npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01958-8

Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour

Sujeong Hur[1,2,11], Yura Lee[3,11], Joongheum Park[2,4,5], Yeong Jeong Jeon[6], Jong Ho Cho[6], Duck Cho[7], Dobin Lim[8], Wonil Hwang[8], Won Chul Cha[1,9,10] & Junsang Yoo[1]

Clinical decision-making substantially impacts patients' lives and their quality of life. However, the black-box nature of AI-powered clinical decision support systems (CDSSs) complicates the interpretation of how decisions are derived. Explainable AI (XAI) improves acceptance and trust with explanations, but the effectiveness of different methods remains uncertain. We compared the acceptance, trust, satisfaction and usability of various explanatory methods among clinicians. We also explored the factors associated with acceptance levels for each item using trust, satisfaction and usability score questionnaires. Surgeons and physicians ($N = 63$), who had prescribed blood products before surgery, made decisions before and after receiving one of three CDSS explanation methods, each comprising six vignettes, in a counterbalanced design. We found empirical evidence, which indicates that providing a clinical explanation enhances clinicians' acceptance than presenting 'results only' or 'results with SHapley Additive exPlanations (SHAP)'. Additionally, trust, satisfaction and usability were correlated with acceptance. This study suggests best practices for the strategic application of the XAI–CDSS in the medical field.

A clinical decision support system (CDSS) is defined as 'software that is designed to be a direct aid to clinical decision-making, in which the characteristics of an individual patient are matched to a computerised clinical knowledge base and patient-specific assessments or recommendations are then presented to the clinician or the patient for a decision'[1]. As considerable medical data accumulate, various medical artificial intelligence (AI)-powered CDSSs are being introduced for tasks, such as diagnosis, prognosis prediction and process optimisation. Their performance is often comparable to, or even exceeds, that of human experts[2,3]. AI-powered CDSSs have outperformed traditional statistical methods in identifying nonlinear relationships, which are common in many clinical problems[4]. However, concerns have been raised regarding the lack of interpretability in how these results are derived[5].

Clinical decision-making affects patients' lives and their quality of life. Healthcare professionals are trained to adhere to evidence-based practices when making decisions about their patients for good reasons[6]. However, the black-box nature of AI-powered CDSSs complicates the interpretation of how decisions are derived[7]. This lack of interpretability has been identified as a critical barrier to realising the full potential of medical AI in healthcare, despite its strong performance[8,9].

Since the introduction of explainable AI (XAI), various methodologies based on mathematical and technical principles have been proposed. SHapley Additive exPlanations (SHAP) and local interpretable model agnostic explanations (LIME) are the most extensively used XAI methods [10,11]. SHAP demonstrates the contribution of individual predictions and can provide global surrogate explanations for the overall behaviour of the model. LIME provides a local explanation for any classifier or regressor near a specific instance using interpretable models. These techniques have become the new standards for AI adoption in the healthcare field[12].

[1]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea. [2]AvoMD, New York, NY, USA. [3]Department of Information Medicine, Asan Medical Center, Seoul, Republic of Korea. [4]Assistant Professor, Chobanian & Avedisian School of Medicine, Boston, MA, USA. [5]Associated Harvard Medical Faculty Physician, Beth Israel Deaconess Medical Center, Boston, MA, USA. [6]Department of Thoracic and Cardiovascular Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [7]Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [8]Department of Industrial and Information Systems Engineering, Soongsil University, Seoul, Republic of Korea. [9]Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [10]Digital Innovation Center, Samsung Medical Center, Seoul, Republic of Korea. [11]These authors contributed equally: Sujeong Hur, Yura Lee. e-mail: junnsang@skku.edu

Most AI-powered CDSSs provide clinicians with SHAP plots to visualise how model predictions are derived. Some studies that utilised SHAP with AI prediction models have concluded that it can be helpful in clinical decision-making by providing objective results[13,14]. However, research on how clinicians respond to AI systems based on the way information is presented to them is limited[15]. This highlights the need for further studies to expand our understanding of the explanation methods provided by the system, which can influence clinicians' decision-making processes.

This study aimed to: (1) compare the different explanation methods, from AI-powered CDSS to clinicians and their effects, on acceptance, trust, satisfaction and usability; and (2) explore factors related to clinicians' acceptance of each item on the trust questionnaire, satisfaction questionnaire, and usability score.

Results

Overall, 63 physicians participated in the study. Among them, 37 (58.7%) were females, 20 (31.7%) were surgeons, 24 (38.1%) worked in internal medicine, 14 (22.2%) worked in emergency medicine departments and 5 (7.9%) worked in other departments. A total of 43 (68.3%) were residents, 11 (17.5%) were faculty members and 9 (14.3%) were fellows. Two-thirds of the participants reported no prior experience in using medical AI, and approximately three-fourths reported having limited prior knowledge. After randomisation to minimise order effects, the participants' characteristics showed no statistically significant differences, except for caregivers' specialities. The general characteristics of the participants are presented in Table 1.

Acceptance: weight of advice

Overall, 238 cases were excluded because the initial estimates and AI advice were identical. The AI results-only (RO) group comprised 69 cases (6.08%), whereas the AI results with SHAP plot (RS) group encompassed 85 cases (7.50%), and the AI results with SHAP plot and clinical explanation (RSC) group included 84 cases (7.41%). The average weight of advice (WOA)[16] for the RSC type of AI advice was 0.73 (standard deviation [SD] = 0.26), which was statistically higher than the RS (mean = 0.61, SD = 0.33) and the RO group (mean = 0.50, SD = 0.35) in the Friedman test and subsequent Conover post-hoc analysis (Fig. 1).

Trust, satisfaction and usability of explanation

Across all metrics—Trust in AI Explanation, Explanation Satisfaction and System Usability Scale—the RSC group presented significant benefits compared to RO and RS groups (Fig. 2)

Figure 2a presented that the Trust Scale Recommended for XAI progressively increased across the three conditions: RO (mean = 25.75, SD = 4.50), RS (mean = 28.89, SD = 3.72) and RSC (mean = 30.98, SD = 3.55). A Friedman test confirmed a statistically significant difference among conditions ($p < 0.001$), with post-hoc Conover test indicating significant pairwise differences ($p < 0.001$). These results suggest that SHAP-based explanations enhance user trust, with additional improvements when clinical interpretation is provided. Statistically significant differences were observed for all items on both the Friedman and post hoc Conover test, the RSC group showed statistically higher scores than the RS and RO groups across the following constructs of the Trust Scale Recommended for XAI: confidence, predictability, reliability, safety, wariness, comparison with novice humans and preference ($p < 0.001$). The detailed responses to each item are shown in Supplementary Fig. 1.

