A Prospective Evaluation of Clinical HEART Score Agreement, Accuracy, and Adherence in Emergency Department Chest Pain Patients

William E. Soares III, MD, MS*; Alex Knee, MS; Seth R. Gemme, MD; Ruth Hambrecht, MD; Stacy Dybas, BS; Kye E. Poronsky, MS; Shelby C. Mader, BS; Timothy J. Mader, MD

*Corresponding Author. E-mail: william.soaresmd@baystatehealth.org.

Study objective: The HEART score is a risk stratification aid that may safely reduce chest pain admissions for emergency department patients. However, differences in interpretation of subjective components potentially alters the performance of the score. We compared agreement between HEART scores determined during clinical practice with research-generated scores and estimated their accuracy in predicting 30-day major adverse cardiac events.

Methods: We prospectively enrolled adult ED patients with symptoms concerning for acute coronary syndrome at a single tertiary center. ED clinicians submitted their clinical HEART scores during the patient encounter. Researchers then independently interviewed patients to generate a research HEART score. Patients were followed by phone and chart review for major adverse cardiac events. Weighted kappa; unweighted Cohen's kappa; prevalence-adjusted, bias-adjusted kappa (PABAK); and test probabilities were calculated.

Results: From November 2016 to June 2019, 336 patients were enrolled, 261 (77.7%) were admitted, and 30 (8.9%) had major adverse cardiac events. Dichotomized HEART score agreement was 78% (kappa 0.48, 95% confidence interval [CI] 0.37 to 0.58; PABAK 0.57, 95% CI 0.48 to 0.65) with the lowest agreement in the history (72%; WK 0.14, 95% CI 0.06 to 0.22) and electrocardiogram (85%; WK 0.4, 95% CI 0.3 to 0.49) components. Compared with researchers, clinicians had 100% sensitivity (95% CI 88.4% to 100%) (versus 86.7%, 95% CI 69.3% to 96.2%) and 27.8% specificity (95% CI 22.8% to 33.2%) (versus 34.6%, 95% CI 29.3% to 40.3%) for major adverse cardiac events. Four participants with low research HEART scores had major adverse cardiac events.

Conclusion: ED clinicians had only moderate agreement with research HEART scores. Combined with uncertainties regarding accuracy in predicting major adverse cardiac events, we urge caution in the widespread use of the HEART score as the sole determinant of ED disposition. [Ann Emerg Med. 2021;78:231-241.]

Please see page 232 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

https://doi.org/10.1016/j.annemergmed.2021.03.024

INTRODUCTION

At 6.5 million visits per year, chest pain is the second most common presenting symptom to US emergency departments, with hospital costs exceeding $10 billion annually.[1] Despite the high cost of hospitalization, the incidence of major adverse cardiac events, including percutaneous coronary intervention, myocardial infarction, coronary artery bypass graft surgery, and death in patients with nonischemic electrocardiogram (ECG) results and normal serum cardiac troponins is low.[2] Yet, provider overestimation of risk combined with personal and legal fears of misdiagnosis continue to reinforce potentially unnecessary low-risk chest pain admissions.[3]

The HEART score is a risk stratification aid designed to estimate the probability of major adverse cardiac events within 30 days for patients who present to the ED with symptoms concerning for acute coronary syndrome. The HEART score uses a mix of subjective and objective variables to stratify patients into a low-risk category (with an estimated 30-day major adverse cardiac events rate between 0.9% and 1.7%) or moderate-to-high risk (with an estimated 30-day major adverse cardiac events rate between 12% and 65%).[4,5] The components used in calculating a HEART score include patient history, past medical risk factors, age, ECG interpretation, and serum troponin results. The HEART score has been validated and

Editor's Capsule Summary

What is already known on this topic

The HEART score is widely used to risk stratify emergency department patients with chest pain.

What question this study addressed

What is the interrater reliability of the HEART score?

What this study adds to our knowledge

In this prospective study of 336 ED chest pain patients, clinician-calculated HEART scores were compared with those independently assessed by researchers. Measures of agreement were moderate at best, and poor for the history component. The most frequent disagreement (41% of the time) was at the 3 versus 4 threshold advocated for hospitalization/observation versus discharge.

How this is relevant to clinical practice

The HEART score appears insufficiently reliable to determine a decision as important as chest pain discharge.

endorsed by clinical practice guidelines, and its use has expanded throughout the United States and beyond.[6-10]

However, the utility of the HEART score when used by ED clinicians outside a research setting remains unclear. Specifically, because the HEART score is composed of components that require subjective interpretation, it is unknown whether clinicians are calculating HEART scores in a manner consistent with previous validation studies or whether persistent overestimation of risk, uncertainty, or lack of knowledge may alter the calculation and subsequent accuracy of HEART scores in predicting major adverse cardiac events. The primary objective of our study was to evaluate the agreement between HEART scores calculated during clinical practice (clinician scores) and scores generated using standardized research methods similar to previous validation studies (research scores). In addition, we estimated the accuracy of clinician and research scores in predicting 30-day major adverse cardiac events.

METHODS

We conducted a prospective, observational study at a single tertiary, academic, ST-elevation myocardial infarction (STEMI) receiving center with an annual ED volume of more than 115,000 visits and a 3-year emergency medicine residency program. We enrolled patients presenting to the ED with symptoms concerning for acute coronary syndrome for whom clinicians reported using the HEART score as part of their usual care, and we compared the clinician scores with research-generated HEART scores. The study was approved by the Baystate Medical Center Institutional Review Board. The Strengthening the Reporting of Observational Studies in Epidemiology guidelines were used to ensure the reporting of this observational study.

