A Methodological Appraisal of the HEART Score and Its Variants

Steven M. Green, MD*; David L. Schriger, MD, MPH

*Corresponding Author. E-mail: Steve@SteveGreenMD.com.

We performed a methodological appraisal of the history, electrocardiogram, age, risk factors, and troponin (HEART) score and its variants in the context of *Annals of Emergency Medicine*'s methodological standards for clinical decision rules. We note that this chest pain risk stratification tool was not formally derived, omits sex and other known predictors, has weak interrater reliability, and its 0, 1, and 2 score weightings do not align with their known predictivities. Its summary performance (pooled sensitivities of 96% to 97% with lower confidence interval bounds of 93% to 94%) is below that which emergency physicians state a willingness to accept, below the 98% sensitivity exhibited by baseline practice without the score, and below the 1% to 2% acceptable miss threshold specified by the American College of Emergency Physicians chest pain policy. Two variants (HEART Pathway, HEART-2) have the same inherent structural limitations and demonstrate slightly better but still suboptimal sensitivity. Although a simple prediction tool for chest pain outcomes is appealing, we believe that the widespread use of the HEART score and its variants should be reconsidered. [Ann Emerg Med. 2021;78:253-266.]

Continuing Medical Education exam for this article is available at http://www.acep.org/ACEPeCME/.

0196-0644/$-see front matter

https://doi.org/10.1016/j.annemergmed.2021.02.007

INTRODUCTION

The history, electrocardiogram (ECG), age, risk factors, and troponin (HEART) score was created in 2008 as a tool to risk-stratify emergency department (ED) patients with chest pain.[1] This clinical decision rule (Figure 1) rates 5 items using the mnemonic "HEART" into 0, 1, or 2 points each, and the summed score is used to classify patients into broad categories of low (score 0 to 3), intermediate (score 4 to 6), and high (score 7 to 10) cardiac risk. Numerous follow-up studies confirmed the general association of an ascending score with more frequent adverse cardiac outcomes, and in particular address the safety and potential cost-efficacy of discharging low-risk (score$\leq$3) patients without ED observation or further cardiac testing.[2-5]

Although clinical decision rules such as the HEART score have gained in popularity over the past 20 years, there is almost no evidence in support of their superiority over clinical gestalt, ie, that they improve care above and beyond baseline practice without the rules.[6,7] Although such evidence would be the optimal endorsement of their use, in the absence of such evidence, it seems prudent to appraise the rigor of each rule's development, methodology, and structure. One such metric is the *Annals of Emergency Medicine* methodological standards for decision rules, a document created by gathering the evidence for best practices in decision rule development.[7]

In this article, we critically evaluate the HEART score and its variants in the context of *Annals'* methodological standards, and discuss how such factors might reasonably impact their appraisal, interpretation, and clinical application.[7] The question headers in the article following mirror the topic order used in the *Annals'* methodological standards document, with items of general compliance discussed in the Appendix E1 (available at http://www.annemergmed.com) and items of substantial noncompliance discussed in the main text. We summarize our results in the Figure 2.

SEARCH STRATEGY

A medical librarian searched PubMed, Web of Science, Cochrane Database, and EMBASE for any occurrence of the phrase "HEART score" from 2008 (the year the HEART score was first described) through 2020, limiting to articles in the English language and where possible to human subjects. The librarian removed duplicates according to the method of Bramer et al,[8] yielding 657 results.

We reviewed the titles and abstracts from this search and omitted 265 abstracts, 52 veterinary studies, 11 trial registrations, 1 erratum, and 189 publications clearly unrelated to our objective. We performed a full-text review of the remaining 139 articles and searched the references of the most pertinent of these for additional relevant papers.

Item		Score
History	Slightly suspicious	0
	Moderately suspicious	1
	Highly suspicious	2
ECG	Normal	0
	Nonspecific repolarization disturbance	1
	Significant ST depression	2
Age	<45 years	0
	45-65 years	1
	≥65 years	2
Risk factors	No risk factors known	
	1 or 2 risk factors	1
	≥3 risk factors or history of atherosclerotic disease	2
Troponin	≤ normal limit	0
	1-3x normal limit	1
	>3x normal limit	2

Total score:

0-3 points	Low risk
4-6 points	Intermediate risk
7-10 points	High Risk

Figure 1. HEART score. *ECG*, electrocardiograph.

DEFINITIONS

We herein refer to the "HEART score" in its common 10-point configuration (Figure 1), and note variant modifications that have been described:

- The "HEARTS3 score" is the HEART score plus sex plus a repeat ECG and troponin measurement 2 hours after the first.[9]
- The "HEAR score" (ie, HEART without the troponin) is the HEART score omitting the "T" component, ie, excluding any consideration of troponin measurement.[10,11]
- The "HEART Pathway" refers to the HEART score plus the added requirement of a second troponin measurement 3 hours after the first.[12]
- The "HEART-2 score" is the HEART score in which at least 2 troponin measurements were considered regardless of their timing.[13]
- The "modified low-risk HEART score" refers to a HEART score of 0 to 3 when the troponin scoring element is 0.[2]

In our methodological commentary following, we use the terms "HEART score" and "HEART" to concurrently refer to the original score and its variants when the issue at hand pertains similarly to all versions. When one or more of the variants differ and the distinction is not intuitive, we have highlighted the applicable differences.

IS THE METHODOLOGY OF THE RULE SOUND?

How Were Patients Selected? (Suboptimal compliance)

An optimal clinical decision rule is derived and tested in the population to which it is intended to apply, ie, those for whom there is uncertainty about their diagnosis. We do not need a decision rule to identify patients obviously at minimal risk or patients whose ECG is diagnostic for acute ischemia or infarction.

The original HEART score reports by Backus et al[1,14-17] studied undifferentiated chest pain, ie, all patients with chest pain while excluding only ST-segment elevation myocardial infarctions. Most subsequent studies have similarly included patients with any chest pain, or any chest pain when at least one troponin measurement was obtained.[2-5] This yields a broad chest pain population for whom a standard clinical evaluation will readily identify a substantial subset of patients whose pain is almost certainly noncardiac (eg, of traumatic, gastroesophageal, or musculoskeletal origin), and another substantial subset whose pain is almost certainly cardiac (eg, typical history, ischemic ECG). Such anchoring of both ends of the HEART score with outliers—meaning patients who aren't in the target population for the test—dilutes the intended sample for whom it is a clinical challenge to distinguish a cardiac versus noncardiac etiology. Decision rules tested on samples weighted with such outlier cases might reasonably demonstrate deteriorated performance when applied to the intended, non-straightforward subset. In some EDs, a troponin measurement is reflexively ordered at triage for most or all chest pain prior to physician evaluation, weakening this laboratory test as a proxy for clinical suspicion.

