Clinical and Experimental Emergency Medicine
Clin Exp Emerg Med 2023;10(4):354-362
https://doi.org/10.15441/ceem.23.145

Review Article

eISSN: 2383-4625

Explainable artificial intelligence in emergency medicine: an overview

Yohei Okada[1,2], Yilin Ning[3], Marcus Eng Hock Ong[1,4]

[1]Health Services and Systems Research, Duke-NUS Medical School, Singapore
[2]Preventive Services, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[3]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore
[4]Department of Emergency Medicine, Singapore General Hospital, Singapore

Artificial intelligence (AI) and machine learning (ML) have potential to revolutionize emergency medical care by enhancing triage systems, improving diagnostic accuracy, refining prognostication, and optimizing various aspects of clinical care. However, as clinicians often lack AI expertise, they might perceive AI as a "black box," leading to trust issues. To address this, "explainable AI," which teaches AI functionalities to end-users, is important. This review presents the definitions, importance, and role of explainable AI, as well as potential challenges in emergency medicine. First, we introduce the terms explainability, interpretability, and transparency of AI models. These terms sound similar but have different roles in discussion of AI. Second, we indicate that explainable AI is required in clinical settings for reasons of justification, control, improvement, and discovery and provide examples. Third, we describe three major categories of explainability: pre-modeling explainability, interpretable models, and post-modeling explainability and present examples (especially for post-modeling explainability), such as visualization, simplification, text justification, and feature relevance. Last, we show the challenges of implementing AI and ML models in clinical settings and highlight the importance of collaboration between clinicians, developers, and researchers. This paper summarizes the concept of "explainable AI" for emergency medicine clinicians. This review may help clinicians understand explainable AI in emergency contexts.

Keywords Artificial intelligence; Machine learning; Resuscitation; Emergency medicine

Received: 9 October 2023
Revised: 6 November 2023
Accepted: 16 November 2023

Correspondence to: Yohei Okada
Health Services and Systems Research, Duke-NUS Medical School, National University of Singapore, 8 College Rd, Singapore 169857
Email: yohei_ok@duke-nus.edu.sg

Capsule Summary

What is already known

Artificial intelligence (AI) and machine learning have potential to revolutionize emergency medical care, enhance triage, diagnosis, prognostication, and many aspects of clinical care. However, as clinicians often lack AI expertise, they might perceive AI as a "black box", leading to trust issues. To address this, "explainable AI", which makes AI functionalities comprehensible to end-users, is important. This review introduces the definitions, importance, and role of explainable AI, and potential challenges in emergency medicine.

What is new in the current study

This paper summarizes the concept of "explainable AI" for emergency medicine clinicians. This review may help clinicians understand explainable AI in emergency contexts.

How to cite this article:
Okada Y, Ning Y, Ong MEH. Explainable artificial intelligence in emergency medicine: an overview. Clin Exp Emerg Med 2023;10(4):354-362. https://doi.org/10.15441/ceem.23.145

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) are powerful technologies that have the potential to improve medical care [1]. AI refers to the broader concept of technology being able to carry out tasks in an autonomous and smart way, encompassing a variety of technologies, while ML is a subset of AI focused on the idea that machines can learn from data, identify patterns, and make decisions with minimal human intervention [1–4]. Particularly in emergency medicine, AI and ML are expected to play critical roles in accelerating triage, diagnosis, and prognostication to optimize individual patient care through the input of clinical information and/or image recognition [2,4–8]. Furthermore, streamlined clinical documentation or recording using natural language processing is expected to make these tasks more efficient [9–11]. These technologies will also contribute to drug discovery, patient monitoring, resource allocation, and epidemiological surveillance [12–15].

Despite expectations that emergency physicians will become general users of AI and ML in the near future, critics are doubtful whether they can trust and rely on AI and ML models [16]. Physicians are usually not experts in AI and may not have an in-depth understanding of it or ML. When an AI model outputs a medical classification or prediction, without necessarily "explaining" the underlying process or showing the variables and weights driving the prediction, physicians who are not familiar with AI algorithms may perceive an AI model as a "black box." Such a situation may lead to doubt and mistrust in AI output, a major challenges for implementation of AI and ML tools in clinical settings [17].

To address these concerns, the concept of "explainable AI" has been highlighted as a possible solution for successful implementation of AI and ML in medical practice [18–20]. Explainable AI aims to teach the functioning of AI systems to end-users, researchers, or developers [18–20]. As more accurate, complex models are developed, it may become increasingly difficult to understand how they work. This review introduces the definitions, importance, and role of explainable AI and covers potential related challenges in emergency medicine.

