Protocol

Applying AI and Guidelines to Assist Medical Students in Recognizing Patients With Heart Failure: Protocol for a Randomized Trial

Hyeon Joo[1], MS, MHI; Michael R Mathis[2], MD; Marty Tam[3], MD; Cornelius James[1,4,5], MD; Peijin Han[6], MBBS, MHS; Rajesh S Mangrulkar[1,5], MD; Charles P Friedman[1,7], PhD; VG Vinod Vydiswaran[1,7], PhD

[1]Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI, United States
[2]Department of Anesthesiology, University of Michigan, Ann Arbor, MI, United States
[3]Department of Internal Medicine, Cardiology, University of Michigan, Ann Arbor, MI, United States
[4]Department of Pediatrics, University of Michigan, Ann Arbor, MI, United States
[5]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States
[6]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States
[7]School of Information, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Hyeon Joo, MS, MHI
Department of Learning Health Sciences
University of Michigan
1111 East Catherine Street
Ann Arbor, MI, 48109
United States
Phone: 1 7349361644
Email: thejoo@umich.edu

Abstract

Background: The integration of artificial intelligence (AI) into clinical practice is transforming both clinical practice and medical education. AI-based systems aim to improve the efficacy of clinical tasks, enhancing diagnostic accuracy and tailoring treatment delivery. As it becomes increasingly prevalent in health care for high-quality patient care, it is critical for health care providers to use the systems responsibly to mitigate bias, ensure effective outcomes, and provide safe clinical practices. In this study, the clinical task is the identification of heart failure (HF) prior to surgery with the intention of enhancing clinical decision-making skills. HF is a common and severe disease, but detection remains challenging due to its subtle manifestation, often concurrent with other medical conditions, and the absence of a simple and effective diagnostic test. While advanced HF algorithms have been developed, the use of these AI-based systems to enhance clinical decision-making in medical education remains understudied.

Objective: This research protocol is to demonstrate our study design, systematic procedures for selecting surgical cases from electronic health records, and interventions. The primary objective of this study is to measure the effectiveness of interventions aimed at improving HF recognition before surgery, the second objective is to evaluate the impact of inaccurate AI recommendations, and the third objective is to explore the relationship between the inclination to accept AI recommendations and their accuracy.

Methods: Our study used a 3 × 2 factorial design (intervention type × order of prepost sets) for this randomized trial with medical students. The student participants are asked to complete a 30-minute e-learning module that includes key information about the intervention and a 5-question quiz, and a 60-minute review of 20 surgical cases to determine the presence of HF. To mitigate selection bias in the pre- and posttests, we adopted a feature-based systematic sampling procedure. From a pool of 703 expert-reviewed surgical cases, 20 were selected based on features such as case complexity, model performance, and positive and negative labels. This study comprises three interventions: (1) a direct AI-based recommendation with a predicted HF score, (2) an indirect AI-based recommendation gauged through the area under the curve metric, and (3) an HF guideline-based intervention.

Results: As of July 2023, 62 of the enrolled medical students have fulfilled this study's participation, including the completion of a short quiz and the review of 20 surgical cases. The subject enrollment commenced in August 2022 and will end in December 2023, with the goal of recruiting 75 medical students in years 3 and 4 with clinical experience.

Conclusions: We demonstrated a study protocol for the randomized trial, measuring the effectiveness of interventions using AI and HF guidelines among medical students to enhance HF recognition in preoperative care with electronic health record data.

International Registered Report Identifier (IRRID): DERR1-10.2196/49842

(JMIR Res Protoc 2023;12:e49842) doi: 10.2196/49842

KEYWORDS

medical education; clinical decision support systems; artificial intelligence; machine learning; heart failure; evidence-based medicine; guidelines; digital health interventions

Introduction

Computer-based diagnostic systems have played a critical role in both clinical practice and medical education, enhancing diagnostic accuracy and fostering the development of necessary knowledge and skills [1]. According to the Institute of Medicine's report [2], the use of informatics for clinical decision-making is an essential educational competency required for all health care professionals. However, the informatics landscape has undergone rapid advancement with the advent of deep learning in artificial intelligence (AI) and has revolutionized disease diagnosis, treatment delivery, and patient care [3-5]. Subsequently, AI-based tools become increasingly prevalent in health care, allowing health care providers to make informed clinical decisions for high-quality care and optimal patient outcomes [6-8].

As AI continues to be integrated into medical practice, it becomes inevitable for medical students, residents, and professionals to acquire the necessary skills for effective medical practice [9]. In 2021, Lomis et al [10] highlighted the importance of incorporating AI training across health care professions to maximize its benefits while mitigating the potential drawbacks in routine patient care. Mckinsey's report [11], titled "Transforming Healthcare with AI," emphasizes the need to implement this transformation within the realm of education and training. Moreover, AI training within the medical school curriculum is an active area of discussion and investigation [10,12,13]. With AI in the medical program, students have the opportunity to learn about the use of apps and limitations of AI in clinical practice and improve their ability to use AI-based tools to enhance patient care and decision-making.

In 1989, Iliad [14,15], a computer-aided diagnosis, was introduced to enhance diagnostic abilities, encompassing over 6300 disease manifestations and addressing 1300 diseases related to internal medicine. Iliad offered a unique diagnostic learning experience by simulating cases and guiding users through a series of decision-making processes encountered in clinical workups. It also provided tailored feedback at each decision-making of the clinical workups based on the user's performance. Lincoln et al [16] and Lange et al [17] demonstrated the improvement of diagnostic reasoning and a diagnostic error reduction among medical students and nurse practitioner students. Similarly, Friedman et al [18] confirmed enhanced diagnostic accuracy before and after using diagnostic consultation systems among medical students, residents, and physicians.

Despite the benefits of diagnostic systems, their integration into educational programs has been limited and has insufficient results [19,20]. Berner and McGowan [1] also pointed out that 1 contributing factor to limited educational usage is the stand-alone nature of these systems, which needs to be better integrated into existing clinical workflows. Conversely, Tolsgaard et al [21] observed that commercially available AI systems used in clinical settings are predominantly tailored to address specialized clinical challenges rather than assisting health care providers in skill enhancement. As a result, there is a gap in the development of health care professionals' ability to use such decision support systems to use algorithmically generated recommendations in their clinical decision-making responsibly.

