Evaluating Clinical Decision Tools: Can We Optimize Use Before They Turn Us Into Fools?

Stephen M. Schenkel, MD, MPP*; Peter C. Wyer, MD

*Corresponding Author. E-mail: sschenkel@som.umaryland.edu.

0196-0644/$-see front matter

https://doi.org/10.1016/j.annemergmed.2019.04.013

SEE RELATED ARTICLE, P. 60.

[Ann Emerg Med. 2019;74:69-71.]

One of the biggest changes in acute care during the past 20 years has been the generation and application of risk scores and algorithms to guide care: the Pneumonia Severity Index, CURB-65 Score for Pneumonia Severity, Wells criteria, Daniels criteria, pulmonary embolism rule-out criteria, Drug Resistance in Pneumonia, and more. These are accompanied by nationally mandated treatment and time-to-treatment guidelines (sometimes rescinded), a move toward computed tomography diagnosis over radiographs, and ever-increasing acronyms.

Because we are faced with so many tools, it is easy to think all evidence-based medicine is the development of clinical tools. Massive clinical databases combined with data parsing hasten this development, culling relationships drawn from large groups of people and observations. We encounter a new challenge whenever we apply these tools in the clinical setting, an action that requires aptitude in the evaluation of a risk score[1] and facility in on-the-fly integration into care. We think this is best achieved with a tool to aid use of the tools, ideally a simple algorithm or structured approach.

Ilg et al,[2] the authors of the current *Annals* article assessing a pneumonia decision support tool, allow an opportunity to consider our proposed "tool to help assess a tool." The investigative team evaluated a previously developed score, the CURB-65,[3] using a new outcome, critical care interventions, instead of the original outcome the tool targeted, mortality. They chose to recast the usage of a previous tool to parallel how the score was being applied in their own emergency department (ED), seeking to address potential doubts in regard to its ability to distinguish patients who might benefit from hospitalization.

Let's take the current authors' work and ask a few questions to evaluate the clinical decision tool. We seek a clinically useful assessment of clinical decision tools. The following 3 questions are posited in a specific order. Answering any one of them in the negative suggests the tool is not ready for clinical use. They are asked in order of difficulty. The first should be answerable with a quick thought; the second and third, with a reading of the original article.

QUESTION ONE: HOW DOES THE OUTCOME INFORM MY DECISIONMAKING?

This fundamental clinical question requires knowing what the tool predicts and assessing the need for predicting that outcome. In the case of the CURB-65, the original outcome was mortality. The use of the score to then support a decision of outpatient versus inpatient therapy seems a sleight of hand, maybe not what was really needed to best inform a decision. It is not evident that because a patient may die that hospitalization will reduce that likelihood, nor is it evident that a patient likely to live will not benefit from hospital care. To make the translation to wider use in an ED more problematic, every patient in the original derivation of CURB-65 was an inpatient; all potentially received benefit from hospital care. This was precisely the challenge that led Ilg et al to reevaluate CURB-65. Unfortunately, the more proximal outcomes of critical care interventions are also more slippery; central line placement, hospitalization in the ICU, and intubation are all subject to the variability in practice of the physicians involved. This swap from prediction tool to decision tool reveals another intellectual swap. We move rapidly from using the tool to predict the need for intervention to allowing it to make the decision to intervene. Almost without intent, the tool that aids prediction becomes a tool that makes a decision.[4]

QUESTION TWO: IS THE TOOL USEABLE IN PRACTICE?

Any tool that does not make sense within the clinical environment runs the risk of poor adoption, fomenting uncertainty and creating misunderstanding. The most

useful tool should serve simultaneously to align patient care and simplify the work of the clinician. This means that the tool should be applied with ease either as a simple and readily recalled mental calculation or as a reasonably available electronic medical record assessment. The data needed are best routinely collected as part of a normal history and physical examination and without inherent variation in assessing their presence or absence, something that is often a problem with subtler physical findings. The usable tool is understandable and carries face validity, allowing it to garner the trust of the physician and to make sense to a patient when explained. The best tools show improvement in care or efficiency compared with clinician gestalt or the common actions of providers.[5]

The CURB-65 has logical attraction. It relies on basic clinical indicators such as confusion, respiratory rate, age, blood pressure, and blood urea nitrogen (BUN). Even in this there are multiple challenging elements: confusion is not necessarily readily or consistently defined, and the BUN requires a blood test that will be omitted for low-risk patients. The use of *International Classification of Diseases, Ninth Revision* coding to mimic clinical assessment of confusion may be the best the authors could do retrospectively, but it cannot match the assessment of confusion at the bedside, which may or may not have found its way to clinical documentation. The need for a blood test to determine BUN illustrates that a substantial decision was made early in the course of care, before the point at which the tool could be developed or applied: to draw blood. As a result, CURB-65 still requires a blood test and, logically, an assessment that the patient needs blood drawn, and so it may not be as practically useable for *all* patients as might be wished.