Does the Outcome Matter? (Suboptimal compliance)

What is it that we want the HEART score to predict? The HEART score most commonly uses the traditional cardiology research outcome of "major adverse cardiac events" within 6 weeks of the index visit—defined as any combination of acute myocardial infarction, percutaneous coronary intervention, coronary artery bypass graft surgery, death, and in some studies also significant coronary stenosis with conservative treatment.[2-5] This outcome is cardiologist-centric in that it is crafted to identify patients

Standards	*Compliant / Noncompliant*	*Comment*
Derivation methodology		
What is the population and setting?	Compliant	
How were patients selected?	Suboptimal compliance	All chest pain patients rather than just those with diagnostic uncertainty
Does the outcome matter?	Suboptimal compliance	Major cardiac adverse events at 6 weeks rather than does the patient require admission or consultation
Was the outcome routinely assessed?	Compliant	
Were all potentially important predictor variables included?	Noncompliant	Does not include specific predictive historical features, sex, or timeliness of follow-up; incorporation bias from the inclusion of troponin and ECG
Are the predictor variables objective or collected prospectively?	Originally noncompliant	Subjective elements derived from chart review
Were the predictor variables recorded before knowledge of the outcome?	Compliant	
Are the potential predictor variables reliable?	Noncompliant	Risk factors, ECG, and particularly history demonstrate frequently poor and highly variable interrater reliability.
How were the predictor variables coded?	Noncompliant	Variable thresholds arbitrary rather than derived
Was the derivation sample large enough?	Noncompliant	Derived with insufficient sample size
Is the analytic technique appropriate?	Noncompliant	No quantitative analytic technique used
Was the goal of the rule explicitly specified a priori?	Noncompliant	No quantitative threshold for success pre-specified
Is the need for external validation before clinical application acknowledged?	Noncompliant	Not specified
Can the final rule itself be reliably assessed?	Noncompliant	Frequently poor and highly variable interrater reliability
Can the rule be refined?	Compliant	
Reporting		
Are sensitivity and specificity both emphasized?	Originally noncompliant	Neither reported
Does the rule improve on baseline clinical practice?	Noncompliant	No clear evidence that it improves upon unstructured clinical judgment
Is the rule's performance sufficiently precise?	Noncompliant	Heterogenous research with suboptimal performance
How is the rule interpreted?		
Is the rule successful?	See discussion	
Is the rule sensible?	Compliant	
Is the rule easy to apply?	Compliant	
Does the rule provide a 2-way course of action?	Suboptimal compliance	1-way in that it specifies what to do if low risk but omits specific guidance for the various higher score levels
Postvalidation research		
Can the rule be successfully implemented into clinical practice?	Compliant	
Is the rule cost-effective?	Compliant	
How can the rule be widely integrated into practice?	Compliant	

Figure 2. Compliance of the HEART score with *Annals'* methodological standards for clinical decision rules in emergency medicine. Each item is discussed in the text.

who, in the coming weeks, would likely benefit from percutaneous coronary intervention or anti-ischemic pharmaceutical therapies.

Are major adverse cardiac events what we most care about in the ED? In contrast to cardiologists, the typical focus of emergency physicians is shorter-term, ie, to identify patients who require same-day admission or urgent consultation. Although we of course value the extended health of each patient, our assigned task and most important metric is the interim health and safety of our patient between the ED visit and their next physician evaluation. Unfortunately, the HEART score was not

designed for, and has not yet been tested for, this critical bridge to meaningful follow-up. Nevertheless, some emergency physicians may regard 6-week major adverse cardiac events as a reasonable general proxy for whether or not a patient can be discharged without consultation. One HEART study used a somewhat more ED-centric endpoint: major adverse cardiac events at 15 days.[18]

Neither HEART nor major adverse cardiac events account for disparities in follow-up. In some settings, it is regularly possible to arrange cardiology or primary care clinic visits within 1 to 2 days—as well as prompt outpatient objective cardiac testing—whereas for other settings or patients a prompt, effective reevaluation is at best uncertain. Janes et al[19] found that 32% of their discharged, low-risk patients did not receive follow-up within 6 weeks—and this was in a military system with free care and a dedicated HEART follow-up clinic. Such follow-up is almost certainly worse without free care or special clinics. Multiple studies reported physician noncompliance with HEART directives related to concerns about inadequate follow.[10,17,20] The ideal ED chest pain clinical decision rule would specifically predict short-term safety for discharge and should account for differences in timeliness, likelihood, and quality of meaningful physician follow-up.

Were All Potentially Important Predictor Variables Included? (Noncompliant)

To avoid missing valid predictors, the creators of clinical decision rules should optimally derive them only after scrutiny of all relevant factors with a biologically plausible association with the outcome, such as demographics, history, signs, symptoms, and testing. The creators of the HEART score did not do this, but rather personally selected 5 items based on their intent to mirror the Apgar score with the HEART mnemonic.[1]

What potentially important variables might the HEART score have overlooked? Multiple specific historical features predict higher or lower cardiac risk, such as radiation to both arms (positive likelihood ratio LR+ 2.6), pain similar to prior ischemia (LR+ 2.2), change in pattern over the past 24 hours (LR+ 2.0), and pain reproducible on chest wall palpation (LR+ 0.28).[21] The regression-derived Emergency Department Assessment of Chest Pain score, for example, included other independent predictors—both positive (diaphoresis, radiation, sex) and negative (pain worse with inspiration, pain reproduced on palpation).[22,23]

HEART omits consideration of patient sex, despite men being at greater cardiac risk[9,21,22,24,25] and despite women with cardiac ischemia less frequently reporting chest pain as a symptom.[26] Indeed, men have a higher prevalence of major adverse cardiac events across all HEART score categories relative to women, with the overall male risk approximately twice as high.[24,25] Women are more likely to be classified as low-risk by HEART.[25,27]

HEART also omits consideration of the promptness and quality of outpatient follow-up which, as noted earlier, can vary substantially by patient and setting and should reasonably impact the magnitude of the workup during the ED visit.

A challenge to the HEART score is its inclusion of an elevated troponin measurement and an ECG with "significant ST depression." These features are highly specific for a cardiac etiology, and few clinicians will be comfortable that a patient is "low risk" when their score of 3 includes an elevated troponin measurement or an ischemic ECG. The inclusion of troponin measurement and an ischemic ECG in the score also imparts incorporation bias. It is not at all surprising that these features predict acute myocardial infarction—because they are key elements of how acute myocardial infarction is diagnosed. Such circular logic within HEART can only exaggerate its apparent predictive value. Finally, the troponin element of the score takes no account of baseline elevated values because of renal failure—particularly when stable over time. How are clinicians to apply the score in this circumstance?

Are the Predictor Variables Objective or Collected Prospectively? (Originally noncompliant)

Many clinical findings cannot be reliably assessed using chart review, and the derivation of an optimal clinical decision rule relies on prospective data collection.[28] The original HEART study and its first validation were retrospective[1,14]; however, many of the subsequent score evaluations were prospective.[2-4] There should be little concern with the investigator coding of age and troponin; however, subjective elements like history, risk factors, and ECG may be difficult to reliably determine from medical records (see the reliability discussion following).

Are the Potential Predictor Variables Reliable? (Noncompliant)

Even with prospective collection, many history and physical examination findings are sufficiently subjective

that they are not reliably assessed between clinicians. Nonreproducible variables should not be inserted into clinical decision rules, as unreliable input will logically ensure unreliable output.

Although age and troponin are objective, the other HEART score elements are not—most notably whether the history is "suspicious." We summarize reports that have assessed score interrater reliability in Table 1[12,29-40] and note problems with both the observed agreement and interstudy variation.

Each of the 3 subjective variables exhibit agreement (Table 1) that ranges from poor to moderate, with history more frequently closer to random than perfect agreement. Perhaps surprisingly, age and troponin do not consistently show perfect agreement. Even the strongest of the studies in Table 1 are unable to support any contention of excellent reliability for the HEART score and its components. History seems indisputably below any threshold of acceptable reliability. Oliver et al[32] note that "The most frequent disagreement was between a score of 3 and 4, and it was most frequently due to the history."