Moreover, an overreliance on AI systems can lead to unintended errors due to automation bias [22], which refers to the tendency for individuals to place excessive trust in automated systems and disregard their own judgment, such as the failure of autopilot technology [23]. Without proper training on the responsible use of AI, health care professionals may unknowingly commit mistakes by relying too heavily on AI-based systems [24,25]. The lack of understanding and awareness of AI's limitations and potential biases can lead to errors in decision-making and patient care [26]. Health care professionals need the knowledge and skills to critically evaluate and interpret AI-generated recommendations, taking into account contextual factors and individual patient needs [27].

To mitigate the adverse effects and ensure the responsible use of AI, Tolsgaard et al [21] emphasized the importance of integrating learning sciences with clinical science and data science while developing new AI systems. This interdisciplinary approach aims to combine insights from educational research, clinical practice, and data analysis to design AI systems [28]. This approach enhances performance and fosters continuous learning and professional development. By integrating these disciplines, AI systems can be designed to provide appropriate feedback, facilitate reflective practice, and support the acquisition of clinical reasoning skills, promoting a balance between clinical practice and education in health care.

This paper is centered on demonstrating a study protocol, including a study design, a systematic procedure for case selection from electronic health record (EHR) data, and

intervention designs to improve the recognition of patients with heart failure (HF) before surgery in a preoperative care clinical setting. HF is a common and serious chronic condition characterized by an abnormality in the structure and function of the heart, leading to reduced blood circulation throughout the body [29,30]. However, identifying patients with HF remains challenging due to its subtle and concurrent progression with other conditions and the lack of a single gold standard diagnostic test for HF [31,32]. Several algorithmic approaches have been recently published to improve HF detection, such as convolutional neural network with ECG [33-35], logistic regression [36], recurrent neural network [37], and transformer [38] models with EHR data.

The primary objective of this study is to measure the effectiveness of interventions aimed at improving HF recognition in preoperative care through a web-based clinical decision support (CDS) tool. The interventions are grounded in our previous research using traditional machine learning (ML) algorithms [39] and guideline-based diagnostic factors reviewed by clinical experts [40,41]. These interventions are algorithmically and manually synthesized from a large amount of EHR data for identifying HF. The secondary objective is to evaluate the impact of inaccurate ML recommendations in recognizing patients with HF, as compared to HF guideline-based recommendations. The third objective is to explore the relationship between the inclination to accept AI recommendations and the ability to accurately recognize HF using ML-based interventions.

Methods

Study Setting and Participants

Participants in this study are medical students to simulate the clinical task of preoperative surgical screening, targeting the recognition of HF as a specific educational objective using surgical cases performed at a tertiary hospital. This study's participants review surgical cases on our internally developed web-based CDS educational tool that displays deidentified EHR data, including clinical visits, diagnoses, procedures, vitals, laboratory, medications, imaging or studies, and clinical notes up to the day of surgery. To maximize the efficiency of case reviews, clinical data included in the CDS tool was tailored to preoperative care and cardiovascular diseases for screening HF. Irrelevant EHR data to this study (eg, administrative data, non-HF-related imaging) was excluded. This CDS tool was developed and deployed in a secure internal environment where study participants can access and review surgical cases.

The eligible students for participation in this study are medical students in their third (M3) or fourth (M4) year of medical school, who completed some coursework requirements and underwent clinical rotations. In addition, medical school graduates who have not yet started their residency program are eligible for inclusion. The rationale behind including senior medical students in this study is that accurate identification of HF using real-world EHR data needs a comprehensive understanding of clinical knowledge and skills acquired through coursework and clinical rotations.

This study is entirely on the internet and comprises two components as follows:

1. The first component entails an e-learning module, including an instructional video about the intervention and a short quiz, with an estimated completion time of 30 minutes.
2. The second component involves reviewing 20 surgical cases to evaluate the presence or absence of HF using EHR data, with an estimated completion time of 90 minutes.

Prior to reviewing the surgical cases, participants are required to complete a 5-question short quiz with all correct answers. They have the option to attempt the quiz multiple times if necessary. As this study is conducted on the internet, participants can complete this study at their own pace and at their own convenient time. Upon successful completion of this study, participants receive US $50 as a gift of appreciation.

Ethics Approval

This research study involving human subjects has received ethical approval from the institutional review board at the University of Michigan (HUM00207646). This approval indicates that this study's protocol, including research design, data collection methods, and informed consent, has been reviewed and considered to be in compliance with ethical standards for human subject research.

Study Design

The study design of this HF recognition study is a 3 × 2 factorial randomized trial that involves 3 intervention groups and the order exchange of 2 sets of surgical cases during the pre- and posttests. This study design allows us to measure the effectiveness of 3 interventions without introducing bias from surgical case selection in the pre- and posttests, summarized in Multimedia Appendix 1.

The rationale behind exchanging the order of 2 sets is to minimize the potential bias arising from differences in individual surgical cases. For example, if set A in the pretest contains 10 surgical cases that are more difficult than those in set B in the posttest, it would be challenging to attribute any improvements in accuracy solely to the interventions. Therefore, exchanging the order of the sets between the pre- and posttests ensures the measurement of unbiased interventions' effectiveness.

Furthermore, the use of random selection for the surgical cases is not suitable in this study due to the implementation of specific case selection criteria. The selection criteria aim to provide medical students with targeted learning opportunities to recognize HF through the review of surgical cases. The selection process is designed to ensure that the chosen cases offer valuable educational experiences and align with the intended learning objectives. Further details about the case selection are discussed in the later section dedicated to the systematic procedure for surgical case selection.

Prior to implementing the 3 × 2 factorial study design, a pilot study was conducted using a 3-arm study design. In this preliminary investigation, participants were asked to determine the presence or absence of HF before intervention access and then reassess the same surgical case after intervention access. However, this study design demonstrated an "anchoring effect,"

where participants tended to adhere to their initial clinical judgment. Therefore, it underestimated the effectiveness of the intervention. Consequently, the decision was made to adopt the 3 × 2 factorial design to overcome this limitation and provide a more accurate evaluation of the effectiveness of interventions.