A similar trend was observed for Explanation Satisfaction Scale. Satisfaction scores increased from RO (mean = 18.63, SD = 7.20) to RS (mean = 26.97, SD = 5.69), reaching the highest level with RSC

Table 1 | General characteristics of the study population

	RO-RS-RSC (*N* = 10)	RO-RSC-RS (*N* = 11)	RS-RO-RSC (*N* = 11)	RS-RSC-RO (*N* = 11)	RSC-RO-RS (*N* = 10)	RSC-RS-RO (*N* = 10)	*P* value
Age	29.8 ± 3.9	33.5 ± 4.7	32.9 ± 3.6	35.4 ± 6.4	32.6 ± 2.9	33.1 ± 5.4	0.292
Sex							0.832
- Female	4 (40.0%)	7 (63.6%)	7 (63.6%)	6 (54.5%)	6 (60.0%)	7 (70.0%)	
- Male	6 (60.0%)	4 (36.4%)	4 (36.4%)	5 (45.5%)	4 (40.0%)	3 (30.0%)	
Department							0.039
- Surgery	3 (30.0%)	2 (18.2%)	4 (36.4%)	5 (45.5%)	3 (30.0%)	3 (30.0%)	
- Internal medicine	6 (60.0%)	3 (27.3%)	6 (54.5%)	2 (18.2%)	2 (20.0%)	5 (50.0%)	
- Emergency medicine	1 (10.0%)	1 (9.1%)	1 (9.1%)	4 (36.4%)	5 (50.0%)	2 (20.0%)	
- Miscellaneous	0 (0.0%)	5 (45.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Designation							0.923
- Faculty	0 (0.0%)	2 (18.2%)	1 (9.1%)	4 (36.4%)	2 (20.0%)	2 (20.0%)	
- Fellow	1 (10.0%)	2 (18.2%)	2 (18.2%)	0 (0.0%)	3 (30.0%)	1 (10.0%)	
- Resident	9 (90.0%)	7 (63.6%)	8 (72.7%)	7 (63.6%)	5 (50.0%)	7 (70.0%)	
Professional experience (Months)	49.4 ± 21.9	60.5 ± 42.9	63.1 ± 28.7	96.9 ± 78.0	79.4 ± 43.0	74.6 ± 61.7	0.537
Number of transfusions Prescriptions (Pack/Month)	4.0 ± 3.4	5.4 ± 8.3	6.5 ± 5.4	3.6 ± 4.2	7.1 ± 7.7	6.0 ± 4.9	0.536
Artificial intelligence algorithm usage experience							0.516
- Exists	1 (10.0%)	3 (27.3%)	3 (27.3%)	4 (36.4%)	5 (50.0%)	4 (40.0%)	
- Never	9 (90.0%)	8 (72.7%)	8 (72.7%)	7 (63.6%)	5 (50.0%)	6 (60.0%)	
Prior knowledge regarding medical artificial intelligence							0.560
- Very high	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (20.0%)	0 (0.0%)	
- High	7 (70.0%)	5 (45.5%)	8 (72.7%)	6 (54.5%)	5 (50.0%)	6 (60.0%)	
- Moderate	0 (0.0%)	3 (27.3%)	2 (18.2%)	2 (18.2%)	1 (10.0%)	2 (20.0%)	
- Low	0 (0.0%)	1 (9.1%)	1 (9.1%)	2 (18.2%)	0 (0.0%)	0 (0.0%)	
- Very low	3 (30.0%)	2 (18.2%)	0 (0.0%)	1 (9.1%)	2 (20.0%)	2 (20.0%)	

RO result only, *RS* result with SHAP, *RSC* result with SHAP and clinical interpretation.