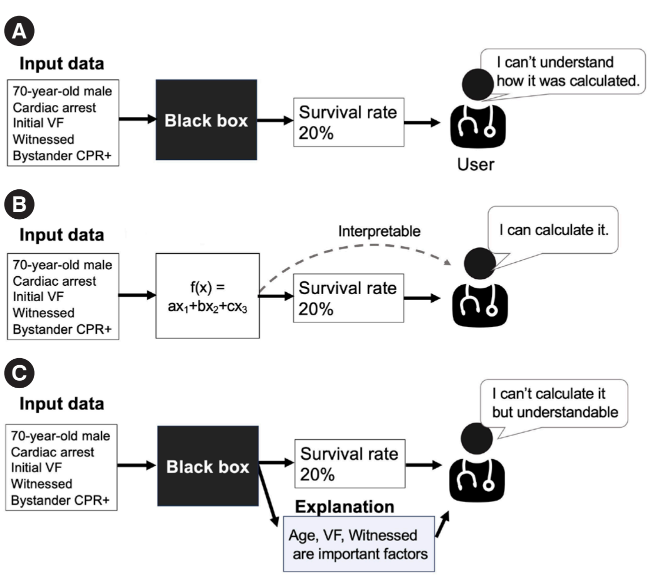

Fig. 1. Black box, transparency, and post hoc explainability. (A) Black-box model. (B) Transparent model. (C) Post hoc explainability. VF, ventricular fibrillation; CPR, cardiopulmonary resuscitation.

WHAT IS TRANSPARENCY, INTERPRETABILITY, AND EXPLAINABILITY IN AI?

Before discussing the concept of "explainability," we introduce the terms "interpretability" and "transparency" [18–20]. While these terms have distinct meanings, they are sometimes mistaken for explainability. AI models are often labeled as a "black box," suggesting a lack of transparency, as shown in Fig. 1A [21]. In an opaque model, the conversion of an input to an output is invisible, making it challenging for users to understand the process. In contrast, transparent/interpretable models allow users to understand how inputs are processed to produce outputs to observe operation of a model, as shown in Fig. 1B. As shown in Fig. 2, straightforward models like linear regression with a limited number of variables or a decision tree with a few branches are easier to understand [21].

Explainability is different conceptually. Arrieta et al. [19] suggested that, "Given a certain audience, explainability refers to the ability of a model to show details and make its internal functioning clear or easy to understand." This definition emphasizes the audience's perspective, which can vary by background, experiences, and capacities, resulting in different needs of explainability in ML. While explainability is inherent in transparent models, it is also tied to post hoc explainability, as shown in Fig. 1C. This concept refers to techniques that provide a rationale or explanation

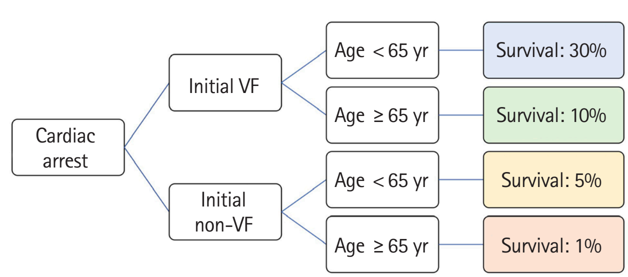

Fig. 2. Decision tree model to predict the possibility of survival. You can trace the algorithm to generate the output without any computers or devices. VF, ventricular fibrillation.

to support user understanding of model operation even if the model itself is noninterpretable (often referred to as a "black box"); this may be achieved by providing text or a visual explanation, etc. [22].

WHY IS EXPLAINABLE AI NEEDED IN EMERGENCY MEDICINE?

There are four main reasons why explainable AI is required in clinical settings: justification, control, improvement, and discovery of novel ideas [23]. These four reasons may appear to overlap, but from a clinical perspective, they capture different motivations.

First, explainable AI is useful to justify AI model outputs, to enhance trust, and to support clinical decision-making [23]. Generally, clinicians need to explain a medical condition, treatment plan, and expected outcomes to patients and their families. Despite medical complexity, clinicians need to make the effort to communicate to the patients and their families to facilitate shared decision-making and trust [24]. Thus, explainability is important in a medical decision-making process, and the results generated from clinical AI or ML need to be handled appropriately to build trust. In the absence of a clear and satisfactory explanation, clinicians will likely be hesitant to trust AI. For example, a recent randomized control trial investigated the efficacy of an AI model that aims to detect cardiac arrest cases at the dispatch center using voice data analysis during an emergency call [25,26]. The study was unable to demonstrate the effectiveness of the AI model, with the researchers suggesting that some dispatchers might not have trusted the output from the AI model due to the absence of a reasonable explanation. Explainable AI can produce an auditable and testable way to defend algorithmic decisions as fair and ethical [27,28].