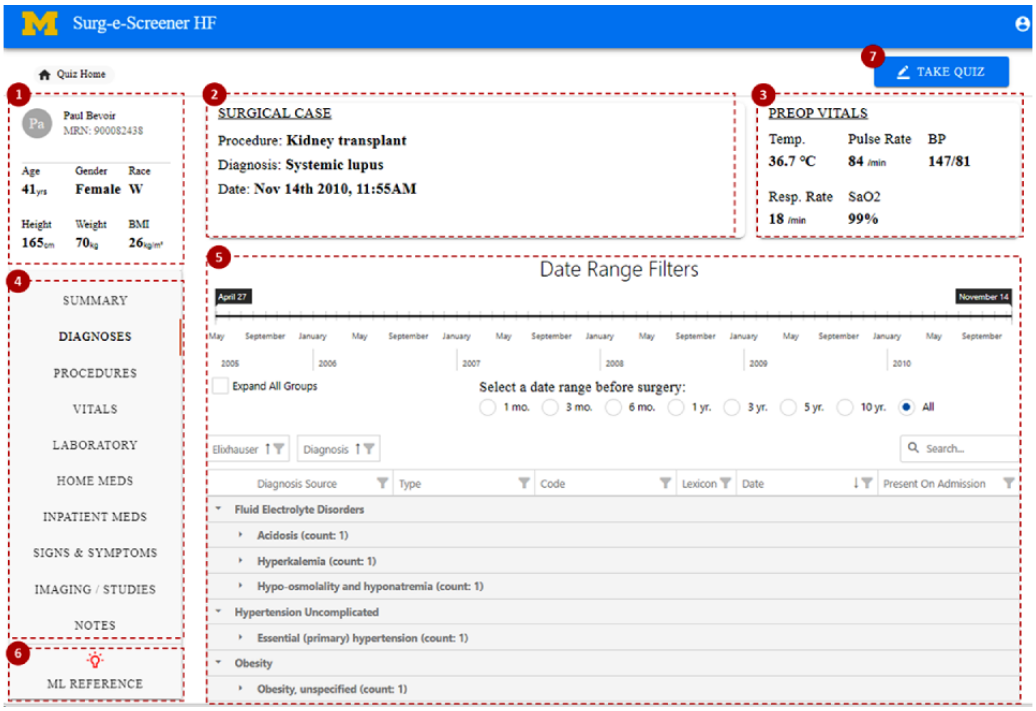
Interventions: ML_{DR}, ML_{IR}, and EB References

In order to measure the effectiveness of AI and guideline-based educational methods with 3 distinct interventions, the interventions need to provide essential information specific to the methods to enhance the ability to accurately recognize HF using patient data from the EHR.

The essential information is consolidated into a single *reference* page, which compiles risk factors associated with HF and method-specific recommendations (eg, HF yes or no, risk scores) to augment medical students' ability to recognize HF accurately. The risk factors are categorized into nine sections based on consultations with HF experts to improve the readability and comprehensibility of the interventions. These sections include (1) signs and symptoms, (2) past HF history, (3) past medical history, (4) surgical history, (5) medications, (6) physical exams, (7) test labs, (8) imaging or studies, and (9) ECG. A detailed description of the risk factors within each section can be found in Multimedia Appendix 2.

During the posttest phase, the intervention, represented by the reference page, is available to medical students who completed the first 10 surgical cases using only their clinical judgment during the pretest. In the posttest, the reference page is accessible through an ML or EB reference tab (6) in Figure 1 and can be accessed at any time while medical students review a surgical case. Further information about developing the ML model used in ML reference is included in Multimedia Appendix 3, and the details of the intervention differences are outlined as follows.

Figure 1. A web-based educational tool to review surgical cases, including (1) demographics (top left), (2) surgical case information (top middle), (3) preoperative vitals (top right), (4) subject domains (left side), (5) EHR data of a subject domain (center), (6) ML and EB intervention (left bottom). The intervention tab is only available during the posttest. After reviewing surgical cases, answer survey questions by clicking the (7) "TAKE QUIZ" button. EB: evidence based; EHR: electronic health record; ML: machine learning.



ML_{DR} Reference

The ML_{DR} reference provides a direct recommendation (DR), which is an output from the ML algorithm in the form of the presence or absence of HF, a dichotomous response. To achieve this, an optimal threshold is set to discriminate between positive and negative cases based on the estimated HF probability. The threshold configuration is critical in determining the sensitivity and specificity of the system. In this study, the optimal threshold was determined to be 38% based on the assumption that sensitivity and specificity are equally important. The ML_{DR} reference includes a brief explanation of the dichotomous recommendation, such as "this patient has a 98% chance of having HF, which is above the 38% threshold," to allow medical students to incorporate this into their clinical decision-making. It is important to note that the diagnosis of HF is a complex process that requires a thorough examination, and ML_{DR} reference offers a simple and efficient method to recommend the presence or absence of HF to medical students, along with a justification of the recommendation. The screenshot of ML_{DR} reference is in Multimedia Appendix 4.

ML_{IR} Reference

ML_{IR} reference offers an indirect recommendation (IR), which is a proxy of carrying HF risk, using the area under the receiver operating characteristic (AUROC), which comprises the true-positive rate (TPR) and false-positive rate (FPR). The AUROC plot, with TPR and FPR plotted along its axes, provides

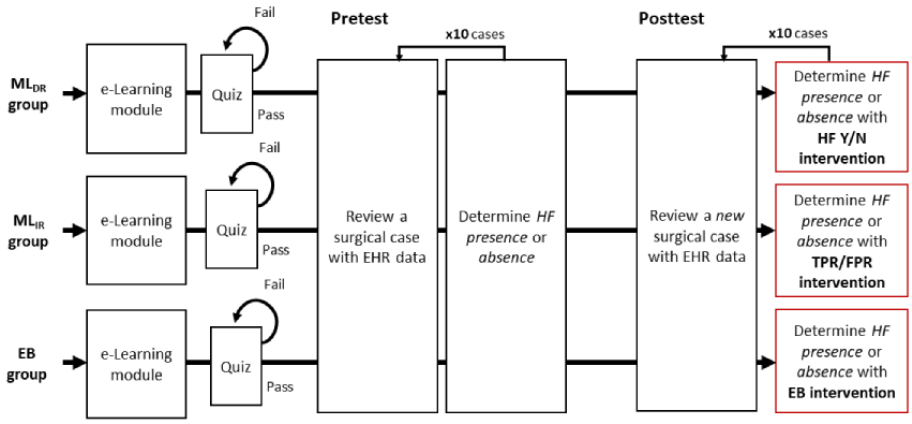
a comprehensive view of the model's performance across various thresholds. In contrast to ML_{DR} reference, which is limited to a single threshold, ML_{IR} reference shows the position of TPR and FPR on the AUROC plot when the threshold is defined as the predicted probability of HF for individual cases. For example, in the case of a surgical case with a high predicted HF probability, ML_{IR} reference displays the corresponding TPR or FPR position on the AUROC plot, reflecting a high-level threshold equivalent to the predicted HF probability. This approach enables medical students to examine surgical cases with different TPR and FPR positions on the AUROC plot, thereby avoiding a *decision bias* toward a single threshold. The screenshot of ML_{IR} reference is in Multimedia Appendix 5. In short, ML_{DR} and ML_{IR} references adopt 2 distinct approaches to present direct and indirect recommendations to medical students while using the same underlying ML model.