Equally problematic is the substantial variability in agreement between studies. It is well known that variable assignment and prediction model performance can differ by setting.[41] In any given clinical environment, one cannot know if the local interrater reliability ranks with the best or worst of the studies in Table 1.

To remedy the problematic history element, Marchick et al[42] attempted to create an altogether separate clinical decision rule just to assign this subscore within HEART. None of their 3 tested models were successful.[42]

How Were the Predictor Variables Coded? (Noncompliant)

Clinical variables may be coded into clinical decision rules as binary elements (present versus absent), graded elements (ranked score), or as continuous numerical elements (eg, WBC count). The decision regarding how to best code such variables is typically driven by both derivation analytics (ie, which format and threshold is the most predictive?) and what will demonstrate face validity to clinicians. The creators of the HEART score coded their variables (ie, age thresholds, number of risk factors, troponin thresholds) using investigator arbitrary preference without the benefit of quantitative analysis. Some of their choices and point weightings can be questioned.

Should age more than 65 years or having 3 risk factors contribute the same number of points (2 each) as having an acutely ischemic ECG or a markedly elevated troponin measurement?[9] Is it sensible that a patient aged 46 receives the same age risk score as a 64-year-old patient?[13] Are emergency physicians half as alarmed when the troponin

Table 1. Studies of the interrater reliability of the HEART score and its components.

Study	Overall Score	Score ≤3	History	ECG	Age	Risk Factors	Troponin
Retrospective							
Plewa et al, 2014[29]	29%, κ=0.30	69%, κ=0.41	κ=0.36	κ=0.34	κ=0.96	κ=0.73	NA
Wu et al, 2017[30]	30%, κ=0.13	70%, κ=0.24	45%, κ=0.13	76%, κ=0.51	85%, κ=0.72	NA	NA
Oliver et al, 2018[32]	κ=0.73	NA	κ=0.66	κ=0.92	κ=0.98	κ=0.91	κ=1.00
Tesson et al, 2018[58]	NA	κ=0.72	NA	NA	NA	NA	NA
Parenti et al, 2019[33]	κ=0.63	κ=0.72	κ=0.37	κ=0.42	κ=0.94	κ=0.71	κ=0.92
Mixed							
Ras et al, 2017[34]	NA	NA	NA	NA	97%	90%	96%
Mahler et al, 2013[35]	NA	75%	NA	NA	NA	NA	NA
Prospective							
Mahler et al, 2015[12]	NA	κ=0.63	NA	NA	NA	NA	NA
Gopal et al, 2018[31]	ICC=0.89	NA	ICC=0.62	ICC=0.51	NA	NA	NA
Niven et al, 2018[36]	ICC=0.91	NA	ICC=0.41	ICC=0.64	ICC=1	ICC=0.84	ICC=1
Soares et al, 2018[37]	25%, κ=0.31	74%	44%, κ=0.10	NA	93%, κ=0.89	67%, κ=0.43	NA
Gershon et al, 2019[38]	46%, κ=0.68	84%, κ=0.68	69%, κ=0.52	69%, κ=0.46	NA	70%, κ=0.67	98%, κ=0.83
Van Dongen et al, 2020[39]	ICC=0.78	74%	NA	NA	NA	NA	NA
Van Meerten et al, 2020[40]	κ=0.51	23%, κ=0.59	κ=0.17	κ=0.31	κ=0.93	κ=0.42	κ=1.00

κ, Kappa; *NA*, not applicable; *ICC*, intraclass correlation coefficient.

level is 1 to 3 times normal (1 point) versus when it is 3 or more times normal (2 points), or are we always alarmed when a troponin elevation accompanies chest pain unless that patient has a documented long-standing troponin leak?

Simply counting up the number of cardiac risk factors is misleading given that these risk factors are not each similarly predictive. Their positive likelihood ratios range from unhelpful (LR+ 1.0 for family history) to potentially helpful (LR+ 3.1 for an abnormal prior stress test).[21] The decision for HEART to grade a "suspicious" history might have more objectively focused on the specific signs and symptoms already established as predictive, as discussed under predictor variable inclusion above.

An additional HEART coding oddity is that chest pain patients more than 65 years of age with 3 or more risk factors have an intermediate-risk score of 4 before they receive any medical evaluation, and strict HEART score adherence would stipulate admission or observation even when their chest pain is clinically judged to almost certainly be noncardiac or stable cardiac.[43]

Was the Derivation Sample Large Enough? (Noncompliant)

The original HEART score was described in 122 patients with chest pain, of whom 29 had adverse cardiac outcomes. There was no quantitative justification for this sample size. The authors' conclusion that patients with a score of 3 or less had only a 2.5% risk of adverse cardiac outcomes was based on 1 such outcome out of the 39 low-risk patients. The upper bounds of the 95% confidence intervals of this proportion (0% to 13%) are not compatible with the declaration of safety.

Is the Analytic Technique Appropriate? (Noncompliant)

The most common techniques used to develop rigorous decision rules are multivariable logistic regression and forms of binary recursive partitioning. As described by its authors, "HEART was not developed from a database as modern scores often are. The HEART score was based on clinical experience and medical literature and designed to be as easy to use as the Apgar score for newborns."[15] As noted previously regarding variable coding, some of the authors' point weightings have questionable face validity.

When the HEART authors later retroactively applied regression techniques, they found regression coefficients inconsistent with the "1" and "2" scores assigned in their rule (Table 2).[16] Half of the weighting errors were 44% or more with the highest 76%. History, troponin measurement, and age 45 to 65 years are underweighted in the rule, whereas ECG, risk factors, and age 65 years or more are substantially overweighted. A second regression study focused on a low- to mid-risk patients also noted substantial discordance between assigned scores and their predictivity.[44]

Table 2. Variance between HEART score elements and their associated multivariable regression coefficients taken from Backus et al 2016.[16]

Item	Score	Regression Coefficient	Difference
History			
Slightly suspicious	0	Reference	
Moderately suspicious	1	1.27	−0.27 (27%)
Highly suspicious	2	2.06	−0.06 (3%)
ECG			
Normal	0	Reference	
Nonspecific repolarization disturbance	1	0.35	+0.65 (65%)
Significant ST depression	2	1.04	+0.96 (48%)
Age			
<45 years	0	Reference	
45 to 65 years	1	1.24	−0.24 (24%)
≥65 years	2	1.12	+0.88 (44%)
Risk factors			
No risk factors known		Reference	
1 or 2 risk factors	1	0.33	+0.67 (67%)
≥3 risk factors or history of atherosclerotic disease	2	0.49	+1.51 (76%)
Troponin			
≤ normal limit	0	Reference	
1 to 3 time the normal limit	1	1.17	−0.17 (17%)
>3 times the normal limit	2	2.26	−0.26 (13%)

The inherent mathematical assumption when adding 5 independent subscales into a total score is that each component point is of similar clinical impact relative to each of the others. Table 2 indicates that this is not the case for HEART.[16] Additive scoring of disparately weighted elements is statistically unsound and will almost certainly produce unequally predictive permutations. A known example of this problem is the Glasgow Coma scale (GCS), whose 13 scoring options (ie, 3 to 15) can be calculated 120 (ie, 4×5×6) different ways. A summed score of 4 predicts a mortality rate of 48% if calculated 1+1+2 for eye, verbal, and motor, a mortality rate of 27% if calculated 1+2+1, but a mortality rate of only 19% if calculated 2+1+1.[45,46] HEART's 10 scoring options can be

Table 3. Pooled sensitivities from meta-analyses.