Selection of Participants

Patients eligible for inclusion were adults (aged 18 years or older) who presented to the ED with a chief complaint of chest pain, pressure or discomfort, or other symptom for which the treating emergency clinician considered acute coronary syndrome among their top 3 diagnoses and for whom the treating emergency clinician was using the HEART score for risk stratification. Exclusion criteria included patients the treating ED clinician determined were clinically unstable, altered, or unable to complete an interview; patients with a STEMI or other dynamic ECG changes concerning for active ischemia; patients who were pregnant; and patients with an alternate diagnosis confirmed by the treating clinician through objective testing, including but not limited to aortic dissection, pneumothorax, pneumonia, esophageal rupture, pulmonary embolism, congestive heart failure, or arrhythmia.

Emergency medicine clinicians were included in the study if they were attending physicians, senior residents (years 2 and 3), or advanced practitioners who self-identified that they used the HEART score as part of their regular clinical practice. Given their relative inexperience with the HEART score, emergency medicine interns, medical students, and off-service rotating providers were excluded. Because no special training in the HEART score was offered during the study, emergency clinicians who stated they were unfamiliar with the HEART score or did not use it in clinical practice were excluded. Finally, emergency clinicians who inherited the patient as a sign out, who were not the provider of record for the patient, or who were study authors were excluded.

Study Design

Trained research associates, available Monday to Friday, 7 AM to 9 PM with occasional weekend coverage, screened eligible patients by chief complaint and the presence of a laboratory fourth-generation troponin order on the ED tracking board. When a patient was identified, the research associate approached the treating ED clinician to determine if both the provider and patient were eligible for the study according to the protocol. The patient screening form can

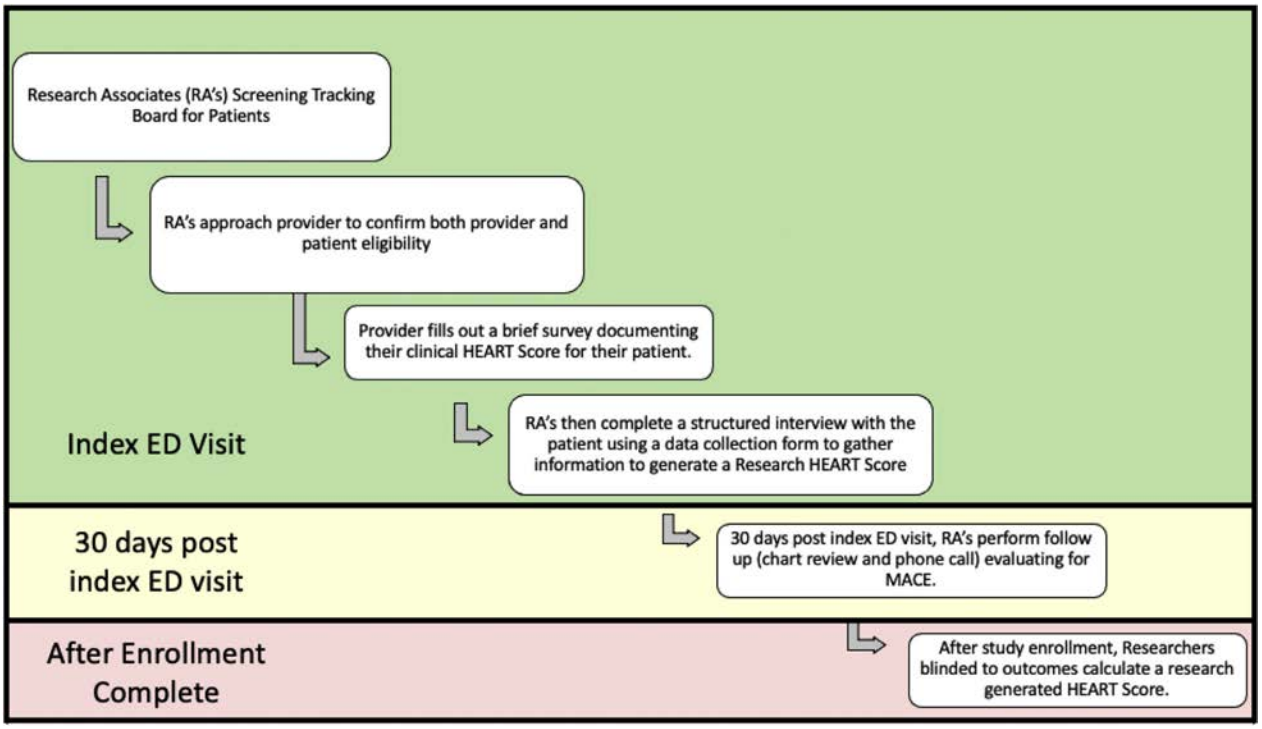

Figure 1. ED HEART score study design. *MACE*, major adverse cardiovascular events.

be found in Appendix E1 (available at http://www.annemergmed.com/). If both were eligible, the ED clinician was asked to provide the HEART score used during the patient encounter on a brief data collection form.

The research associate then approached the patient for enrollment. Following written consent, the patient completed an in-person structured interview with the research associate. The interview collected information necessary to calculate the history and risk factor components of the research-generated HEART score, which included a detailed description of the patient's current symptoms, past medical history, family history, and history of smoking. The results of the patient interview were not discussed with the treating ED clinician and did not lead to any changes in the care of the patient (Figure 1).