EB Reference

The third intervention, *EB reference*, includes risk factors described in HF guidelines [42,43] and shows the presence or absence of these risk factors explicitly documented in EHR and clinical experts' impressions while reviewing the case from EHR. In our previous study [40], the 20 surgical cases were adjudicated by a consensus panel of HF experts (cardiologists, cardiac anesthesiologists, and critical care physicians) who indicated the explicit documentation of the risk factors in EHR as well as their impression of presence or absence of the risk factors. Unlike the risk factors from EHR data in ML references, this intervention uses the highest-quality evidence in the medical literature from various clinical studies and a consensus of domain experts to incorporate evidence-based practice into HF diagnostic recommendations. Further, because the HF recommendations in the guidelines [42,43] are intended to provide evidence-based recommendations for health care practitioners, they are presented in a way that is easy to *understand* and *acceptable* for clinicians. As such, the risk factors related to heart diseases are incorporated into medical education and training, resulting in less friction in clinical reasoning and decision-making. The screenshot of the EB reference is in Multimedia Appendix 6.

Each group in the HF recognition study is given access to one of the interventions during the posttest case reviews. After assignment to an intervention group, all participants reviewed the same 20 surgical cases, 10 surgical cases during the pretest phase, and then 10 new surgical cases during the posttest phase. The intervention is only accessible during the posttest. Figure 2 shows an overview of this study's participants for the three intervention groups: (1) ML_{DR} group, which receives direction recommendations indicating the presence or absence of HF; (2) ML_{IR} group, which receives indirect recommendations in the form of TPR, FPR, and AUROC plot; and (3) EB group, which receives expert-reviewed evidence-based risk factors from HF guidelines.

Figure 2. An overview of study participation using a 3 × 2 factorial study design. The participation includes an e-learning module, a short quiz, and 20 surgical case reviews. The order of 2 sets of 10 surgical cases is predetermined before the participant begins this study's activities. ML_{DR} group: a group of students who receive ML-generated risk factors and a direct recommendation of HF Y/N. ML_{IR} group: a group of students who receive ML-generated risk factors and an indirect recommendation of the likelihood of having HF from TPR, FPR, and AUROC. EB group: a group of students who receive HF expert-reviewed risk factors extracted from HF guidelines. AUROC: area under the receiver operating characteristic; EB: evidence based; EHR: electronic health record; FPR: false-positive rate; HF: heart failure; ML: machine learning; TPR: true-positive rate; Y/N: yes or no.



E-Learning Module

Prior to starting the surgical case review, participants are required to engage in a series of preparatory activities, including a web-based e-learning module that includes a 20-minute prerecorded instructional video, and a short quiz with 5 multiple-choice questions. The purpose of the e-learning module is to provide participants with the necessary knowledge and understanding to recognize patients with HF using the web-based tool. Given the diversity of participants' backgrounds and lack of familiarity with the new tool, this module offers educational materials and essential concepts that are fundamental to comprehending the interventions used to facilitate clinical decision-making.

The e-learning module focuses explicitly on facilitating participants' comprehension of machine learning concepts, including predicted risk scores, important features, thresholds, and evaluation metrics (eg, sensitivity, specificity, TPR, and FPR) for ML intervention groups. Additionally, the module covers the identification of risk factors contributing to HF, as outlined in the American Heart Association [42,44] and the

European Society of Cardiology [43] HF guidelines, and Framingham criteria [45] for the EB intervention group.

In an attempt to confirm that subjects have viewed and comprehended the material presented in the prerecorded video, participants are required to complete a short quiz, which must be completed with a score of 100% (5 out of 5 questions). When questions are answered incorrectly, participants are instructed to review the material and revise their responses prior to proceeding to the next step. This e-learning module reduces variations in participants' prior knowledge and background to develop a standardized basis for evaluating surgical cases and measuring the effectiveness of interventions.

Surgical Case Selections

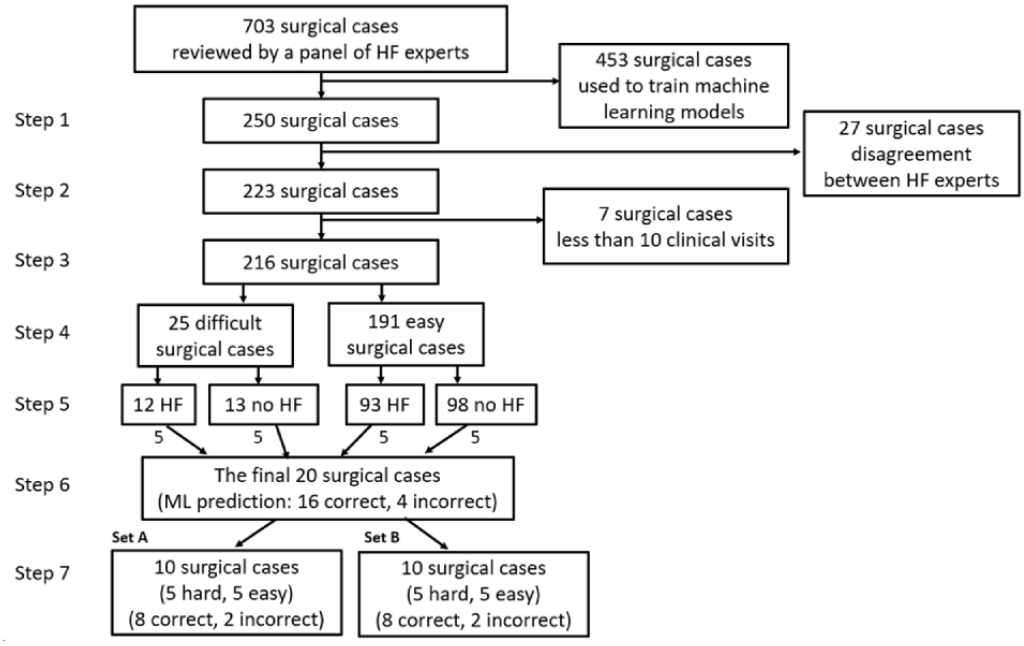
In our previous study [40,41], a consensus panel of HF experts, comprising cardiologists, cardiac anesthesiologists, and critical care physicians, assessed 1018 surgical cases, of which 703 were available during this study. These cases were part of a stratified subsample of 40,659 adult noncardiac surgical procedures between 2015 and 2019 at Michigan Medicine. Using a feature-based sampling process [46], 20 surgical cases were deliberately selected, considering factors, such as case complexity, expert consensus, ML performance, and HF outcomes. The 20 cases were divided into 2 sets, each consisting of 10 cases, for the use of pre- and posttests. Each set included an equal number of easy and difficult cases, as well as an equal representation of patients with and without HF.

