NEJM AI

NEJM AI 2025;2(11)
DOI: 10.1056/AIcs2401170

CASE STUDY

A Case Study of AI-Enabled Software as a Medical Device Cleared by the FDA for Assessing Hemorrhage Risk Index (APPRAISE-HRI) after Trauma

Andrew Frock, B.S.,[1,2] Jeffrey T. Robbins, B.S.,[1,2] Francisco G. Vital-Lopez, Ph.D.,[1,2] Valmik Desai, M.S.,[1,2] Gheorghe Doros, Ph.D., M.B.A.,[3] Barry E. Sands, B.S.B.M.E., M.B.A.,[3] Arunkumar Prabhakaran, Pharm.D., M.S.-R.A.,[3] Christopher Nemeth, Ph.D., C.H.F.P.,[4] Gregory T. Rule, M.S.E., P.E.,[4] Jason L. Sperry, M.D., M.P.H.,[5] Francis X. Guyette, M.D., M.P.H.,[6] Stephen R. Wisniewski, Ph.D.,[7] Ernest E. Moore, M.D.,[8] Martin Schreiber, M.D.,[9] Bellal Joseph, M.D.,[10] Chad T. Wilson, M.D.,[11] Bryan Cotton, M.D.,[12] Daniel Ostermayer, M.D.,[13] Brian G. Harbrecht, M.D.,[14] Mayur B. Patel, M.D., M.P.H.,[15] Suzanne Tamang, Ph.D.,[16,17] Sanjay Malunjkar, B.E.,[17] David A. Spain, M.D.,[18] Andrew T. Reisner, M.D.,[19] Jonathan D. Stallings, Ph.D.,[20] and Jaques Reifman, Ph.D.[1]

Received: December 2, 2024; Revised: July 11, 2025; Accepted: August 25, 2025; Published: October 16, 2025

Abstract

Hemorrhage is the leading cause of preventable death on the battlefield, yet combat medics lack clinical decision support systems to help stratify hemorrhage risk in trauma casualties. We previously trained the Automated Processing of the Physiological Registry for Assessment of Injury Severity — Hemorrhage Risk Index (APPRAISE-HRI) software to associate patterns in vital signs (heart rate and blood pressure) collected from trauma patients with three HRI levels: I (low), II (average), or III (high). To independently validate APPRAISE-HRI and obtain U.S. Food and Drug Administration (FDA) clearance, we collected trauma registry and continuous vital sign data from 5895 trauma patients (543 with hemorrhagic injuries and 5352 controls) in an emergency department or during prehospital transport to one of eight medical centers. The study outcome was hemorrhagic injury, defined by documented injuries and blood transfusion. Using the likelihood ratio to assess the ability of APPRAISE-HRI to stratify hemorrhage risk, we found that hemorrhagic patients were 6.88 times as likely as controls to be at level III, strongly suggesting the presence of hemorrhage at this level. Similarly, hemorrhagic patients were 0.18 times as likely as controls to be at level I, suggesting the absence of hemorrhage at this level. Hemorrhagic patients were almost as likely as controls to be at level II (0.70 times as likely). Subsequently, the U.S. Department of Defense obtained FDA 510(k) clearance for the artificial intelligence–enabled APPRAISE-HRI Class II device (K233249), the first software as a medical device approved for assessing hemorrhage risk in trauma patients, allowing for triage and identification of casualties who need immediate attention and evacuation. (Funded by the U.S. Army Medical Materiel Development Activity and the Combat Casualty Care Program Area Directorate (CCCPAD) of the U.S. Army Medical Research and Development Command (USAMRDC), Fort Detrick, MD and others.)

The author affiliations are listed at the end of the article.

Jaques Reifman can be contacted at jaques.reifman.civ@health.mil or at Department of Defense Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research and Development, Medical Research and Development Command, attn: FCMR-TT, 504 Scott Street, Fort Detrick, MD 21702-5012.

Introduction

Hemorrhage remains the leading cause of preventable death on the battlefield.[1-5] Over the last decade, several artificial intelligence (AI)–enabled clinical decision support systems (CDSSs) have been proposed to triage trauma casualties for lifesaving interventions.[6-9] However, until now, no CDSS has been cleared by the U.S. Food and Drug Administration (FDA) for assessing hemorrhage risk in patients after trauma.

In 2024, the U.S. Department of Defense (DoD) obtained FDA 510(k) clearance for the Automated Processing of the Physiological Registry for Assessment of Injury Severity — Hemorrhage Risk Index (APPRAISE-HRI)[10,11] as a Class II device (K233249). The APPRAISE-HRI is a software designed to help military health care providers in triaging service members for hemorrhage risk after a physically traumatic event and stratifying casualties who need immediate attention and emergency evacuation from those who are at low risk for hemorrhage. The APPRAISE-HRI is also the first AI-enabled software as a medical device (SaMD) cleared by the FDA from the DoD.[12] SaMD, which is defined as software intended for medical purposes that is independent of a hardware medical device, is becoming increasingly important and common in health care.[13] APPRAISE-HRI met this definition and required FDA oversight because it processed data from a signal-acquisition system.[14] This case study describes the process of obtaining FDA clearance and the performance characteristics of the device.