Second, explainability helps users maintain control of a complex technology. Indeed, a deeper understanding of AI models increases awareness of vulnerabilities and flaws [23], which can help to quickly identify and correct (debug) errors in critical situations. Thus, user controls can be strengthened. If the AI generates unexpected results and unreasonable explanations inconsistent with clinical experience or show potential risk of bias/discrimination, clinicians can bypass the AI and review results for errors or bias [29]. For example, if a patient is unexpectedly evaluated as having low likelihood of a favorable outcome, mainly driven by ethnicity or socioeconomic status of the patient, clinicians may suspect hidden discrimination or bias in the AI training data [30,31].

Third, explainability of AI models is needed to allow continual improvement. If AI models can discuss the process used to produce specific results, the information can be used for further improvements [32]. For example, when an AI model does not accurately predict an outcome, its process can be determined using explainability features. If a certain predictor highly contributes to a model but carries a risk of measurement bias due to the absence of standardized definitions in clinical settings, the model might be improved by excluding the variable or standardizing the input. In this way, explainable AI can lay the groundwork for continuous iteration and improvement.

Furthermore, explainable AI may allow development of new ideas, hypotheses, and knowledge [22]. For example, if an explanation from an AI model shows an unexpected contribution of a certain risk factor to the prediction of outcomes, a novel hypothesis might be developed regarding this factor and its association with outcomes. In an AI study investigating clinical subgroups of cardiac arrest patients treated effectively with extracorporeal cardiopulmonary resuscitation (ECPR), creatinine value was associated with outcome. This led to the development of a novel score for indications of ECPR that included creatinine [33]. The importance of explainable AI is increasing with the more critical role of AI in clinical settings.

HOW DOES EXPLAINABLE AI WORK?

Explainable AI encompasses three main approaches [18–20]. The first is pre-modeling explainability, such as data visualization, summarization, and transformation [34]. Before deploying AI in clinical settings, it is essential to grasp the data structure, patients' characteristics, time trends, and proportion of the outcome for an appropriate understanding of the AI. This may include simple descriptions, such as mean, standard deviation, and range, and missing data using data visualization or summarization. Data transformation is also crucial to change row data into a usable format or structure [34]. For example, when developing ML models, clinical data such as the date of incidence, time of emergency call, and hospital arrival are generally transformed to the month or day of an incident or the duration between the emergency call and the time of hospital arrival, allowing easier analysis. Data transformation is more applicable to development and understanding of models.

The second approach is to develop an interpretable model with inherent understandability or a blend of different model types [18–20]. Models exhibit various levels of interpretability and transparency: at the level of the training algorithm (referred to as "algorithmic transparency"), at the component level (known as "decomposability"), and at the level of the model itself (or "simulatability") [22]. For example, as shown in Fig. 2, a tree model might pose a human-understandable question, such as whether the patient is younger than 65 years to clarify the prediction process and enhance algorithmic transparency. Such a model can be broken down into individual segments, like patients with or without initial ventricular fibrillation (VF) (Fig. 2). This allows users to identify the contribution of each segment to the overall output, showcasing decomposability. In an example of a 40-year-old cardiac arrest patient with initial VF, we can follow the entire prediction pathway of the model, estimating a survival probability of 30%, without specialized mathematical tools. This demonstrates simulatability, with model transparency and user-friendly interpretation.

A hybrid interpretable model approach has also been proposed [18–20]. It includes a set of methods that attempts to combine a complex black-box model with an inherently interpretable model to build an interpretable model that achieves comparable performance to the black-box model. The AutoScore framework is an example of this hybrid interpretable model approach [35,36]. In this framework, development of an ML model is complicated, but the final result is familiar to users [5,35,37].

The last method is called post-modeling explainability [18]. It helps break down complex developed AI models so that they are easier to understand. These techniques were created based on human understanding.

POST-MODELING EXPLAINABILITY

In this section, we introduce some examples of post-modeling explainability. Fig. 3 shows several categories of post-modeling explainability, including visualization, textual justification, simplification, and feature relevance [22].

Text explanations improve the understanding of ML models by generating text-based explanations in the form of phrases or sentences using natural language generation methods. Examples include AI models used to classify pathological images, which attempt to provide user-friendly explanations [38]. These models can generate sentences such as: "The input image is diagnosed as tissue A type for sure because it could not be misclassified to any other tissue types"; "The input image is suspected as tissue B type, and there is a low possibility that it could be tissue C type, D type, or E type"; or "The input image is tissue A type. However, there is a possibility that it could be tissue F type." These explanations about the possibilities of misclassification provide rationale for predictions and help clinicians with their decision-making (Fig. 3).