Meta-Analysis	No. of Studies	No. of Patients	Pooled Sensitivity for Major Adverse Cardiac Events	Risk of Bias	Statistical Heterogeneity
Van Den Berg and Body, 2018[4]	12	11,217	97% (95% CI 94%, 98%)	Acceptable	High
Fernando et al, 2019[3]	30	44,202	96% (95% CI 93%, 98%)	High	High
Laureano-Phillips et al, 2019[2]	25	25,266	96% (95% CI 93%, 98%)	High	High

calculated 243 (ie, 3^5) different ways. A HEART score of 3 alone can represent 30 different subscale permutations and a score of 4 can have 45. In a large database study, Ioannides et al[44] noted that HEART "points obtained from different components of the score are associated with different risk elevations." The frequency and magnitude of outcome risk differences between such same-score permutations require further study but may, like the GCS, also include substantial, clinically important variations.

Was the Goal of the Rule Explicitly Specified A Priori? (Noncompliant)

The optimal clinical decision rule is derived to predict a specific outcome and to predict that outcome within a predefined quantitative threshold of success. Later validation studies can then test the rule using this benchmark. When deriving the pulmonary embolism rule-out criteria (PERC), for example, Kline et al[47] specified a priori that the rule should lower the risk below 1.8% to be clinically worthwhile. The HEART authors did not begin with any specific outcome or quantitative threshold for success, but rather designed their score to predict broad categories of low, intermediate, or high cardiac risk. Subsequent studies have not used an optimal outcome for emergency medicine, as discussed under "outcome" previously.

Is the Need for External Validation Before Clinical Application Acknowledged? (Noncompliant)

The need for external validation was not specified when the HEART score was first presented.[1] Subsequent studies have corroborated the general association of a progressive score with increased cardiac risk but that is not the same as validating the score for a specific ED decision or action.[2-4]

Can the Final Rule Itself Be Reliably Assessed? (Noncompliant)

The interrater reliability for the overall HEART score has varied widely between studies, with percent agreement ranging from 29% to 46% and kappa statistics from 0.13 to 0.73 (Table 1). The corresponding reliability for assigning low versus higher risk (ie, score 3 or less versus more than 3) ranges from 23% to 84% percent agreement and kappa statistics from 0.24 to 0.72 (Table 1) Such frequent disagreement in summary score measures is not surprising given the limited reliability of the subjective components discussed earlier. Meta-analyses have noted that rule performance heterogeneity is high (Table 3),[2-4] and this is compatible with an underlying instability of rule application between clinicians and settings.

ARE RESULTS OF THE RULE CLEARLY REPORTED?

Are Sensitivity and Specificity Both Emphasized? (Originally noncompliant)

The original HEART score report did not report either sensitivity or specificity for any of its score thresholds but instead emphasized the general correlation of an increasing score with an increased frequency of the adverse endpoints, a metric without specific value to a single provider caring for a single patient.

Does the Rule Improve on Baseline Clinical Practice? (Noncompliant)

The fundamental purpose of a decision rule is to improve clinical care, not just to predict what we are already doing. Does applying the rule improve diagnosis or decrease test use compared with unstructured clinical judgment, ie, gestalt? Unfortunately, many decision rules and scores do not compare their performance with gestalt, and when they do they are seldom superior.[6]

The sensitivity for emergency physician clinical judgment in detecting acute myocardial infarction or adverse cardiac risk appears to be about 98%, ie, about 2% are missed.[5,48,49] Thus, to improve gestalt, a clinical decision rule should either demonstrate greater sensitivity or match this sensitivity while decreasing resource use.

The HEART score was not derived[1] or initially validated[14] with any reference to baseline clinical practice or unstructured clinical judgment. Subsequent studies contrasting HEART with gestalt or baseline clinical practice show similar frequencies of major adverse cardiac events (Table 4),[11-13,17,50-55] although most were

Table 4. Studies contrasting the HEART score or its variants with gestalt or baseline clinical practice.

Study	Study Format	Score	Size Without/ With HEART	Safety Outcomes	Efficacy Outcomes
Mahler et al, 2013[50]	Database contrast of Likert scale gestalt vs calculated HEART	HEART Pathway	1,005 with both measures	Similar 30-day major adverse cardiac events	More frequent early discharge
Mahler et al, 2015[12]	Randomized controlled trial	HEART Pathway	141 vs 141	Similar 30-day major adverse cardiac events	More frequent early discharge, less later cardiac testing
Visser et al, 2015[51]	Prospective contrast of physician low / medium / high gestalt vs HEART	HEART	255 with both measures	Similar 6-week major adverse cardiac events	Not studied
Poldervaart et al, 2017[17]	Prospective before-and-after	HEART	1,827 vs 1,821	Similar MACE at 6 weeks	Similar early discharge, hospitalization
Singer et al, 2017[13]	Database contrast of low/ medium/high gestalt vs calculated HEART	HEART-2	434 with both measures	Similar acute myocardial infarction	Not studied
Mahler et al, 2018[10]	Prospective before-and-after	HEART Pathway	3,713 vs 4,761	Similar death or acute myocardial infarction	More frequent early discharge, fewer hospitalizations
Ljung et al, 2019[52]	Prospective before-and-after	HEART-2	612 vs 621	Similar 30-day major adverse cardiac events	Fewer admissions
Sharp et al, 2019[55]	Prospective before-and-after	HEART Pathway	30,522 vs 34,871	Similar death or acute myocardial infarction	Fewer hospitalizations and objective testing
Stopyra et al, 2020[11]	Prospective before-and-after	HEART Pathway	3,713 vs 4,761	Similar death or acute myocardial infarction at 12 months	Fewer hospitalizations over 12 months
Trent et al, 2020[53]	Prospective before-and-after	HEART	521 vs 649	Similar major adverse cardiac events at 6 weeks	More frequent hospitalizations and stress testing
Wang et al, 2020[54]	Theoretical application of HEART to usual care sample	HEART	2,185 with both measures	Similar 30-day major adverse cardiac events	Less frequent early discharge

underpowered for this comparison. The impact of the score on resource utilization is mixed, with some studies reporting improvements, no change, and worsening (Table 4). Thus, there is not compelling evidence that HEART improves unstructured physician judgment.

Is the Rule's Performance Sufficiently Precise? (Noncompliant)

The HEART score is widely described as "validated"; however, this is misleading in that the score was not originally designed to meet a specific outcome or a specific outcome threshold, as discussed earlier. For HEART, validation has instead meant that subsequent studies have confirmed the general association of a progressive score with increased cardiac risk—as would readily be expected from any summation of preestablished risk factors.

