# Structured Clinical Decision Aids Are Seldom Compared With Subjective Physician Judgment, and Are Seldom Superior

David L. Schriger, MD, MPH*; Joshua W. Elder, MD, MPH; Richelle J. Cooper, MD, MSHS

*Corresponding Author. E-mail: schriger@ucla.edu.

**Study objective:** We determine how often studies that evaluate the performance of an aid for decisionmaking, be it a simple laboratory or imaging test or a complex multielement decision instrument, compare the aid's performance to independent, unaided physician judgment.

**Methods:** This was a cross-sectional survey of all Original Research and Brief Research Report articles in *Annals of Emergency Medicine* from 1998 to 2015. We included all articles that evaluated the performance of an aid for decisionmaking in assisting a physician with a decision about testing, treatment, diagnosis, or disposition. Two authors independently characterized the intent and purpose of each aid for decisionmaking, determined whether each study had a comparison to unaided physician judgment within the article or in a separate article, and recorded the result of that comparison.

**Results:** One hundred seventy-one (8.3%) of 2,060 research articles studied the performance characteristics of an aid for decisionmaking, 48 of which were formal clinical decision instruments. Forty of the 171 studies retrospectively analyzed existing databases and therefore could not assess physician judgment. Investigators compared the aid for decisionmaking to physician judgment in 11% (15/131) of the prospective studies, including 15% (6/41) of studies that evaluated a formal clinical decision instrument. For 9 articles that had no comparison to physician judgment, we found 6 unique external publications that compared that aid to physician clinical judgment. The decision aid was superior to clinical judgment in 2 of the 21 studies that contained a comparison.

**Conclusion:** Physician judgment is infrequently assessed when the performance of an aid for decisionmaking is evaluated, and, when reported, the decision aid seldom outperformed physician judgment. [Ann Emerg Med. 2017;70:338-344.]

Please see page 339 for the Editor's Capsule Summary of this article.

Readers: click on the link to go directly to a survey in which you can provide **feedback** to *Annals* on this particular article.
A **podcast** for this article is available at www.annemergmed.com.

0196-0644/$-see front matter

http://dx.doi.org/10.1016/j.annemergmed.2016.12.004

**SEE EDITORIAL, P. 345.**

## INTRODUCTION

### Background

Aids for decisionmaking are so commonplace in emergency medicine that they often define care, frame medical education, and sculpt the lexicon of everyday practice. Young physicians have difficulty imagining emergency practice in a time when they did not exist. Since Stiell et al[1] published the Ottawa Ankle Rules in 1993, there has been a rush to develop rules to advise or assist clinicians on a panoply of decisions. There are even guidelines on how to make and publish decision rules.[2-4] The recent emergency medicine literature is replete with research of the form "can test A predict which patients will have a positive (or negative) finding on test B, or a need for intensive care, or a specific outcome."

### Importance

Many tests and treatments that logically should be helpful are not. For this reason, effectiveness research is conducted to distinguish what benefits patients from what does not. Yet, in general, aids for decisionmaking have not been subjected to the same scrutiny. In our experience, the typical article of this kind describes the diagnostic or discriminatory performance of the decision aid but fails to compare the aid's performance with the independent clinical judgment of an unaided physician. Implicit in this omission is the idea that an unaided physician could not possibly perform as well as a

**Editor's Capsule Summary**

*What is already known on this topic*

Clinical decision rules or other decision aids must be superior to baseline clinical judgment to improve care.

*What question this study addressed*

How often does decision aid research include a performance comparison with clinical judgment?

*What this study adds to our knowledge*

In this analysis of 171 consecutive *Annals* articles evaluating decision aids, only 21 contrasted their performance with judgment, and of these, only 2 found the decision aid superior.

*How this is relevant to clinical practice*

Most clinical decision rules and other decision aids have not been established to improve on baseline clinical judgment and thus appear clinically unhelpful.

diagnostic test or a decision aid.[5,6] This assumption is part of a general trend in medicine in which the appearance of objectivity is valued over subjectivity, even if the "objective" judgments are, on closer examination, subjective themselves.[7-9]

The necessity of subjective judgment is acknowledged by some aids for decisionmaking. At least 3 points (likely alternative diagnosis) and arguably 6 (signs of deep venous thrombosis) of the 12.5 points in the Wells criteria require such judgment, and the pulmonary embolism rule-out criteria (PERC) rule is designed to be used only for patients the clinician subjectively believes have less than a 15% chance of having a pulmonary embolism.[10,11]

### Goals of This investigation

We sought to determine how often studies that evaluate the performance of an aid for decisionmaking, be it a simple laboratory test, imaging test, or complex multielement decision instrument, compare the aid's performance to independent unaided physician judgment. Our experience and some published evidence suggested that physician judgment is seldom measured in such studies.[12] We wanted to quantify the behavior so that future investigators can consider the wisdom of including a comparison to unaided physician judgment in their studies. We also sought to determine the outcome of such comparisons when they did occur and whether the frequency of comparisons with unaided physician judgment has changed.