All research associates involved in the study received a 1-hour training on the HEART score and the data collection forms. Further, research associates were required to observe 5 interviews performed by study authors and then complete 5 proctored interviews with study authors before independently conducting patient interviews. To ensure consistency, research associates were periodically observed at least once a semester during interviews by study authors KEP and SCM throughout the study period.

Patient Follow-up

Research associates conducted patient follow-up 6 to 8 weeks after the index ED visit using both clinical chart review and structured phone follow-up. Patients were screened for troponin laboratory values, hospitalizations, future health care visits, and major adverse cardiac events, including myocardial infarction, percutaneous coronary intervention, coronary artery bypass graft, and death. Myocardial infarction was defined by final hospital discharge diagnosis and a rise in fourth-generation cardiac troponin with at least one value above the 99th percentile as described in previous literature.[11] Percutaneous coronary intervention was defined as coronary catherization with documented balloon angioplasty and/or stent placement; diagnostic coronary catheterizations without intervention were not included. Any major adverse cardiac event diagnoses discovered by phone interview were confirmed in the clinical chart.

Research associates were required to attend mandatory chart review training by study authors and abstract at least 5 charts in parallel with authors prior to independent data abstraction. Phone interviews and chart reviews were abstracted into standardized data entry forms. The 30-day chart review template can be found in Appendix E2 (available at http://annemergmed.com/). To ensure consistency, completed chart reviews of research associates

were periodically audited at least once a semester during interviews by study authors KEP and SCM throughout the study period. At the completion of follow-up, study author WES abstracted and confirmed all occurrences of major adverse cardiac events as well as an additional 5% of randomly selected charts; no changes or missed occurrences of major adverse cardiac events or hospitalization were discovered.

Instruments

The ED clinician data collection form was created by the research team and included basic demographic information (sex, role), the provider HEART score (components and final score), the method the ED clinician used to calculate the HEART score (online resource, phone application, hospital-specific protocol, no external resource, other), and a self-assessment of ED workload (1 indicating "not busy" and 10 indicating "busiest you have ever been"). The ED clinician was provided no instructions or prompts by the research team on how to determine their HEART score. The ED clinician data form can be found in Appendix E3 (available at http://www.annemergmed.com).

The research-generated HEART score data collection form used to guide the research associate interview was modeled after previously published HEART score validation data collection forms and included detailed "yes/no" questions to elicit the pertinent presenting symptoms, past medical risk factors, previous surgeries, and smoking history required for HEART score calculation.[12] The research participant interview form can be found in Appendix E4 (available at http://www.annemergmed.com/).

Research heart Score Derivation

After study enrollment and follow-up were complete, research team members not involved in patient interviews or primary data collection (WS, SG, TM, and AK) convened to calculate a research-generated HEART score for each study participant. Age and index troponin were confirmed in the clinical chart. Risk factors were scored per primary HEART score literature using information gathered in the structured patient interview.

The history score was calculated by assigning scores to low- and high-risk chest pain features previously described.[8,12] High-risk features were each scored as 1 point and included the following: middle- or left-sided pain, heavy chest pain, diaphoresis, radiation, nausea or vomiting, exertional pain, and relief of symptoms using sublingual nitrates. Low-risk features were each scored as −1 point and included the following: well-localized pain, sharp pain, nonexertional pain, no diaphoresis, and no nausea or vomiting. Scores were totaled and assigned a HEART score value as follows: −5 to −2 points (mostly low-risk features) scored as 0; −1 to 3 points (mix of high- and low-risk features) scored as 1; and 4 to 7 points (mostly high-risk features) scored as 2. Further details on derivation of the research-generated score can be found in Appendix E5 (available at http://www.annemergmed.com).

All ECGs were deidentified and independently scored over 2 rounds by research team members and content experts WS and TM, who were blinded to participant outcomes. Final interrater agreement was 87% (weighted kappa [WK] 0.76). Remaining discrepancies were solved through consensus discussion. Further details on the ECG scoring can be found in Appendix E6 (available at http://www.annemergmed.com).

Outcomes

The primary outcome was the agreement between the clinician's HEART score and the research-generated dichotomized HEART score, defined as low risk (HEART score 0 to 3) or moderate-to-high risk (HEART score 4 to 15). We also compared agreement of HEART scores as a continuous score (0 to 15), agreement between HEART score components (history, age, ECG, risk factors, troponin), and agreement of scores stratified by provider position.

As a secondary study outcome, we evaluated the diagnostic accuracy of clinical and research-generated HEART scores on 30-day major adverse cardiac events, overall and stratified by provider role.

Analysis

Descriptive statistics (frequencies and percentages) were used to summarize differences in HEART scores overall and stratified by predefined groups. The frequency of ED clinician participation was summarized to ensure that a minority of ED providers were not overrepresented in the study population.

Cohen's kappa was used to evaluate the primary outcome, agreement between clinician and research dichotomous HEART scores, with a score of 0.01 to 0.2 representing poor agreement, 0.21 to 0.4 representing fair agreement, 0.41 to 0.6 representing moderate agreement, 0.61 to 08 representing substantial agreement, and 0.81 to 1 representing near-perfect agreement.[13] The intraclass correlation coefficient (ICC) was used to evaluate agreement between clinician and research continuous HEART scores, and WK was used to evaluate HEART score components with more than 2 categories.