The more traditional and useful validation format is when a decision rule can accurately predict the chosen outcome in a new patient sample within a predefined threshold of clinical importance. Many subsequent studies have applied the HEART score (or one of its multiple variants) in this fashion—using a low-risk assessment to exclude major adverse cardiac events.[2-5] Success for this objective depends on how many rule failures emergency physicians are willing to accept. When deriving PERC, as discussed earlier, Kline et al[47] specified a priori that the rule should miss pulmonary embolism less than 1.8% of the time to be clinically useful. One survey found that emergency physicians might accept missing 0.5% to 1% of short-term major adverse cardiac events as a routine practice expectation.[56] The American College of Emergency Physicians (ACEP) 2018 Policy selected a 1%

to 2% frequency of missed major adverse cardiac events as "acceptable."[5]

Although some HEART studies (or variants such as the HEART pathway) reported high sensitivity, there are also many with lesser performance—with pooled sensitivities from 3 meta-analyses (Table 3) of 96% to 97%, ie, compatible with missing 3% to 4% of major adverse cardiac events.[2-4] The lower bounds of the 95% confidence intervals for these pooled sensitivities are 93% to 94%, ie, compatible with missing 6% to 7% of major adverse cardiac events. Said another way, these data suggest that the HEART score could, within 95% confidence, miss 1 in 14 occurrences of major adverse cardiac events.

The observed performance of the HEART score (Table 3) thus appears well below that which emergency physicians state a willingness to accept,[5,49,56] below the 98% sensitivity exhibited by baseline practice without the score,[5,48,49] and below the 1% to 2% acceptable threshold specified by ACEP in their chest pain policy.[5] Indeed, the ACEP policy describes a HEART score threshold of 2 or less as "more acceptable" than 3 or less because of its higher sensitivity, but with the downside of worse specificity.[5]

Does bundling HEART with a second troponin measurement—as codified by score variants HEART Pathway and HEART-2—overcome the performance limitations of HEART alone? This has not yet been the subject of meta-analysis; however, we describe the reports from our search that detail sensitivity in Table 5.[10,12,13,18,32,50,57-61] Most are underpowered, and the 2 with the strongest lower confidence interval bounds report these values as 96%, ie, compatible with missing 4% of major adverse cardiac events and outside of the 1% to 2% desirable threshold discussed previously. We note limitations to these 2 studies as follows.

In the first of the 2 studies, Mahler et al[50] applied the HEART pathway post hoc to a 2006 to 2007 data set limited to chest pain patients for whom objective cardiac testing was planned. Because these data lacked the HEART historical assessment, the investigators transposed Likert scale data to approximate it. The data set also lacked the HEART ECG ranking, and the investigators could only approximate a "0" and a "2", but not a "1." Their outcome was acute coronary syndrome, not major adverse cardiac events. Thus, this analysis differs in multiple ways from any real-world, clinical application of the HEART pathway.

In the second study, Thiruganasambandamoorthy et al[18] reported a protocol of 2 troponin measurements 3 to 6 hours apart using 15-day major adverse cardiac events as the endpoint. The HEART score elements were assigned by investigator chart review; however, they do not report their retrospective methodology or abstraction reliability.[28] Single troponin measures led to the exclusion of 21% of subjects. Again, this analysis differs in multiple ways from any real-world, clinical application of HEART-2.

Thus, the performance of the HEART pathway and HEART-2 is only marginally better than that of HEART alone, with the best-performing studies for each demonstrating key limitations. Some readers may recall seeing more optimistic appraisals elsewhere for HEART and its variants, and we attribute this difference to the form of reporting rather than true performance.

Table 5. HEART studies with 2 troponin measurements*.

Study	HEART Coding	Cardiac outcome prevalence	Sensitivity (95% Confidence Intervals)
HEART Pathway (2 Troponin Measurements 3 Hours Apart).			
Mahler et al, 2013[50]	Processed, see text	22%	198/200=99% (96%, 100%)
Mahler et al, 2015[12]	Prospective	6%	8/8=100% (63%, 100%)
Mahler et al, 2018[10]	Prospective	7%	341/353=97% (94%, 98%)
Hyams et al, 2018[57]	Chart review	6%	25/25=100% (86%, 100%)
Oliver et al, 2018[32]	Chart review	6%	25/25=100% (86%, 100%)
Tesson et al, 2018[58]	Chart review	8%	64/68=94% (86%, 98%))
HEART-2 (2 Troponin Measurements Regardless of Their Timing)			
Mahler et al, 2011[59]	Chart review	1%	12/12=100% (74%, 100%)
Singer et al, 2017[13]	Prospective	18%	347/374=93% (90%, 95%)
Thiruganasambandamoorthy et al, 2020[18]	Chart review	5%	88/88=100% (96%, 100%)

*All studies used major adverse cardiac events, except as follows: Mahler et al, 2013[50] acute coronary syndrome, Mahler et al, 2018[10] death or acute myocardial infarction, Singer et al, 2017[13] acute myocardial infarction. We have recalculated the confidence intervals using the binomial exact method (Stata 15.1) and rounded to integers, and accordingly these results may differ slightly from that reported in the original studies. This table excludes studies missing sensitivity data, eg, Halder et al, 2021,[60] Allen et al, 2018.[61] Hyams et al, 2018,[57] and Oliver et al, 2018[32] are studies from the same authors of the same data with the same results.

When the negative predictive value is emphasized over sensitivity, the apparent performance will in most cases be exaggerated. For example, in the latest meta-analysis, the pooled negative predictive value of 99% corresponded to a pooled sensitivity of 96%.[2] The negative predictive value, the percentage of negative tests that are true negatives, can be misleading because it varies with disease prevalence—a factor that varies widely among HEART score populations. In their meta-analysis, Van Den Berg et al[4] noted that the prevalence of major adverse cardiac events in their included studies ranged from 7.3% to 29.1%. In Table 5, we report a study in which the prevalence of major adverse cardiac events was just 1%, and thus the negative predictive value was effectively already 99% (because, at most, there can be 1 false negative per 100 patients) before one even begins to apply a test, score, or evaluation, ie, a negative predictive value below 99% is impossible. Accordingly, sensitivity rather than negative predictive value is the appropriate benchmark for judging how well HEART identifies adverse cardiac outcomes, as sensitivity is independent of prevalence and is thus more stable.

The distinction between negative predictive value and sensitivity is easy to miss when reviewing the medical literature, as some authors use ambiguous terminology such as "miss rate."[13,58] For example, one study concludes that "the HEART pathway was associated with … death and myocardial infarction rates well below 1% among low-risk patients."[10] Readers may erroneously assume this means that less than 1% of all adverse cardiac outcomes were missed, ie, near-perfect sensitivity. However, in this circumstance, the denominator of the "miss rate" is instead the number of patients judged low risk.

Another confusing and potentially misleading measure of HEART performance is the area under the receiver operating characteristic curve (ie, the c-statistic).[15,62,63] This measure simultaneously considers all scale thresholds along the curve and is thus useful for applications in which this occurs or is preferred. Clinical decisionmaking—particularly in the context of decision rules—does not consider all scale thresholds simultaneously, but rather applies a single selected dichotomous threshold to make a decision. Accordingly, sensitivity at the chosen threshold is the more reasonable and important measure of HEART performance.

A final observation about HEART performance is that clinicians do not consistently comply with its recommendations. Pena et al[20] noted that their emergency physicians were uncomfortable with early discharge for 55% of their HEART score 3 or less patients, citing poor follow-up or concerning clinical factors. Poldervaart et al[17] noted similar noncompliance with 41% of their low-risk patients, and Mahler et al[10] found the same in 16%. Janes et al[19] noted that 26% of their low-risk patients had objective testing ordered as follow-up despite it being deemed unnecessary by the HEART score.

