



This is a repository copy of *Do all HEART Scores beat the same: evaluating the interoperator reliability of the HEART Score*.

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/136004/>

Version: Published Version

Article:

Niven, W.G.P., Wilson, D., Goodacre, S. orcid.org/0000-0003-0803-8444 et al. (3 more authors) (2020) Do all HEART Scores beat the same: evaluating the interoperator reliability of the HEART Score. *Emergency Medicine Journal*, 35 (12). pp. 732-738. ISSN 1472-0205

<https://doi.org/10.1136/emered-2018-207540>

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. More information and the full terms of the licence here:
<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>



OPEN ACCESS

Do all HEART Scores beat the same: evaluating the interoperator reliability of the HEART Score

William G P Niven,¹ David Wilson,¹ Steve Goodacre,² Alexander Robertson,¹ Samira Jane Green,¹ Tim Harris³

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/emmermed-2018-207540>).

¹Accident and Emergency Department, Homerton University Hospital, London, UK
²Health Services Research, School of Health and Related Research, University of Sheffield, Sheffield, UK
³Department of Emergency Medicine, Royal London Hospital, Barts Health, London, UK

Correspondence to

Dr William G P Niven, Accident and Emergency Department, Homerton University Hospital, London E9 6SR, UK; william.niven@nhs.net

Received 16 April 2018

Revised 16 August 2018

Accepted 23 August 2018

ABSTRACT

Background Patients presenting with chest pain represent a significant proportion of attendances to the ED. The History, ECG, Age, Risk Factors and Troponin (HEART) Score is validated for the risk stratification of suspected ischaemic chest pain within the ED. The goal of this research was to establish the interoperator reliability of the HEART Score as performed in the ED by different grades of doctor and nurse.

Methodology Patients with suspected ischaemic chest pain presenting to the ED of an inner city, London Hospital, were recruited prospectively between January and May 2016. Patients that had been enrolled in the study were interviewed by clinicians from four different categories: senior doctor, junior doctor, senior nurse and junior nurse. Clinicians, blinded to other raters' results, calculated the HEART Scores for each patient with the assistance of a pocket-sized HEART Score card. The intraclass correlation coefficient (ICC) was calculated as the primary measure of reliability. 120 patients were required to achieve a desired power of 80%.

Results 88 complete comparisons were obtained. There were no significant differences between the distributions of HEART Scores for each clinician group ($p=0.95$). The ICC for the overall HEART Score was 0.91 (95% CI 0.87 to 0.93). The ICC for troponin and age were '1', for 'history' 0.41 (95% CI 0.30 to 0.52), 'ECG' 0.64 (95% CI 0.54 to 0.73) and 'risk factors' 0.84 (95% CI 0.79 to 0.89).

Conclusion This study demonstrates very strong overall interoperator reliability between the four groups of clinicians studied. This suggests that the HEART Score is reproducible when used by different professional groups and grade of clinician.

INTRODUCTION

The History, ECG, Age, Risk Factors and Troponin (HEART) Score has become a popular chest pain risk stratification tool in the emergency medicine community and is a simple, structured and practical approach to chest pain risk stratification. With its five parameters, the HEART Score, outlined in [table 1](#), enables emergency physicians to assess the 30-day risk of developing a major adverse cardiac event.^{1–3} The scoring system was developed and validated in an emergency medicine population which sets it apart from other commonly used chest pain risk tools such as the Thrombolysis in Myocardial Infarction⁴ and Global Registry of Acute Cardiac Events⁵ scores which were derived from a less representative cohort of patients with established acute coronary syndrome.

Key messages

What is already known on this subject

- The History, ECG, Age, Risk Factors and Troponin (HEART) Score is a well-established tool for chest pain risk stratification in cases of suspected acute coronary syndrome.
- The interoperator characteristics of the HEART Score have been analysed retrospectively and as a secondary outcome in four publications.
- To date, no research has prospectively evaluated the interoperator agreement for the HEART Score in a diverse population of both doctors and nurses.

What this study adds

- In this prospective, cross-sectional study, senior and junior doctors and nurses assessed HEART Scores on the same patients.
- We found excellent inter-rater reliability suggesting that the HEART Score can be used by both doctors and nurses irrespective of grade or level of experience.

The diagnostic accuracy of the HEART Score for the identification of adverse events has been validated in several studies^{1–3} and most recently by a comprehensive and well-designed, step wedge trial conducted by Poldevaart *et al.*⁶ However, of all the studies to date, only one study by Plewa *et al.*⁷ has had, as its primary outcome, the interoperator reliability of the HEART Score. This study was retrospective and relied on information extrapolated from chart reviews. A recent study by Oliver *et al.*⁸ analysed the interoperator variability of the HEART Score as a secondary outcome but was, once again, based on retrospective chart reviews. One prospective and one retrospective study by Mahler *et al.*^{9 10} analysed interoperator agreement as a secondary outcome but only among relatively senior doctors; residents and attending physicians. Thus to date no study has prospectively measured the reproducibility of the HEART Score as a primary outcome among both nurses and junior doctors.