Because variations in prevalence influence kappa, and because the troponin component of the HEART score had

a very low prevalence of abnormal values, we conducted a sensitivity analysis and estimated a prevalence-adjusted, bias-adjusted kappa (PABAK) to help understand the range of kappa based on differences in prevalence of scores.[14,15]

The assessment of diagnostic accuracy included sensitivity, specificity, predictive values, and likelihood ratios, which were calculated for clinician and research-generated HEART scores based on 30-day major adverse cardiac events outcomes.

Study data were collected and managed using REDCap (version 10.6.16, Vanderbilt University)[16] electronic data capture tools. All analyses were performed using Stata MP (version 15.1).

Sample Size

Based on prior literature demonstrating an ICC statistic of 0.6 between 2 attending physicians using the thrombolysis in myocardial infarction (TIMI) cardiac risk stratification score (12% absolute difference in scores with 2 or more points), we assumed a similar baseline ICC for 2 providers using the HEART score of 0.6.[17] We estimated that a sample of 260 patients would be needed to able to calculate an ICC with 95% confidence intervals within 0.1 (half-width CI) using an alpha of 0.05 and a power of 0.8. Given the frequency of chest pain ED visits, uncertainty surrounding the enrollment of both providers and patients, and a desire to include major adverse cardiac events rates, we included an additional 20% enrollment for a final sample size of 312 participants.

RESULTS

From November 2016 to June 2019, 3,335 patients were screened for inclusion, for which 815 patients were approached for enrollment and 336 were included in the study (Figure 2).

Participants

Fifty-three unique ED clinicians used the HEART score to evaluate the 336 patients included in the study, with each ED clinician evaluating a median of 10 patients during the study period (interquartile range [IQR] 7 to 12). Of the participants enrolled in the study, 53% (178 of 336) were men, and the median age was 59 years (IQR 52 to 68). In comparison, of the 479 patients who were approached but not enrolled in the study, 44% (209 of 479) were men, and the median age was 60 years (IQR, 51 to 72).

Participants' past medical history included diabetes (27.7%, n=93 of 336), hypertension (61.3%, n=206), and hypercholesterolemia (57.1%, n=192). A diagnosis of atherosclerotic disease, which included a reported history of myocardial infarction, percutaneous coronary intervention, coronary artery bypass graft, or stroke, was present in 36.6% (n=123) of patients.

Over half of participants (51.8%, n=174) were seen by a woman ED provider, with 27% (n=91) seen primarily by an attending emergency physician, 31.3% (n=105) by an advanced practitioner, 22% (n=74) by a third-year resident, and 19.6% (n=66) by a second-year resident. Of the participants, 78% (n=261) were admitted to the hospital on the index ED visit (Table 1).

Primary Outcome

Dichotomous HEART score agreement between clinicians and researchers was 78% (n=263) with a kappa of 0.48 (95% CI 0.37 to 0.58; PABAK 0.57, 95% CI 0.48 to 0.65). ED clinicians scored 49 patients as high-risk who had low-risk research HEART scores and 24 patients as low-risk who had high-risk research HEART scores (Table 2). ICC, of the continuous HEART score (0 to 15) among ED clinicians and researchers was 0.65 (95% CI 0.59 to 0.71).

Evaluating components of the HEART score, agreement was highest for age (agreement 96.7%; WK 0.89, 95% CI 0.85 to 0.94; PABAK 0.93, 95% CI 0.90 to 0.96) and troponin (agreement 98%; WK 0.46, 95% CI 0.26 to 0.67; PABAK 0.96, 95% CI 0.93 to 0.98). Agreement was lowest for history (agreement 72%; WK 0.14, 95% CI 0.06 to 0.22; PABAK 0.37, 95% CI 0.30 to 0.43). Compared with research HEART scores, ED clinicians assigned higher mean history scores for participants (1.2, standard deviation [SD] 0.7; versus 1.0, SD 0.6) (Table 3).

Of the 73 participants with discordant HEART scores, 59% (43 of 73) differed by 1 point, 37% (27 of 73) by 2 points, and 4% (3 of 73) by 3 points. The most common difference in discordant scores was between a threshold score of 3 to 4 (58.9%, 43 of 73), followed by a score of 3 to 5 (20.5%, 15 of 73). The most common HEART score disagreement in discordant scores was history, at 44.7% (51 of 114), followed by risk factors at 32.5% (37 of 114) and ECG interpretation at 19.3% (22 of 114). More details on discordant HEART scores can be found in Tables E1 and E2 (available at http://www.annemergmed.com/).

Heart Score Accuracy in Predicting 30-day Major Adverse Cardiac Events

In addition to the structured chart review, 74.2% (n=262) of participants were able to be reached for phone follow-up. During the study period, the 30-day participant major adverse cardiac events rate was 8.9%