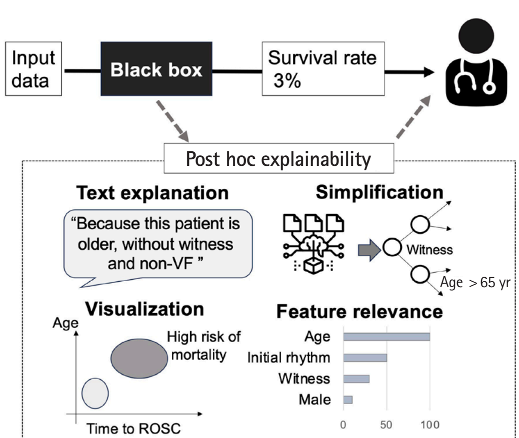

Fig. 3. Methods of post-explainability. VF, ventricular fibrillation; ROSC, return of spontaneous circulation.

Visual explanations describe models by applying techniques that aim to visualize the model behavior. Popular literature makes use of dimensionality reduction techniques to produce simple visualizations that can be easily interpreted by humans. Visual explanation is particularly effective in conveying complex interactions between variables [18–20]. For instance, when describing black-box models to predict the probability of favorable outcomes for cardiac arrest patients, the interactions between probability and some factors (such as age and transport time to the hospital) are difficult to recognize. In explainable AI, these variables can be

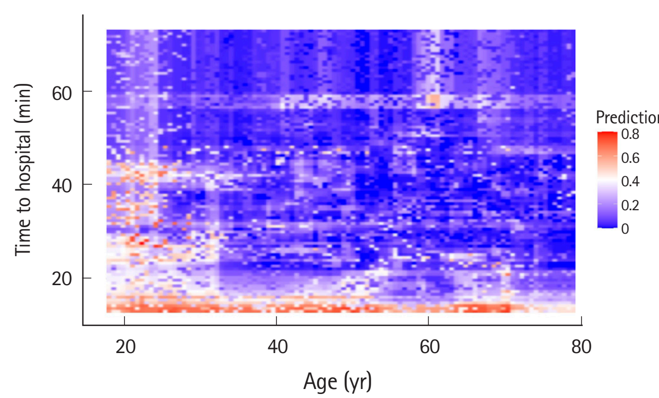

Fig. 4. Visualization of interactions between factors and outputs. This plot illustrates the relationships between the factors (age and time to hospital) and the predictions of the random forest machine-learning model. This random forest model was constructed to predict the survival probability using simulated data from cardiac arrest patients, including patient age and gender, whether the event was witnessed, provision of bystander cardiopulmonary resuscitation, and the time from the call to arrival at a hospital. The x-axis represents age, while the y-axis denotes the time from the call to arrival at a hospital. Blue dots indicate cases with a low probability of survival (<25%), red dots are cases with a high likelihood of survival (>35%), and white dots signify intermediate cases (around 30%). While users might not grasp the intricacies of artificial intelligence model prediction, they can broadly infer that patients who are younger and have a shorter time to reach a hospital are predicted to have higher survival rates. Conversely, older patients with a longer time to reach a hospital are estimated to have lower survival probability. Moreover, users can observe the interactions between factors and predictions.

plotted along the x-axis and y-axis, respectively, to create a scatter plot (Fig. 4) of the distribution of the possibilities [39]. This approach visualizes the predicted probability of interaction between the factors.

Simplification creates a straightforward, less complicated interpretable model from a black-box model [22]. One example of simplification is selection of a single decision tree as the representative of a random forest ensemble of numerous decision tree models [40,41]. A simplified model could aggregate predictions from the individual trees to produce a final output, as shown in Fig. 3 [42]. Although this approach is commonly utilized in medical research, the results can be challenging to interpret due to the ensemble nature. Identifying a single tree that captures the primary patterns and behaviors of the entire forest allows a bal-