Methods

STUDY DESIGN AND OVERSIGHT

Obtaining FDA approval through the 510(k) clearance pathway relied on the demonstration of substantial equivalence with a predicate device, which continuously monitors electrocardiogram (ECG) waveforms to identify patients with hemodynamic instability.[15] The FDA also required an independent clinical validation of the APPRAISE-HRI using a "prospective retrospective" study design, where we prospectively validated its performance on two independent, retrospectively collected samples of real-world trauma patient data not used for training:[10] an in-hospital study at the emergency department (ED) of the Stanford University Hospital (Stanford) and a prehospital study from the Linking Investigations in Trauma and Emergency Services (LITES) Consortium.[16,17]

The U.S. Army Institutional Review Board (IRB) and the Office of Human Research Oversight (OHRO), Fort Detrick, MD, provided a determination for the use of deidentified data from these two studies to establish our study protocol. The Stanford and LITES studies received approval from their respective IRBs under a waiver of consent and from OHRO.

STUDY PROTOCOL

From Stanford, the APPRAISE-HRI study obtained clinical records from patient beds in the ED by linking various electronic medical records within the hospital. The provided deidentified records included demographics, clinical procedures and outcomes, and continuous vital sign data (ECG-derived heart rate [HR] at 1 Hz, and cuff-based systolic blood pressure [SBP] and diastolic blood pressure [DBP] at multiminute intervals) collected during the first hour in the ED.

LITES is an ongoing multicenter observational study of moderate to severe traumatic injuries in the United States.[16,17] The University of Pittsburgh leads the study, with vital sign data collected during ground- or air-ambulance transport from the point of injury to eight receiving hospitals (see the Supplementary Appendix). From LITES, we obtained deidentified clinical records similar to those from Stanford, with ECG- or pulse oximeter–derived HR and cuff-based SBP and DBP, each recorded at varying multiminute intervals.

The data analyses involved trauma patients who met demographic and clinical eligibility requirements, including patients between 18 and 90 years of age who had penetrating or blunt injuries, 24-hour packed red blood cells (PRBCs) transfusion information, and at least one of the following: clinical notes, *International Statistical Classification of Diseases and Related Health Problems,* Tenth Revision (ICD-10) codes, or Abbreviated Injury Scale (AIS) codes. Inclusion and exclusion criteria are available in the Supplementary Appendix.

THE APPRAISE-HRI SYSTEM

The APPRAISE-HRI software resided in an Android smartphone and continuously pulled and processed vital sign data (HR, SBP, and DBP) from a ZOLL Propaq M monitor via Bluetooth to generate an output every 1 minute (Fig. 1A). The output consisted of one of three possible HRI levels: low (I), average (II), or high (III). The software consisted of three modules previously described in Stallings et al.[10] (Fig. 1B), which we fixed before this independent validation. Briefly, the first module assessed the quality of the vital sign data every 1 minute to provide internal controls and

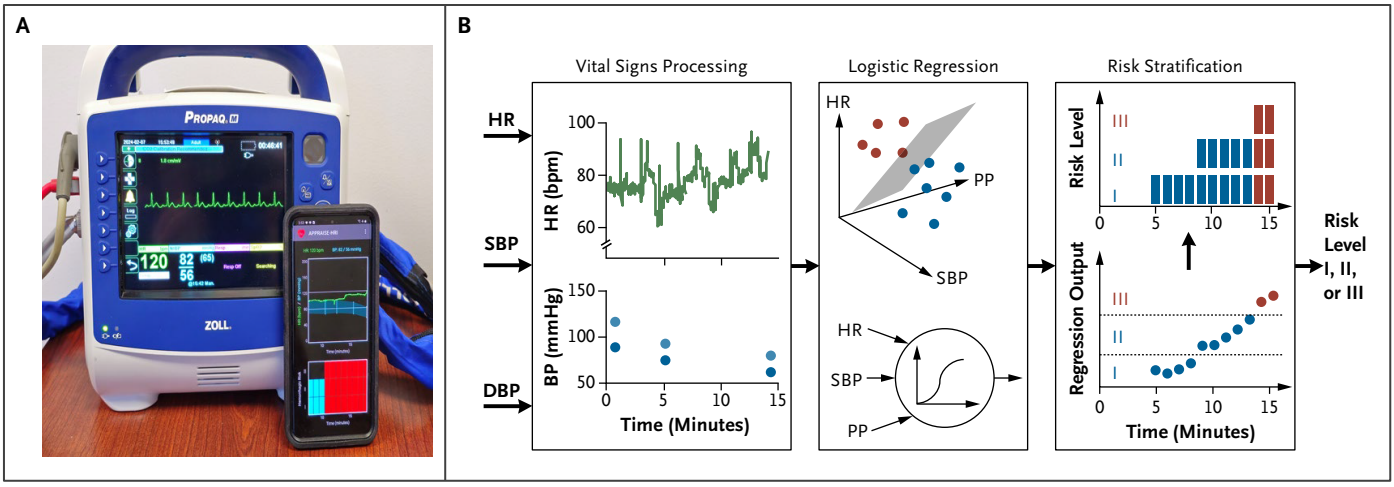

Figure 1. APPRAISE-HRI Software as a Medical Device.

Panel A shows the Automated Processing of the Physiological Registry for Assessment of Injury Severity — Hemorrhage Risk Index (APPRAISE-HRI) software as a medical device (SaMD) cleared by the U.S. Food and Drug Administration (FDA), which resides in a smartphone using the Android operating system version 9 or higher. The SaMD continuously pulls and processes data from the FDA-cleared ZOLL Propaq M vital sign monitor to generate an output every 1 minute. The SaMD displays two graphs as a function of time, one (top) showing the vital sign values as displayed by the monitor and the other (bottom) showing the output of the device, that is, the HRI levels (I and II in blue and III in red). APPRAISE-HRI consists of three modules, as shown in Panel B. Every 1 minute, the vital sign processing module identifies and discards invalid heart rate (HR) and blood pressure (BP) values and computes the pulse pressure (PP; the difference between systolic BP [SBP] and diastolic BP [DBP]). The artificial intelligence algorithm in the second module takes the valid vital signs as input and, through a multivariate logistic regression model, generates an output corresponding to the likelihood of hemorrhage. Lastly, the risk stratification module uses two thresholds established during model training to categorize the hemorrhage risk level for the trauma patient at the current time. Adapted from Stallings et al.[10]

assurance that the downstream algorithm only used valid data as input. The second module consisted of a multivariate logistic regression model trained to map the three vital signs into a continuous output ranging from 0.0 (control) to 1.0 (hemorrhagic). Finally, the third module provided hemorrhage risk stratification based on two fixed cutoff values on the output of the logistic regression model, separating the three risk levels.[10]