In the UK and in many other countries, patients presenting to an ED are primarily assessed by nurses and junior doctors, and initial care is often delivered by junior clinicians under the supervision of consultants. In this regard the care delivered is often described as consultant led rather than consultant delivered. With ED chest pain attendances constituting approximately 6% of all annual attendances



© Author(s) (or their employer(s)) 2018. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Niven WGP, Wilson D, Goodacre S, *et al.* *Emerg Med J* Epub ahead of print: [please include Day Month Year]. doi:10.1136/emmermed-2018-207540



Table 1 The History, ECG, Age, Risk Factors and Troponin (HEART) Score risk stratification tool for chest pain patients

History	Not suggestive	0
	Moderately suggestive	1
	Highly suggestive	2
ECG	Normal	0
	Non-specific repolarisation disturbance, left bundle branch block, paced rhythm, T-wave inversion with no ST changes	1
	Significant ST deviation=1. 0.05 mV in two or more contiguous leads or 2	2
	ST depression of >1 mV	2
Age (years)	<45	0
	45–65	1
	≥65	2
Risk factors	No risk factors	0
	1–2 risk factors (hypertension, diabetes, obesity, smoker, hypercholesterolaemia, family history)	1
	3 risk factors or known atherosclerotic disease	2
Troponin	< normal limit	0
	1–2 times normal limit	1
	>3 times normal limit	2

in the UK,¹¹ it is important that if a risk stratification tool is employed it is both accurate and reproducible, yielding reliable agreement between individuals regardless of their role or seniority. In particular, evidence demonstrating that nurses can also carry out reliable chest pain risk stratification could result in improved and expedited diagnosis and treatment. Despite this, only one study by Carlton *et al*¹² has included nurses in the evaluation of the interoperator characteristics of another chest pain risk stratification tool, the Goldman Score.

We therefore designed a prospective, cross-sectional study to establish the interoperator reliability of the HEART Score performed in the ED by various grades of doctor and nurse.

METHODS

Study design

This was a primary data collection study prospectively conducted in the ED between 31 January and 31 May 2016. The study design analysed the HEART Scores calculated on individual patients by each of the following four categories of rater: senior doctors, junior doctors, senior nurses and junior nurses. Senior doctors were defined as having at least 5 years clinical experience and 3 years emergency medicine training, UK grades senior trainee or consultant. Junior doctors were defined as all who were not senior trainees. Senior nurses were defined as having at least 5 years clinical nursing experience, postgraduate training in emergency nursing and clinical management experience, UK band six and above. Junior nurses were defined as all other qualified nursing staff that were not senior nurses, UK band five or band five star.

Study setting

The study was conducted in a large, London ED (census 120 000 patients a year).

Patient inclusion and exclusion criteria

Included in this study were adults aged ≥18 presenting to the ED with non-traumatic chest pain of possible ischaemic origin. Patients were excluded if the ECG showed significant ST elevation in two contiguous precordial (>2 mm) or limb (>1 mm)

leads, there was a history of chest trauma, the patient was adjudged to be too unwell to take part such that being enrolled might delay treatment or transfer to a specialist cardiac centre, the patient could not speak English and did not have a relative or friend to translate to an acceptable level as adjudged by the enrolling clinician, the patient had a GCS <14 or was unable to provide consent.

Study protocol

Four Good Clinical Practice-certified researchers (one senior doctor, two junior doctors and a senior nurse) working in shifts were involved in approaching patients presenting to the ED with chest pain during the hours of 08:00 and 20:00. Patients were asked if they wished to participate in the study. If the patient consented and was eligible, they were enrolled in the study and were taken to a specified cubicle where they had an ECG performed and blood tests including a lab-based high-sensitivity troponin assay (Abbot Architect Stat High Sensitivity Troponin). The diagnosis, treatment and management decisions for the patients enrolled in the study were undertaken by either the junior or senior doctor that saw the patient as part of the further study protocol outlined below.

After the ECG and blood tests had been performed, two doctors and two nurses interviewed the patient separately and in no prespecified order. Each clinician then proceeded to calculate and enter the HEART Score either via a customised electronic form on the hospital intranet or, if they were unaware of how to access the form, by telling the researcher who would enter the score for them. Participants were blinded to each other's scores and there was no discussion allowed between them, the on-site researcher policed this. As an aide memoire, doctors and nurses were allowed to take a laminated A6 HEART scorecard into the consultation. Clinicians were also allowed to look at the HEART Score posters put up around the department prior to entering their final score.

The troponin result was based on the first available troponin taken in the ED. The time taken for troponin results to come back from the laboratory was variable. If the troponin result had not been uploaded to the electronic patient record system by the time the clinicians had calculated the other four parameters, they were asked to enter '0' on the electronic form. When the result was available, generally within 90 min of blood being taken, the clinicians were asked to score the result and the researcher would update the previously calculated scores accordingly. Calculations were based on a multiple of the absolute cut-offs for the Abbot Architect Stat High Sensitivity Troponin: 34 ng/L for men and 16 ng/L for women. Patients with a troponin value below the cut-off were scored '0', one to two times the cut-off were scored '1' and a value of three or more times the cut-off were scored '2'. If the troponin had been taken <6 hours after the onset of chest pain, a second troponin was sampled 3 hours after the first troponin. Patients with an initial or repeat troponin level above the specified cut-offs were adjudicated to have had an acute myocardial infarction (AMI).