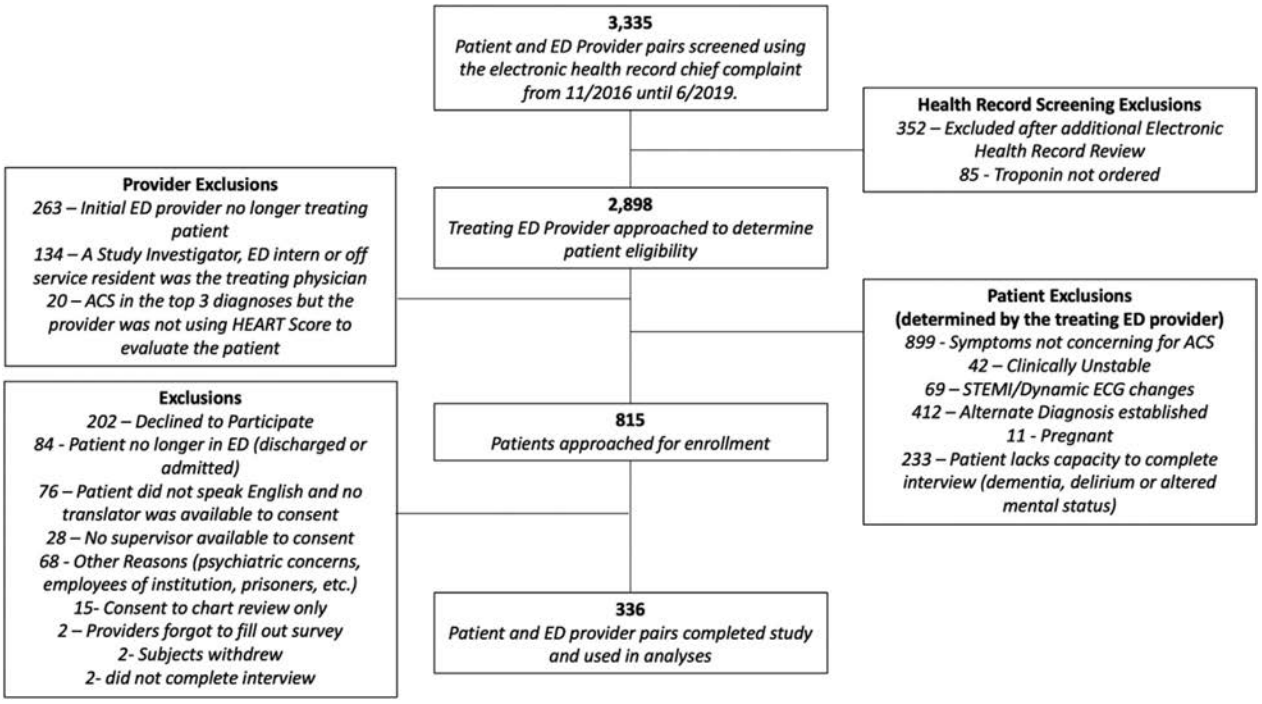

Figure 2. ED HEART score enrollment flow diagram. *ACS*, acute coronary syndrome.

(n=30), which included 1 STEMI, 13 non-STEMIs, 14 percutaneous coronary interventions, 10 coronary artery bypass grafts, and 1 death. All patients with 30-day major adverse cardiac events were admitted during the index ED visit. Four instances of major adverse cardiac events occurred after discharge from the index hospital admission.

Compared with research-generated HEART scores, ED clinicians' HEART scores had 100% sensitivity (95% CI 88.4% to 100%; versus 86.7%, 95% CI 69.3% to 96.2%) and a 27.8% specificity (95% CI 22.8% to 33.2%; versus 34.6%, 95% CI 29.3% to 40.3%) in predicting 30-day major adverse cardiac events. This trend was consistent across all ED clinician roles (Table 4).

Based on index admission to the hospital, ED providers adhered to their own clinical HEART score recommendations in 87.5% of cases (n=294, 95% CI 83.5% to 90.8%). Of the 85 patients with low-risk clinical HEART scores, 33% (26 of 85, 95% CI 23.1% to 44.0%) were admitted to the hospital. None of the patients classified by clinicians as low risk went on to have major adverse cardiac events. Of 251 patients with high-risk clinical HEART scores, 6.4% (16 of 251, 95% CI 3.7% to 10.1%) were discharged home, with 8 leaving against medical advice as documented by the clinical chart.

In contrast, 4 patients classified as low risk by research-generated scores subsequently had major adverse cardiac events, which included 1 percutaneous coronary intervention, 1 coronary artery bypass graft, 2 STEMI events, and 1 death. Case details can be found in Table E3 (available at http://www.annemergmed.com).

LIMITATIONS

Our study has multiple limitations. Sampling bias may impact the generalizability of results. Specifically, the screening of patients, the determination of eligibility by the treating ED clinician, and the high decline-to-participate rate limit generalizability. With respect to screening, owing to available resources, trained undergraduate research associates screened the electronic health record for eligible participants. As such, we were unable to screen during overnights and some weekends. Further, because research associates were not medical providers, screening was based on typical chief complaint symptoms and may have missed atypical presentations of acute coronary syndrome.

After screening, the treating ED clinician determined whether the patient met the study criteria. As our goal was to capture HEART scores used in clinical practice, it was important that the ED clinician determined whether acute coronary syndrome was on their differential and if using the

Table 1. Demographics of the 336 ED patients enrolled in the study.

ED Patient Demographics (N=336)	
Age, median (IQR)	59 (52, 68)
Sex	
Male	178 (53.0%)
Female	158 (47.0%)
Race/ethnicity	
Black	33 (9.8%)
Hispanic	45 (13.4%)
White	243 (72.3%)
Past medical history	
Hypertension	206 (61.3%)
Diabetes	93 (27.7%)
Hypercholesterolemia	192 (57.1%)
Atherosclerotic disease	123 (36.6%)
Current smoker	71 (21.1%)
Emergency severity index level	
2	279 (83.0%)
3	51 (15.2%)
4	1 (0.3%)
Treating ED clinician sex	
Male	162 (48.2%)
Female	174 (51.8%)
ED clinician role	
2nd-year resident	66 (19.6%)
3rd-year resident	74 (22.0%)
Advanced practitioner	105 (31.3%)
Fellow or attending physician	91 (27.1%)
Used to calculate HEART score	
Online calculator (eg, MDCalc)	243 (72.3%)
Phone application	20 (6.0%)
Hospital-specific protocol	1 (0.3%)
Memory	65 (19.3%)
Other	7 (2.1%)
Perceived ED clinician workload (1–10), mean (SD)	5.3 (2.1)
ED Patient Outcomes	
Present within 30 days after index ED visit:	
Hospitalization during index ED visit	261 (77.7%)
Hospitalization after ED discharge	28 (8.3%)
Repeat ED visit	22 (6.5%)
30-day major adverse cardiac events event*	30 (8.9%)
Myocardial infarction	14 (4.2%)
Percutaneous coronary intervention	14 (4.2%)
Coronary artery bypass grafting	10 (3%)
Death	1 (0.3%)

*Each patient may have had multiple major adverse cardiac events qualifying events.