## MATERIALS AND METHODS

### Study Design

This was a cross-sectional survey of all issues of *Annals of Emergency Medicine* from 1998 to 2015. We wanted to identify all research studies that sought to determine whether a single test (be it a laboratory test, radiology study, historical question, or physical examination finding) or a combination of test results (such as in a clinical decision instrument) could guide physician decisionmaking. Throughout this article, we use the term "aid for decisionmaking" to signify any of the aforementioned activities, reserving the term "formal clinical decision instrument" for multifactorial clinical decision rules. No institutional review board approval was sought because study subjects were published articles. We use "physician judgment" to indicate decisions made without the use of such aids, occasionally using "unaided physician judgment" to stress this point.

We examined the tables of contents of 36 randomly selected 2009 issues of the 6 highest-impact-factor general medical journals and found only 3 potentially eligible articles. After finding similarly small numbers in a sample of specialty journals, we decided to focus exclusively on *Annals*, a journal we knew had many articles of this kind.

Each author independently reviewed the tables of contents of the 36 2009 to 2011 issues of *Annals* to determine which articles might be eligible. Through an iterative process, we developed an algorithm (Appendix E1, available online at http://www.annemergmed.com) for identifying and classifying eligible articles. Two authors then independently read the abstract and, if needed, the text of all research articles (Original Research and Brief Research Reports) whose title met the inclusion criteria in 216 issues (18 years' worth) of *Annals*. We included an article if its goal was to determine whether an aid for decisionmaking could help a physician in making a decision or predicting an outcome. We excluded nonhuman studies, studies focused on care provided by nonphysician medical personnel (eg, out-of-hospital care personnel, nurses), meta-analyses, and studies that evaluated changes in practice resulting from implementation of an aid for decisionmaking rather than the aid's diagnostic test characteristics. Data on each rater's performance were retained so that interrater reliability could be assessed.

### Methods of Measurement

Two of the authors independently reviewed each included article. They first classified each study on whether, according to its design, the investigators could have measured each physician's judgment in regard to the clinical question the aid for decisionmaking was designed to

answer. Studies that used preexisting databases generally could not do so and were distinguished from those in which the physician's judgment could have been elicited during the research process. For example, if a study design required the physician to complete a form about patient characteristics before ordering a test, that form could also have contained the question "Do you think the test result will be positive?," whereas a study that used retrospective chart review methodology could not.

Each rater then classified articles in regard to the kind of help the aid for decisionmaking was intended to provide: directive, providing direct advice (eg, "if none of these are present, do not order a computed tomography [CT] scan"); informative (eg, "the probability of a bad outcome in the next 7 days is very low; therefore, you might send the patient home"); or prognostic, providing prognostic information with no implication for decisionmaking (eg, "individuals with a positive test result have a 3-fold higher risk of stroke in the next 90 days").

For each article, the rater noted whether the test was intended to assist the physician with a decision about the ordering of a test, the ordering of a treatment, the assignment of a diagnosis, or the determination of a disposition and also noted whether the article was evaluating a formal clinical decision instrument (in a derivation or validation study) or a single test (eg, a biomarker).

Finally, for each prospective study each rater determined whether the article included a comparison to unaided physician judgment. For articles that did not, we conducted a literature search on both PubMed and Google Scholar, checking first, second, and last author names independently, title words (eg, selection of patients for pulmonary CT angiogram), and key concepts (eg, Ottawa Ankle Rules) in an attempt to find any articles that compared the article's decision aid with physician judgment. We jointly reviewed all candidate articles identified by the search.

For articles that contained a comparison with physician judgment, whether in the original article or in one discovered through the literature search process, we determined whether the evidence favored the aid for decisionmaking or physician judgment. This was done by consensus. We accepted the investigators' determination unless there was compelling evidence to dispute their interpretation of the data.

We noted the percentage of agreement between raters, and all authors jointly adjudicated discrepancies. During initial scoring of the 2009 to 2011 articles, we used discrepancies to modify our scoring manual to improve interrater reliability.

### Outcome Measures

The primary outcome measure was whether the study included an assessment of unaided physician judgment. The secondary outcome was whether such comparisons favored physician judgment or the aid for decisionmaking.