Patients were admitted to the hospital if they had a HEART Score of '>6' or if they were diagnosed with an AMI. The decision to admit and initiate further treatment was made by either the junior or senior doctor that saw the patient as part of the study protocol. Admission and treatment decisions were made after the patient had completed the protocol and did not involve nursing staff.

All medical and nursing staff involved in the study received training prior to the study, in how to perform the HEART Score

and interpret its results through an hour-long face-to-face tutorial. The tutorials were standardised, approximately 1 hour long and carried out in four separate sessions to foundation trainees, junior registrars, senior registrars and consultants. This session was repeated at a mandatory nurse training day. The teaching was interactive, practical and focused specifically on how to calculate the ECG, risk factor and history parameters of the HEART Score. For those who were not present at the teaching sessions, a specially designed, 11 min video about the study was posted on YouTube.¹³ This can be accessed on <https://www.youtube.com/watch?v=w799yRm7hYg>.

Outcomes

The primary outcome measurement was the interoperator agreement as measured by the intraclass correlation co-efficient (ICC) between the four categories of doctor and nurse involved in the study. The diagnosis of AMI was also measured for the cohort and was based on the Third Universal Definition, namely a rise or fall in the troponin value with at least one value above the 99th centile.¹⁴ In this study, the 99th percentile was sex-specific in line with the recommendations of the manufacturers of the Abbot Architect Stat High Sensitivity Troponin.

Analysis

Descriptive statistics were used to describe the demographic data recorded for this study. Age was reported with mean and SD, gender and language proficiency were reported using numbers and relative frequency. Those patients diagnosed with an AMI were expressed as a number and percentage of the total number recruited.

The ICC is a commonly used measure of inter-rater reliability. It is well suited to situations where there are more than two raters and ratings are either discrete ordinal or continuous within specific groups or classes. The ICC is calculated by comparing the variance of ratings within the groups of four with the total variation across all ratings and all patients.¹⁵ There are 10 forms of ICC defined by Shrout and Fleiss¹⁶ of which the most appropriate for an inter-rater reliability study in which the raters are not the same for all subjects is the 'one-way random effects model'. Table 2 illustrates how the various ICC values correlate with level of reliability.

We calculated the ICC to determine the level of overall reliability between the four categories of rater in addition to reliability for each of the individual components of the score. Thereafter we measured reliability when the calculated HEART Scores were stratified first into low (≤ 3), intermediate (4-6) or high (≥ 7) and then again for scores dichotomised into low (≤ 3) or high (> 3) risk. The ICC was used for all calculations and the results were reported with values and 95% CIs.

The Net Reclassification Index (NRI) was then calculated for each of the four categories of rater for the index diagnosis of AMI. The purpose of the NRI calculation was to provide reassurance that the results were not unduly contributed to by one

category of rater. This is included as part of the post-hoc sensitivity analysis in the online supplementary appendix.

We used an assumption of non-normality and calculated the median as the measure of central tendency. The Kruskal-Wallis rank-sum test¹⁷ was used to determine whether there was any statistically significant difference between the distributions of HEART Scores for each of the four categories of rater.

The sample size calculation was based on an estimated width of the CI around a chosen ICC as described by Gireudeau *et al.*¹⁸ The reason for this is that the other method of sample size calculation using null and alternative hypotheses for the correlation coefficients of the population would have been difficult to define and questionable given the lack of knowledge in this area. So, an ICC of 0.8 was chosen indicating good reliability and using four operators with a type 1 error of 0.05, a type 2 error of 0.2 and a 95% CI of 0.1 gave a projected sample size of 120 patients.

The data were analysed using the R-statistical package V.3.3.1 for Windows. The R project for statistical computing is a free, open-source software package.¹⁹

RESULTS

In total, 159 patients were considered for participation in the study of which 54 had exclusion criteria. Thus a total of 105 patients were recruited during the trial period. 4 patients had only one response and 13 had only two or three responses leaving a total of 88 patients who completed the full study protocol. The recruitment process is outlined in figure 1.

The mean age of the participants was 53.3 years (range 24–87, SD; 15.2 years). Of the 105 patients enrolled, 57 were male (54%) and 48 were female (46%). Of patients who completed the study protocol, 50 (57%) were first-language English speakers and 38 (43%) spoke English as a second language or not at all.

In total, 88 patients completed the study protocol of whom 24 (27%) were admitted to hospital and 9 (10.2%) had a diagnosis of myocardial infarction as defined by the third universal definition of myocardial infarction.¹⁴

A total of 107 doctors and nurses participated in the study including 22 senior doctors, 37 junior doctors, 18 senior nurses and 30 junior nurses. The median HEART Score overall was '3'. The range of scores for each patient as calculated by the raters is illustrated in figure 2. There was no significant difference between the distribution of the HEART Scores in the four groups using the Kruskal-Wallis test $p=0.95$.

The average HEART Score was 3.95 for first-language and 3.36 for second-language or non-English-speaking patients. This was not a statistically significant difference ($p=0.13$) using a two-sample t-test where $t=1.531$ and $df=82.83$.