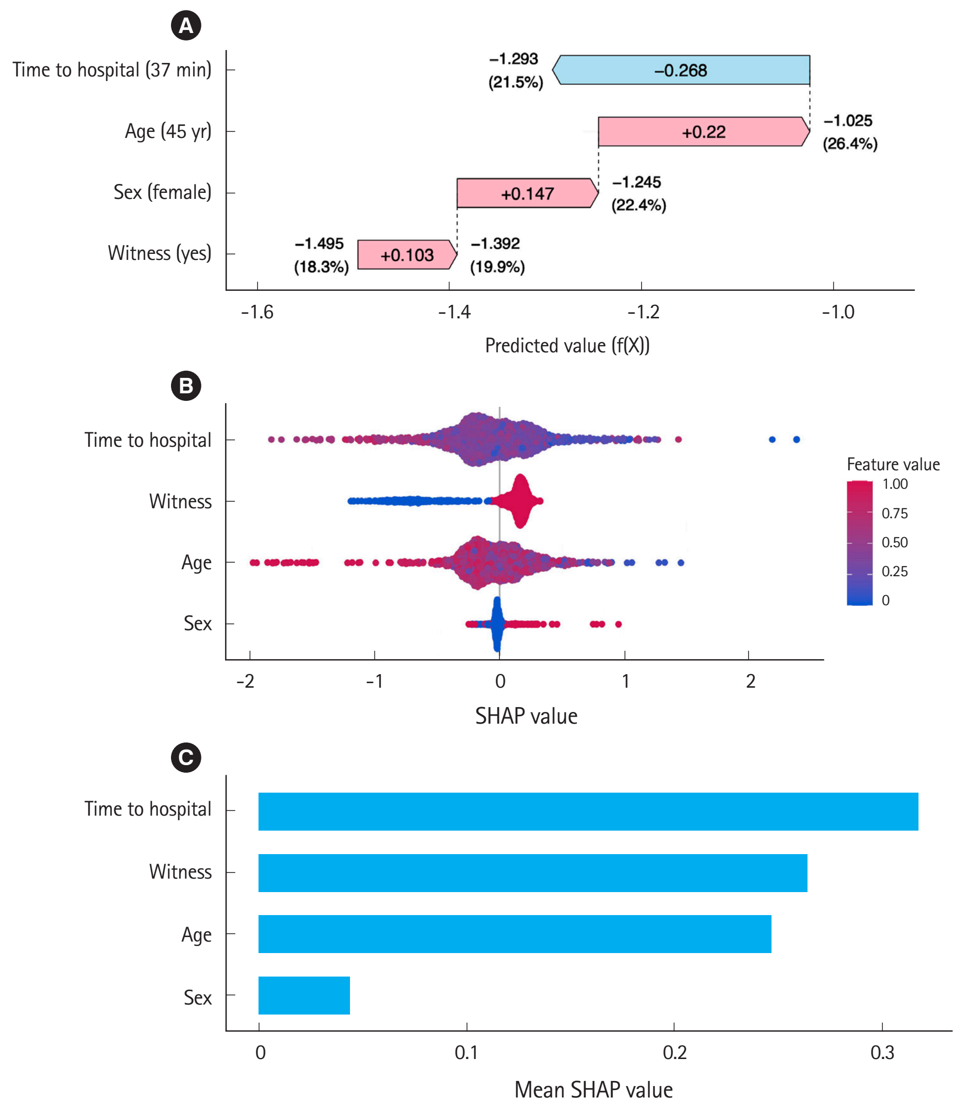

Fig. 5. Example of Shapley Additive Explanations (SHAP) values. (A) The waterfall plot indicates the predicted value (i.e., f(X)) of each factor for a 45-year-old female who suffered cardiac arrest with witnesses and time to the hospital of 37 minutes. The change in predicted value from baseline (−1.495, corresponding to a survival rate of 18.3%) to that for this particular case (−1.293, corresponding to a survival rate of 21.5%) attributed to each factor is the SHAP value of each factor. The survival probability is calculated as the inverse logit function given by $[1/(1+e^{(-f(X))})]$. (B) The beeswarm plot demonstrates the SHAP values of each factor across all cases. A central vertical line (at SHAP value=0) indicates "no influence" on the prediction. If a point is to the right of this line, it means that the factor influences the model predictions in a positive direction (increases the survival rate); if it is to the left, it influences predictions in a negative direction (decreases the survival rate). The color of the dot represents the value of the factor. For example, red and blue correspond to female or male for sex and "witnessed" and "not witnessed" for witness status. Also, red or blue means older or younger in age. (C) The bar plot displays the absolute SHAP values, indicating that factors of time to hospital, witnessed status, and age are more relevant predictors of survival than age across all cases.

ance between interpretability and performance [40,41]. This representative tree can be visualized, providing insights into the decision-making process using the same foundational logic as the original ensemble. Another example is the local interpretable model-agnostic explanations (LIME) approach [43], which approximates a complex black-box model with a simpler and more easily interpretable linear regression model. This is achieved by generating numerous samples of input data, predicting their outputs using the original model, and then training a linear model on these samples with emphasis on those close to the original data point. LIME can identify the feature importance that contributes most to each prediction and helps users determine the factors most crucial in the, as explained below.