### Primary Data Analysis

Our analysis is purely descriptive. We report how often studies assessed unaided physician judgment overall and stratified on the aforementioned study characteristics. Stata (version 14.0; StataCorp, College Station, TX) was used for data management and analysis.

## RESULTS

Of 2,060 research articles in 1998 to 2015 issues of *Annals,* 442 had titles that met our screening criteria and 171 of these were eligible, including 48 that evaluated formal clinical decision instruments (Figure 1 and Figure E1, available online at http://www.annemergmed.com). The 2 authors who evaluated each journal issue disagreed on whether an article should be included on 56 occasions (3%). There was perfect agreement on our primary and secondary outcome measures, but initially there was considerable disagreement in coding whether the test being evaluated was directive, informative, or prognostic, which was reduced with the development and refinement of the coding algorithm (0% discrepancies when 2009 to 2011 data were recoded and 7% [1/15] in a second interrater assessment of 2013 data).

Fifteen of the 171 studies (9%) had an unaided physician judgment arm (Figures 1 and 2). However, for the 40 articles (23%) that used retrospective techniques, authors had no opportunity to introduce a physician judgment arm. Excluding these articles, 15 of 131 (11%) had an unaided physician judgment arm. Comparisons with physician judgment were present in 10 of 75 (13%) directive studies, 4 of 28 (14%) informative studies, and 1 of 28 (4%) prognostic studies.

For 9 prospective studies that did not compare the aid for decisionmaking with physician judgment, we found such a comparison in a separate publication (Figure 2). Four of these 9 articles were on the San Francisco Syncope Rule and all were given credit for a single article that contained a comparison on this topic[13]; the 5 other external articles involved the Ottawa Ankle Rules, the Manchester Self-Harm Rule, 2 related instruments for predicting injury from blunt trauma in children, and a neural network for identifying chest pain of cardiac origin. In 6 of the 9 articles for which we found a comparison in an external article, the index article in *Annals* and the