The systematic sampling procedure for surgical case selection is illustrated in Figure 3. There are seven steps involved: (1) selecting cases from the holdout data set, (2) including cases that are consensus among expert reviewers, (3) including patients who had at least 10 clinical visits, (4) determining the level of difficulty from clinical diagnostic factors presented in EHR, (5) selection priority based on HF outcome, (6) consideration of ML performance (90% AUROC, 82% sensitivity, and 82% of specificity in our model), and (7) dividing cases into two sets for pre- and posttests. The more detailed descriptions of each step are included in Multimedia Appendix 7.

The implementation of a systematic methodology for case selection from EHR is critical, as the validation derived from this study could serve as a foundational cornerstone for scalability across diverse medical conditions. For example, diagnostic tasks using ML with expert-reviewed labels are potential candidates that could be integrated into this CDS tool by following the steps above, requiring only minor adjustments to accommodate unique clinical nuances. Thus, these ML-based tasks could be transformed into additional educational resources and presented to learners through this educational tool.

In the last step, surgical cases in set A and set B were presented to medical students in pre- and posttests to determine the presence or absence of HF. However, it does not guarantee equivalent difficulty levels, which is critical for evaluating intervention effectiveness. To mitigate the issue, we altered the order of sets in the pre- and posttests. Half of the participants received set A in the pretest and set B in the posttest, and vice versa for the rest of the participants, resulting in a 3 × 2 factorial design that ensured a balanced measurement of the interventions' effectiveness.

Figure 3. A systematic sampling procedure for the final 20 surgical cases with respect to the level of difficulty, positive and negative HF cases, and the ML performance from expert-reviewed surgical cases. HF: heart failure; ML: machine learning.



Sample Size and Recruitment Strategies

To assess the effectiveness of interventions with a power of 80% at a significant level of 5%, the total number of subjects needed is 75, with 25 subjects assigned to each group. This power analysis was conducted using a 1-sample *t* test using SPSS (version 28.0; IBM Corp), a statistical software package. The power analysis was based on our preliminary results, which indicated a mean difference of 0.08 and an SD of 0.13.

The recruitment for this study commenced in August 2022 and will continue until December 2023. The primary recruitment method involves sending a group email to year 3 and 4 medical students through faculty and academic curriculum coordinators, posting notices in various institutional newsletters and web-based bulletin boards, and distributing fliers. The recruitment started at a single institution and expanded to participants from other institutions.

Furthermore, snowball sampling has been used, which asks for referrals from enrolled study participants. To encourage referrals, a referral-based lottery has been implemented for individuals who refer their friends and colleagues in year 3 or 4 medical students. This lottery is recurring quarterly, and the winner will be drawn each term and awarded a US $100 prize. In addition, we have implemented a sweepstakes for study participants who will enter for a chance to win US $500 worth of prizes to recruit more students.

Assignment of Interventions

To participate in this study, subjects must first complete a survey form providing information about their school year, clinical experiences, and their level of acceptance toward clinical recommendations made by both clinical experts and AI algorithms. As an individual's behavior and willingness to accept recommendations can significantly impact this study's outcome [47], the level of acceptance was included in the survey form for use in the subject assignment.

Upon registration, subjects are allocated to one of the intervention groups using a permuted block and stratification method [48]. This method involves stratifying subjects based on key characteristics and randomly assigning them across groups within a fixed size of blocks. For this study, the acceptance level of AI recommendations (≥50%) and medical institutions serve as the stratification criteria, with subjects being randomly assigned to intervention groups within each block of size 6. As a result, each block consists of 2 occurrences of 3 distinct interventions.

Data Collection

Upon completing the e-learning module, medical students begin reviewing 20 surgical cases. To assess the accuracy of recognizing patients with HF, medical students respond to a set of 3 survey questions for each case review. This survey is embedded in the web-based tool, which is accessible after reviewing EHR data and intervention information. Medical students have the option to revisit data points if needed to address the survey questions. As shown in Multimedia Appendix 8, 3 survey questions are presented to evaluate the ability of medical students to accurately identify the presence or absence of HF. The first question pertains to the final decision of HF presence or absence at the start of surgery, followed by a question on their level of confidence in their clinical decision. Finally, medical students are requested to check the type of information they use to inform their clinical decision-making process.

Outcomes

The primary outcome of this study is to assess the effectiveness of interventions in improving the accuracy of recognizing patients with HF, by measuring the mean difference of recognition accuracy before and after intervention in pre- and posttests. Specifically, this study will demonstrate the mean difference in HF recognition accuracy before and after accessing ML_{DR} ML_{IR}, and EB reference interventions.

The second outcome of this study is to evaluate the impact of accurate or inaccurate ML recommendations on medical students' clinical judgment to recognize the presence or absence of HF, compared to the EB intervention group. In addition, this study explores the differential effects of direct and indirect ML recommendations on clinical decision-making among students.

Lastly, this study measures the effect of the acceptance level of AI reported in the prescreening survey on the accuracy of recognizing HF using ML references. This exploratory outcome will provide valuable insights for enhancing the acceptability of AI tools.

Results

As of July 2023, 62 of the enrolled medical students have fulfilled this study's participation, including the completion of a short quiz and the review of 20 surgical cases. The recruitment process for these study participants began in August 2022 and will end in December 2023.