This assessment criterion is a high hurdle despite the ability to create new clinical decision tools through artificial intelligence working on massive databases. These frequently appear as black boxes, necessitating an enormous amount of faith in the electronic medical record to produce a tool and a result that have meaning in the absence of an understanding of the data points used and the means of calculation. The obvious catch is that the electronic medical record can make a tool more available while hiding what it is doing. Additionally, such an approach, necessarily based on retrospective review of previous assessments, bakes all previous judgments into the tool, normalizing the foibles and inconsistencies of the clinicians who made the original decisions.

QUESTION THREE: IS THE TOOL VALID IN A POPULATION LIKE MINE?

Creation and demonstration of the usefulness of a tool fall into 2 natural categories: derivation and validation. The rub is in the selection of these populations, and any number of approaches may be taken, including publication of the derivation before any validation, use of the same data set for both processes, derivation and validation in different data sets drawn from the same clinical setting such as one department or one set of hospitals, or derivation and validation in entirely separate populations. Because outcome prediction has been maximized in the population used to create the tool, most tools will prove to be most effective in that initial population; predictive accuracy should naturally decrease in subsequent use.

Prospective validation in a new population, using the same algorithm and examining the same outcomes, provides the best evidence to support adoption of a tool; it is also the most difficult to carry out and therefore the least likely evidence to find. Evaluating the validation often requires examining other published work.

Here the current work of Ilg et al is best seen as a beginning because they have altered the use of CURB-65 by selecting a new outcome. Ironically, the original CURB-65 mortality prediction was not well validated either for its chosen outcome of mortality or for its interpolated usage in predicting safety for discharge from the ED. (An easy way to look for the evidence behind many clinical tools is with a quick trip to http://www.mdcalc.com, where an "evidence" key at the bottom of the page provides a summary listing of references, including a specific section on validation.[6]) Neither CURB-65 nor this newer reassessment of CURB-65 for critical care intervention has this important feature of evidentiary support. The current authors' finding of limited sensitivity and specificity suggests that the effort might not be worthwhile.

One might also ask whether there is another tool that already serves the purpose the authors sought. Here, the Pneumonia Severity Index exists. Validated in a broad population, it suffers from age and a perceived complexity, although online calculators ease this task. It may be the more meaningful predicter of need to hospitalize in a broad population of pneumonia patients.[7]

A TOOL TO ASSESS THE TOOLS?

Our goal is describing an approach that is useful for the clinician and that aligns with recommendations about tools submitted for peer review and publication.[8] More detailed approaches are left to the statistician and the guideline maker. As acronyms proliferate and combine to form guidelines and inform care, we need a way to assess what we, and they, are doing. These 3 questions might help the harried clinician and the patient.

Supervising editor: Donald M. Yealy, MD. Specific detailed information about possible conflict of interest for individual editors is available at https://www.annemergmed.com/editors.

Author affiliations: From the Department of Emergency Medicine, University of Maryland School of Medicine, Baltimore, MD (Schenkel); and Columbia University Medical Center, New York, NY (Wyer).

Authorship: All authors attest to meeting the four ICMJE.org authorship criteria: (1) Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND (2) Drafting the work or revising it critically for important intellectual content; AND (3) Final approval of the version to be published; AND (4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

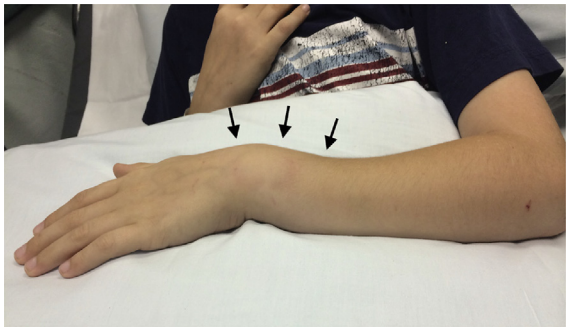
Funding and support: By Annals policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships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

REFERENCES

1. 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;284:79-85.
2. Ilg A, Moskowitz A, Konanki V, et al. Performance of the CURB-65 score in predicting critical care interventions in patients admitted with community-acquired pneumonia. *Ann Emerg Med*. 2019;74:60-68.
3. Lim WS, van der Eerden MM, Laing R, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax*. 2003;58:377-382.
4. O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown Publishing Group; 2016.
5. Schriger DL, Elder JW, Cooper RJ. Structured clinical decision aids are seldom compared with subjective physician judgment, and are seldom superior. *Ann Emerg Med*. 2017;70:338-344.
6. MDCalc. CURB-65 Score for Pneumonia Severity. Available at: https://www.mdcalc.com/curb-65-score-pneumonia-severity#evidence. Accessed November 29, 2018.
7. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*. 1997;336:243-250.
8. Green SM, Schriger DL, Yealy D. Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. *Ann Emerg Med*. 2014;64:286-291.

Images in Emergency Medicine

The *Annals* Web site (*www.annemergmed.com*) contains a collection of hundreds of emergency medicine-related images, complete with brief discussion and diagnosis, in 18 categories. Go to the Images pull-down menu and test your diagnostic skill today. Below is a selection from the Trauma Images.



"Child With Dinner Fork Deformity" by Kardouni, February 2016, Volume 67, #2, pp. 165, 188.