in certain clinical situations, rules that fail to provide clinicians with clear direction may not achieve widespread adoption because they do not remove the uncertainty that created the need for the rule.

The ideal decision rule will demonstrate both high sensitivity and high specificity. When only 1 of these is achieved, developers may be tempted to offer their decision instrument as "1-way."[11] For example, when clinicians are faced with high sensitivity but low specificity, a 1-way application might be recommended, eg, they should not order the radiograph if all elements are absent but may or may not order it if 1 or more elements are present. When sensitivity is low but specificity is high, the converse recommendation might be to order the radiograph if any element is present but not necessarily order it when all are absent. In both cases, clinicians are asked to apply the rule clinically if it suggests one course of action, but then if it indicates the opposite, they are asked to ignore the rule altogether and apply some separate and independent manner of decisionmaking.

Such 1-way interpretations were not typically part of the investigators' original rule creation and testing, but were proposed post hoc. The pitfall of such an approach is the natural human propensity to apply the rule regardless in a 2-way fashion. It is possible that the net effect of applying the rule will increase rather than reduce the undesired outcome (eg, radiograph ordering). The capability of clinicians to successfully implement a 1-way rule should be demonstrated before clinical application.

WHAT POSTVALIDATION RESEARCH IS HELPFUL?

After successful validation, additional research may be warranted.

Can the Rule Be Successfully Implemented Into Clinical Practice?

It is one thing to develop and validate a successful clinical decision rule and much different to show that clinicians can and will use it. Will physicians find the rule ungainly or be uncomfortable using it because of perceived medicolegal risk? Follow-up impact research is warranted to assess the potential postdissemination benefit in actual users.

Is the Rule Cost-effective?

Implementation of a decision rule typically results in a tradeoff between altered testing and altered rate of diagnosis. A full assessment of costs is needed, including settings in which rule use increases resources or later care needs. A cost-effectiveness evaluation quantifies the estimated impact and should ideally include sensitivity analyses to simulate the performance of the rule in multiple settings.

Cost-effectiveness studies should not be performed on derivation sets because the optimal performance and potential overfit of the model can exaggerate the results. Validation or impact data sets are better used, especially the latter because they best reflect actual use.

How Can the Rule Be Widely Integrated Into Practice?

When a validated rule is successful and has been shown to improve clinical care, dissemination may still represent a challenge. Research may then be warranted on how best to advertise the rule and encourage clinical adoption. Strategies may include training sessions, mnemonics, pocket cards, posters, mobile telephone applications, and customized prompts within electronic medical records.

THE FATE OF RULES—PUBLICATION AND POSTPUBLICATION

The pool of clinical decision rules is increasing and few are being formally discarded. Some rules never met initial quality targets; for others, the disease or clinical approach has changed over time and may alter the need or utility. Journal peer review before publication does not always identify when a rule initially fails to fulfill the scientific rigor needed for clinical use. As in any study, authors may at times portray their results in the most positive light and omit or miss important rule limitations. Peer reviewers are human and may miss important limitations, and journals may lack the resources to sufficiently critique rule methodology; this may allow unduly optimistic results to be published without corresponding limitations. Readers who take study conclusions at face value will be misled. Postpublication peer evaluation is a potential correcting force but is underused.[12]

In some situations, rule performance may require periodic reevaluation. Although a rule guiding extremity imaging will likely retain validity over time, one that predicts outcomes in an infection or a cardiac illness may not as population, illness burden, and care patterns shift.

Currently, there is no agreed-on approach for the postreporting assessment of clinical decision rules. We suggest that rules failing to meet our success criteria (Figure) be labeled as unsuccessful in review articles and textbook chapters, and when

possible be removed altogether. Rules will fail. Not every clinical condition can be predicted by mathematical modeling or reduced to an algorithm, even when attempted with enormous, multicenter studies.

Studies reporting unsuccessful clinical decision rules can still be worthwhile additions to the medical literature. Documentation of such failure may prevent other researchers from pursuing the same investigative course, and it is hoped that it will steer them to formulate alternative approaches. Furthermore, even when rules are not successful clinicians may still benefit from their descriptions of predictor variables, using them in unstructured fashion to hone their gestalt. They might discount clinical variables found to be unreliable or inaccurate.

LIMITATIONS

We seek to help readers decide whether a clinical decision rule is likely to be valid and helpful. Our recommendations are based on a combination of evidence and experience, with the latter predominating. As such, they are subject to the foibles of all experiential knowledge.

Each rule development effort is different and must be evaluated independently. Some of our guidance may be inapplicable or insufficient for the evaluation of specific rules. Nonetheless, we believe that these recommendations will help readers assess the possibility that some rules are invalid or not useful.

CONCLUSION

Clinical decision rules abound and the quality varies. We present an approach to assist in the critical appraisal of clinical decision rules and in judging when to use such rules.

The authors acknowledge the Annals of Emergency Medicine *panel of methodology and statistics editors for their critique of this article and many valuable comments.*

Supervising editor: Michael L. Callaham, MD

Dr. Callaham was the supervising editor on this article. Drs. Green, Schriger, and Yealy did not participate in the editorial review or decision to publish this article.

Author affiliations: From the Loma Linda University Medical Center and Children's Hospital, Loma Linda, CA (Green); the University of California, Los Angeles, Los Angeles, CA (Schriger); and the University of Pittsburgh School of Medicine, Pittsburgh, PA (Yealy).

Funding and support: By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships exist.

Publication dates: Received for publication December 26, 2013. Accepted for publication January 14, 2014.

Address for correspondence: Steven M. Green, MD, E-mail steve@viridissimo.com.

REFERENCES

1. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med*. 1999;33:437-447.
2. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules: a review and suggested modifications of methodological standards. *JAMA*. 1997;277:488-494.
3. McGinn TG, Guyatt GH, Wyer PC, et al. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. *JAMA*. 2000;284:79-84.
4. Gilbert EH, Lowenstein SR, Koziol-McLain J, et al. Chart reviews in emergency medicine research: where are the methods? *Ann Emerg Med*. 1996;27:305-308.
5. Cooper RJ. NHAMCS: does it hold up to scrutiny? *Ann Emerg Med*. 2012;60:722-725.
6. Gill MR, Reiley DG, Green SM. Interrater reliability of Glasgow Coma Scale scores in the emergency department. *Ann Emerg Med*. 2004;43:215-223.
7. Day FC, Schriger DL. Journal Club: the measurement of reliability. *Ann Emerg Med*. 2009;54:9-11.
8. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med*. 1993;118:201-210.
9. Snead GR, Wilbur LG. Can the San Francisco Syncope Rule predict short-term serious outcomes in patients presenting with syncope? *Ann Emerg Med*. 2013;62:267-268.
10. Schriger DL, Newman DH. Medical decisionmaking: let's not forget the physician. *Ann Emerg Med*. 2012;59:219-220.
11. Green SM. When do clinical decision rules improve patient care? *Ann Emerg Med*. 2013;62:132-135.
12. Schriger DL, Altman DG. Inadequate post-publication review of medical research. *BMJ*. 2010;341:c3803.