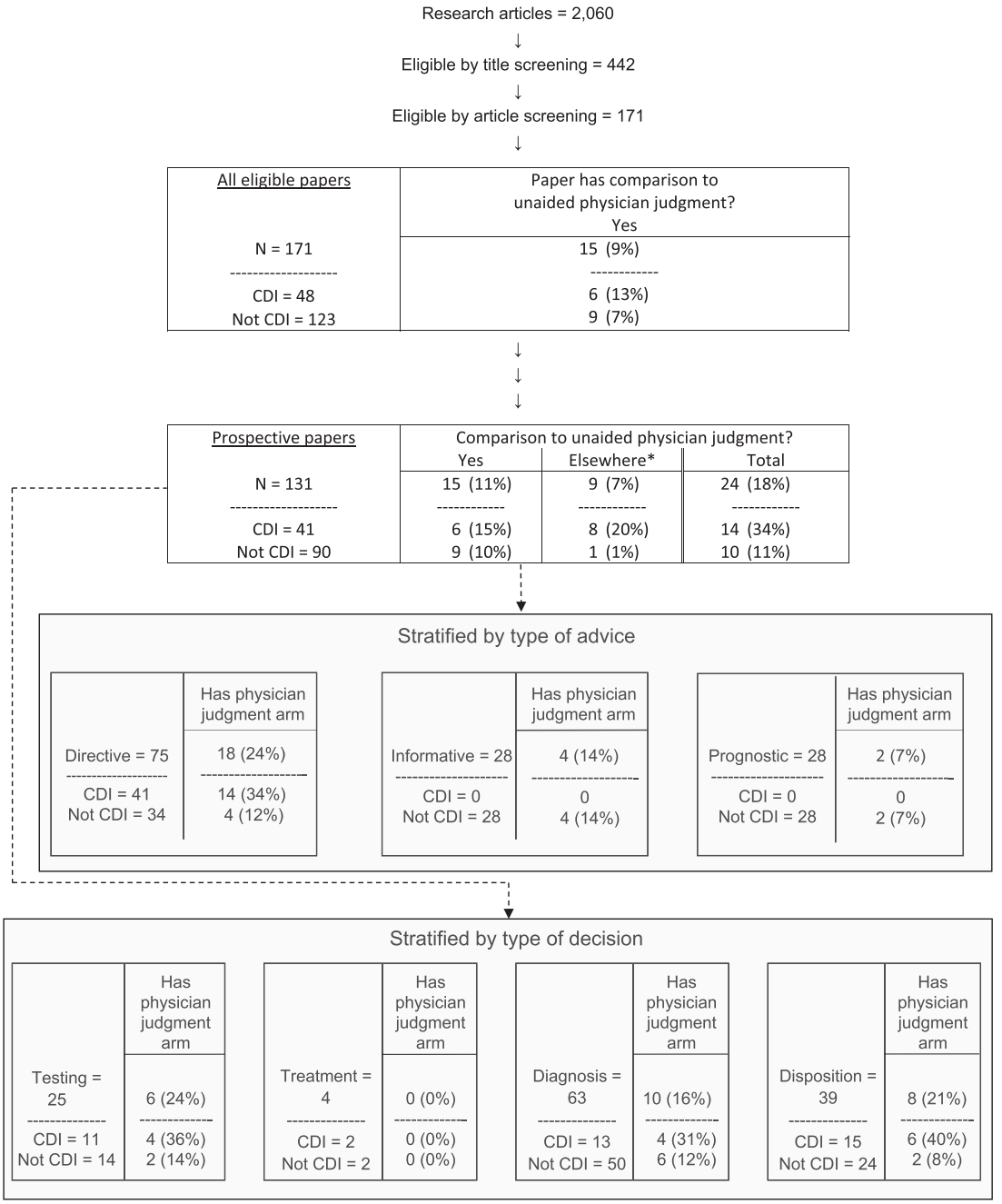


**Figure 1.** Study flow diagram with results. This figure depicts how we identified and selected the 131 prospective evaluations of aids for decisionmaking in the 2,060 research articles in 1998 to 2015 *Annals* articles. It also depicts what percentage of these articles compared the aid's performance with independent physician judgment overall and stratified by when the aid to decisionmaking was a formal clinical decision instrument or not. Finally, it examines these data stratified on the intent of the rule (top shaded box) and the type of decision being aided (bottom shaded box). *CDI*, Clinical decision instrument. *We found 6 unique articles that contained a comparison with physician judgment and were relevant to 9 articles in our sample that did not contain such a comparison.

article that contained the comparison of the aid for decisionmaking with physician judgment were by the same group of authors. In 2 instances, the external comparison was published before the *Annals* article was published; in the other 7, the comparison article was published between 1 and 6 years later.

In total, there were 15 articles that had an internal comparison of the aid for decisionmaking with physician judgment and 6 external articles that did so for 9 of the 171 *Annals* articles. Of the 21 unique articles with a comparison with an aid for decisionmaking, physician judgment was found superior in 6 (29%), results were tied

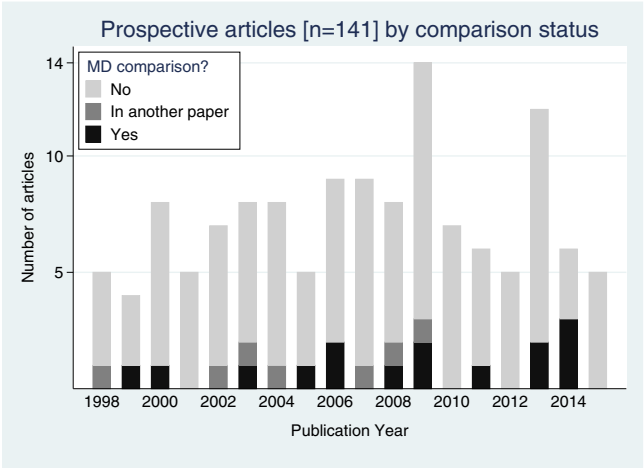


**Figure 2.** For each year, the graph depicts the total number of studies about decision aids and indicates how often the aid was compared with physician judgment either in the article (black bars) or in "another article" (light gray bars). There is no obvious trend over time.

or mixed in 10 (46%) (eg, sensitivity better with one test but specificity better with the comparator), the decision aid was superior in 2 (10%), and it was impossible to tell or not applicable in 3 (15%) (Table and Appendix E2 [available online at http://www.annemergmed.com]). Articles were deemed not applicable when physician judgment was compared with a criterion standard directly rather than with a specific decision aid. For example, Chinnock et al[14] investigated whether physicians could identify patients with spontaneous bacterial peritonitis but did not attempt to establish whether a single laboratory test or combination of laboratory tests was a better predictor of positive culture result. The 2 instances in which the aid was superior were a neural network for chest pain and an aid for decisionmaking for obtaining cervical spine radiographs, with only the latter reported in the same article.[15,16]

## LIMITATIONS

Our classification taxonomy and algorithm for sorting articles into that taxonomy are new and have not been formally validated. We have no doubt that if we repeated the classification effort, results would vary slightly. We are confident, however, that discrepancies would be insufficiently large to alter conclusions. It is possible that articles published in 2016 and beyond will contain comparisons to physician judgment relevant to articles in our database.