OUTCOME

The study outcome was hemorrhagic injury defined by documented records from at least one of three sources (i.e., clinical notes, ICD-10, or AIS codes) and transfusion of one or more units of PRBCs within 24 hours of hospital admission. Documented records included hemorrhage control procedures or injuries consistent with a hemorrhagic outcome. We categorized all other trauma patients as controls.

DEVICE ASSESSMENT

We assessed the diagnostic usefulness of the software by performing a primary analysis, where we computed the likelihood ratio[18] (LR) of hemorrhagic injury for APPRAISE-HRI output levels I, II, and III based on its first output for each patient. The FDA concurred with the use of LR as the primary statistic to assess device effectiveness because APPRAISE-HRI has three possible outputs and LR is a powerful measure of the accuracy of a diagnostic test; it indicates how much the test results raise (or lower) the probability of disease (i.e., the posttest probability) compared with the prevalence of the disease (i.e., the pretest probability). Given the LR and the prevalence of hemorrhage in the population, we can estimate the posttest probability. We also performed a secondary analysis where we computed the LR over time and for three population subgroups (i.e., age, mode of injury, and study site).

USABILITY TESTING

As part of the FDA clearance process, we performed formative and summative usability tests of the APPRAISE-HRI, which the Army's OHRO determined to be exempt from regulatory oversight. Per the FDA's recommendations, we followed their guidance document to establish the usability framework for the tests.[19] This document provided the overarching principles — rather than prescribing specific validated models — to guide human factors and usability engineering processes, maximizing the likelihood that the device

is safe and effective for its intended users, uses, and environments. Using these principles, we constructed the tests to characterize the intended user population, identify and assess risks of misuse, and gauge the usability of the device. For both tests, we recruited DoD medics (the intended end users) stationed at Fort Detrick through word of mouth. We conducted the formative test at the beginning of the process to determine usability requirements and functionalities, which involved open-ended and five-point Likert scale questionnaires, as well as a review of a mock-up device interface. After the development of the device, we conducted the summative test to determine whether the device met the medics' needs and whether the outputs were easy to interpret. This evaluation involved answers to questionnaires and a cognitive walkthrough, including multiple patient scenarios.

STATISTICAL ANALYSIS

Assuming the same performance of the APPRAISE-HRI algorithm in the independent validation sample as in the sample used to train the algorithm,[10] we calculated a sample size of 2400 trauma patients (including 400 hemorrhagic patients). This calculation was based on the following success criteria submitted to the FDA prior to data analysis: (1) the lower bound of the 95% confidence interval (CI) for the LR of hemorrhagic injury in HRI level III was greater than 2.00; (2) the upper bound of the 95% confidence interval for the LR in HRI level I was less than 0.60; and (3) the 95% confidence intervals for the three LRs did not overlap with each other.

Results

PATIENT CHARACTERISTICS

From Stanford, we obtained data from 1649 consecutive patients, of whom 1464 satisfied the inclusion criteria (Fig. 2). From LITES, we obtained data from 9332 consecutive patients, of whom 4431 satisfied the inclusion criteria (Fig. 3). Table 1 shows the characteristics of the 5895 patient records and their categorization into hemorrhagic (543, or 9.2%) or control patients.

OUTCOMES

Table 2 shows the results of the primary analysis, including the number of hemorrhagic and control patients, LR (95% confidence interval), and posttest probability (95% confidence interval) for each of the three output levels. An LR of 1.00 indicates that hemorrhagic and control patients are equally likely (i.e., they have the same probability) to be at a given risk level, and as the LR deviates from 1.00, the ability to differentiate between the two groups at that level increases. Based on the LR results, hemorrhagic patients were 6.88 times (95% CI, 6.04 to

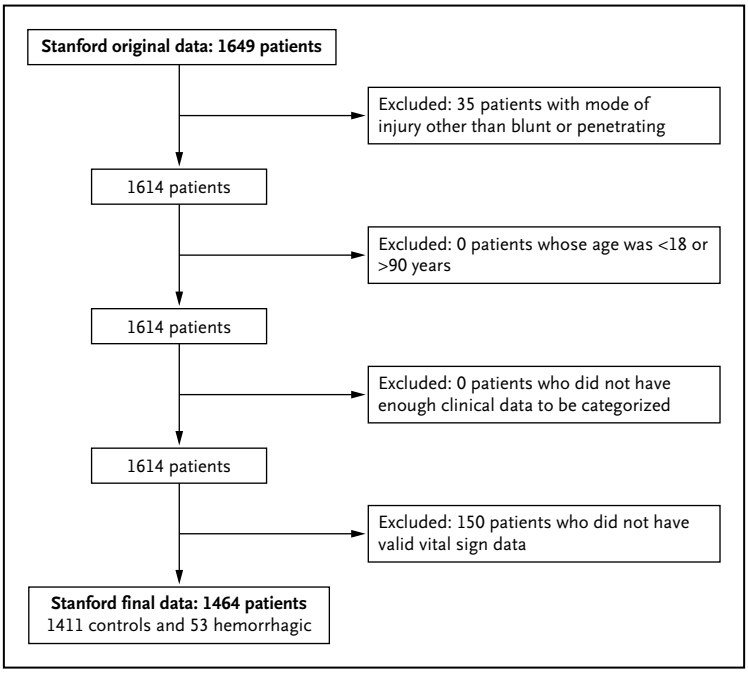

Figure 2. Step-By-Step Exclusion Process of Trauma Patients from the Stanford Study, Where We Collected the Data between August 2020 and August 2021.