Discussion

Principal Findings

In this study protocol paper, we outlined a methodological approach for measuring the effectiveness of ML-based interventions, both direct and indirect recommendations, and EB intervention. The interventions were integrated with EHR data within a web-based educational tool to enhance the recognition of HF before surgery. Specifically, we demonstrated a novel randomized trial using a 3 × 2 factorial design, systematic procedures for surgical case selection, and 3 interventions using ML algorithms and HF guidelines using EHR data.

The demonstration of 3 × 2 factorial design (intervention type × order of prepost sets) represents a unique methodological contribution, enabling the measurement of intervention effectiveness through enrolling students in pre- and posttests. Prior work such as Lincoln et al [16] with medical students and Lange et al [17] with nurse practitioner students adopted a 2 × 2 × 2 mixed factorial design (disease sets × trained or untrained groups × replication) to evaluate diagnostic errors and posterior probability pertaining to the comprehensiveness of clinical workups. In contrast to Iliad, which offers feedback during training via simulated cases, our study provides an e-learning module designed to impart and apply synthesized information, grounded in AI or HF guidelines, to enhance diagnostic accuracy. With our proposed factorial design, the effectiveness of the synthesized information was measured before and after reviewing surgical cases derived from EHR.

Furthermore, this study protocol paper articulates a systematic procedure of selecting HF cases from expert-reviewed surgical cases in EHR, incorporating the difficulty level and positive and negative outcomes. We believe that this systematic procedure of case selection not only mitigates selection bias for this study but also offers a replicable framework for other diseases when expert-reviewed labels and EHR data are available. In addition, our approach has the potential of scalability to present algorithmically or manually synthesized information, along with other data available in EHRs, on other diseases. This is challenging for expert systems like Iliad, equipped with a knowledge base or inference engine [14,49].

Limitations

While this study provides a foundational approach to measure the effectiveness of enhancing HF recognition, it is important to note limitations with respect to a limited scope of findings and generalizability. Specifically, the target sample size of 75 participants (25 in each study arm) is adequate with a statistical power of 0.80 to measure the effectiveness of enhancing HF decision-making before and after intervention access. However, additional research questions, such as comparative effectiveness between ML and EB interventions, are limited in this study. Furthermore, this study was primarily conducted within an academic medical center that could be limited to a representative of medical schools in the United States. Despite this sample size constraint, our proposed study has the potential to inform future studies to broaden our understanding of how AI-based tools can be helpful for medical education and training.

Moreover, this study primarily focuses on intervention effectiveness, neglecting the essential facet of learner engagement and feedback mechanisms within the web-based tool. To enhance the effective learning outcomes in future iterations, the web-based decision support tool could incorporate real-time feedback upon each case completion, such as indicating the accuracy of case reviews and providing preannotated expert commentary on examined surgical cases. Additionally, offering structured guidance on interpreting AI-generated outputs is critical for accurately using predictive scores and in-depth analysis of misclassification instances. Integrating these postreview feedback mechanisms holds considerable promise for aiding learners in effectively leveraging AI-based tools for clinical decision-making.

The ML model used in this study was trained and evaluated on a stratified and manually reviewed data set, comprising patients who had developed HF and patients at high and low risk [40,41]. While stratification ensures the model learns various stages of HF progression, it may not accurately reflect the prevalence of HF estimated between 1.5% and 1.9% in the US population [50]. As a result, the model performance should be interpreted with caution when considering its use in a routine clinical setting and requires careful further examination. In this study, however, the deliberate inclusion of surgical cases for training the model with various complexities serves educational purposes effectively. This approach allows learners to gain experience in clear-cut positive or negative outcomes and also borderline cases, which are inherently challenging yet particularly informative.

Conclusions

In summary, this study protocol demonstrates a study design for measuring the effectiveness of interventions to enhance HF recognition among adult noncardiac surgical patients. Despite limitations of sample size and generalizability, this study serves as a foundational step toward a more comprehensive understanding of how AI techniques and evidence-based medicine can be synergistically used to advance both medical education and patient care outcomes.

Acknowledgments

The deployment of this study was possible with the extensive financial, infrastructure, and resource support provided by the University of Michigan. Several funding sources supported this study, including Research. Innovation. Scholarship. Education. (RISE) initiative, Michigan Institute for Computational Discovery and Engineering, Rackham Graduate Student Research Grant, and Aikens Innovation Academy from Frankel Cardiovascular Center. The Department of Learning Health Sciences offered unique opportunities to engage with professionals in health care, data, and implementation sciences. We thank Dr Karandeep Singh and Dr Zhenke Wu for their critical study reviews, and Dr Gretchen Piatt and Dr Anne Sales for their valuable insights in the realm of implementation science during this study's development. Additionally, during the RISE fellowship, the RISE team offered training sessions, ongoing support, and feedback to enhance my understanding of developing an AI-based tool for medical education. Notably, the core members of the RISE team, namely Dr Paula Thompson and Dr Nikki Bibler Zaidi, facilitated the training sessions and recommended developing relationships with relevant individuals for further guidance in order to obtain essential knowledge regarding the problem. With these supports, necessary connections were established with Dr Brian George, a surgeon whose insights into surgical consultation and workflow proved invaluable, and Dr James Woolliscroft, a professor of internal medicine and former dean of the medical school at Michigan Medicine, who provided valuable feedback on the overall study design and student recruitment processes. Lastly, the development and deployment of the web-based tool were feasible with the tremendous support from the Michigan Anesthesiology Informatics and Systems Improvement Exchange with the necessary IT infrastructure and resources to make this tool easily accessible to researchers and medical students. The use of ChatGPT (OpenAI, Microsoft Corporation) to improve the style of academic writing and grammar checking is disclosed in Multimedia Appendix 9.

Data Availability

Due to patient identification concerns and confidentiality, the data used for this study is not publicly available. Any request for data access may need additional approval from the institute review board at Michigan Medicine.

Conflicts of Interest

HJ is currently an employee of Meta Platforms, Inc (aka Facebook). This research study was developed and deployed when HJ was affiliated with the University of Michigan.