We studied articles in a single journal and results may not apply to other journals. There were only 41 prospective evaluations of clinical decision instruments, so the 95% confidence interval around our 34% estimate of the inclusion of a physician arm in such studies is wide (20% to 50%). However, even if the true value is closer to the upper limit of this confidence interval, the conclusion that a minority of studies of decision instruments compare the rule's performance to unaided physician judgment holds.

## DISCUSSION

Only 11% (15/131) of articles in *Annals* that prospectively evaluated the test characteristics of an aid for decisionmaking compared the aid's performance with unaided physician judgment in the same article, with the percentage increasing to 18% (24/131) when we included outside comparisons. Furthermore, only 2 of the 21 articles that did so found the aid for decisionmaking superior. These are important findings that should guide research on decision instruments and all aids for decisionmaking. The first result shows that decision instruments are typically not tested against physician judgment, and the second shows that the assumption needed to justify such behavior—that almost all aids for decisionmaking outperform physician judgment—is not true. Just as we should not introduce a new medical treatment until there is evidence from well-designed studies that it outperforms current therapy so also we should not advocate clinical decision aids (whether they are a laboratory test or a formal clinical decision instrument) until they are proven superior to physician judgment.

Almost half (46%) of the 21 studies that compared an aid with physician judgment had mixed or inconclusive results. This was often due to its being unclear whether the aid for decisionmaking and clinician judgment were performing differently or were just calibrated differently. For example, many aids for decisionmaking are designed to improve specificity (order fewer radiographs that lead to negative results) while maintaining sensitivity (do not fail to radiograph patients whose radiograph results would be abnormal). When the aid's specificity was higher but sensitivity was lower, it was often impossible to tell whether the differences represented different points on the same performance curve or different performance curves.

Our results are supported by a recent systematic review of aids for ordering diagnostic tests. In a 2015 article in *PLoS One*, Sanders et al[12] used several standard databases to search the medical literature from inception to 2011 and found only 31 studies of 13 medical conditions that conducted a comparison with unaided physician judgment. They found that "the limited studies included in this review show that none of the CPRs [clinical prediction rules] evaluated to date are clearly superior to clinical judgment…."

**Table.** List of articles that had internal or external comparisons with independent physician judgment.

| Reference | First Author | Year | Topic | Purpose | Type | Comparison Favors | | | | Comparison Article (Reference) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Decision Aid | Physician Judgment | Toss-up | Not Applicable | |
| **Comparison within article** | | | | | | | | | | |
| 1 | Chinnock | 2008 | Subacute bacterial peritonitis | Diagnosis | Directive | | | | 1 | |
| 2 | Kline | 2009 | Low-risk chest pain | Disposition | Prognostic | | | | 1 | |
| 3 | Mitchell | 2006 | Acute coronary syndrome | Disposition | Informative | | | 1 | | |
| 4 | Gupta | 2011 | CT in trauma | Test | Directive | | | | 1 | |
| 5 | Tung | 2006 | BNP for heart failure | Diagnosis | Informative | | | 1 | | |
| 6 | Hendey | 2000 | Radiograph in shoulder dislocation | Test | Directive | | 1 | | | |
| 7 | Kline | 2014 | Acute coronary syndrome | Diagnosis | Informative | | | 1 | | |
| 8 | Collins | 2009 | Heart sound S3 in dyspnea | Diagnosis | Informative | | 1 | | | |
| 9 | Morris | 1999 | Chest syndrome in sickle cell | Diagnosis | Directive | | 1 | | | |
| 10 | Stein | 2005 | Testing for influenza | Diagnosis | Directive | | 1 | | | |
| 11 | Meltzer | 2013 | Alvarado score in appendicitis | Diagnosis | Directive | | 1 | | | |
| 12 | Bandiera | 2003 | Canadian C-Spine Rule | Test | Directive | 1 | | | | |
| 13 | Easter | 2014 | Rules in pediatric head injury | Test | Directive | | | 1 | | |
| 14 | Nishijima | 2014 | Mild traumatic intracranial bleeding event | Disposition | Directive | | | 1 | | |
| 15 | Penaloza | 2013 | Rules for pulmonary embolism | Diagnosis | Directive | | 1 | | | |
| **Comparison in external article** | | | | | | | | | | |
| 16 | Baxt | 2002 | Neural network for chest pain | Diagnosis | Prognostic | 1 | | | | Baxt (25) |
| 17 | Cooper | 2006 | Risk of self-harm | Disposition | Directive | | | 1 | | Cooper (26) |
| 18 | Sun | 2007 | San Francisco Syncope Rule | Disposition | Directive | | | 1 | | Quinn (27) |
| 19 | Auleley | 1998 | Ottawa Ankle Rules | Diagnosis | Directive | | | 1 | | Glas (28) |
| 20 | Birnbaum | 2008 | San Francisco Syncope Rule | Disposition | Directive | | | 1* | | Quinn (27) |
| 21 | Quinn | 2006 | San Francisco Syncope Rule | Disposition | Directive | | | 1* | | Quinn (27) |
| 22 | Quinn | 2004 | San Francisco Syncope Rule | Disposition | Directive | | | 1* | | Quinn (27) |
| 23 | Holmes | 2009 | Pediatric blunt trauma | Test | Directive | | | 1 | | Mahajan (29) |
| 24 | Palchak | 2003 | Pediatric blunt head injury | Test | Directive | | | 1 | | Palchak (30) |
| | | | | | | 2 | 6 | 10 | 3 | |