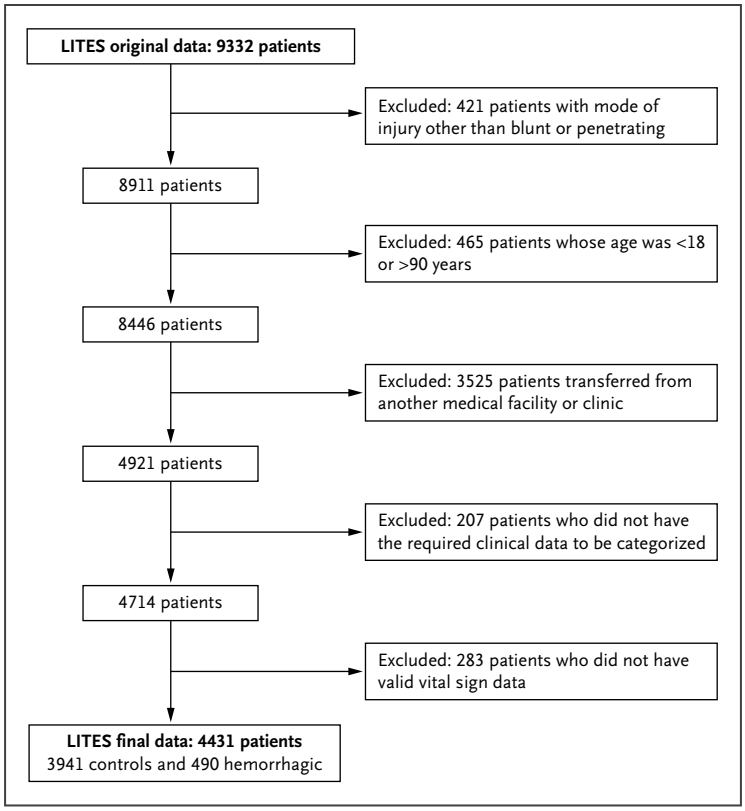

Figure 3. Step-By-Step Exclusion Process of Trauma Patients from the LITES Study, Where We Collected the Data between January 2017 and June 2019.

LITES denotes Linking Investigations in Trauma and Emergency Services.[16,17]

7.84) more likely to be at level III than control patients, and considerably less likely to be at level I (LR of 0.18; 95% CI, 0.12 to 0.26), indicating the effectiveness of APPRAISE-HRI to differentiate between patients at these levels. Hemorrhagic patients were almost as likely as controls to be at level II (0.70 times as likely). In terms of posttest probability, compared with the pretest probability (9.2%), the LR-estimated posttest probability for hemorrhagic injury was substantially higher for level III (41.0%; 95% CI, 38.0 to 44.3%) and lower for level I (1.8%; 95% CI, 1.2 to 2.5%), indicating strong risk stratification. For level II (a gray zone), the posttest probability (6.6%; 95% CI, 6.1 to 7.2%) was close to the pretest probability. We repeated the primary analysis for pretest probabilities (i.e., prevalence of hemorrhage) ranging from 1 to 90% (Table S1). Consistently, a level III categorization increased the posttest probability of hemorrhage relative to the pretest probability, whereas a level I categorization consistently reduced it, showing that APPRAISE-HRI invariably shifted the posttest probabilities in the correct direction, regardless of the value of the pretest probability.

To assess the ability of the device to rule in hemorrhage in trauma patients at level III and rule out hemorrhage in trauma patients at level I, we performed a dichotomized analysis (Tables S2 and S3). For the rule-in evaluation of HRI level III versus the combined levels I or II, hemorrhagic patients were 6.88 times more likely to be at level III than control patients (i.e., they had a positive LR of 6.88; 95% CI, 6.04 to 7.84) and about half as likely to be in levels I or II than control patients (i.e., they had a negative LR of 0.56; 95% CI, 0.51 to 0.60). APPRAISE-HRI placed about half of the hemorrhagic patients at level III (sensitivity of 48.1%; 95% CI, 43.9 to 52.2%) and a small percentage (7.0%) of the control patients (specificity of 93.0%; 95% CI, 92.3 to 93.7%). For the rule-out evaluation of HRI level I versus the combined levels II or III, hemorrhagic patients were considerably less likely to be at level I than control patients (negative LR of 0.18; 95% CI, 0.12 to 0.26) and more likely to be at levels II or III than control patients (positive LR of 1.27; 95% CI, 1.24 to 1.31). APPRAISE-HRI placed most of the hemorrhagic patients at levels II or III (sensitivity of 95.4%; 95% CI, 93.6 to

Table 1. Characteristics and Categorization of the Patients Analyzed in the Study.