The overall ICC for the HEART Score was 0.91 (95% CI 0.88 to 0.94). The ICCs for age and troponin showed perfect agreement of '1'; risk factors 0.86 (95% CI 0.81 to 0.90), ECG was 0.64 (95% CI 0.55 to 0.73) and history 0.41 (95% CI 0.30 to 0.53). The aggregate HEART Score of the subjective elements (history, ECG and risk factors) had an ICC of 0.78 (95% CI 0.71 to 0.84) (figure 3).

Using the original risk categories of the HEART Score, low (≤ 3), intermediate (4-6) or high (≥ 7), the ICC was 0.82 (95% CI 0.77 to 0.87). When the HEART Score was dichotomised into low risk (≤ 3) or high risk (> 3), the ICC was 0.84 (95% CI 0.79 to 0.89).

The results of a post-hoc sensitivity analysis including the NRI are included in the online supplementary appendix. In general, it was observed that there was minimal reclassification between

Table 2 Intraclass correlation coefficient values and reliability²⁰

Intraclass correlation coefficient	Reliability
<0.5	Poor
0.5 to <0.75	Moderate
0.75 to <0.90	Good
>0.90	Excellent

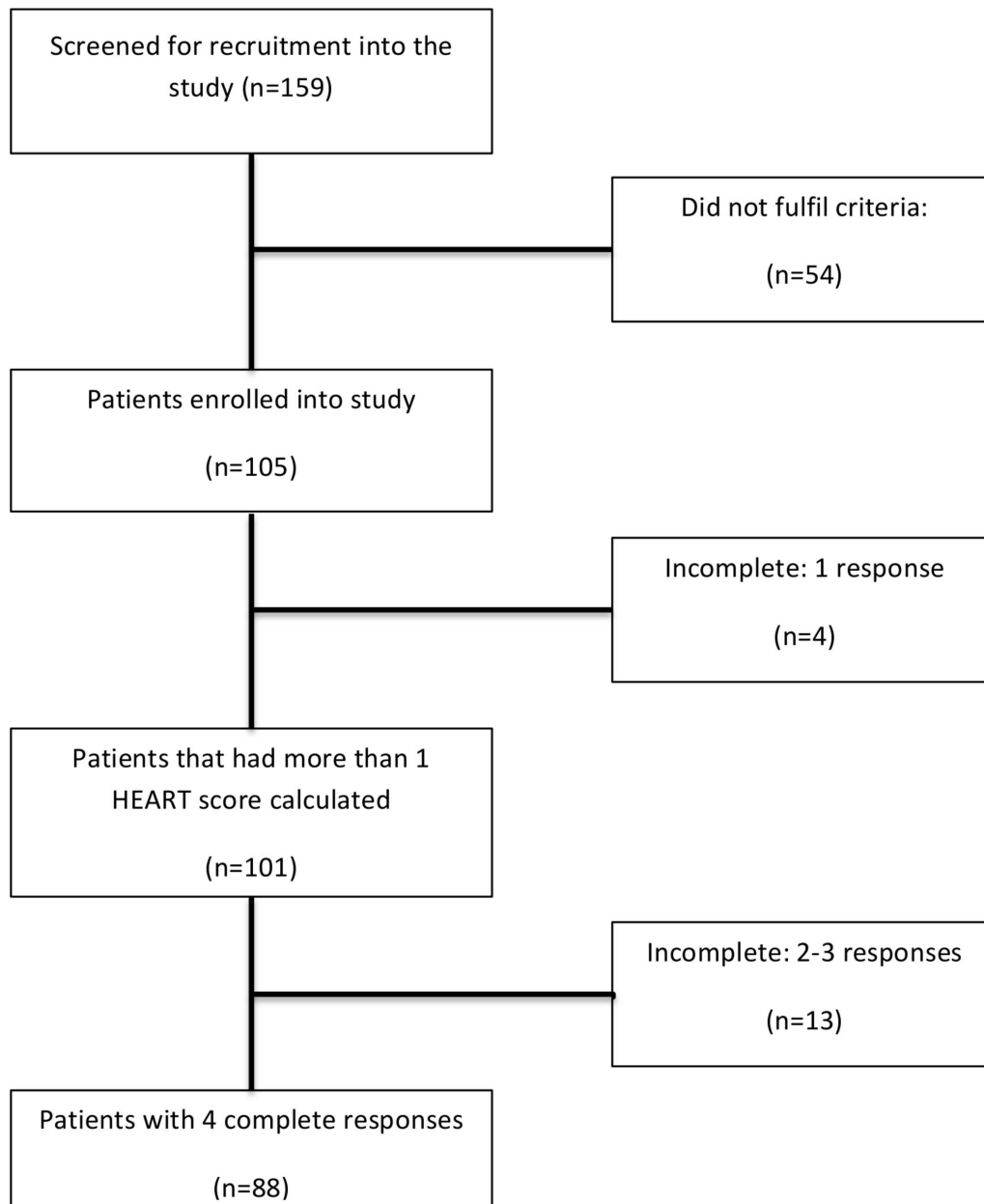


Figure 1 Flow diagram for the recruitment of patients into the trial. HEART, History, ECG, Age, Risk Factors and Troponin.

risk categories on removing one of each of the four categories of rater.