Explanation by feature relevance aims to provide post-modeling explainability by assessing the internal processes of an algorithm. This type of explanation is commonly utilized in ML models in emergency medicine [33,44–47]. It calculates relevance scores for all variables managed by the algorithm to quantify the importance of features critical to model decisions. Shapley Additive Explanations (SHAP) is one of the methods used to evaluate the contribution of each input feature to AI model operation [48,49]. Similar to LIME, SHAP performs local linear approximations to explain the predicted risk for each individual. However, it uses a different approach that allows more desirable properties than LIME in terms of local accuracy and consistency (for details, please see the reference) [49].

SHAP can quantify and visualize how each factor increases or decreases risk from baseline to reach the predicted risk for each individual using a waterfall plot, as shown in Fig. 5. Consider an ML model that predicts the survival rate of cardiac arrest patients using sex, age, witness status (yes/no), and time to the hospital from the emergency call (minutes). For example, consider a 45-year-old female with witnessed arrest and a time to hospital of 37 minutes. This ML model predicts her survival rate as 21.5%. The waterfall plot in Fig. 5A demonstrates how these four factors influence the prediction of ML. In this case, the baseline of the predicted value (f(X)), i.e., the average prediction across all cases, is −1.495, which translates to a baseline survival rate of 18.3% via the inverse logit function [$1/(1+e^{(-f(X))})$]. The witnessed status increases the predicted value by 0.103, which is equivalent to an increase in the survival probability to 19.9% from baseline. This 0.103 increase in the predicted value attributable to the witnessed status is the SHAP value of this factor for this individual. The patient's sex (female) has a SHAP value of 0.147, which further increases the survival probability to 22.4%. Additionally, the patient's age (45 years old) has a SHAP value of 0.22, which increases it further to 26.4%. However, the time taken to reach a hospital, which has a negative SHAP value of −0.268, reduces the survival rate to the final predicted value of 21.5% for this particular case. Through this example, we can understand how each variable impacts the model prediction using a waterfall plot and SHAP values. This demonstrates how SHAP can provide local explanations for individual predictions. When the contributions to the predicted risk of each factor are visualized across all patients in a beeswarm plot (Fig. 5B), the relationships between factor levels and contributions to prediction can be determined, facilitating a straightforward comparison of the impact of each factor on the prediction.

SHAP is also valuable for global explanations to understand how the model behaves overall. This is done by considering all data points using the average absolute SHAP value. For example, the bar plot in Fig. 5C indicates that the average absolute SHAP value of "time to hospital" is highest, while those of "witnessed" and "age" are lower but still considerably high compared to that of "sex." The model suggests time to hospital, witnessed status, and age as strong predictors of survival. This dual capability allows users to understand both specific decisions of the model and the broader trends and behaviors across data points.

CHALLENGES IN IMPLEMENTATION INTO CLINICAL SETTINGS

Even though the explainability of AI has advanced, there remain several challenges to the implementation of AI models in clinical settings. One of the issues is whether the explanation is acceptable and trustworthy enough from the points of view of clinicians and patients [16,18–20]. Previously, an explainable AI model was developed to predict the deterioration of patients with subarachnoid hemorrhage in the intensive care unit. To enhance the implementation of the AI tool, the perception gap between the developers and clinicians was investigated [50]. Through interviews, the study found that the developers believed that clinicians must be able to understand model operation and developed the AI model with explainability by providing SHAP values, as mentioned above. In contrast, from the perspectives of the clinicians, the SHAP value was not sufficiently helpful in understanding or trusting the AI model. Clinicians were more focused on clinical plausibility based on the pathophysiological rationale or clinical experience and a holistic approach referring to the multispectral clinical information. As illustrated in this example, the kind of explainability required depends on the audience and context of use of the AI model [19]. In emergency settings, the contexts and patient conditions change rapidly. Especially during resuscitation, which is an incredibly time-critical situation, clinicians may not have adequate time to try to understand how AI models work. Therefore, a model must be understandable quickly. Furthermore, it is a challenge to assess the quality/effectiveness of explainability. A previous systematic review reported various methods for assessing explainable AI effectiveness, with few established methods [51]. Establishing standardized approaches to measure the effectiveness of explainable AI might increase its integration into clinical settings and act as a tool of communication among clinicians, researchers, and developers [28]. Finally, with increasing emphasis on fair and trustworthy AI-assisted decision-making in clinical settings, the contribution of explainable AI to model development should be determined through a multidisciplinary approach [52]. Considering such situations, collaboration among AI developers, researchers, and clinicians in designing explainable AI systems is imperative for improving their effectiveness, usability, and reliability in healthcare.

CONCLUSION

This paper summarizes the concept of "explainable AI" for clinicians in emergency medicine. With the expected increasing role of AI in medicine, emergency physicians and researchers will need to become knowledgeable about its use. Furthermore, a multidisciplinary approach is essential to develop trustworthy AI for use in clinical emergency medicine. This review will help interpret explainable AI to clinicians working in emergency departments.