HOW IS THE RULE INTERPRETED?

Does the Rule Provide a 2-Way Course of Action? (Suboptimal compliance)

The most useful clinical decision rules are 2-way, ie, they tell us to act one way if the rule is positive and to act in an opposite or different way if the rule is negative. The HEART score has been popularized as a 1-way rule. If the score is 3 or less, then the patient should receive early discharge without further testing. The rule leaves unclear what specific management might be best for individual score gradations past 3. Admit? Observe? Serial troponin measurements? Stress testing?

DISCUSSION

The HEART score and its variants have received appreciable attention because they are simple to remember and apply, and they address an important need. For emergency physicians, HEART provides a relatively clear management pathway, may speed the ED care of low-risk patients, and provides a perception of medicolegal protection. For health payors and policy leaders, the HEART score is seen as a tool to enhance ED efficiency and decrease the costs of observation and objective cardiac testing.

Our methodological appraisal (Figure 2) demonstrates that the HEART score was not formally derived and did not consider all known predictors or other potentially helpful variables, eg, sex, chest pain features. The rule does not differentiate its recommendations based on the rapidity and quality of follow-up, ie, next-day cardiology clinic versus delayed or unlikely reevaluation. The HEART score exhibits weak interrater reliability, ie, it is not consistently calculated between clinicians (Table 1). The 0, 1, and 2 weightings for each scoring element do not match their known predictivities (Table 2), and accordingly the various permutations of any single calculated score may readily signify clinically important differences in risk. There is not compelling evidence that HEART or its variants are superior to or improve the risk judgments that emergency physicians make without the score, ie, their clinical gestalt. The optimal target patient population for the score also remains unclear, as researchers have applied it to patients with varying risks, eg, any chest pain, those receiving a troponin, those for whom the clinicians believe that objective testing is warranted.

The pooled HEART score sensitivities from 3 meta-analyses (Table 3) are 96% to 97% with lower confidence interval bounds of 93% to 94%, ie, compatible with missing up to 7% or 1 in 14 occurrences of major adverse cardiac events. This summary performance is below that which emergency physicians state a willingness to accept,[5,49,56] below the 98% sensitivity exhibited by baseline practice without the score,[5,48,49] and below the 1% to 2% acceptable miss threshold specified by ACEP in their chest pain policy.[5] Some studies of HEART and its variants have emphasized negative predictive value rather than sensitivity for their outcome, thus exaggerating perceived test performance.[10,15,62,63]

The HEART Pathway and HEART-2 are variants that bundle the core HEART score with 2 troponin measurements. Unfortunately, these modified rules perform only marginally better, with 96% lower confidence interval bounds for their sensitivities in the strongest studies (Table 5). In a large study of the HEART pathway, Mahler et al[10] found the paired troponin measurements to be 92% sensitive in isolation and the HEAR score to be 84% sensitive in isolation. Thus, the paired troponin measurements substantially outperform the remainder of the score and appear primarily responsible for any presumed strength of these HEART variants. Given the fundamental and unfixable shortcomings that we have detailed for the HEART score, it is inevitable that any variant built on this unstable rule framework will also be impacted by these inherent limitations.

Some studies have noted decreased resource use with HEART or its variants (Table 4). As commonly implemented, when the score is 3 or less, clinicians are given explicit permission to forego the observation and stress testing otherwise recommended by American College of Cardiology/American Heart Association (ACC/AHA) guidelines.[64] Such cost savings cannot necessarily be specifically attributed to unique features or capabilities of HEART itself, but should similarly apply to any scenario in which clinicians are granted permission to selectively disregard ACC/AHA recommendations. The latter should reasonably also include a low-risk assessment based solely on an emergency physician's clinical judgment. There is no reason to believe that gestalt would be any less safe or cost-effective than HEART or one of its variants.

Although it is certainly plausible that some form of clinical decision rule may improve clinical judgment in the identification of patients with low-risk chest pain, we do not regard the HEART score or its variants as reasonable choices given their problems with methodology and sensitivity. Other more objective clinical decision rules deserve further study.[23] The ideal ED chest pain decision rule would be derived and validated in accordance with optimal standards,[7] would specifically predict short-term safety for discharge, and would account for differences in timeliness, likelihood, and quality of meaningful physician follow-up.

Emergency physicians may perceive that HEART provides medicolegal protection, as they are following a decision rule perhaps approved by their hospital for this purpose. In the event of litigation over an adverse outcome, however, the subjective elements of HEART may be more problematic than protective given their weak interrater reliability. Plaintiff's attorneys may readily identify expert witnesses confident that the patient merited a HEART of 4 rather than 3 and, thus, should have been observed and had further testing and/or cardiology consultation. In their HEART study, Oliver et al[32] noted that "The most frequent disagreement was between a score of 3 and 4, and it was most frequently due to the history."

Trent et al[53] comment that "our HEART score guideline significantly improved satisfaction of both our ED and hospitalist providers by creating a clear pathway for patient disposition." As observed by Wears,[65] clinical scoring systems provide "psychological comfort" in their "appearance of reducing uncertainty, creating order, and rationalizing clinical practice"—even when a score might be inferior to clinical judgment alone.

In summary, despite its need and simplicity, the HEART score and its variants have clinically important weaknesses in creation, structure, interrater reliability, and outcome selection and omit consideration of the timeliness or likelihood of meaningful physician follow-up. Their test sensitivity is below the threshold that emergency physicians state a willingness to accept. We believe that the widespread use of the HEART score and its variants should be reconsidered.

The authors thank Shanalee Tamares, MLIS, for her expert production of the literature search.

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, Loma Linda University Health, Loma Linda, CA (Green); and the Department of Emergency Medicine, University of California, Los Angeles, Los Angeles, CA (Schriger).

Funding and support: Dr. Schriger was supported in part by an unrestricted research grant from The Korein Foundation. There was no other outside funding or support. The authors have no conflicts to report.

Publication dates: Received for publication December 31, 2020. Revision received February 2, 2021. Accepted for publication February 3, 2021. Available online April 29, 2021.

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

APPENDIX E1.

Herein, we present areas in which the HEART score is in general compliance with *Annals of Emergency Medicine*'s methodological standards for clinical decision rules.[7] The question headers below mirror the topic order used in this methodological standards document.

IS THE METHODOLOGY OF THE RULE SOUND?

What Is the Population and Setting? (Compliant)

The HEART score has been appropriately studied in the ED setting to which it is intended to apply. A minority of early studies took place in a dedicated "cardiac" ED[14-16] or "chest pain unit,"[17] which presumably had similar, though perhaps slightly higher risk, chest pain patients as those treated in general EDs.

Was the Outcome Routinely Assessed? (Compliant)

HEART score studies comprise observational series of chest pain patients—a format that facilitates the later chart review assessment of adverse cardiac outcomes. Uncommon situations of error would include death not captured by coroner record searches and subsequent cardiac care at locations other than the original hospitals.

Were the Predictor Variables Recorded Before Knowledge of the Outcome? (Compliant)

Foreknowledge of the anticipated outcome for a clinical decision rule might introduce conscious or subconscious investigator bias in the coding of predictor variables. This limitation was almost certainly avoided for later prospective HEART research but might have impacted the original retrospective studies.