HEART Score was appropriate. It is possible that this method of selection resulted in a slightly healthier patient population, reflected in the low 30-day major adverse cardiac events rate.

Finally, many eligible patients declined to participate in the study. Patients were often approached toward the end of their ED workup, given the need for a troponin level to calculate the HEART score, and we were told many did not want to spend extra time to participate in the study. Additionally, although similar in age, a higher ratio of women declined to participate compared with those in our study population.

To address sampling bias, we attempted to capture and present inclusion and exclusion criteria for all screened patients. Of patients who consented to participate in the study, all but 4 completed the study. Further, we found no additional major adverse cardiac events on phone follow-up that were not already discovered on chart review. Therefore, although the characteristics of our study population may be slightly different from those of the general population, our adherence to rigorous study methods and transparency in reporting strengthens the validity of our study.

Second, if ED clinicians had become aware of the study aims, it could have changed their approach to calculating the HEART score. In order to avoid bias secondary to the Hawthorne effect, we maintained multiple safeguards. First, our provider data entry sheet did not contain any prompts or guidelines on how to calculate the HEART score. Second, ED clinicians were not present during the patient interview to generate data for the research HEART score. Third, research HEART scores were not calculated until the end of the study, in part so that providers could not compare the research score with their own clinical score. Finally, the study took place over 3 years involving 53 unique ED clinicians, limiting the likelihood that ED clinicians adapted their practice owing to participating multiple times in the study.

Finally, our study was not powered to detect a significant difference in major adverse cardiac events. As a secondary outcome, we hope that our major adverse cardiac events results help to inform the creation of future studies to determine the performance of the HEART score in clinical practice.

DISCUSSION

Understanding the real-world application of risk stratification aids outside of a research setting is important, especially when components of the tool require clinician interpretation. Poor agreement, differences in accuracy, or

Table 2. 2×2 contingency table of the frequency of ED clinician and research-generated high and low HEART score risk stratification with total percent for each cell.

		Research-Generated Score		
		Low-Risk HEART Score (0–3)	High-Risk HEART Score (4–15)	
ED clinician score	Low-risk HEART score (0–3)	61 (18%)	24 (7%)	85 (25%)
	High-risk HEART score (4–15)	49 (15%)	202 (60%)	251 (75%)
		110 (33%)	226 (67%)	336 (100%)

misapplication may change the performance of the tool and the subsequent impact on patient care.

Previous studies examining agreement among providers using the HEART score have yielded mixed results. An evaluation of 88 patients with suspected ischemic chest pain found excellent agreement among emergency physicians and nurses, with an ICC of 0.91 (95% CI 0.87 to 0.93). However, clinicians had access to educational initiatives, including posters, YouTube videos, and pocket-sized HEART score reference cards not normally available to the larger emergency medicine community.[18] Additionally, a prospective evaluation of 311 ED patients with chest pain found good agreement on dichotomized HEART score ratings between attending and resident emergency physicians (kappa 0.68, 95% CI 0.60 to 0.77), though it was unclear how much communication about patient presentation occurred prior to the HEART score calculation.[19] By contrast, studies comparing agreement between emergency physicians and out-of-hospital nursing and cardiology attending physicians found substantially lower rates of agreement (kappa 0.514 and 0.13, respectively).[20,21] In all studies, interrater reliability was worse for the subjective components of the HEART score.

In order to evaluate HEART scores in practice, we prospectively enrolled a diverse group of ED clinicians who were using the HEART score in real time during actual patient encounters. We provided no study-specific education or aid to the ED providers to avoid influencing their HEART score calculations and subsequent patient care. For comparators, we generated standardized research scores independent of the clinicians' HEART scores. Although simple agreement was relatively high at 78%, interrater agreement was only moderate for the dichotomized HEART scores, even when adjusting for

Table 3. Agreement between ED clinician and research-generated HEART scores. Interrater reliability (IRR) includes ICC for continuous variables, kappa (K) for dichotomous variables, WK and PABAK for ordinal variables.

	ED Provider Score	Research-Generated Score				
	Mean (SD)	Mean (SD)	Agreement	IRR Test	IRR	95% CI
HEART score (continuous)	4.3 (1.4)	4.1 (1.4)	n/a	ICC	0.65	0.59-0.71
HEART score (dichotomous)	n/a	n/a	78%	K	0.48	0.37-0.58
				PABAK	0.57	0.48-0.65
Components						
History	1.2 (0.7)	1 (0.6)	72%	WK	0.14	0.06-0.22
				PABAK	0.37	0.30-0.43
ECG	0.3 (0.5)	0.4 (0.6)	84.8%	WK	0.40	0.30-0.49
				PABAK	0.66	0.60-0.72
Age	1.2 (0.6)	1.1 (0.6)	96.7%	WK	0.89	0.85-0.94
				PABAK	0.93	0.90-0.96
Risk factors	1.5 (0.6)	1.5 (0.6)	85.1%	WK	0.51	0.42-0.59
				PABAK	0.67	0.61-0.72
Troponin	0 (0.2)	0 (0.2)	98.1%	WK	0.46	0.26-0.67
				PABAK	0.96	0.93-0.98

Table 4. Diagnostic accuracy of the HEART score for predicting 30-day major adverse cardiac events. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (+LR), and negative likelihood ratio (–LR) with 95% CI.