ETHICS STATEMENT

Not applicable.

CONFLICT OF INTEREST

Yohei Okada has received a research grant from the Zoll Foundation and overseas scholarships from the Japan Society for Promotion of Science, the Fukuda Foundation for Medical Technology, and the International Medical Research Foundation. Marcus Eng Hock Ong reports grants from the Laerdal Foundation, Laerdal Medical, and the Ramsey Social Justice Foundation for funding of the Pan-Asian Resuscitation Outcomes Study and an advisory relationship with Global Healthcare SG, a commercial entity that manufactures cooling devices. Funding from Laerdal Medical was received for an observation project of the Community CPR Training Centre Research Program in Norway. Marcus Eng Hock Ong is a Scientific Advisor to TIIM Healthcare SG and Global Healthcare SG. Marcus Eng Hock Ong is also an Editorial Board member of *Clinical and Experimental Emergency Medicine*, but was not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflict of interest relevant to this article was reported.

FUNDING

This study was supported by a scientific research grant from the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. JP22K21143/23K16253) and by the Zoll Foundation.

ACKNOWLEDGMENTS

The authors thank Professor Liu Nan (Health Services and Systems Research, Duke-NUS Medical School, Singapore) for his support and helpful advice.

AUTHOR CONTRIBUTIONS

Conceptualization: YO, MEHO; Funding acquisition: YO, MEHO; Methodology: YN; Writing–original draft: YO; Writing–review & editing: YN, MEHO. All authors read and approved the final manuscript.

ORCID

Yohei Okada https://orcid.org/0000-0002-2266-476X
Yilin Ning https://orcid.org/0000-0002-6758-4472
Marcus Eng Hock Ong https://orcid.org/0000-0001-7874-7612

REFERENCES

1. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol 2019;28:73–81.
2. Kuhl N, Schemmer M, Goutier M, Satzger G. Artificial intelligence and machine learning. Electron Mark 2022;32:2235–44.
3. Chang H, Cha WC. Artificial intelligence decision points in an emergency department. Clin Exp Emerg Med 2022;9:165–8.
4. Kang CY, Yoon JH. Current challenges in adopting machine learning to critical care and emergency medicine. Clin Exp Emerg Med 2023;10:132–7.
5. Liu N, Xie F, Siddiqui FJ, et al. Leveraging large-scale electronic health records and interpretable machine learning for clinical decision making at the emergency department: protocol for system development and validation. JMIR Res Protoc 2022;11:e34201.
6. Goto T, Camargo CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for chil-