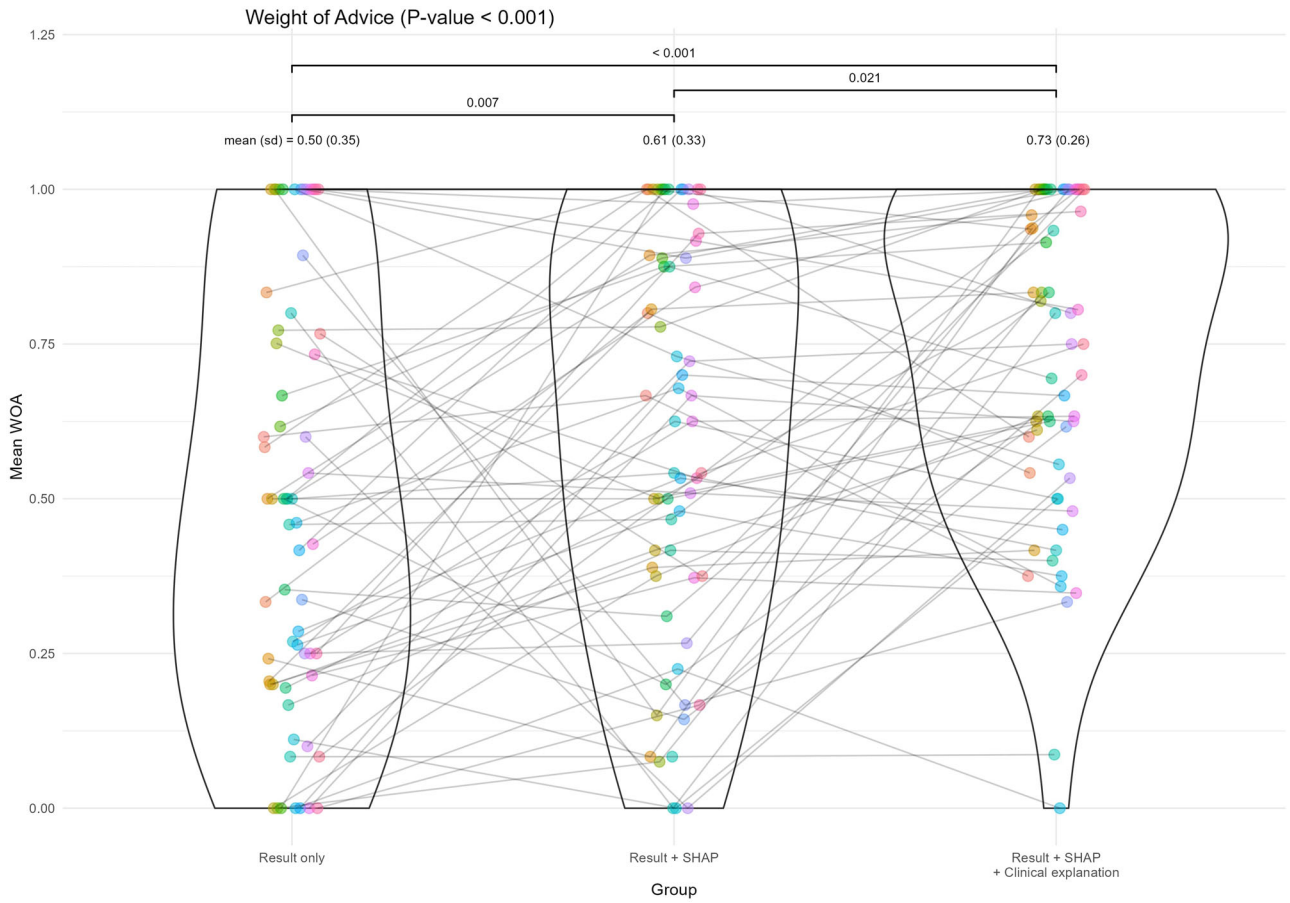

Fig. 1 | Friedman and Conover's post-hoc analysis of the weight of advice. Mean WOA is shown for each of three conditions with violin outlines indicating the distribution across participants, coloured dots representing individual means and grey lines connecting each participant's scores across conditions. Abbreviations: WOA weight of advice, SHAP SHapley Additive exPlanations.

(mean = 31.89, SD = 5.14). A Friedman test confirmed a significant effect ($p < 0.001$), with post-hoc analyses revealing significant differences across all conditions ($p < 0.001$) (Fig. 2b). On the Explanation Satisfaction Scale, the RSC group presented significantly higher scores than the other advice types for all items, including understanding the algorithm's operation, satisfaction with the explanation, appropriateness of detailed information, completeness of the explanation, understanding the usage method, utility of the explanation, understanding the accuracy and trustworthiness of the explanation (Supplementary Fig. 2).

The mean System Usability Scale (SUS)[17] score for the RSC group 72.74 (SD = 11.71), indicating good and acceptable usability. This score was significantly higher than that of the RS (mean = 68.53, SD = 14.68) and RO (mean = 60.32, SD = 15.76), both classified as marginally acceptable (Fig. 2c). A Friedman test confirmed a significant difference among groups ($p < 0.001$), with post-hoc analysis indicating significant improvements from RO to RS ($p < 0.001$) and from RS to RSC ($p = 0.025$). Supplementary Table 1 provides a detailed breakdown of item-level responses.

Correlation analysis

In the correlation between participants' acceptance of the CDSS recommendations and the trust, satisfaction and usability of the explanation methods, statistically significant correlations were observed across all items except wariness[18]. We found a moderate correlation for the construct of trust in 'predictability ($r = 0.463$)', 'comparison with novice human ($r = 0.432$)', 'preference ($r = 0.431$)', 'confidence ($r = 0.429$)', 'efficiency ($r = 0.427$)' and for the construct of explanation satisfaction in 'appropriateness of detailed information ($r = 0.431$)', 'trustworthiness of the explanation ($r = 0.414$)', 'utility of the explanation ($r = 0.412$)', 'completeness of the explanation ($r = 0.41$)', 'satisfaction of the explanation ($r = 0.41$)', 'understanding of the algorithm's operation ($r = 0.408$)'; as well as 'SUS score ($r = 0.434$)'. Weak correlations were found for 'safety ($r = 0.396$)', 'reliability ($r = 0.353$)' and 'wariness ($r = -0.183$)' for trust; and 'understanding of the usage method ($r = 0.397$)' and 'understanding of accuracy ($r = 0.328$)' for explanation satisfaction (Fig. 3).

Post-hoc analysis

We performed subgroup analyses to evaluate whether clinician experience levels and departmental differences influenced the primary and secondary outcomes. Across clinician experience levels and departments, the overall pattern was consistent with the primary findings, demonstrating a clear directional trend (RSC > RS > RO) in WOA, trust, explanation satisfaction and usability scores (Supplementary Tables 2 and 3). Although subgroup sizes were limited, this consistency supports the robustness of our main results regardless of clinician experience level or department. However, small sample sizes in certain subgroups resulted in analyses not reaching statistical significance; therefore, caution is necessary when interpreting these subgroup-specific results.

We additionally conducted the mean absolute decision change to include all previously excluded cases in WOA analysis. There was a significant difference among the three explanation method ($p = 0.014$) with the largest mean absolute decision change observed in the RSC (mean = 1.43, SD = 0.89), followed by RO (mean = 1.23, SD = 1.01) and RS (mean 1.21, SD = 0.93). These findings further support the robustness of our main findings.