DISCUSSION

The aim of this study was to find out whether or not emergency doctors and nurses could reliably perform the HEART Score irrespective of grade and experience. The principal finding of an ICC of 0.91, defined as 'excellent',²⁰ strongly suggests this to be true and that neither grade nor experience should preclude any doctor or nurse from being able to apply the scoring system.

The ICC statistics demonstrated by this study were higher than the measures of agreement previously published for the HEART Score by Plewa⁷ (0.41), Oliver⁸ (0.72) and Mahler^{9,10} (0.67, 0.81). This could possibly be explained by the various initiatives carried out prior to commencing data collection as well as the differing methodologies. The contrast between the studies was most evident with the ECG (Plewa 0.34 vs our study 0.64) and risk factors (Plewa 0.73 vs our study 0.86). Both ECG and risk

factor scoring were clearly defined by Backus and therefore relatively easy to teach. The presence of posters, HEART Score cards and the YouTube video that was circulated through social media may also have further reinforced the learning sessions.

It is worth noting, however, that the objective elements of troponin and age, which show perfect correlation, increase the overall reliability statistic beyond that of the subjective components of the HEART Score alone, namely the ECG, risk factors and history elements. The level of reliability for these combined subjective elements that we observed was still 'good', however, at 0.78 (95% CI 0.71 to 0.84). Of the subjective elements, the history component demonstrated the lowest level of reliability and was 'poor' at 0.41 (95% CI 0.3 to 0.53) and this was despite the various educational initiatives carried out prior to the study. The ICC value for history was similar to the weighted kappa of 0.36 (95% CI 0.23 to 0.48) achieved by Plewa but significantly lower than the value of 0.6 recorded by Oliver *et al.*

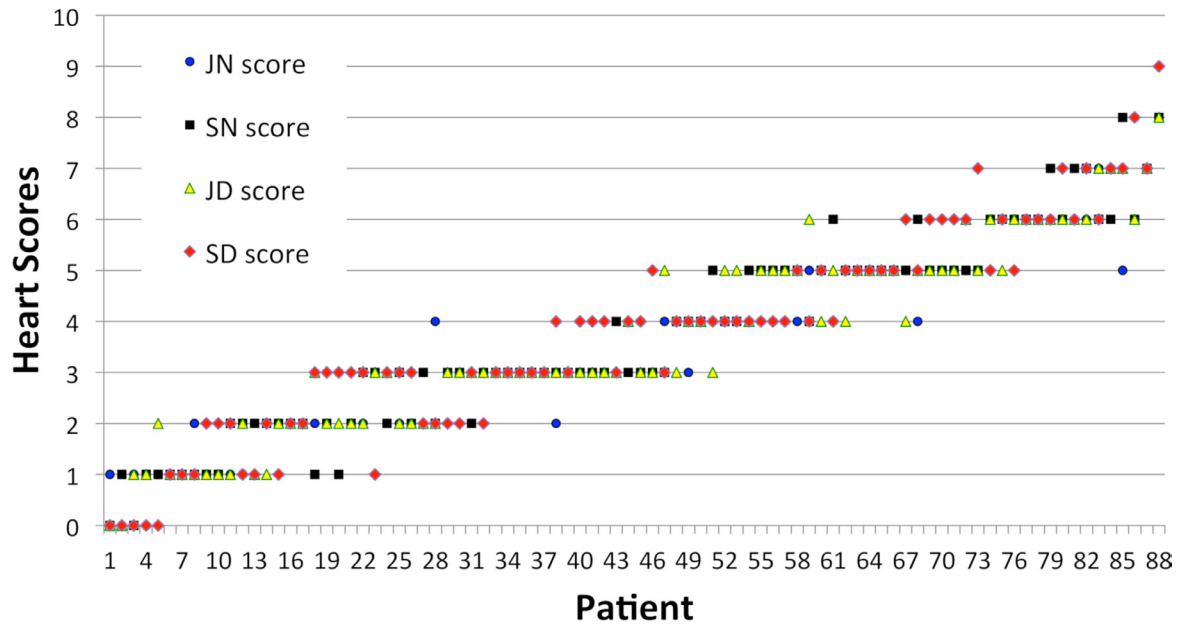


Figure 2 History, ECG, Age, Risk Factors and Troponin (Heart) Score for each grade of rater (patients ordered by average score). JD, junior doctor; JN, junior nurse; SD senior doctor; SN, senior nurse.