Can the Rule Be Refined? (Compliant)

As noted in the "definitions" section previously, multiple investigators have identified and explored opportunities to potentially improve the HEART score. Variations include the HEART pathway,[12] the HEAR score,[11] the modified low-risk HEART score,[2] HEART-2,[13] and the HEARTS3 score.[9] Each of these appear prompted by perceived deficiencies in the standard HEART score.

HOW IS THE RULE INTERPRETED?

Is the Rule Successful? (See discussion)

The HEART score is successful in that it has been widely researched and disseminated.[2-5] The discussion section addresses the challenges to such success based on noncompliance with these methodological standards.

Is the Rule Sensible? (Compliant)

The components of the HEART score are sensible and have face validity.

Is the Rule Easy to Apply? (Compliant)

The foremost strength of HEART is that it is relatively easy to remember and calculate.

WHAT POSTVALIDATION RESEARCH IS HELPFUL?

Can the Rule Be Successfully Implemented Into Clinical Practice? (Compliant)

Many hospitals have incorporated it into chest pain protocols and have embedded it into their electronic medical record.

Is the Rule Cost-Effective? (Compliant)

The cost-effectiveness of the HEART score is presumed based on its association in many (but not all) studies with decreased resource use (Table 4). The latest ACC/AHA guideline recommends that ED patients with chest pain suspicious for ischemia receive repeated troponin testing and noninvasive stress testing within 72 hours.[64] The very premise of a HEART score 3 or less is its explicit permission to forego such observation and follow-up testing. Such cost savings cannot necessarily be specifically attributed to unique features or capabilities of HEART itself, but might similarly apply to any other low-risk rule or marker for which clinicians are similarly instructed to selectively disregard ACC/AHA recommendations.

Can the Rule be Widely Integrated into Practice? (Compliant)

As previously, the HEART score appears to be used in a variety of hospitals.[2-5]

REFERENCES

1. Six A, Backus B, Kelder J. Chest pain in the emergency room: value of the HEART score. *Neth Heart J*. 2008;16:191-196.
2. Laureano-Phillips J, Robinson RD, Aryal S, et al. HEART score risk stratification of low-risk chest pain patients in the emergency department: a systematic review and meta-analysis. *Ann Emerg Med*. 2019;74:187-203.
3. Fernando SM, Tran A, Cheng W, et al. Prognostic accuracy of the HEART score for prediction of major adverse cardiac events in patients presenting with chest pain: a systematic review and meta-analysis. *Acad Emerg Med*. 2019;26:140-151.
4. Van Den Berg P, Body R. The HEART score for early rule out of acute coronary syndromes in the emergency department: a systematic review and meta-analysis. *Eur Heart J Acute Cardiovasc Care*. 2018;7:111-119.
5. Tamaszewski CA, Nestler D, Shah KH, et al. Clinical policy: critical issues in the evaluation and management of emergency department patients with suspected non-ST-elevation acute coronary syndromes. *Ann Emerg Med*. 2018;72:e65-e106.
6. 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 e3.
7. 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.
8. Bramer WM, Giustini D, de Jonge GB, et al. De-duplication of database search results for systematic reviews in EndNote. *J Med Libr Assoc*. 2016;104:240-243.
9. Fesmire FM, Martin EJ, Cao Y, et al. Improving risk stratification in patients with chest pain: the Erlanger HEARTS3 score. *Am J Emerg Med*. 2012;30:1829-1837.
10. Mahler SA, Lenoir KM, Wells BJ, et al. Safely identifying emergency department patients with acute chest pain for early discharge. *Circulation*. 2018;138:2456-2468.
11. Stopyra JP, Snavely AC, Lenoir KM, et al. HEART Pathway implementation safely reduces hospitalizations at one year in patients with acute chest pain. *Ann Emerg Med*. 2020;76:555-565.
12. Mahler SA, Riley RF, Hiestand BC, et al. The HEART Pathway randomized trial: identifying emergency department patients with acute chest pain for early discharge. *Circ Cardiovasc Qual Outcomes*. 2015;8:195-203.
13. Singer AJ, Than MP, Smith S, et al. Missed myocardial infarctions in ED patients prospectively categorized as low risk by established risk scores. *Am J Emerg Med*. 2017;35:704-709.
14. Backus BE, Six AJ, Kelder JC, et al. Chest pain in the emergency room: a multicenter validation of the HEART Score. *Crit Pathw Cardiol*. 2010;9:164-169.
15. Backus BE, Six AJ, Kelder JC, et al. A prospective validation of the HEART score for chest pain patients at the emergency department. *Int J Cardiol*. 2013;168:2153-2158.
16. Backus BE, Six AJ, Doevendans PA, et al. Prognostic factors in chest pain patients: a quantitative analysis of the HEART score. *Crit Pathw Cardiol*. 2016;15:50-55.
17. Poldervaart JM, Reitsma JB, Backus BE, et al. Effect of using the HEART score in patients with chest pain in the emergency department: a stepped-wedge, cluster randomized trial. *Ann Intern Med*. 2017;166:689-697.
18. Thiruganasambandamoorthy V, Stiell IG, Chaudry H, et al. Use of conventional cardiac troponin assay for diagnosis of non-ST-elevation myocardial infarction: 'the Ottawa troponin pathway. *PLoS One*. 2020;15:e0226892.
19. Janes JL, Streitz MJ, Hyams JM, et al. Are patients discharged on the HEART pathway following up? *Mil Med*. 2020;185:e2110-e2114.
20. Pena M, Tsao J, Falaiye T, et al. Why won't emergency physicians discharge patients with a low HEART score from an observation unit without further evaluation? *Crit Pathw Cardiol*. 2020;19:195-199.
21. Fanaroff AC, Rymer JA, Goldstein SA, et al. Does this patient with chest pain have acute coronary syndrome?: the rational clinical examination systematic review. *JAMA*. 2015;314:1955-1965.
22. Than M, Flaws D, Sanders S, et al. Development and validation of the emergency department assessment of chest pain score and 2 h accelerated diagnostic protocol. *Emerg Med Australas*. 2014;26:34-44.
23. Boyle RS, Body R. The diagnostic accuracy of the emergency department assessment of chest pain (EDACS) score: a systematic review and meta-analysis. *Ann Emerg Med*. 2021;. https://doi.org/10.1016/j.annemergmed.2020.10.020.
24. Bank IEM, de Hoog VC, de Kleijn DPV, et al. Sex-based differences in the performance of the HEART score in patients presenting to the emergency department with acute chest pain. *J Am Heart Assoc*. 2017;6:e005373.
25. Preciado SM, Sharp AL, Sun BC, et al. Evaluating sex disparities in the emergency department management of patients with suspected acute coronary syndrome. *Ann Emerg Med*. 2021;. https://doi.org/10.1016/j.annemergmed.2020.10.022.
26. Canto JG, Goldberg RJ, Hand MM, et al. Symptom presentation of women with acute coronary syndromes: myth vs reality. *Arch Intern Med*. 2007;167:2405-2413.
27. Snavely AC, Hendley N, Stopyra JP, et al. Sex and race differences in safety and effectiveness of the HEART pathway accelerated diagnostic protocol for acute chest pain. *Am Heart J*. 2021;232:125-136.
28. Kaji AH, Schriger D, Green S. Looking through the retrospectoscope: reducing bias in emergency medicine chart review studies. *Ann Emerg Med*. 2014;64:292-298.
29. Plewa M, Gibson D, Bingham S, et al. Inter-observer reliability and test characteristics of the HEART score in predicting low-risk admissions to an observation unit for cardiac evaluation. *Acad Emerg Med*. 2014;21(Suppl 1):S185-S186.
30. Wu WK, Yiadom MY, Collins SP, et al. Documentation of HEART score discordance between emergency physician and cardiologist evaluations of ED patients with chest pain. *Am J Emerg Med*. 2017;35:132-135.
31. Gopal S, Backus B, Kelder H, et al. Interobserver variability of the HEART score. *Eur J Emerg Med*. 2018;25:e21.
32. Oliver JJ, Streitz MJ, Hyams JM, et al. An external validation of the HEART pathway among emergency department patients with chest pain. *Intern Emerg Med*. 2018;13:1249-1255.
33. Parenti N, Lippi G, Bacchi Reggiani ML, et al. Multicenter observational study on the reliability of the HEART score. *Clin Exp Emerg Med*. 2019;6:212-217.
34. Ras M, Reitsma JB, Hoes AW, et al. Secondary analysis of frequency, circumstances and consequences of calculation errors of the HEART (history, ECG, age, risk factors and troponin) score at the emergency departments of nine hospitals in the Netherlands. *BMJ Open*. 2017;7:e017259. https://doi.org/10.1136/bmjopen-2017-017259.
35. Mahler SA, Riley RF, Burke GL, et al. Real-time use of a chest pain risk stratification clinical decision rule: a cautionary tale [abstract]. *Ann Emerg Med*. 2013;62(Suppl):S112.
36. Niven WGP, Wilson D, Goodacre S, et al. Do all HEART Scores beat the same: evaluating the interoperator reliability of the HEART Score. *Emerg Med J*. 2018;35:732-738.
37. Soares W, Poronsky K, Mader S, et al. Interrater agreement of HEART scores in emergency department patients with chest pain [abstract]. *Acad Emerg Med*. 2018;25(Suppl 1):S35.
38. Gershon CA, Yagapen AN, Lin A, et al. Inter-rater reliability of the HEART score. *Acad Emerg Med*. 2019;26:552-555.
39. van Dongen DN, Badings EA, Fokkert MJ, et al. Pre-hospital versus hospital acquired HEART score for risk classification of suspected non ST-elevation acute coronary syndrome. *Eur J Cardiovasc Nurs*. 2020;. https://doi.org/10.1177/1474515120927867; 1474515120927867.