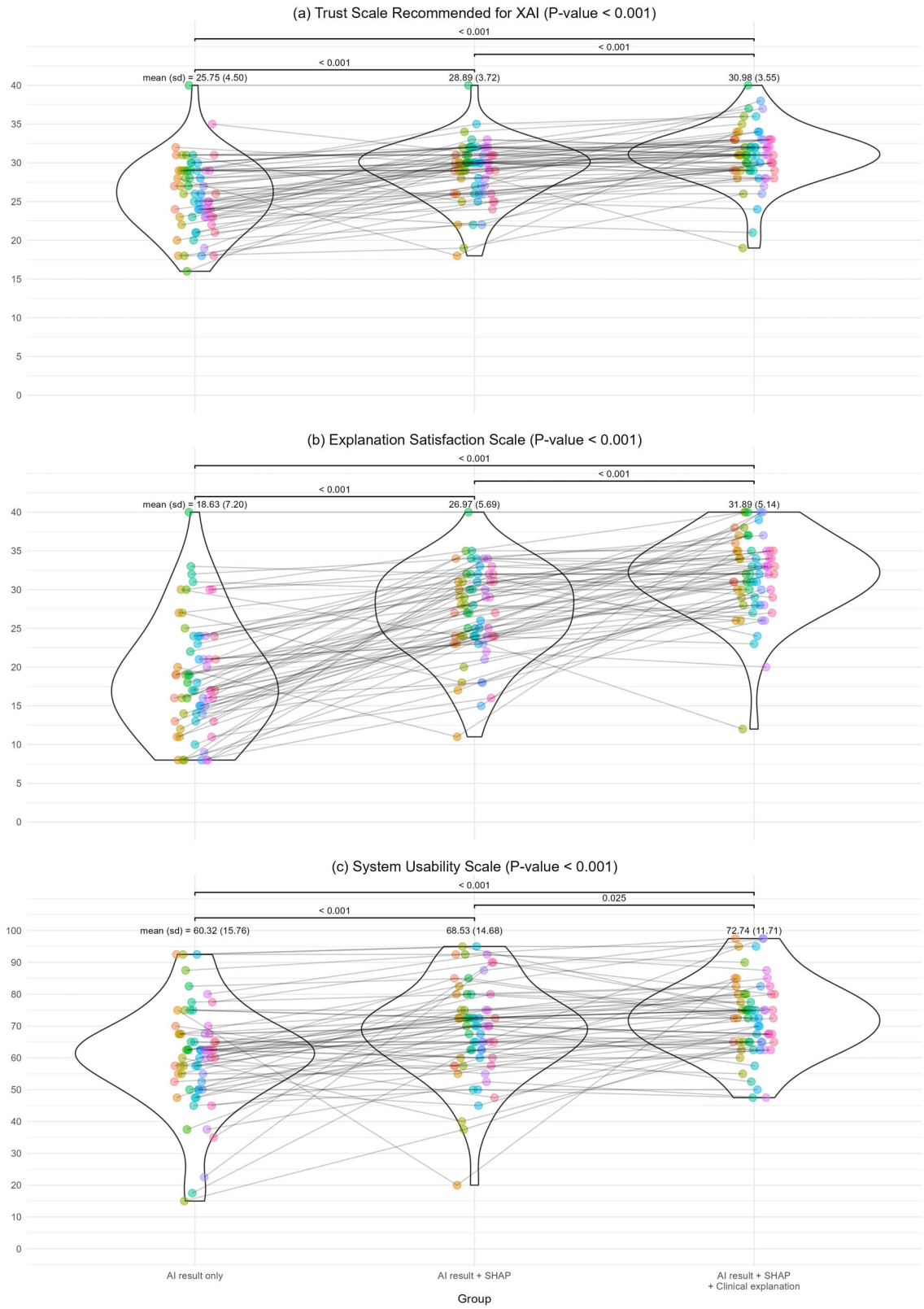

Fig. 2 | Comparison of clinicians' trust, satisfaction, and usability across explanation formats. Friedman and Conover's post-hoc analysis of the '(**a**) Trust Scale Recommended for XAI', '(**b**) Explanation Satisfaction Scale', '(**c**) System Usability Scale'. Abbreviations: XAI explainable artificial intelligence, SHAP SHapley Additive exPlanations.

Discussion

XAI plays a crucial role in enhancing trust and transparency, enabling users to accept AI recommendation[19]. However, an uncharted gap remains between the explanations provided by AI-based CDSS developers and clinicians' understanding. Moreover, there is no consensus on how explanation formats affect clinicians' acceptance to AI suggestions[13,20]. This study evaluated how different explanation formats—RO, RS, RSC—influence clinicians' acceptance and potential factors that can affect acceptance using a CDSS designed to predict perioperative blood transfusion requirements. Our results showed that the RSC condition significantly improved