The accuracy of a chest pain history is potentially influenced by many factors including gender,²¹ language, culture²² and clinician variation. Backus, one of the original developers of the HEART Score (Backus B, 2015 personal communication June 20), found that unless directly prompted by a case report form clinicians only recorded about 50% of the relevant information required for a thorough chest pain history. The lack

of determinative power of the chest pain history is corroborated by a number of studies that have questioned the value of unstructured clinician assessment. Carlton *et al*²³ found that when physicians were asked to decide whether a patient's chest pain was 'typical' or 'atypical' based on the history and a non-diagnostic ECG the determination of AMI was limited with a receiver operating curve (ROC) of 0.54 (95% CI 0.40 to

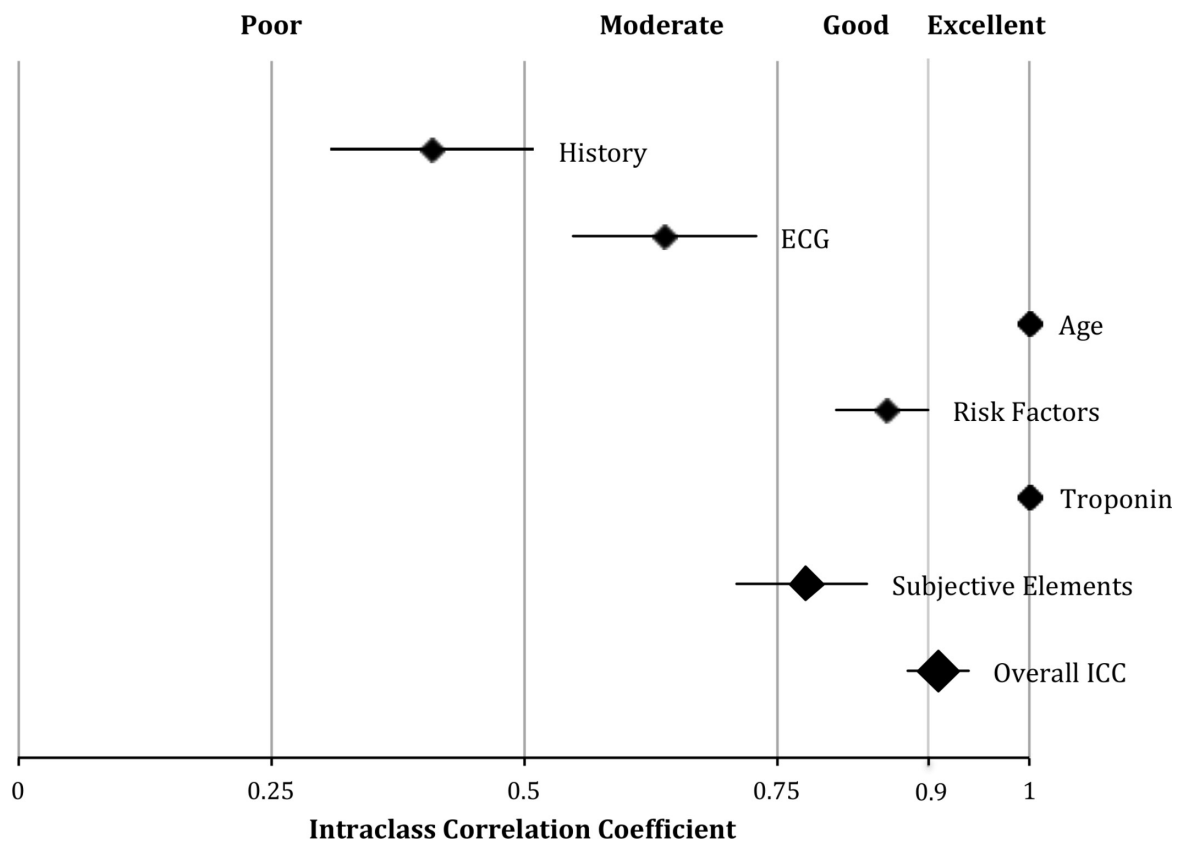


Figure 3 Intra-class correlation coefficients (ICCs) for the History, ECG, Age, Risk Factors and Troponin score.

0.67). Body *et al*²⁴ found an improved ROC of 0.76 (95% CI 0.7 to 0.82) when chest pain history was classified as either 'definitely not', 'probably not', 'not sure', 'probably' and 'definitely'. Both studies suggested that while increased clinician experience correlated with improved chest pain assessment unstructured chest pain assessment cannot be relied on in isolation to 'rule in' or 'rule out' the diagnosis recommending that chest pain assessment should incorporate the use of risk stratification tools.

However, clinicians still rely heavily on a description of the chest pain to help filter the list of potential differential diagnoses and then decide which decision aid to apply. The unique problem of the HEART Score is that while the chest pain history constitutes 20% of the final score it is also crucial in deciding whether the pain is potentially ischaemic and thus the appropriate tool to use in the first place. This means that even if, as this study shows, the HEART Score can be reproduced reliably despite the history, the problem of correctly selecting the patients on which to perform the score remains.

While this study does not, unfortunately, address the question of initial patient selection, the index diagnosis of myocardial infarction in this cohort was 9/88 (10.2%), which was not dissimilar to that of the studies by Body (17.7%) and Carlton (12.5%). Given that the researchers represented a broad range of clinical experience and included a nurse, two junior and one senior doctor in the recruitment process, this may suggest that the initial selection of patients on whom to perform the HEART Score need not necessarily be a senior doctor. Nevertheless, this study highlights how complex and subjective the assessment of chest pain history is and should remind clinicians of the importance of thorough and comprehensive chest pain assessment as well as of the dangers of over-reliance on clinical judgement alone.