Multimedia Appendix 1

A 3 x 2 factorial study design with three interventions and the set order of surgical cases. A study subject reviews a total of 20 surgical cases, 10 surgical cases in each pre- and post-tests. The mean difference in accuracy before and after accessing an intervention will be measured in each cell.
[DOCX File , 14 KB-Multimedia Appendix 1]

Multimedia Appendix 2

ML and EB Reference tables. The nine sections of heart failure–related risk indicators are categorized in the sequence of informed necessity for diagnosis.
[DOCX File , 22 KB-Multimedia Appendix 2]

Multimedia Appendix 3

A summary of developing and adopting a machine learning model in ML References.
[DOCX File , 13 KB-Multimedia Appendix 3]

Multimedia Appendix 4

ML_{DR} Reference: This intervention includes a direct recommendation based on heart failure (HF) probability, like HF or No HF. It also consists of the top 5 risk factors and the entire list in the reference table.
[DOCX File , 185 KB-Multimedia Appendix 4]

Multimedia Appendix 5

ML_{IR} Reference: This intervention includes the true-positive rate, false-positive rate, and area under the receiver operating characteristic plot. These will give a proxy of heart failure risk level but no direct recommendation. It also consists of the top 5 risk factors and the entire list in the reference table.
[DOCX File , 124 KB-Multimedia Appendix 5]

Multimedia Appendix 6

EB Reference: This intervention includes heart failure (HF) expert-reviewed presence or absence of risk factors listed in HF guidelines and HF expert's impression on the presence or absence of risk factors as reviewing a surgical case.
[DOCX File , 191 KB-Multimedia Appendix 6]

Multimedia Appendix 7

This is a detailed 7-step procedure to systematically select two sets of surgical cases from electronic health records taking into account factors such as case complexity, expert agreement, machine learning performance, and heart failure outcome.
[DOCX File , 14 KB-Multimedia Appendix 7]

Multimedia Appendix 8

The questionnaire that study participants are asked to answer after reviewing each case via the web-based educational tool.
[DOCX File , 15 KB-Multimedia Appendix 8]

Multimedia Appendix 9

The disclosure of the use of ChatGPT or other generative language models.
[DOCX File , 13 KB-Multimedia Appendix 9]

References

1. Berner ES, McGowan JJ. Use of diagnostic decision support systems in medical education. Methods Inf Med 2010;49(4):412-417 [doi: 10.3414/ME9309] [Medline: 20405092]
2. Committee on the Health Professions Education Summit, Board on Health Care Services, Institute of Medicine. In: Greiner AC, Knebel E, editors. Health Professions Education: A Bridge to Quality. Washington, D.C: National Academies Press; 2003.
3. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
4. Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open 2021;4(3):e211276 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.1276] [Medline: 33704476]
5. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 2016;25(Suppl 1):S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]
6. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
7. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56 [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
9. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
10. Lomis K, Jeffries P, Palatta A, Sage M, Sheikh J, Sheperis C, et al. Artificial intelligence for health professions educators. NAM Perspect 2021;2021:1-14 [FREE Full text] [doi: 10.31478/202109a] [Medline: 34901780]
11. Spatharou A, Hieronimus S, Jenkins J. Transforming healthcare with AI: the impact on the workforce and organizations. McKinsey & Company. 2020. URL: https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai [accessed 2023-06-03]
12. Ötleş E, James CA, Lomis KD, Woolliscroft JO. Teaching artificial intelligence as a fundamental toolset of medicine. Cell Rep Med 2022;3(12):100824 [FREE Full text] [doi: 10.1016/j.xcrm.2022.100824] [Medline: 36543111]
13. James CA, Wheelock KM, Woolliscroft JO. Machine learning: the next paradigm shift in medical education. Acad Med 2021;96(7):954-957 [FREE Full text] [doi: 10.1097/ACM.0000000000003943] [Medline: 33496428]
14. Warner HR. Iliad: moving medical decision-making into new frontiers. Methods Inf Med 1989;28(4):370-372 [Medline: 2695788]
15. Guo D, Lincoln MJ, Haug PJ, Turner CW, Warner HR. Comparison of different information content models by using two strategies: development of the best information algorithm for Iliad. Proc Annu Symp Comput Appl Med Care 1992:465-469 [FREE Full text] [Medline: 1482918]
16. Lincoln MJ, Turner CW, Haug PJ, Warner HR, Williamson JW, Bouhaddou O, et al. Iliad training enhances medical students' diagnostic skills. J Med Syst 1991;15(1):93-110 [doi: 10.1007/BF00993883] [Medline: 1748852]
17. Lange LL, Haak SW, Lincoln MJ, Thompson CB, Turner CW, Weir C, et al. Use of Iliad to improve diagnostic performance of nurse practitioner students. J Nurs Educ 1997;36(1):36-45 [doi: 10.3928/0148-4834-19970101-09] [Medline: 8986960]
18. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA 1999;282(19):1851-1856 [FREE Full text] [doi: 10.1001/jama.282.19.1851] [Medline: 10573277]
19. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med 1994;330(25):1792-1796 [FREE Full text] [doi: 10.1056/NEJM199406233302506] [Medline: 8190157]
20. Lincoln MJ. Medical education applications. In: Berner ES, editor. Clinical Decision Support Systems: Theory and Practice. New York, NY: Springer; 1999:105-137
21. Tolsgaard MG, Pusic MV, Sebok-Syer SS, Gin B, Svendsen MB, Syer MD, et al. The fundamentals of artificial intelligence in medical education research: AMEE Guide No. 156. Med Teach 2023;45(6):565-573 [FREE Full text] [doi: 10.1080/0142159X.2023.2180340] [Medline: 36862064]
22. Goddard K, Roudsari A, Wyatt JC. Automation bias—a hidden issue for clinical decision support system use. Stud Health Technol Inform 2011;164:17-22 [Medline: 21335682]
23. Boudette NE. Tesla's autopilot technology faces fresh scrutiny. The New York Times. 2021. URL: https://www.nytimes.com/2021/03/23/business/teslas-autopilot-safety-investigations.html [accessed 2021-03-23]
24. Alberdi E, Povykalo A, Strigini L, Ayton P. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Acad Radiol 2004;11(8):909-918 [FREE Full text] [doi: 10.1016/j.acra.2004.05.012] [Medline: 15354301]