40. van Meerten KF, Haan RMA, Dekker IMC, et al. The interobserver agreement of the HEART-score, a multicentre prospective study. *Eur J Emerg Med*. 2020; https://doi.org/10.1097/MEJ.0000000000000758.
41. Pajouheshnia R, van Smeden M, Peelen LM, et al. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol*. 2019;105:136-141.
42. Marchick MR, Setteducato ML, Revenis JJ, et al. Comparison of 3 symptom classification methods to standardize the history component of the HEART score. *Crit Pathw Cardiol*. 2017;16:102-104.
43. Webb JL, Streitz M, Hyams J, et al. HEART score of four for age and risk factors: a case series. *Cureus*. 2020;12:e9576.
44. Ioannides KLH, Sun BC, Baecker AS, et al. Not all HEART scores are created equal: identifying "low-risk" patients at higher risk. *J Am Coll Emerg Physicians Open*. 2020;1:1161-1167.
45. Healey C, Osler TM, Rogers FB, et al. Improving the Glasgow Coma scale score: motor score alone is a better predictor. *J Trauma*. 2003;54:671-678.
46. Green SM. Cheerio, laddie! Bidding farewell to the Glasgow Coma scale. *Ann Emerg Med*. 2011;58:427-430.
47. Kline J, Mitchell A, 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.
48. Pope JH, Aufderheide TP, Ruthazer R, et al. Missed diagnoses of acute cardiac ischemia in the emergency department. *N Engl J Med*. 2000;342:1163-1170.
49. Mitchell AM, Garvey JL, Chandra A, et al. Prospective multicenter study of quantitative pretest probability assessment to exclude acute coronary syndrome for patients evaluated in emergency department chest pain units. *Ann Emerg Med*. 2006;47:447.e1.
50. Mahler SA, Miller CD, Hollander JE, et al. Identifying patients for early discharge: performance of decision rules among patients with acute chest pain. *Int J Cardiol*. 2013;168:795-802.
51. Visser A, Wolthuis A, Breedveld R, et al. HEART score and clinical gestalt have similar diagnostic accuracy for diagnosing ACS in an unselected population of patients with chest pain presenting in the ED. *Emerg Med J*. 2015;32:595-600.
52. Ljung L, Lindahl B, Eggers KM, et al. A rule-out strategy based on highsensitivity troponin and HEART score reduces hospital admissions. *Ann Emerg Med*. 2019;73:491-499.
53. Trent SA, Stella S, Skinner A, et al. Improving atraumatic chest pain evaluation in an urban, safety-net hospital through incorporation of a modified HEART score. *Crit Pathw Cardiol*. 2020;19:173-177.
54. Wang G, Zheng W, Wu S, et al. Comparison of usual care and the HEART score for effectively and safely discharging patients with low-risk chest pain in the emergency department: would the score always help? *Clin Cardiol*. 2020;43:371-378.
55. Sharp AL, Baecker AS, Shen E, et al. Effect of a HEART care pathway on chest pain management within an integrated health system. *Ann Emerg Med*. 2019;74:171-180.
56. Than M, Herbert M, Flaws D, et al. What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the emergency department?: a clinical survey. *Int J Cardiol*. 2013;166:752-754.
57. Hyams JM, Streitz MJ, Oliver JJ, et al. Impact of the HEART Pathway on admission rates for emergency department patients with chest pain: an external clinical validation study. *J Emerg Med*. 2018;54:549-557.
58. Tesson A, Abdo R, Kamath A, et al. HEART failure? Episodes of missed major cardiac events when applying the HEART Pathway to an observation unit population. *Crit Pathw Cardiol*. 2018;17:88-94.
59. Mahler SA, Hiestand BC, Goff DC Jr, et al. Can the HEART score safely reduce stress testing and cardiac imaging in patients at low risk for major adverse cardiac events? *Crit Pathw Cardiol*. 2011;10:128-133.
60. Halder D, Mathew R, Jamshed N, et al. Utility of HEART pathway in identifying low-risk chest pain in emergency department. *J Emerg Med*. 2021;. https://doi.org/10.1016/j.jemermed.2020.12.004.
61. Allen BR, Simpson GG, Zeinali I, et al. Incorporation of the HEART score into a low-risk chest pain pathway to safely decrease admissions. *Crit Pathw Cardiol*. 2018;17:184-190.
62. Sun B, Laurie A, Fu R, et al. Comparison of the HEART and TIMI risk scores for suspected scute coronary syndrome in the emergency department. *Crit Pathw Cardiol*. 2016;15:1-5.
63. Six AJ, Backus BE. The HEART score for the assessment of patients with chest pain in the emergency department: a multinational validation study. *Crit Pathw Cardiol*. 2013;12:121-126.
64. Amsterdam EA, Wenger NK, Brindis RG, et al. 2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2014;130:2354-2394.
65. Wears RL. Lessons from the Glasgow Coma scale. *Ann Emerg Med*. 2012;59:338.