Patient Categorization or Characteristic	Overall (N=5895)	Stanford (N=1464)	LITES (N=4431)
Categorization* — n			
Hemorrhagic	543	53	490
Control	5352	1411	3941
Characteristic			
Age (years)			
Mean±SD	49±20 (5895)	52±21 (1464)	48±20 (4431)
Median (Q1, Q3)	48 (31, 65)	53 (33, 70)	46 (30, 63)
(Minimum, maximum)	(18, 90)	(18, 89)	(18, 90)
Age ≥35 years — %			
No	31.5% (1858/5895)	27.0% (395/1464)	33.0% (1463/4431)
Yes	68.5% (4037/5895)	73.0% (1069/1464)	67.0% (2968/4431)
Sex — %			
Male	67.4% (3971/5895)	61.1% (895/1464)	69.4% (3076/4431)
Female	32.6% (1920/5895)	38.6% (565/1464)	30.6% (1355/4431)
Unknown	0.1% (4/5895)	0.3% (4/1464)	0.0% (0/4431)
Race† — %			
Asian	4.0% (233/5822)	12.8% (188/1464)	1.0% (45/4358)
Black	10.6% (627/5822)	5.7% (83/1464)	12.3% (544/4358)
Native American	0.1% (7/5822)	0.1% (2/1464)	0.1% (5/4358)
Pacific Islander	0.4% (25/5822)	1.2% (18/1464)	0.2% (7/4358)
White	67.9% (4003/5822)	47.5% (696/1464)	74.6% (3307/4358)
Unknown	0.7% (43/5822)	2.7% (39/1464)	0.1% (4/4358)
Other	15.0% (884/5822)	29.9% (438/1464)	10.1% (446/4358)
Length of vital sign recording (seconds)			
Mean±SD	2285±1590 (5895)	3024±476 (1464)	2040±1746 (4431)
Median (Q1, Q3)	2100 (1440, 2981)	3113 (2827, 3344)	1794 (1320, 2400)
(Minimum, maximum)	(120, 56,640)	(699, 3660)	(120, 56,640)
Mode of injury — %			
Blunt	90.8% (5354/5895)	95.1% (1394/1464)	89.4% (3960/4431)
Penetrating	9.2% (541/5895)	4.8% (70/1464)	10.6% (471/4431)

* We categorized patients as having a hemorrhagic injury if they had transfusion of one or more units of packed red blood cells within 24 hours of hospital admission and documented records indicative of hemorrhage-control procedures (e.g., packing or suture of an artery) or injuries consistent with a hemorrhagic outcome (e.g., major laceration of internal organs or vessels or hemothorax). We categorized all other trauma patients as controls. For patients who did not survive at least 24 hours, we used blood transfusion information up to the time of death. LITES denotes Linking Investigations in Trauma and Emergency Services; Q1, first quartile; Q3, third quartile; and SD, standard deviation.
† Race obtained from patient electronic medical records.

Table 2. Primary Outcome Using the Device's First Output for Each Patient.*

HRI Level	Hemorrhagic, N	Control, N	Total, N	Likelihood Ratio (95% CI)†	Posttest Probability % (95% CI)
I	25	1347	1372	0.18 (0.12 to 0.26)	1.8 (1.2 to 2.5)
II	257	3631	3888	0.70 (0.63 to 0.76)	6.6 (6.1 to 7.2)
III	261	374	635	6.88 (6.04 to 7.84)	41.0 (38.0 to 44.3)
Total	543	5352	5895	–	–

* A useful diagnostic test would ideally have a low likelihood ratio (LR less than 1.00) or a high LR (greater than 1.00). As the LR approaches 1.00, the utility of the test decreases to zero because the posttest probability would be equal to the pretest probability.[18] Hemorrhage risk index (HRI) level I is enriched with control patients, while HRI level III is enriched with hemorrhagic patients. CI denotes confidence interval.
† Confidence intervals are based on Monte Carlo simulations of 10,000 samples with replacement from the total population.

97.1%); however, it only placed a quarter of the control patients at level I (specificity of 25.2%; 95% CI, 24.0 to 26.3%).

For the population subgroups in the secondary analysis, we assessed the LR of the first output for each patient by age (under 35 years vs. 35 years and over), mode of injury (blunt vs. penetrating), and study site (Stanford vs. LITES). With three exceptions (out of 18 tests), these secondary analyses also met the device's success criteria (Tables S4–S6). The exceptions were for the age under 35 years subgroup and the penetrating injury subgroup, where the 95% confidence intervals for the LRs at HRI levels I and II overlapped. In the penetrating injury subgroup, the upper bound of the 95% confidence interval for the LR at HRI level I was greater than or equal to 0.60. Thus, regarding the mode of injury, the performance in the blunt injury subgroup was consistent with the primary analysis results, whereas APPRAISE-HRI did not meet the success criteria in the penetrating injury subgroup. In addition, we repeated the secondary analysis for the dichotomized outcomes discussed above, which allowed us to assess the device using additional statistical metrics (Tables S7–S12). For the analysis over time, to determine if later data recordings resulted in changes in test performance, we assessed the LRs for each of six consecutive outputs over time and for the last output of each patient record. Although the number of patient records decreased with time because patients left the ED or arrived at the receiving hospital, each analysis met the device's success criteria (Table S13). We also assessed the stability of the HRI outputs over time by determining whether patients switched levels compared with their first output. By the last output, 66.0% retained the same level, and only 0.6% changed from HRI I to III or from HRI III to I.

FORMATIVE AND SUMMATIVE TESTING

We enrolled a representative cross section of potential users, including five medics for the formative test, which allowed us to discover usability issues, and 15 medics for the summative test, where we assessed their ability to use and interpret the device results. Feedback from the formative test resulted in modifications to the graphical user interface shown in Figure 1A (e.g., plot size, color, and numerical values), and feedback from the summative study allowed us to confirm that medics found that it was relatively easy to determine a patient's hemorrhage risk and that the device was helpful, with 93.3% (14 out of 15) correctly interpreting the outputs of the device.