25. Ibrahim SA, Pronovost PJ. Diagnostic errors, health disparities, and artificial intelligence: a combination for health or harm? JAMA Health Forum 2021;2(9):e212430 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.2430] [Medline: 36218658]
26. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4(1):31 [FREE Full text] [doi: 10.1038/s41746-021-00385-9] [Medline: 33608629]
27. James CA, Wachter RM, Woolliscroft JO. Preparing clinicians for a clinical world influenced by artificial intelligence. JAMA 2022;327(14):1333-1334 [doi: 10.1001/jama.2022.3580] [Medline: 35311917]
28. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33(7):1163-1170 [FREE Full text] [doi: 10.1377/hlthaff.2014.0053] [Medline: 25006142]
29. Braunwald E. Heart failure. JACC Heart Fail 2013;1(1):1-20 [FREE Full text] [doi: 10.1016/j.jchf.2012.10.002] [Medline: 24621794]
30. Wagner S, Cohn K. Heart failure: a proposed definition and classification. Arch Intern Med 1977;137(5):675-678 [doi: 10.1001/archinte.137.5.675] [Medline: 856090]
31. Ramani GV, Uber PA, Mehra MR. Chronic heart failure: contemporary diagnosis and management. Mayo Clin Proc 2010;85(2):180-195 [FREE Full text] [doi: 10.4065/mcp.2009.0494] [Medline: 20118395]
32. Hobbs FDR, Doust J, Mant J, Cowie MR. Heart failure: Diagnosis of heart failure in primary care. Heart 2010;96(21):1773-1777 [doi: 10.1136/hrt.2007.139402] [Medline: 20956495]
33. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019;394(10201):861-867 [doi: 10.1016/S0140-6736(19)31721-0] [Medline: 31378392]
34. Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pellikka PA, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. J Cardiovasc Electrophysiol 2019;30(5):668-674 [doi: 10.1111/jce.13889] [Medline: 30821035]
35. Porumb M, Iadanza E, Massaro S, Pecchia L. A convolutional neural network approach to detect congestive heart failure. Biomed Signal Process Control 2020;55:101597 [doi: 10.1016/j.bspc.2019.101597]
36. Blecker S, Katz SD, Horwitz LI, Kuperman G, Park H, Gold A, et al. Comparison of approaches for heart failure case identification from electronic health record data. JAMA Cardiol 2016;1(9):1014-1020 [FREE Full text] [doi: 10.1001/jamacardio.2016.3236] [Medline: 27706470]
37. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017;24(2):361-370 [FREE Full text] [doi: 10.1093/jamia/ocw112] [Medline: 27521897]
38. Rao S, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Cleland J, et al. An explainable transformer-based deep learning model for the prediction of incident heart failure. IEEE J Biomed Health Inform 2022;26(7):3362-3372 [FREE Full text] [doi: 10.1109/JBHI.2022.3148820] [Medline: 35130176]
39. Mathis MR, Engoren MC, Joo H, Maile MD, Aaronson KD, Burns ML, et al. Early detection of heart failure with reduced ejection fraction using perioperative data among noncardiac surgical patients: a machine-learning approach. Anesth Analg 2020;130(5):1188-1200 [FREE Full text] [doi: 10.1213/ANE.0000000000004630] [Medline: 32287126]
40. Golbus JR, Joo H, Janda AM, Maile MD, Aaronson KD, Engoren MC, Michigan Congestive Heart Failure Investigators. Preoperative clinical diagnostic accuracy of heart failure among patients undergoing major noncardiac surgery: a single-centre prospective observational analysis. BJA Open 2022;4:1-13 [FREE Full text] [doi: 10.1016/j.bjao.2022.100113] [Medline: 36643721]
41. Kamyszek RW, Newman N, Ragheb JW, Sjoding MW, Joo H, Maile MD, Michigan Congestive Heart Failure Investigators, et al. Differences between patients in whom physicians agree versus disagree about the preoperative diagnosis of heart failure. J Clin Anesth 2023;90:111226 [doi: 10.1016/j.jclinane.2023.111226] [Medline: 37549434]
42. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, American College of Cardiology Foundation, American Heart Association Task Force on Practice Guidelines. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. J Am Coll Cardiol 2013;62(16):e147-e239 [FREE Full text] [doi: 10.1016/j.jacc.2013.05.019] [Medline: 23747642]
43. Ponikowski P, Voors A, Anker S, Bueno H, Cleland J, Coats A, ESC Scientific Document Group. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J 2016;37(27):2129-2200 [FREE Full text] [doi: 10.1093/eurheartj/ehw128] [Medline: 27206819]
44. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Colvin MM, et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines and the Heart Failure Society of America. J Card Fail 2017;23(8):628-651 [FREE Full text] [doi: 10.1016/j.cardfail.2017.04.014] [Medline: 28461259]
45. McKee PA, Castelli WP, McNamara PM, Kannel WB. The natural history of congestive heart failure: the Framingham study. N Engl J Med 1971;285(26):1441-1446 [doi: 10.1056/NEJM197112232852601] [Medline: 5122894]

46. Friedman CP, Wyatt JC, Ash JS. Evaluation Methods in Biomedical and Health Informatics. Cham, Switzerland: Springer International Publishing; 2022.
47. Kelly S, Kaye SA, Oviedo-Trespalacios O. What factors contribute to the acceptance of artificial intelligence? A systematic review. Telemat Inform 2023;77:101925 [FREE Full text] [doi: 10.1016/j.tele.2022.101925]
48. Broglio K. Randomization in clinical trials: permuted blocks and stratification. JAMA 2018;319(21):2223-2224 [doi: 10.1001/jama.2018.6360] [Medline: 29872845]
49. Tso GJ, Tu SW, Oshiro C, Martins S, Ashcraft M, Yuen KW, et al. Automating guidelines for clinical decision support: knowledge engineering and implementation. AMIA Annu Symp Proc 2016;2016:1189-1198 [FREE Full text] [Medline: 28269916]
50. Roger VL. Epidemiology of heart failure: a contemporary perspective. Circ Res 2021;128(10):1421-1434 [FREE Full text] [doi: 10.1161/CIRCRESAHA.121.318172] [Medline: 33983838]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic
CDS: clinical decision support
EHR: electronic health record
FPR: false-positive rate
HF: heart failure
ML: machine learning
TPR: true-positive rate

Edited by A Mavragani; submitted 19.06.23; peer-reviewed by E Kamana, A Hadianfard; comments to author 18.08.23; revised version received 16.09.23; accepted 20.09.23; published 24.10.23

Please cite as:
Joo H, Mathis MR, Tam M, James C, Han P, Mangrulkar RS, Friedman CP, Vydiswaran VGV
Applying AI and Guidelines to Assist Medical Students in Recognizing Patients With Heart Failure: Protocol for a Randomized Trial
JMIR Res Protoc 2023;12:e49842
URL: https://www.researchprotocols.org/2023/1/e49842
doi: 10.2196/49842
PMID: 37874618