*BNP*, *β*-Natriuretic peptide.

*Not counted in the total because they refer to an external article that has already been counted. For definitions of entries in columns "Purpose" and "Type," see text. The "comparison favors" section is generally based on the article's conclusion unless results strongly contradict that conclusion. In general, when sensitivity went up with the decision aid and specificity went down (or vice versa), we considered that a toss-up because it was impossible to tell whether this was due to different points on the same performance curve or different performance curves. Alternatively, if sensitivity and specificity were both higher for one method, that method was approved as long as the magnitude of the improvement was clinically important. Otherwise, it too was called a toss-up. Our goal was not to make an absolute tally of what was better but to show that the perspective "we do not need to compare new decision aids with physician judgment because the overwhelming evidence shows that the decision aid is always superior" is unfounded. References to articles and comparison articles are numbered here and links are provided in Appendix E2, available online at http://www.annemergmed.com.

Gallagher[7] reported there are occasions in which aids for decisionmaking prove superior to physician judgment. However, this does not mean that unaided physician assessment is always less objective or inferior to a clinical decision aid. Wears[8] argued that the "…proliferation of decision rules, the desire for guidelines, the quest for standardization and the aversion to variation or heterogeneity, the faith in 'evidence-based medicine,' the yearning for quantitative measurement, the fascination with templates and checklists, and the magical thinking about information technology" are all part of creating order and attempting to rationalize clinical practice. The truth likely lies somewhere in between: some aids for decisionmaking outperform physician judgment and others do not.

The recently published 22-item Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis reporting guideline focuses on technical aspects of model development and does not consider whether comparison with physician judgment is desirable.[17] Further studies should be directed at understanding how to accurately assess physician judgment and how to assess the combination of an aid for decisionmaking with physician judgment.

In summary, we found that articles that report on the performance of aids for physician decisionmaking seldom compare the aid with clinical judgment, and the few that did failed to demonstrate that the aids are consistently superior.

*The authors acknowledge Cheri-Ann Parris, BA, and Felix German Contreras-Castro, AA, AS, for assisting with the literature searches to find outside comparative articles.*

*Supervising editor:* Steven M. Green, MD

*Author affiliations:* From the Department of Emergency Medicine, University of California, Los Angeles, CA (Schriger, Cooper); and the Robert Wood Johnson Clinical Scholars Program, Yale University School of Medicine, New Haven, CT (Elder).

*Author contributions:* DLS conceived the study. All authors participated in the study design, development of study protocol, and data abstraction and analysis. DLS and JWE drafted parts of the article, and all authors participated in the revision process. DLS takes responsibility for the paper as a whole.

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). Dr. Schriger was funded in part by an unrestricted grant from the Korein Foundation. Drs. Schriger and Cooper receive monthly stipends for the editorial services to *Annals of Emergency Medicine*.

*Publication dates:* Received for publication September 22, 2016. Revisions received November 15, 2016, and November 22, 2016. Accepted for publication November 29, 2016.