Discussion

This report describes the clinical and usability testing undertaken to obtain FDA clearance of the APPRAISE-HRI, the first SaMD for triage of trauma casualties for hemorrhage risk. The strengths of our study include the use of patient data collected from nine sites across the United States, a relatively large number of patients, and population subgroup analyses.[20] Our primary analysis showed that the device met the predefined LR success criteria, effectively and consistently stratifying trauma patients between hemorrhage risk levels. We found that the device output was stable, with 66.0% of the patients staying at the same assigned HRI level over time. It was rare (less than 1.0% probability) for a patient to increase or decrease by two levels of hemorrhage risk. End users found the device relatively easy to use and agreed on its utility for hemorrhage risk detection.

The secondary subgroup analyses showed that the performance of APPRAISE-HRI across age groups and study sites yielded similar trends as those of the overall study population in the primary analysis (Table 2, and Tables S4 and S6). However, for the mode of injury, while the performance of the blunt injury subgroup was consistent with that of the primary analysis, APPRAISE-HRI did not meet the success criteria in the penetrating injury subgroup (Table S5). Specifically, although the device was able to differentiate between patients at level III, it did not clearly distinguish between patients at levels I or II, because the confidence intervals for these levels overlapped. This is likely attributable to the small sample size in this subgroup, which reduced the precision of the estimates and led to a wide confidence interval for level I. As a result, the device's discriminatory capability within this subgroup is inconclusive. This analysis is relevant for two main reasons. First, the prevalence of hemorrhage was substantially higher in the penetrating injury subgroup (25.5%) than that of the blunt injury subgroup (7.6%). Second, penetrating trauma is the predominant mode of injury in combat, accounting for around 70.0% of battlefield wounds,[21] which is considerably higher than in our study (9.2%) and the civilian population (9.7%).[22]

The entire clearance process included two presubmissions, a 510(k) submission, and a resubmission. In the first presubmission, we provided a summary of the device, intended use and indications for use, proposed predicate device, and a series of questions. By far, the FDA's answers to our questions were the most useful because they provided specific

guidance on what to report in terms of algorithm development, device testing, and data analysis. In the second presubmission, we sought feedback on our 510(k) submission plan, which included the clinical validation protocol and the statistical analysis plan, including the success criteria (see Statistical Analysis). We completed the first 510(k) submission in 9 months, and within 60 days received the FDA's letter of Additional Information Request, including a detailed description of major and minor deficiencies of our submission. Most importantly, the letter consistently referenced the Special Controls of the Title 21 of the Code of Federal Regulations 870.2220 for Cardiovascular Monitoring Devices,[23] which provided specific guidance on what to report to address the identified deficiencies and ensured that the assessment of the device was consistent with "the intended use population and relevant use conditions in the intended use environment." Within 90 days, we addressed all the deficiencies and provided the final submission. Overall, it took 7 months to obtain FDA clearance from the initial 510(k) submission. Throughout the entire process, we had multiple face-to-face meetings and direct email communications with the FDA, which were very helpful. Being naive to the regulatory process, we hired an experienced FDA consultant who helped us navigate through the required comprehensive documentation and considerably expedited the approval process.

We learned several lessons in the clearance process that may help guide future applicants. As scientists, research on APPRAISE-HRI focused primarily on principles related to algorithm development and performance. We investigated different types of AI algorithms, combinations of vital signs,[24] and variations of the definition of the study outcome.[25] In sharp contrast, the FDA's primary focus is on clinical benefit (i.e., device effectiveness) and patient risk.[26] Initially, this balance between benefit versus risk was not part of our mindset. Through the clearance process, we learned how to categorize, quantify, and mitigate risk, including end-user misuse of the device or misinterpretation of the device's output; erroneous or unphysiological vital signs provided to the smartphone; software technical risks related to communication, computation, or display errors; and cybersecurity concerns related to data confidentiality and system integrity.

A better understanding of the importance of the device labeling (i.e., the user manual) would have allowed us to draft a more comprehensive document from the start. The original software[10] included robust methods for artifact rejection to control for invalid vital signs. In retrospect, a more detailed description of this functionality from the onset would have reduced the number of iterations with the FDA. Cybersecurity was a major concern because APPRAISE-HRI is a SaMD. Ensuring that the software only had access to hardware resources or services critical for its functionality allowed us to reduce potential cybersecurity vulnerabilities. Finally, it was imperative that the submission directly "connected the dots" and did not make inferences based on external information not included in the submitted documents (e.g., the technical specifications of monitors used to collect the vital sign data). To this end, interactive communications with the FDA were constructive.

Separate from what was submitted to the FDA, we compared diagnostic test characteristics of APPRAISE-HRI with the shock index (SI), defined as HR/SBP, which has been proposed as a marker for significant injury and critical bleeding in trauma patients.[27,28] We selected two commonly used SI cutoff values (a SI greater than 1.0 and a SI greater than 1.4)[29,30] for comparison against APPRAISE-HRI (level III vs. combined levels I or II). For the overall population, there was a nonsignificant trend favoring APPRAISE-HRI with higher sensitivity than SI greater than 1.0 (P=0.07), but similar specificity, whereas a SI greater than 1.4 better differentiated between hemorrhagic cases (positive LR of 15.22 [95% CI, 11.44 to 20.66] vs. 6.88 [95% CI, 6.04 to 7.84]) at the expense of identifying 60.0% fewer such cases at level III than APPRAISE-HRI (Table S14). The differences in test characteristics for the LITES cohort were similar to those of the overall population. In sharp contrast, the differences were more pronounced in the Stanford cohort, where APPRAISE-HRI had significantly higher sensitivity than both SI >1.0 and SI >1.4 (P<0.01). This is quite likely because we computed the APPRAISE-HRI and SI results for the Stanford cohort using raw data from the vital sign monitor, whereas we computed the results for the LITES cohort using medic-documented vital signs, which presumably involved filtering out spurious measurements.