	Accuracy	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	+LR (95% CI)	–LR (95% CI)
Research-generated HEART score	39.3%	86.7% (69.3%-96.2%)	34.6% (29.3%-40.3%)	11.5% (7.7%-16.4%)	96.4% (91%-99%)	1.33 (1.13-1.56)	0.39 (0.15-0.97)
Clinician HEART score	34.2%	100.0% (88.4%-100%)	27.8% (22.8%-33.2%)	12.0% (8.2%-16.6%)	100% (95.8%-100%)	1.38 (1.29-1.48)	0
ED provider position							
Senior resident (n=140)	36.4%	100% (75.3%-100%)	29.9% (22.1%-38.7%)	12.7% (7%-20.8%)	100% (90.7%-100%)	1.43 (1.278-1.6)	0
Advanced practitioner (n=105)	32.4%	100% (63.1%-100%)	26.8% (18.3%-36.8%)	10.1% (4.5%-19%)	100% (86.8%-100%)	1.37 (1.21-1.54)	0
Attending physician (n=91)	33%	100% (66.4%-100%)	25.6% (16.6%-36.4%)	12.9% (6.1%-23%)	100% 83.9%-100%)	1.34 (1.18-1.53)	0

differences in prevalence. Our results are similar to those of Mahler et al,[12] who demonstrated moderate agreement (kappa 0.63) between treating ED clinicians and physician study investigators in 282 patients randomized to the HEART Pathway.

Further, there were multiple concerning trends in the 22% of discordant clinician and research HEART scores. While most discordant scores differed by only 1 point, it was common to differ at the critical HEART score test threshold between 3 and 4 points. Additionally, differences in scores were most frequently due to the history component, which had a low kappa value. These results raise concerns that whatever methods our ED clinicians were using to score the history component of the HEART score (and, to a lesser extent, risk factors and ECG interpretation) do not appear to be consistent with the methods described in prior validation studies. The poor agreement in subjective components may subsequently change the overall performance of the HEART score in predicting major adverse cardiac events. It remains unclear whether the differences in agreement between the subjective components of the HEART score could be improved with education, clinical decision support, or other aids.

With respect to accuracy, though not significant, ED clinicians had higher raw sensitivities and lower specificities for major adverse cardiac events compared with research scores, with no observed occurrences of major adverse cardiac events in patients whom ED clinicians assigned as low risk. ED clinicians also had higher overall HEART and history component scores, suggesting that our clinicians may have been more risk averse, potentially reducing the stated benefit of safely discharging low-risk chest pain patients. By contrast, the research-generated HEART score sensitivity was low, with confidence intervals between 68% and 96%. Although wide, our confidence intervals were consistent with those from prior meta-analyses, with pooled HEART score sensitivities of 96% (95% CI 93% to 98%).[22,23] Our results raise concerns that the HEART score, even when collected in a standardized fashion, may miss major adverse cardiac events at a rate too high for broad clinical implementation as standard of care. Large-scale clinical outcome studies are needed to assess the accuracy of the HEART score in clinical practice compared with clinical gestalt in predicting major adverse cardiac events.

Finally, we were surprised to find a high rate of nonadherence by clinicians in the application of their own HEART scores. Specifically, 33% of the clinician-assigned low-risk HEART score patients were admitted on the index visit, with none having subsequent occurrences of major

adverse cardiac events. While adherence is likely determined by hospital culture and policies, it does raise questions of variability in implementation and adherence to the HEART score. In the HEART Pathway, Mahler et al[24] discovered ED provider nonadherence in 20% (28 of 141) of patients, including overtesting in 13% (19 of 141) and 10 additional admissions. Few studies have addressed the prevalence of nonadherence to the HEART score and the subsequent influence on outcomes.[25]

In conclusion, ED clinicians demonstrated only moderate agreement with research-generated dichotomized HEART scores, with most discrepancies occurring at the test threshold and involving the history component, which had low agreement. Even when using standardized methods to generate research HEART scores, sensitivity was low. With uncertainties in agreement, accuracy, and adherence, we urge caution in the widespread use of the HEART score in isolation as a standard of care to determine the disposition of ED patients with chest pain.

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

Author affiliations: From the Institute of Healthcare Delivery and Population Science, University of Massachusetts Medical School-Baystate, Springfield, MA (Soares, Dybas, Mader T); the Department of Emergency Medicine, University of Massachusetts Medical School–Baystate, Springfield, MA (Soares, Gemme, Dybas, Poronsky, Mader S, Mader T); the Department of Medicine, University of Massachusetts Medical School–Baystate, Springfield, MA (Knee); the Epidemiology/Biostatistics Research Core, Office of Research, Baystate Medical Center, Springfield, MA (Knee); and the Department of Emergency Medicine, Advent Health, Tampa, FL (Hambrecht).