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

## REFERENCES

1. Stiell IG, Greenberg GH, McKnight RD, et al. Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. *JAMA*. 1993;269:1127-1132.
2. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med*. 1999;33:437-447.
3. Wasson JH, Sox HC, Neff RK, et al. Clinical prediction rules. Applications and methodological standards. *N Engl J Med*. 1985;313:793-799.
4. Green SM, Schriger DL, Yealy DM. Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. *Ann Emerg Med*. 2014;64:286-291.
5. Grove WM, Zald DH, Lebow BS, et al. Clinical versus mechanical prediction: a meta-analysis. *Psychol Assess*. 2000;12:19-30.
6. Schriger DL. Some thoughts on the stability of decision rules. *Ann Emerg Med*. 2007;49:333-334.
7. Gallagher EJ. The intrinsic fallibility of clinical judgment. *Ann Emerg Med*. 2003;42:403-404.
8. Wears RL. Lessons from the Glasgow Coma Scale. *Ann Emerg Med*. 2012;59:338.
9. Schriger DL, Newman DH. Medical decisionmaking: let's not forget the physician. *Ann Emerg Med*. 2012;59:219-220.
10. Wells PS, Ginsberg JS, Anderson DR, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Ann Intern Med*. 1998;129:997-1005.
11. Kline JA, Mitchell AM, Kabrhel C, et al. Clinical criteria to prevent unnecessary diagnostic testing in emergency department patients with suspected pulmonary embolism. *J Thromb Haemost*. 2004;2:1247-1255.
12. Sanders S, Doust J, Glasziou P. A systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment. *PLoS One*. 2015;10:e0128233.
13. Quinn JV, Stiell IG, McDermott DA, et al. The San Francisco Syncope Rule vs physician judgment and decision making. *Am J Emerg Med*. 2005;23:782-786.
14. Chinnock B, Afarian H, Minnigan H, et al. Physician clinical impression does not rule out spontaneous bacterial peritonitis in patients undergoing emergency department paracentesis. *Ann Emerg Med*. 2008;52:268-273.
15. Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet*. 1996;347:12-15.
16. Bandiera G, Stiell IG, Wells GA, et al. The Canadian C-Spine Rule performs better than unstructured physician judgment. *Ann Emerg Med*. 2003;42:395-402.
17. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162:W1-W73.

## APPENDIX E1

### Coding rules

Process for determining eligibility and scoring articles

1) Study the title of each Original Contribution and Brief Research Report.

If the title contains:

- the name of a clinical decision instrument (eg, NEXUS, Canadian C-Spine Rule, PERC, Well's, CURB-65, Centor, San Francisco Syncope Rule)
- a radiology or laboratory study
- a statistical technique used in decision research, eg, "neural networks"
- a key word: "predicting," "validation," "risk," "scoring," "discharge," "outcomes," "prognostic," "death," "diagnosis," "hospitalization," "accuracy," "sensitivity," "specificity," "physician judgment," "physician impression," "incidence," "confirmation," "detect"

then go to 2). If not, move to the next article in that issue.

2) Review the abstract and (if needed) article and ask:

Does this study attempt to use a biomarker, radiology study, clinical decision instrument, or any other test to direct patient care or predict outcomes?

If yes, ask:

Would it have been desirable to have this study compare the aid's performance with unaided physician judgment?

If the answer to both of the above questions is yes, go to 3). If not, do not include the article. If maybe, flag the article for discussion among authors.

3) Retrospective vs prospective

Given the study design, could the authors have measured unaided physician judgment?

If yes, code as "prospective." If no, code as "retrospective." Go to 4).

4) Code the aid as directive, informative, or prognostic:

a) Directive intent.

Does the decision aid provide specific advice about a future action (eg, "If the rule is negative, then do not order a CT," "If the B-HCG is >1,500, then order ultrasonography")?

If yes, code as "directive."

If no, go to b.

b) Does the decision aid provide information about prognosis at a time ≤31 days from the evaluation (eg, "the probability of death in the next 7 days is <0.001%)?

If no, code as "prognostic."

If yes, go to c.

c) Is there a direct link from the prognostic information to a clinical action (eg, "Because the probability of death in the next 7 days is low, discharge the patient from the ED")?

If no, code as "prognostic."

If yes, code as "informative."

5) Code each article according to the decision that the aid attempts to help. Choices are:

- Testing: help with decision to perform/not perform diagnostic tests
- Therapeutic: help with decision to use/not use a treatment
- Disposition: help with decision about whether to admit or discharge patient or where to admit patient
- Diagnosis: help with determining what diagnosis the patient has received

6) For all prospective studies, code whether there is comparison to unaided physician judgment.

7) If 6) is yes, determine which performed better, the aid or physician judgment. In general, defer to the article authors' determination unless there is compelling evidence that that determination is erroneous. Choices are gestalt, the aid, mixed results/inconclusive results (a wash), impossible to tell/not applicable (eg, no criterion standard).