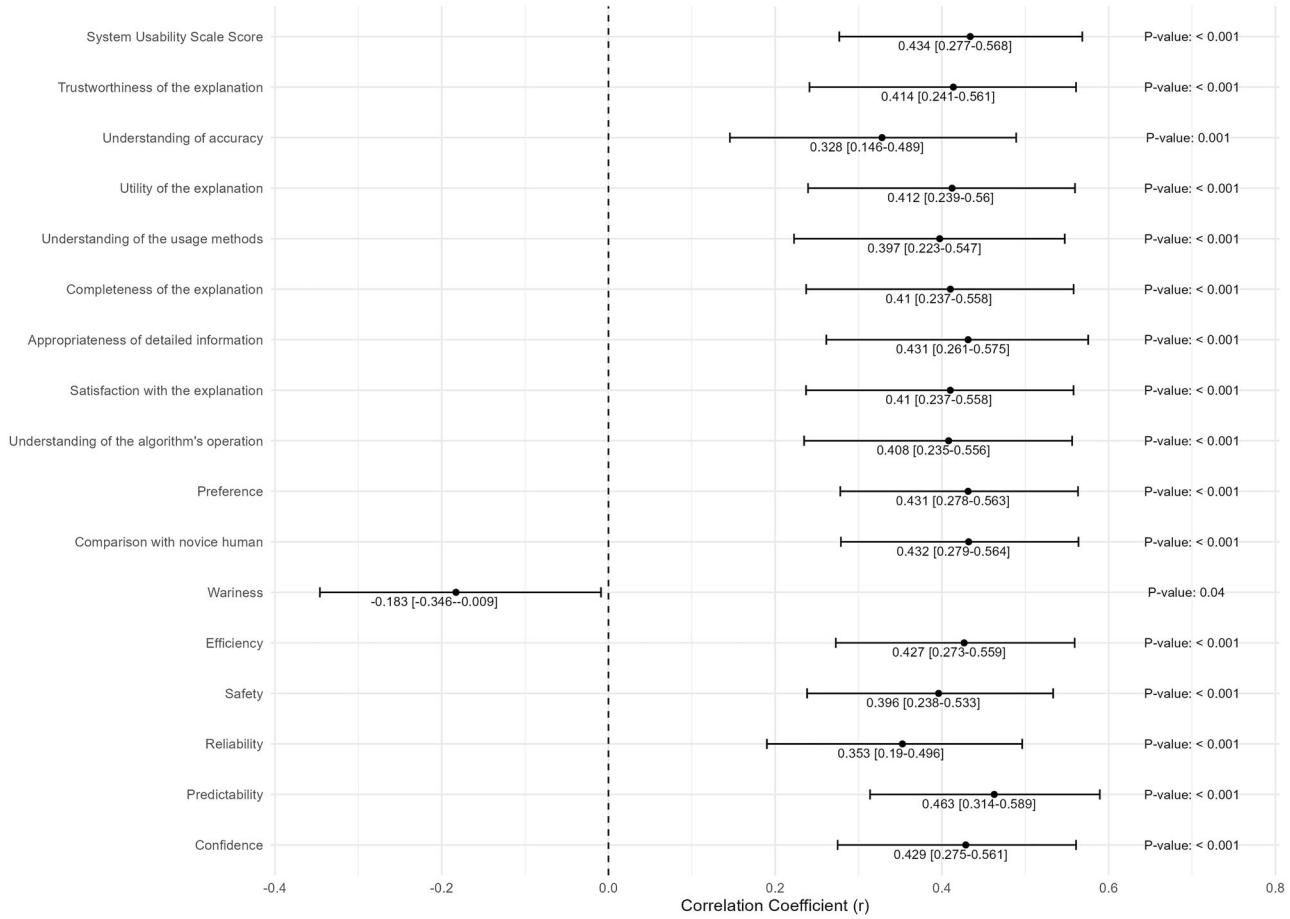

Fig. 3 | Correlation between mean weight of advice and scores from the System Usability Scale, Trust Scale Recommended for XAI and Explanation Satisfaction Scale. Repeated-measures correlation coefficients (with 95% confidence intervals) between mean weight of advice and scores on each item of the Trust Scale Recommended for XAI and the Explanation Satisfaction Scale, as well as the overall score of the System Usability Scale, are shown as horizontal bars. A vertical dashed line indicating no association.

WOA, trust, explanation satisfaction and usability compared to the RO and RS conditions. These findings offer actionable insights for CDSS developers by providing practical guidance on selecting explanation formats to support clinical implementation.

Prior studies have reported mixed findings on the effects of explanation formats. In a CDSS predicting gestational diabetes risk, providing either feature contribution-based or example-based explanations did not lead to significant differences in WOA[13]. Similarly, in a medication prescribing scenarios, clinicians' acceptance did not differ significantly between AI recommendations presented alone and those accompanied by explanations[21]. This suggests that simple or fragmented explanations to AI recommendations are insufficient to influence expert level decision-making. Meanwhile, a study has reported that explanation format could influence clinicians' behaviour to AI-based CDSS[22]. In this multi-centre study conducted in the radiology field[22], local explanations that highlighted X-ray image regions led to faster agreement and reduced decision time compared to case-based explanations referencing similar cases. However, the study focused on time and diagnostic accuracy, rather than directly assessing acceptance behaviour using measures such as the WOA, which limits the generalisability of its findings to acceptance-related outcomes.

The potential of CDSS cannot be realised without clinician acceptance[23]. Our study provides empirical evidence that presenting SHAP-based explanations in a clinical narrative format significantly improves clinicians' acceptance, trust, explanation satisfaction and usability of AI-based CDSS. The RSC condition, which integrates SHAP visualisations with natural language explanations, consistently outperformed both the RO and RS conditions across all outcomes and subgroups. Trust, satisfaction and usability scores were moderately correlated with WOA, suggesting that these factors may mediate clinicians' acceptance behaviour. However, as this study is correlational, causal interpretations are limited. To address this, more sophisticated mental models are needed to explain how clinicians interpret AI explanations and apply them to clinical decision-making. Such models may serve as design blueprints for future CDSS, supporting both clinical adoption and system effectiveness. The explanation approach proposed in this study aligns with the DoReMi framework[24], which emphasises user-centred requirements and reusable design patterns for XAI.

The AI-CDSS used in our study generated explanations by selecting the top three SHAP values in a rule-based manner. In some real-world clinical cases, this method could result in explanations that are unnatural or poorly contextualised. To overcome these limitations, recent approaches have incorporated large language models (LLMs) to generate more human-centred narrative explanations from SHAP outputs. LLMs excel at generating natural, conversational language that closely resembles human communication, making them well-suited for improving the clarity and accessibility of AI-based CDSS explanations[25]. A study attempted to preserve explanation quality by using LLMs to translate SHAP results into natural language[26]. Similarly, another study approach used GPT-4 to generate narrative explanations by prompting it with model predictions, input features, and SHAP values[27]. In that study, 90% of lay users evaluated the explanations positively, and 83% of data scientists believed the output would be helpful for non-experts[27]. Although these studies were conducted in non-medical domains such as finance, their outcomes suggest that this approach

is worth exploring in clinical contexts. Comparative studies of clinician responses to explanations from LLM-based versus rule-based CDSS remain a key direction for future research.

Our study had several limitations. First, the eligibility criteria and participant characteristics in this study may limit the generalisability of our findings. The simulation was conducted in a tertiary academic hospital, where 30% of participants had prior experience with AI in clinical practice. To preserve realism in perioperative transfusion decisions, we excluded clinicians without experience prescribing blood products. However, it may have reduced the applicability of our findings to novice clinicians and settings without AI-based CDSS. Although subgroup analyses showed trends consistent with the main findings, small sample sizes in each subgroup limited the statistical power to confirm significance across all groups. In summary, future studies should broaden inclusion criteria and diversify settings to better assess the applicability of these findings.

Second, due to experimental complexity constraints, we could not include the following conditions in our study design: (1) an independent 'clinical explanation alone' group, as clinical explanations in this study were directly derived from SHAP outputs; and (2) alternative widely used XAI methods such as LIME or counterfactual explanations. Further research is necessary to comprehensively assess the relative effectiveness and generalisability of these different XAI approaches.

Third, we cannot rule out the possibility that financial incentives may have influenced response quality.

In conclusion, our study demonstrates that integrating natural language-based clinical explanations into AI-based CDSSs can significantly improve clinicians' acceptance, trust, satisfaction and usability. Using a CDSS developed to predict personalised red blood cell transfusion needs in thoracic surgery patients, we found that the RSC condition consistently outperformed both RO and RS formats across all evaluation metrics. These findings underscore the importance of delivering AI outputs in clinically meaningful and understandable formats. The observed correlations between trust, satisfaction, usability and acceptance suggest that enhancing the quality and clarity of explanations can drive greater clinician engagement with AI systems. To translate these insights into real clinical practice, AI-based CDSS developers should design explanation formats that are clinically relevant and easy for clinicians to interpret. In parallel, hospitals should provide training to help clinicians understand and apply these explanations effectively, rather than relying solely on technical visualisations.