This study used high-sensitivity troponin with sex-specific cut-offs both for the calculation of the HEART Scores and in the determination of the diagnosis of AMI. This is a deviation from the original HEART studies,¹⁻³ but the substitution of high-sensitivity troponin for conventional troponin has been shown to be both safe and evidence based.²⁵ The use of sex-specific cut-offs used in this study was in line with the manufacturer's recommendations and was locally validated prior to the commencement of the study.

The inclusion of both doctors and nurses in this study was predicated on the assumption that if the interoperator reliability was shown to be good then the HEART Score could serve as the basis of a 'common language' with which to communicate chest pain risk within the ED. As such, the findings of the study are encouraging because it suggests that the HEART Score 'language' is, indeed, sufficiently robust to allow nurses to potentially risk stratify patients early on in their ED journey. Whether early risk stratification by nurses would confer any health resource allocation benefits in terms of expedited discharge for 'low risk' patients is still unclear. In the stepped wedge trial by Poldevaart *et al*,⁶ the introduction of the HEART Score did not result in any significant differences in early ED discharge compared with usual practice. However, Poldevaart suggested that this finding may have been due to clinicians failing to follow the dispositional recommendations made for each of the three risk categories, namely early discharge for low risk, observation and further testing for the intermediate group and admission for the high-risk category. Clearly, further research is required to ascertain if the inclusion of nurses as chest pain risk stratifiers early on in the patient journey would ultimately result in improved safety and efficiency gains for EDs.

This study contrasts with previous attempts to measure the interoperator characteristics of the HEART Score in the method of defining the risk categories. Plewa⁷ and Mahler^{8,9} dichotomised the HEART Score into low risk (HEART Score of '3' or less) versus high risk (HEART Score more than '3') which is probably an oversimplification given that a significant percentage of patients fall into an intermediate category of risk, requiring further observation, serial ECG and troponin testing. In the original studies by Backus *et al*,¹⁻³ 'intermediate risk' was defined as a HEART Score of '4' to '6'. This study, by contrast, not only measured the interoperator agreement across Backus' three, originally specified risk categories: low, intermediate and high but also included nurses in the four categories of rater.

A further strength of this study was that it was the first prospective analysis of interoperator agreement for the HEART Score lending greater validity to the results. This study was carried out in a typical, inner city, secondary care ED which is likely to have similarities to others throughout the UK and the world, therefore increasing the generalisability of the findings.

LIMITATIONS

While it is not possible to define the impact that the educational programme had on the agreement of the aggregate HEART Scores, it would be reasonable to assume that it did have an overall, positive effect on interoperator agreement. However, the educational programme did not teach clinicians how to appropriately identify that cohort of chest pain patients with suspected acute coronary syndrome (ACS) on whom the HEART Score's application would be appropriate. Whether all grades of doctor and nurse could apply the HEART Score to the correct patient cohort is still unanswered by this paper.

Another limitation of this study as regards the 'history' component was the potential for miscommunication and subsequent under-reporting of the history in patients who were not native English speakers. One of the exclusion criteria for this study was lack of ability to speak English or absence of a family member to translate and this may have resulted in a recruitment bias towards subjects with better English proficiency. This is doubtful, however, given that a substantial percentage of the patients recruited (43%) were not first-language speakers. Nevertheless, the lack of a standardised approach to advocacy and translation for the recruited patients with limited English could have contributed to the poor reliability statistic for the history and future studies should address this methodological shortcoming.

Unfortunately there was no randomisation in the order in which clinicians saw the patients. There was a trend towards junior and senior nurses carrying out the first HEART Score as part of the initial assessment process and this may have led to history 'priming' by the time the patient was interviewed by the fourth clinician. There were also some clinicians in each group who performed more HEART Scores than others. These clinicians could have been more engaged with the study, better versed in how the HEART Score worked and this may have resulted in a degree of selection bias.

Finally, the study did not enrol the 120 patients initially calculated on the sample size estimation. While the desired CI width used for the sample size calculation was achieved, this was likely due to the fact that the actual ICC was higher than the one initially used in the sample size estimation. It is known that ICCs that are very high or low tend to have less variability and thus narrower CIs.²⁶

Future research

Questions remain about the reliability of the HEART Score that could be answered by future studies. While this study confirms that both doctors and nurses can reliably perform the HEART Score, this study did not address the interoperator agreement between clinicians within each of the four categories used in this study.

The healthcare profession is increasingly using information technology in the form of electronic patient records and decision support software. In the hospital where this study was conducted, all patient records are electronic and translating the HEART Score into an electronic form with built in checklists, particularly for the 'history' component, would be possible. How such an innovation would affect reliability, consistency and the selection of patients on whom to calculate the HEART Score or any other risk stratification tools is an area of direct relevance to emergency medicine and could form the focus of future research.

CONCLUSION

There is a high degree of reproducibility in HEART Score assessment between two grades of seniority of doctor and two grades of seniority of nurse. The closest agreement in a subjective variable was risk factors and the poorest was history.