The ability of the AI algorithm to use only valid HR and BP measurements allowed us to extract maximum information from these data, while obtaining a practical and effective solution for the pre-hospital environment. Reassuringly, the trends in these vital signs as the HRI levels increased from I to III (Table S15) are in alignment with the trends used by the American College of Surgeons to categorize classes of hemorrhage of increasing severity.[31] The deterioration in hemodynamic status, as assessed by the SI, also increased with higher HRI levels.

The major limitation of this study is that the independent validation of the APPRAISE-HRI did not involve a prospective side-by-side comparison of a medic's performance

with and without the device.[20] Such a prospective validation study should also provide additional insights into the device's performance in patients with penetrating injuries. However, we partially mitigated potential bias and overfitting concerns by using a sample of real-world trauma patients that considerably exceeded the sample size calculation and that was collected from nine geographically distinct sites.

In conclusion, as an FDA-cleared SaMD, the APPRAISE-HRI is now available for combat medics to triage U.S. service members for hemorrhage risk after a physically traumatic event and stratify casualties who need immediate attention and emergency evacuation from those who may not be at risk for hemorrhage.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

This work was supported by the U.S. Army Medical Materiel Development Activity and the Combat Casualty Care Program Area Directorate (CCCPAD) of the U.S. Army Medical Research and Development Command (USAMRDC), Fort Detrick, MD. The Henry M. Jackson Foundation was supported by the USAMRDC under contract number W81XWH20C0031. The Linking Investigations in Trauma and Emergency Services study was supported by the USAMRDC under contract number W81XWH16D0024 and managed by the CCCPAD. We also acknowledge financial support from the Medical Technology Enterprise Consortium, Research Project Award number 02; MTEC-20-01-Hemorrhage-032, under contract number W81XWH1590001.

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Defense Health Agency, the U.S. Department of Defense, or The Henry M. Jackson Foundation for the Advancement of Military Medicine. This article has been approved for public release with unlimited distribution.

We acknowledge the USAMRDC Office of Regulated Activities (Fort Detrick, MD) for regulatory support. We thank Sergeant First Class Jeremy Trapier, Telemedicine and Advanced Technology Research Center, for helping recruit medics for the formative and summative testing studies, and the medics for their participation in the tests. We thank Paolo Giacometti, Jon Jaeb, and Brian Robey from ZOLL Medical for constructive discussions and technical support for the Propaq M vital sign monitor.

Author Affiliations

[1] Department of Defense Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research and Development, Medical Research and Development Command, Fort Detrick, MD

[2] The Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD

[3] RQMIS, Amesbury, MA

[4] Applied Research Associates, Albuquerque, NM

[5] Department of Surgery, University of Pittsburgh, Pittsburgh

[6] Department of Emergency Medicine, University of Pittsburgh, Pittsburgh

[7] School of Public Health, University of Pittsburgh, Pittsburgh

[8] Department of Surgery, Ernest E. Moore Shock Trauma Center at Denver Health, Denver

[9] Donald D. Trunkey Center for Civilian and Combat Casualty Care, Oregon Health and Science University, Portland

[10] Department of Surgery, Division of Trauma, Surgical Critical Care, Burns, and Acute Care Surgery, University of Arizona, Tucson

[11] Department of Surgery, Baylor College of Medicine, Houston

[12] Department of Surgery, McGovern Medical School at the University of Texas Health Science Center, Houston

[13] Department of Emergency Medicine, McGovern Medical School at the University of Texas Health Science Center, Houston

[14] Department of Surgery, University of Louisville, Louisville, KY

[15] Section of Surgical Sciences, Department of Surgery, Division of Acute Care Surgery, Vanderbilt University Medical Center, Nashville

[16] Department of Veterans Affairs, Palo Alto Health Care System, Menlo Park, CA

[17] Department of Medicine, Stanford University School of Medicine, Stanford, CA

[18] Department of Surgery, Stanford University School of Medicine, Stanford, CA

[19] Department of Emergency Medicine, Massachusetts General Hospital, Boston

[20] Joint Trauma System, Defense Health Agency, Fort Sam Houston, San Antonio, TX

References

1. Gurney JM, Spinella PC. Blood transfusion management in the severely bleeding military patient. Curr Opin Anaesthesiol 2018;31:207-214. DOI: 10.1097/ACO.0000000000000574.

2. Ryan KL. Walter B. Cannon's World War I experience: treatment of traumatic shock then and now. Adv Physiol Educ 2018;42:267-276. DOI: 10.1152/advan.00187.2017.

3. Kisat M, Morrison JJ, Hashmi ZG, Efron DT, Rasmussen TE, Haider AH. Epidemiology and outcomes of non-compressible torso hemorrhage. J Surg Res 2013;184:414-421. DOI: 10.1016/j.jss.2013.05.099.

4. Eastridge BJ, Mabry RL, Seguin P, et al. Death on the battlefield (2001–2011): implications for the future of combat casualty care. J Trauma Acute Care Surg 2012;73:S431-S437. DOI: 10.1097/TA.0b013e3182755dcc.

5. Mazuchowski EL, Kotwal RS, Janak JC, et al. Mortality review of US Special Operations Command battle-injured fatalities. J Trauma Acute Care Surg 2020;88:686-695. DOI: 10.1097/TA.0000000000002610.