Methods

Introduction to the 'pMSBOS-TS' CDSS

In our previous work, we introduced an AI algorithm named 'precision maximum surgical blood ordering schedule-thoracic surgery (pMSBOS-TS)'[28]. This extreme gradient-boosting model precisely predicts the demand for red blood cells (RBC) in thoracic surgery patients. This algorithm demonstrates promising performance in improving both the efficient use of blood products and patient safety by reflecting patient-specific decision processes. We developed an AI-powered CDSS using this model as the reasoning engine. The output of the pMSBOS-TS CDSS includes the predicted blood requirement AI outcome value, SHAP and a clinical explanation of the top three features from the SHAP value (Fig. 4).

Study setting

The simulation study was conducted at a tertiary academic hospital in Seoul, Republic of Korea. The inclusion criteria were surgeons and physicians who had prescribed any blood product before surgery within the past 5 years. This recruitment criterion was intentionally set to ensure participants could make realistic and clinically meaningful judgements in perioperative transfusion scenarios presented through the vignettes, based on their direct clinical experience. Clinicians without experience in prescribing blood products were excluded.

The sample size was calculated using G*Power software (version 3.1.9.7)[29], considering that participants would repeatedly evaluate all three methods. An initial sample size of 55 participants was calculated based on an moderate effect size (F) of 0.20, a significance level (α) of 5%, a power (β) of 0.9, correlation among repeated measures of 0.50 and assuming perfect sphericity ($\varepsilon = 1.0$). Considering a dropout rate of 20%, the final target sample size was 66.

Study protocol and vignette experiment

Eighteen distinct clinical vignettes were created, simulating patients scheduled for thoracic surgery. Our study employed a repeated-measures, within-subject experimental design, wherein each participant experienced all three experimental conditions: (1) Results Only (RO), (2) Results with SHAP visualisations (RS) and (3) Results with SHAP visualisations and Clinical explanations (RSC). Specifically, each participant reviewed six distinct vignettes per condition, totalling 18 vignettes per participant. For each vignette, participants initially estimated the number of RBC packs required, then made a final decision after reviewing the advice from the CDSS presented in one of the three explanation formats.

To minimise potential order effects[30], such as learning or fatigue, we employed a systematically counterbalanced design (Fig. 5). Participants were randomly assigned to one of six groups, each experiencing the three explanation conditions in a different sequence. After reviewing each condition (block of six vignettes), participants completed questionnaires measuring trust, satisfaction and usability related to the CDSS explanation method they had just experienced. This process was repeated until all three conditions were completed.

Upon completing the study, participants were compensated with 50,000 Korean won (approximately $36). The compensation was determined based on our institutions' standard practice and the average hourly rate for clinicians' time.

Outcomes

The primary outcome was the WOA, which measures the degree of acceptance of decision-making before and after receiving CDSS suggestions[16].

The formula for calculating WOA is as:

$$WOA = \frac{|Final\ estimate - Initial\ estimate|}{|Advice - Initial\ estimate|}$$

In this calculation, if the initial estimate equals the value obtained by the algorithm, the denominator becomes zero and the WOA cannot be calculated; therefore, it is excluded from the analysis and the average for each part is derived. The WOA ranges from 0 to 1, where 0 implies that the initial and final estimates are identical, indicating that the CDSS does not influence the decisions. In cases where the WOA value was >1 (e.g. when the initial estimate was 3, the AI's advice was 5, while the clinicians' final estimation was 7, resulting WOA of 2), the WOA is capped at 1, indicating full acceptance of the AI's advice.

The secondary outcomes were trust, satisfaction and usability, measured using the 'Trust Scale Recommended for XAI', 'Explanation Satisfaction Scale' and SUS, respectively[17,31]. All instruments used in this study were rated on a 5-point Likert scale.

An additional exploratory outcome, termed the 'mean absolute decision change' was proposed as a post-hoc measure to evaluate decision changes even in cases where WOA was not calculable. This index was defined as the mean absolute difference between each participant's final and initial decisions.

Statistical analysis

First, for each type of AI advice, the mean WOA was calculated for each participant. Trust and satisfaction with the explanations were scored based on individual items using the respective instruments. Additionally, the SUS score was computed for each type of AI advice.

The differences between the types of AI advice were tested accordingly. Considering that all participants were required to respond to