Author contributions: WES, TJM, KEP, and SCM significantly contributed to the conception and design of the study. KEP, SCM, SD, and RH acquired the data. WES, SG, and AK analyzed the data. WES, SG, and TJM drafted the initial manuscript. All authors were involved in data interpretation, in revising the manuscript, and in approving the final version submitted for publication. WES takes responsibility for all aspects of the work.

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: This work is supported by a grant from the Agency for Healthcare Research and Quality (AHRQ) Health Effectiveness and Outcomes Research (5R03HS024815-02) as well as the Tufts REDCap grant UL1TR002544. WES is supported by grant 5K08DA045933-03 from the National Institute on Drug Abuse.

Results were presented at the 2019 SAEM National Meeting in Las Vegas, NV. Interim results were presented at the 2018 SAEM National Meeting in Indianapolis, IN.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Publication dates: Received for publication December 14, 2020. Revision received March 15, 2021. Accepted for publication March 16, 2021. Available online June 18, 2021.

REFERENCES

1. Owens PL, Barrett ML, Gibson TB, et al. Emergency department care in the United States: a profile of national data sources. *Ann Emerg Med*. 2010;56:150-165.
2. Weinstock MB, Weingart S, Orth F, et al. Risk for clinically relevant adverse cardiac events in patients with chest pain at hospital admission. *JAMA Intern Med*. 2015;175:1207-1212.
3. Lin GA, Redberg RF. Addressing overuse of medical services one decision at a time. *JAMA Intern Med*. 2015;175:1092-1093.
4. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. *Neth Heart J*. 2008;16:191-196.
5. 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.
6. 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.
7. American College of Emergency Physicians Clinical Policies Subcommittee (Writing Committee). on Suspected Non–ST-Elevation Acute Coronary Syndromes, Tomaszewski CA, Nestler D, 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.
8. Six AJ, Cullen L, Backus BE, et al. 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.
9. 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.
10. 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.
11. Luepker RV, Apple FS, Christenson RH, et al. Case definitions for acute coronary heart disease in epidemiology and clinical research studies: a statement from the AHA Council on Epidemiology and Prevention; AHA Statistics Committee; World Heart Federation Council on Epidemiology and Prevention; the European Society of Cardiology Working Group on Epidemiology and Prevention; Centers for Disease Control and Prevention; and the National Heart, Lung, and Blood Institute. *Circulation*. 2003;108:2543-2549.
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. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22:276-282.
14. Gwet KL. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. 4th edition. Advanced Analytics, LLC; 2014.

15. Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. *Educ Psychol Meas*. 1981;41:377-381.
16. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377-381.
17. Taylor B, Mancini M. Discrepancy between clinician and research assistant in TIMI score calculation (TRIAGED CPU). *West J Emerg Med*. 2015;16:24-33.
18. 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.
19. Gershon CA, Yagapen AN, Lin A, et al. Inter-rater reliability of the HEART score. *Acad Emerg Med*. 2019;26:552-555.
20. van Meerten KF, Haan RMA, Dekker IMC, et al. The interobserver agreement of the HEART-score, a multicentre prospective study. Eur J Emerg Med. Published online October 2020. https://doi.org/10.1097/MEJ.0000000000000758
21. Wu WK, Yiadom MYAB, 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.
22. 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.
23. 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.
24. Mahler SA, Riley RF, Russell GB, et al. Adherence to an accelerated diagnostic protocol for chest pain: secondary analysis of the HEART Pathway randomized trial. *Acad Emerg Med*. 2016;23:70-77.
25. Westafer LM, Kunz A, Bugajska P, et al. Provider perspectives on the use of evidence-based risk stratification tools in the evaluation of pulmonary embolism: a qualitative study. *Acad Emerg Med*. 2020;27:447-456.

IMAGES IN EMERGENCY MEDICINE

(continued from p. 229)

DIAGNOSIS:

Atypical femoral fracture secondary to prolonged use of bisphosphonates. Extended bisphosphonate use is associated with atypical femoral fractures. Bisphosphonates alter collagen cross-linking and maturity, disrupting the optimal biological state for stability. Patients with a history of prolonged bisphosphonate use may have increased susceptibility to microfractures from minimal trauma.[1]

Significant diagnostic criteria include: a history of minimal trauma, including falls from standing height; "beaking" of the fracture site; fracture location away from the femoral neck; noncomminuted fracture pattern; and nonunion.[2,3]

This patient was treated with intramedullary nail internal fixation. At 15 months follow-up, there was nonunion of the fracture with moderate callus formation and small areas of bridging (Figure 4).

From the Department of Trauma and Orthopaedics (M Ali), Royal Free Hospital, Royal Free London NHS Foundation Trust, London, United Kingdom; and the Department of Breast Surgery (F Ali), Northwick Park Hospital, London North West University Healthcare NHS Trust, Harrow, United Kingdom.

REFERENCES

1. Rizzoli R, Akesson K, Bouxsein M, et al. Subtrochanteric fractures after long-term treatment with bisphosphonates: a European Society on Clinical and Economic Aspects of Osteoporosis and Osteoarthritis, and International Osteoporosis Foundation Working Group Report. *Osteoporos Int*. 2011;22:373-390. https://doi.org/10.1007/s00198-010-1453-5.
2. Black DM, Geiger EJ, Eastell R, et al. Atypical femur fracture risk versus fragility fracture prevention with bisphosphonates. *N Engl J Med*. 2020;383:743-753. https://doi.org/10.1056/NEJMoa1916525.
3. Haworth AE, Webb J. Skeletal complications of bisphosphonate use: what the radiologist should know. *Br J Radiol*. 2012;85:1333-1342. https://doi.org/10.1259/bjr/99102700.