Protocol for finding outside comparison studies:

1) For prospective studies that do not have a comparison to unaided physician judgment in the article:

a) Perform a PubMed search on the first author. If too many results are obtained, add key words based on the MeSH terms of the original article. Scan this output for articles that could contain a comparison of the decision aid to physician judgment. If one is found, stop. If not, go through the following steps until one is found or options are exhausted.

b) Repeat this process for the second author.

c) Repeat the process for the last author.

d) Repeat the process with key title words.

e) Repeat the process with Google Scholar.

## APPENDIX E2

**References for Appendix E1, available online at http://www.annemergmed.com.**

1) Physician clinical impression does not rule out spontaneous bacterial peritonitis in patients undergoing emergency department paracentesis

2) Randomized trial of computerized quantitative pretest probability in low-risk chest pain patients: effect on safety and resource use

3) Prospective multicenter study of quantitative pretest probability assessment to exclude acute coronary syndrome for patients evaluated in emergency department chest pain units

4) Selective use of computed tomography compared with routine whole body imaging in patients with blunt trauma

5) Amino-terminal pro-brain natriuretic peptide for the diagnosis of acute heart failure in patients with previous obstructive airway disease

6) Necessity of radiographs in the emergency department management of shoulder dislocations

7) Clinician gestalt estimate of pretest probability for acute coronary syndrome and pulmonary embolism in patients with chest pain and dyspnea

8) S3 detection as a diagnostic and prognostic aid in emergency department patients with acute dyspnea

9) Clinician assessment for acute chest syndrome in febrile patients with sickle cell disease: is it accurate enough?

10) Performance characteristics of clinical diagnosis, a clinical decision rule, and a rapid influenza test in the detection of influenza infection in a community sample of adults

11) Poor sensitivity of a modified Alvarado score in adults with suspected appendicitis

12) The Canadian C-Spine Rule performs better than unstructured physician judgment

13) Comparison of PECARN, CATCH, and CHALICE rules for children with minor head injury: a prospective cohort study

14) Derivation of a clinical decision instrument to identify adult patients with mild traumatic intracranial hemorrhage at low risk for requiring ICU admission

15) Comparison of the unstructured clinician gestalt, the Wells score, and the revised Geneva score to estimate pretest probability for suspected pulmonary embolism

16) A neural computational aid to the diagnosis of acute myocardial infarction

17) A clinical tool for assessing risk after self-harm

18) External validation of the San Francisco Syncope Rule

19) Validation of the Ottawa Ankle Rules in France: a study in the surgical emergency department of a teaching hospital

20) Failure to validate the San Francisco Syncope Rule in an independent emergency department population

21) Prospective validation of the San Francisco Syncope Rule to predict patients with serious outcomes

22) Derivation of the San Francisco Syncope Rule to predict patients with short-term serious outcomes

23) Validation of a prediction rule for the identification of children with intra-abdominal injuries after blunt torso trauma

24) A decision rule for identifying children at low risk for brain injuries after blunt head trauma

25) Use of an artificial neural network for the diagnosis of myocardial infarction

26) A comparison between clinicians' assessment and the Manchester Self-Harm Rule: a cohort study

27) The San Francisco Syncope Rule vs physician judgment and decision making

28) Comparison of diagnostic decision rules and structured data collection in assessment of acute ankle injury.

29) Comparison of clinician suspicion versus a clinical prediction rule in identifying children at risk for intra-abdominal injuries after blunt torso trauma

30) Clinician judgment versus a decision rule for identifying children at risk of traumatic brain injury on computed tomography after blunt head trauma

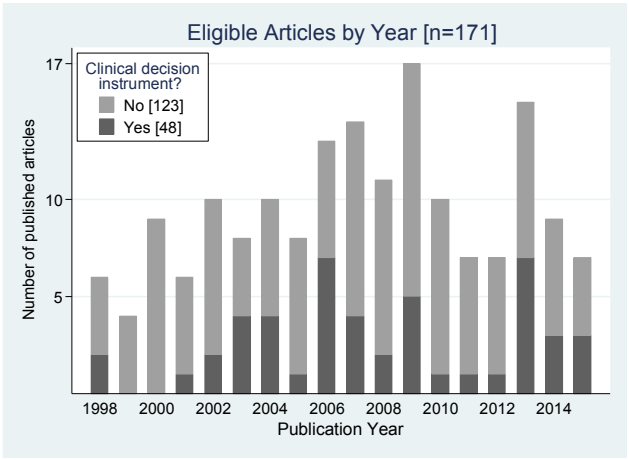


**Figure E1.** This graph depicts the number of research articles in each year's *Annals* issues that assessed the performance of any decision aid (heights of lighter bars) or, specifically, a clinical decision instrument (darker bars). There is no evidence of a trend over time.