dren during emergency department triage. JAMA Netw Open 2019;2:e186937.
7. Fujiwara G, Okada Y, Shiomi N, Sakakibara T, Yamaki T, Hashimoto N. Derivation of coagulation phenotypes and the association with prognosis in traumatic brain injury: a cluster analysis of nationwide multicenter study. Neurocrit Care; 2023.
8. Hirano Y, Kondo Y, Hifumi T, et al. Machine learning-based mortality prediction model for heat-related illness. Sci Rep 2021;11:9501.
9. Goto T, Hara K, Hashimoto K, et al. Validation of chief complaints, medical history, medications, and physician diagnoses structured with an integrated emergency department information system in Japan: the Next Stage ER system. Acute Med Surg 2020;7:e554.
10. Haupt CE, Marks M. AI-generated medical advice: GPT and beyond. JAMA 2023;329:1349–50.
11. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233–9.
12. Walters WP, Barzilay R. Critical assessment of AI in drug discovery. Expert Opin Drug Discov 2021;16:937–47.
13. Poncette AS, Mosch L, Spies C, et al. Improvements in patient monitoring in the intensive care unit: survey study. J Med Internet Res 2020;22:e19091.
14. Goldstein BA, Cerullo M, Krishnamoorthy V, et al. Development and performance of a clinical decision support tool to inform resource utilization for elective operations. JAMA Netw Open 2020;3:e2023547.
15. MacIntyre CR, Chen X, Kunasekaran M, et al. Artificial intelligence in public health: the potential of epidemic early warning systems. J Int Med Res 2023;51:3000605231159335.
16. Fujimori R, Liu K, Soeno S, et al. Acceptance, barriers, and facilitators to implementing artificial intelligence-based decision support systems in emergency departments: quantitative and qualitative evaluation. JMIR Form Res 2022;6:e36501.
17. Floridi L. Establishing the rules for building trustworthy AI. Nat Mach Intell 2019;1:261–2.
18. Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 2022;55:3503–68.
19. Arrieta AB, Diaz-Rodriguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.
20. Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives [Preprint]. Posted 2017 Oct 2. arXiv:171000794. https://doi.org/10.48550/arXiv.1710.00794
21. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2018;51:1–42.
22. Belle V, Papantonis I. Principles and practice of explainable machine learning. Front Big Data 2021;4:688969.
23. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–60.
24. Beach MC, Sugarman J. Realizing shared decision-making in practice. JAMA 2019;322:811–2.
25. Zicari RV, Brusseau J, Blomberg SN, et al. On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Front Hum Dynamics 2021;3:673104.
26. Blomberg SN, Christensen HC, Lippert F, et al. Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: a randomized clinical trial. JAMA Netw Open 2021;4:e2032320.
27. Reddy S. Explainability and artificial intelligence in medicine. Lancet Digit Health 2022;4:e214–5.
28. Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20:310.
29. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns (N Y) 2021;2:100347.
30. Okada Y, Mertens M, Liu N, Lam SS, Ong ME. AI and machine learning in resuscitation: ongoing research, new concepts, and key challenges. Resusc Plus 2023;15:100435.
31. Mertens M, King OC, van Putten MJ, Boenink M. Can we learn from hidden mistakes? Self-fulfilling prophecy and responsible neuroprognostic innovation. J Med Ethics 2022;48:922–8.
32. Bento V, Kohler M, Diaz P, Mendoza L, Pacheco MA. Improving deep learning performance by using explainable artificial intelligence (XAI) approaches. Discov Artif Intell 2021;1:9.
33. Okada Y, Komukai S, Kitamura T, et al. Clinical phenotyping of out-of-hospital cardiac arrest patients with shockable rhythm: machine learning-based unsupervised cluster analysis. Circ J 2022;86:668–76.
34. Kamath U, Liu J. Explainable artificial intelligence: an introduction to interpretable machine learning. Springer; 2021.
35. Xie F, Chakraborty B, Ong ME, Goldstein BA, Liu N. AutoScore: a machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records. JMIR Med Inform 2020;8:e21798.
36. Xie F, Ning Y, Liu M, et al. A universal AutoScore framework

to develop interpretable scoring systems for predicting common types of clinical outcomes. STAR Protoc 2023;4:102302.
37. Raith EP, Udy AA, Bailey M, et al. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. JAMA 2017;317:290–300.
38. Sabol P, Sincak P, Hartono P, et al. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. J Biomed Inform 2020;109:103523.
39. Paluszynska A, Biecek P, Jiang Y. randomForestExplainer: explaining and visualizing random forests in terms of variable importance. R ver. 0.9. R Foundation; 2017.
40. Laabs BH, Westenberger A, Konig IR. Identification of representative trees in random forests based on a new tree-based distance measure. Adv Data Anal Classif 2023 Mar 16 [Epub]. https://doi.org/10.1007/s11634-023-00537-7
41. Weinberg AI, Last M. Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification. J Big Data 2019;6:23.
42. Parmar A, Katariya R, Patel V. A review on random forest: an ensemble classifier. In: Hemanth J, Fernando X, Lafata P, Baig Z, editors. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018; 2018 Aug 7–8; Coimbatore, India. Springer; 2018. p. 758–63.
43. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal C, Shen D, Rastogi R, editors, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. Association for Computing Machinery; 2016. p. 1135–44.
44. Okada Y, Shahidah N, Ng YY, et al. Outcome assessment for out-of-hospital cardiac arrest patients in Singapore and Japan with initial shockable rhythm. Crit Care 2023;27:351.
45. Okada Y, Matsuyama T, Morita S, et al. Machine learning-based prediction models for accidental hypothermia patients. J Intensive Care 2021;9:6.
46. Okada Y, Komukai S, Kitamura T, et al. Clustering out-of-hospital cardiac arrest patients with non-shockable rhythm by machine learning latent class analysis. Acute Med Surg 2022; 9:e760.
47. Ning Y, Ong ME, Chakraborty B, et al. Shapley variable importance cloud for interpretable machine learning. Patterns (N Y) 2022;3:100452.
48. Ekanayake IU, Meddage DP, Rathnayake U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). Case Stud Constr Mater 2022;16: e01059.
49. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, et al., editors. Advances in neural information processing systems. 31st Annual Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. Neural Information Processing Systems; 2017. p. 4766–75.
50. Bienefeld N, Boss JM, Luthy R, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digit Med 2023;6:94.
51. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: a systematic review. Heliyon 2023;9:e16110.
52. Liu M, Ning Y, Teixayavong S, et al. A translational perspective towards clinical AI fairness. NPJ Digit Med 2023;6:172.