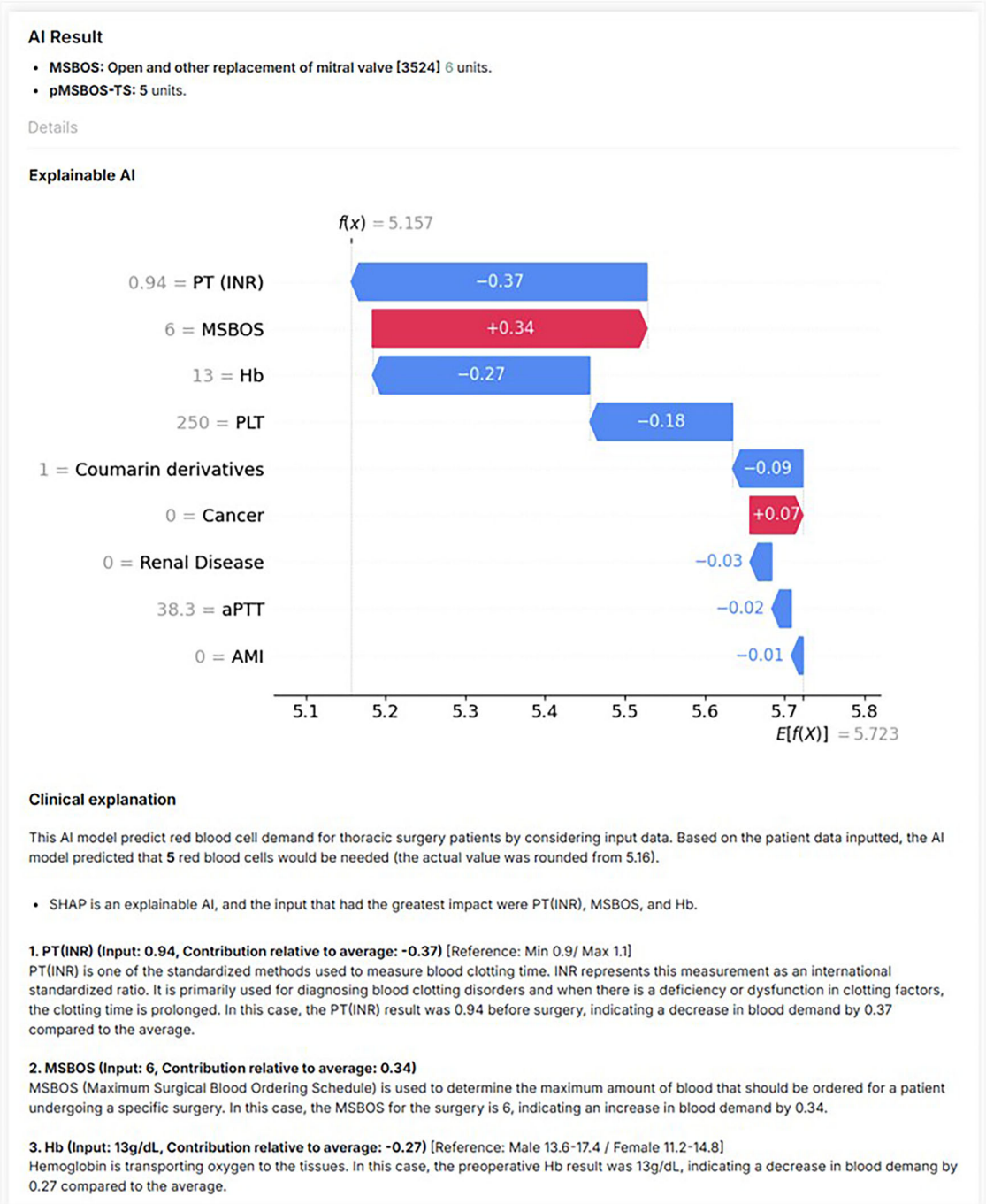

Fig. 4 | Configuration of the output section of the pMSBOS-TS CDSS (Reused with permission from the authors' prior work: Hur et al.[28]). Abbreviations: AMI acute myocardial infarction, aPTT activated partial thromboplastin time, Hb haemoglobin, MSBOS maximum surgical blood order schedule, PT (INR) prothrombin time-international normalised ratio.

all three types of AI advice, the data comprised repeated measures. Given that the assumptions of normality and homogeneity of variance were not met for all variables, the Friedman test was applied as a nonparametric alternative to assess the differences. When significant differences were confirmed, a post-hoc Conover test was conducted to identify specific group differences. A p value of 0.05 was used to determine the statistical significance of all analyses.

Then, we applied a repeated-measures correlation (RMCORR) to examine the relationships between acceptance and trust, explanation satisfaction and usability[18].

Explorative descriptive subgroup post-hoc analysis were conducted. Participants were stratified according to the clinician experience level (resident, fellow, faculty member) and the clinical department (surgical department, internal medicine, emergency medicine and miscellaneous department). Within each subgroup, we re-examined the WOA as well as trust, explanation satisfaction and usability. Additionally, the mean absolute decision change was analysed as a post-hoc measure, allowing inclusion of previously excluded cases due to identical initial estimates and AI advice.

All statistical analyses were performed using R software (version 4.1.1; R Core Foundation, Vienna, Austria)[32].

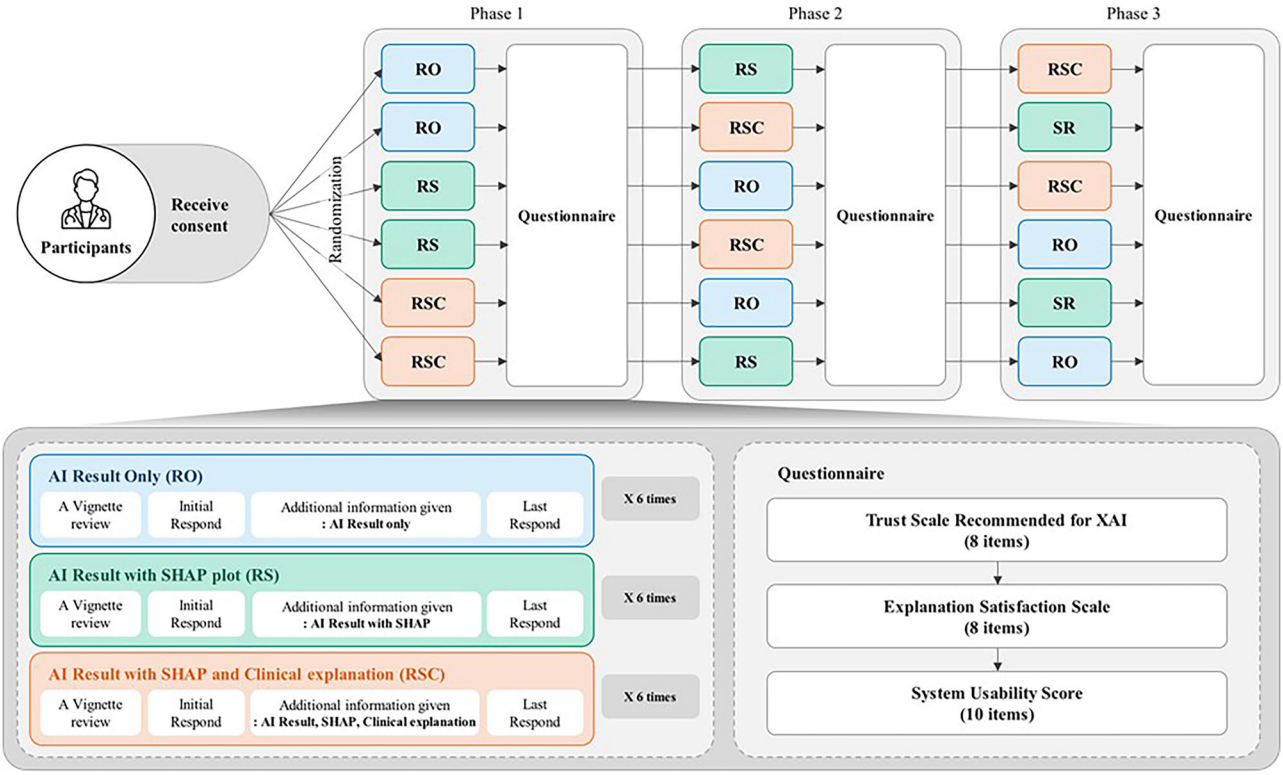

Fig. 5 | Counterbalancing study design to control order effect. Participants were randomly assigned to one of six groups, each experiencing the three explanation conditions in a different sequence. After reviewing each condition (block of six vignettes), participants completed questionnaires measuring trust, satisfaction and usability related to the CDSS explanation method they had just experienced.