6. Liu NT, Holcomb JB, Wade CE, et al. Development and validation of a machine learning algorithm and hybrid system to predict the need for life-saving interventions in trauma patients. Med Biol Eng Comput 2014;52:193-203. DOI: 10.1007/s11517-013-1130-x.

7. Mackenzie CF, Gao C, Hu PF, et al. Comparison of decision-assist and clinical judgment of experts for prediction of lifesaving interventions. Shock 2015;43:238-243. DOI: 10.1097/SHK.0000000000000288.

8. Hodgman EI, Cripps MW, Mina MJ, et al. External validation of a smartphone app model to predict the need for massive transfusion using five different definitions. J Trauma Acute Care Surg 2018;84:397-402. DOI: 10.1097/TA.0000000000001756.

9. Lammers D, Marenco C, Morte K, et al. Machine learning for military trauma: novel massive transfusion predictive models in combat zones. J Surg Res 2022;270:369-375. DOI: 10.1016/j.jss.2021.09.017.

10. Stallings JD, Laxminarayan S, Yu C, et al. APPRAISE-HRI: an artificial intelligence algorithm for triage of hemorrhage casualties. Shock 2023;60:199-205. DOI: 10.1097/SHK.0000000000002166.

11. United States Food and Drug Administration. K233249: APPRAISE-HRI, 2024. September 15, 2025 (https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K233249).

12. United States Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. July 10, 2025 (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices).

13. International Medical Device Regulators Forum Software as a Medical Device Working Group. Software as a medical device (SaMD): key definitions. December 9, 2013 (https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf).

14. United States Food and Drug Administration. Clinical decision support software. Guidance for industry and Food and Drug Administration staff. September 28, 2022 (https://www.fda.gov/media/109618/download).

15. United States Food and Drug Administration. DEN200022: analytic for hemodynamic instability (AHI), 2021. March 1, 2021 (https://www.accessdata.fda.gov/cdrh_docs/pdf20/DEN200022.pdf).

16. Beiriger J, Silver D, Lu L, et al. The geography of injuries in trauma systems: using home as a proxy for incident location. J Surg Res 2023;290:36-44. DOI: 10.1016/j.jss.2023.04.004.

17. Silver DS, Sperry JL, Beiriger J, et al. Association between emergency medical service agency volume and mortality in trauma patients. Ann Surg 2024;279:160-166. DOI: 10.1097/SLA.0000000000006087.

18. Hayden SR, Brown MD. Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. Ann Emerg Med 1999;33:575-580. DOI: 10.1016/s0196-0644(99)70346-x.

19. United States Food and Drug Administration, Center for Devices and Radiological Health, Office of Device Evaluation. Applying human factors and usability engineering to medical devices. Guidance for industry and Food and Drug Administration staff (February 3, 2016). June 21, 2011 (https://www.fda.gov/media/80481/download).

20. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021;27:582-584. DOI: 10.1038/s41591-021-01312-x.

21. Eastridge BJ, Costanzo G, Jenkins D, et al. Impact of joint theater trauma system initiatives on battlefield injury outcomes. Am J Surg 2009;198:852-857. DOI: 10.1016/j.amjsurg.2009.04.029.

22. Tomas C, Kallies K, Cronn S, Kostelac C, deRoon-Cassini T, Cassidy L. Mechanisms of traumatic injury by demographic characteristics: an 8-year review of temporal trends from the National Trauma Data Bank. Inj Prev 2023;29:347-354. DOI: 10.1136/ip-2022-044817.

23. Code of Federal Regulations. 870.2220 Adjunctive hemodynamic indicator with decision point. December 27, 2022 (https://www.ecfr.gov/current/title-21/chapter-I/subchapter-H/part-870/subpart-C/section-870.2220).

24. Chen L, McKenna TM, Reisner AT, Gribok A, Reifman J. Decision tool for the early diagnosis of trauma patient hypovolemia. J Biomed Inform 2008;41:469-478. DOI: 10.1016/j.jbi.2007.12.002.

25. Liu J, Khitrov MY, Gates JD, et al. Automated analysis of vital signs to identify patients with substantial bleeding before hospital arrival: a feasibility study. Shock 2015;43:429-436. DOI: 10.1097/SHK.0000000000000328.

26. Clark P, Kim J, Aphinyanaphongs Y. Marketing and US Food and Drug Administration clearance of artificial intelligence and machine learning enabled software in and as medical devices: a systematic review. JAMA Netw Open 2023;6:e2321792. DOI: 10.1001/jamanetworkopen.2023.21792.

27. King RW, Plewa MC, Buderer NM, Knotts FB. Shock index as a marker for significant injury in trauma patients. Acad Emerg Med 1996;3:1041-1045. DOI: 10.1111/j.1553-2712.1996.tb03351.x.

28. Gianola S, Castellini G, Biffi A, et al. Accuracy of risk tools to predict critical bleeding in major trauma: a systematic review with meta-analysis. J Trauma Acute Care Surg 2022;92:1086-1096. DOI: 10.1097/TA.0000000000003496.

29. Koch E, Lovett S, Nghiem T, Riggs RA, Rech MA. Shock Index in the emergency department: utility and limitations. Open Access Emerg Med 2019;11:179-199. DOI: 10.2147/OAEM.S178358.

30. Mutschler M, Nienaber U, Munzberg M, et al. The Shock Index revisited — a fast guide to transfusion requirement? A retrospective analysis on 21,853 patients derived from the TraumaRegister DGU. Crit Care 2013;17:R172. DOI: 10.1186/cc12851.

31. Henry S. ATLS 10th edition offers new insights into managing trauma patients. Bull Am Coll Surg 2018;103:15-22.