Acknowledgements The authors acknowledge and thank the doctors and nurses of the Homerton University Accident and Emergency Department for their participation, enthusiasm and hard work during the conduct of this study. In addition, they thank Christine Mitchell-Inwang for her assistance in navigating the IRAS system for a first-time user and providing local research resources during the data collection phase.

Contributors WGNP was the lead author of this paper and generated the initial idea and subsequent design and planning for the study. WN wrote the IRAS application for the research ethics committee and, once approved was the chief investigator for the study; recruited the majority of the patients to the study and was responsible for the write up. DW helped with the planning of the study and was responsible for calculating the statistics used in this article. In addition, DW helped with the editing of the final document. SG was the supervisor for the thesis that was submitted as a master dissertation to the University of Sheffield and provided expertise in terms of study design and planning of the study and then helped with the reporting of the study in both this article and the aforementioned dissertation. ACER designed and delivered the teaching video that is referenced in the article and was one of the investigators that helped to recruit patients to the study. SJG was responsible for recruiting patients and analysing the data set for the outcome of acute coronary syndrome. TH helped refine the patient protocol for the initial study after an initial rejection from the research ethics committee. He was involved in helping to clarify the outcomes of interest and later in the editing of the manuscript.

Funding This work was supported by the Royal College of Emergency Medicine: grant number G/2014/3. Receipt of a Royal College of Emergency Medicine Grant of £4000 to fund this research project and collaboration with one of the original developers of the HEART Score to develop a teaching video for doctors and nurses on how to use the tool.

Competing interests Dr Niven reports grants from Royal College of Emergency Medicine, during the conduct of the study; and I have collaborated with Dr Barbra Backus, one of the original developers of the HEART Score, in creating a teaching video for the HEART Score.

Patient consent Not required.

Ethics approval NRES Committee London - City and East.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement There are no additional unpublished data from this study.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which

permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

REFERENCES

- 1 Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. *Neth Heart J* 2008;16:191–6.
- 2 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;168:2153–8.
- 3 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;9:164–9.
- 4 Antman EM, Cohen M, Bernink PJLM, et al. The TIMI risk score for unstable Angina/Non–ST elevation MI. *JAMA* 2000;284:835–42.
- 5 Fox KA, Dabbous OH, Goldberg RJ, et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). *BMJ* 2006;333:1091–4.
- 6 Poldervaart JM, Reitsma JB, Koffijberg H, et al. The impact of the HEART risk score in the early assessment of patients with acute chest pain: design of a stepped wedge, cluster randomised trial. *BMC Cardiovasc Disord* 2013;13:77.
- 7 Plewa MC, 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):S185–6.
- 8 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 [Epub ahead of print 6 Mar 2018].
- 9 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.
- 10 Mahler SA, Hiestand BC, Goff DC, 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;10:128–33.
- 11 Goodacre S, Cross E, Arnold J, et al. The health care burden of acute chest pain. *Heart* 2005;91:229–30.
- 12 Carlton EW, Khattab A, Greaves K. Beyond triage: the diagnostic accuracy of emergency department nursing staff risk assessment in patients with suspected acute coronary syndromes. *Emerg Med J* 2016;33:99–104.
- 13 Robertson ACE. 2016. HEART Score study: testing the inter-operator reliability of the HEART Score (video file). <https://www.youtube.com/watch?v=w799yRm7hYg> (Cited 12 Sept 2006).
- 14 Thygesen K, Alpert JS, Jaffe AS, et al. Third universal definition of myocardial infarction. *Circulation* 2012;126:2020–35.
- 15 Anon. 2018. Intraclass correlation and variance component methods. <http://www.john-uebersax.com/stat/icc.htm> (cited 11 Apr 2018).
- 16 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
- 17 Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. *J Am Stat Assoc* 1952;47:583–621.
- 18 Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Stat Med* 2001;20:3205–14.
- 19 R Development Core Team. 2015. R: a language and environment for statistical computing: R found stat comput. 1:409 <http://www.r-project.org/nhttp://www.r-project.org>
- 20 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
- 21 Milner KA, Funk M, Richards S, et al. Gender differences in symptom presentation associated with coronary heart disease. *Am J Cardiol* 1999;84:396–9.
- 22 Farmer SA, Roter DL, Higginson IJ. Chest pain: communication of symptoms and history in a London emergency department. *Patient Educ Couns* 2006;63(1–2):138–44.
- 23 Carlton EW, Than M, Cullen L, et al. 'Chest pain typicality' in suspected acute coronary syndromes and the impact of clinical experience. *Am J Med* 2015;128:1109–16.
- 24 Body R, Cook G, Burrows G, et al. Can emergency physicians 'rule in' and 'rule out' acute myocardial infarction with clinical judgement? *Emerg Med J* 2014;31:872–6.
- 25 Willems MN, van de Wijngaart DJ, Bergman H, et al. Addition of heart score to high-sensitivity troponin T versus conventional troponin T in risk stratification of patients with chest pain at the coronary emergency rooms. *Neth Heart J* 2014;22:552–6.
- 26 Ionan AC, Polley MY, McShane LM, et al. Comparison of confidence interval methods for an intra-class correlation coefficient (ICC). *BMC Med Res Methodol* 2014;14:121.