Ethics declarations

The study protocol was reviewed and approved by the Institutional Review Board of Samsung Medical Centre (approval No. 2023-09-123). Written informed consent was obtained from all the participants.

Data availability

The data used in this manuscript are available in the supplementary material and on GitHub, and all code is available on GitHub at (URL: https://github.com/junnsang/Explainability).

Code availability

All code used in this manuscript is available online at GitHub (URL: https://github.com/junnsang/Explainability).

Received: 28 February 2025; Accepted: 17 August 2025;
Published online: 26 September 2025

References

1. Sim, I. et al. Clinical decision support systems for the practice of evidence-based medicine. *J. Am. Med. Inform. Assoc.* **8**, 527–534 (2001).
2. Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).
3. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
4. Livne, M. et al. Boosted tree model reforms multimodal magnetic resonance imaging infarct prediction in acute stroke. *Stroke* **49**, 912–918 (2018).
5. Martens, D., Baesens, B. B. & Van Gestel, T. Decompositional rule extraction from support vector machines by active learning. *IEEE Trans. Knowl. Data Eng.* **21**, 178–191 (2009).
6. Lehane, E. et al. Evidence-based practice education for healthcare professions: an expert view. *BMJ Evid. Based Med.* **24**, 103–108 (2019).
7. Hassija, V. et al. Interpreting black-box models: a review on explainable artificial intelligence. *Cogn. Comput.* **16**, 45–74 (2024).
8. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **9**, 1–13 (2019).
9. Durán, J. M. & Jongsma, K. R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J. Med. Ethics* **47**, 329–335 (2021).
10. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst*. https://doi.org/10.48550/arXiv.1705.07874 (Curran Associates, Inc., 2017).
11. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you?': explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Vol. 16, 1135–1144 (Association for Computing Machinery, 2016).
12. Tjoa, E. & Guan, C. A Survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4793–4813 (2021).
13. Du, Y., Antoniadi, A. M., McNestry, C., McAuliffe, F. M. & Mooney, C. The role of XAI in advice-taking from a clinical decision support system: a comparative user study of feature contribution-based and example-based explanations. *Appl. Sci.* **12**, 10323 (2022).
14. Antoniadi, A. M. et al. A Clinical decision support system for the prediction of quality of life in ALS. *J. Pers. Med.* **12**, 435 (2022).

15. Antoniadi, A. M. et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Appl. Sci.* **11**, 5088 (2021).
16. Harvey, N. & Fischer, I. Taking advice: accepting help, improving judgment, and sharing responsibility. *Organ. Behav. Hum. Decis. Process.* **70**, 117–133 (1997).
17. Brooke, J. SUS: A 'quick and dirty' usability scale. in *Usability Evaluation in Industry* 207–212 (CRC Press, 1996).
18. Bakdash, J. Z. & Marusich, L. R. Repeated measures correlation. *Front. Psychol.* **8**, 456 (2017).
19. Nunes, I. & Jannach, D. A systematic review and taxonomy of explanations in decision support and recommender systems. *Use. Model. User Adapt. Interact.* **27**, 393–444 (2017).
20. Du, Y., Rafferty, A. R., McAuliffe, F. M., Wei, L. & Mooney, C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Sci. Rep.* **12**, 1170 (2022).
21. Nagendran, M., Festor, P., Komorowski, M., Gordon, A. C. & Faisal, A. A. Quantifying the impact of AI recommendations with explanations on prescription decision making. *NPJ Digit. Med.* **6**, 206 (2023).
22. Prinster, D. et al. Care to Explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology* **313**, e233261 (2024).
23. Yoo, J., Hur, S., Hwang, W. & Cha, W. C. Healthcare professionals' expectations of medical artificial intelligence and strategies for its clinical implementation: a qualitative study. *Healthc. Inform. Res.* **29**, 64–74 (2023).
24. Hwang, J., Lee, T., Lee, H. & Byun, S. A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: user-centered design and evaluation study. *J. Med. Internet Res.* **24**, e28659 (2022).
25. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
26. Zeng, X. Enhancing the interpretability of SHAP values using large language models. arXiv preprint. https://doi.org/10.48550/arXiv.2409.00079 (2024).
27. Martens, D., Hinns, J., Dams, C., Vergouwen, M. & Evgeniou, T. Tell me a story! narrative-driven XAI with large language models. *Decis. Support Syst* **191**, 114402 (2025).
28. Hur, S. et al. Development, validation, and usability evaluation of machine learning algorithms for predicting personalized red blood cell demand among thoracic surgery patients. *Int. J. Med. Inform.* **191**, 105543 (2024).
29. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).
30. Pollatsek, A. & Well, A. D. On the use of counterbalanced designs in cognitive research: a suggestion for a better and more powerful analysis. *J. Exp. Psychol. Learn. Mem. Cogn.* **21**, 785–794 (1995).
31. Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front. Comput. Sci.* **5**, 1096257 (2023).
32. R. Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. (2017).

Acknowledgements

This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant numbers RS-2022-KH130250 and HI22C1719).

Author contributions

S.H. conceived and designed the study and contributed to the writing and revision of the manuscript. Y.L. contributed to reviewing, editing and conceptualisation. J.P. is the founder who developed the CDSS platform used in this study. Y.J.J. and J.H.C. contributed to recruited subjects and reviewed the manuscript. D.C. and W.C.C. contributed review and editing of the manuscript. D.L. and W.H. contributed to designing the study, reviewing and editing the manuscript. J.Y. contributed to data analysis, writing, revision, manuscript approval, funding acquisition and supervision. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01958-8.

Correspondence and requests for materials should be addressed to Junsang Yoo.