Integrating Explainable AI in Medical Devices: Technical, Clinical and Regulatory Insights and Recommendations

Dima Alattal[1], Asal Khoshravan Azar[2], Puja Myles,
Richard Branson,[3,4†], Hatim Abdulhussein[5], Allan Tucker[6*]

[1,2,6*]Computer Science Department, Brunel University London, UK.
[3,4] Medicine and Healthcare products Regulatory Agency, UK.
[5] NHS England, UK.

*Corresponding author(s). E-mail(s): allan.tucker@brunel.ac.uk;
Contributing authors: dima.alattal@brunel.ac.uk;
asal.khoshravanazar@brunel.ac.uk; puja.myles@mhra.gov.uk,
richard.branson@mhra.gov.uk; hatim.abdulhussein@nhs.net;
†These authors contributed equally to this work.

Abstract

Purpose: There is a growing demand for the use of Artificial Intelligence (AI) and Machine Learning (ML) in healthcare, particularly as clinical Decision Support Systems (CDSS) to assist medical professionals. However, the complexity of many of these models, often referred to as black box models, raises concerns about their safe integration into clinical settings as it is difficult to understand how they arrived at their predictions. Explainable Artificial Intelligence (XAI) offers a potential solution by providing justifications for the decisions produced by these models, thereby enhancing trust and understanding among clinicians. To address the aspects of trust and safety, it is essential to consider AI medical devices from various perspectives, including clinical, technical, and regulatory perspectives.

Methods: This paper discusses insights and recommendations derived from an expert working group convened by the UK Medicines and Healthcare products Regulatory Agency (MHRA). The group consisted of healthcare professionals, regulators, and data scientists, with a primary focus on evaluating outputs from

different AI/XAI algorithms in clinical decision-making contexts. Additionally, the group evaluated findings from a pilot study investigating clinicians' behavior and interaction with XAI methods during clinical diagnoses.

Results: While the data science team provides technical results, the regulators and clinicians highlight their main concerns and recommendations for using AI/XAI methods in CDSS. The study reveals an overall increase in clinicians' trust and diagnostic accuracy when using local explanations, although over-reliance on AI suggestions raises safety concerns from a legal perspective. The study also underscores the importance of other explanation methods in clinical settings from different perspectives, such as global and counterfactual explanations.

Conclusion: Incorporating XAI methods is crucial for ensuring the safety and trustworthiness of medical AI devices in clinical settings. Adequate training for stakeholders is essential to address potential issues, and further insights and recommendations for safely adopting AI systems as CDSS are provided.

Keywords: eXplainable AI, CDSS, medical devices, AI regulation

1 Introduction

1.1 Background

In clinical settings, AI/ML models can be used as decision support systems through supporting healthcare providers and automating routine tasks[1]. These systems assist clinicians in diagnosing diseases and making decisions about treatment. Unlike conventional CDSS, which match patient characteristics to an existing knowledge base, AI/ML based CDSS (AI-CDSS) apply models trained on data from patients with similar conditions. Despite its potential, AI is not a universal solution and brings novel questions and significant challenges. Some are related to the nature of AI models and others are related to regulatory, medical, and patient perspectives[2]. Therefore, a multidisciplinary assessment is essential for the safe introduction of AI-based medical devices in clinical settings.

Clinicians trust in AI-CDSSs is crucial for a safe implementation of AI in clinical settings. However, it can be challenging to build trust in these systems in a setting where clinicians are required to make urgent decisions that could have serious

consequences[1]. XAI has emerged as a potential method for the safe implementation of AI. Accurate diagnosis alone may not suffice; an explanation for the AI's decision-making process is crucial [3]. This necessity for transparency was highlighted from the very early days of AI diagnostic systems, where researchers found that the ability to explain decisions was the most desired feature among clinicians[4]. Recent studies corroborate this, showing that clinicians value understanding the reasoning behind complex AI systems, often referred to as "black boxes", particularly when their recommendation do not align with clinical expectations [2, 5]. Black box models that often use millions of parameters to capture the non-linearity of input features is a major challenge as it undermines trust and hinders interpretation of the models' predictions. Research indicates that XAI can enhance transparency and trust, potentially leading to better healthcare outcomes even if the diagnostic accuracy is not perfect [3].

XAI in general refers to the characteristic of an AI-driven system that allows a person to understand the reasoning for a model's outputs. XAI aims to provide interpretability, explainability, and transparency to support healthcare practitioners in their decision-making. Interpretability involves comprehending the inner workings of the model and understanding how it generates predictions. On the other hand, explainability focuses on providing clear and understandable explanations for particular AI decisions, actions, or recommendations. Transparency in the context of XAI, ensures that all stakeholders have a clear understanding of the functioning of an AI system. This could involve for example providing stakeholders with information about the data used, how it is processed, and the underlying assumptions guiding the development of the AI system. Nevertheless, the debate around XAI extends beyond technical aspects, touching on regulatory and ethical concerns that could impede progress if not adequately addressed. Without thorough consideration of XAI methods, AI technologies might neglect core ethical and professional principles, overlook regulatory

3

concerns, and cause significant harm [6]. Therefore, XAI is expected to enhance decision confidence, generate hypotheses about causality, and increase trustworthiness and acceptability of the system. It could also help uncover historical actions and biases embedded in AI models trained with historical data [7]. More investigation is needed to ensure healthcare professionals can understand and rely on XAI methods in CDSSs[8]. This work adopts a multidisciplinary view to explore how XAI can facilitate the safe introduction of AI in clinical settings, emphasising the importance of this feature for building clinicians trust and ensuring regulatory compliance. This paper reports the findings of the group, focusing on:

- Identifying key concerns and recommendations of regulators regarding using different AI/ML models that vary in their complexity in AI-CDSS.
- Investigating the information needs and main concerns of medical professionals when employing XAI CDSSs in daily clinical situations.
- Evaluating state-of-the-art explanation methods for providing meaningful and helpful explanations in clinical settings.
- Providing a set of recommendations and insights to guide the adoption of AI/XAI in clinical settings.

1.2 Clinical and Regulatory Perspectives

AI systems used in healthcare are deemed medical devices by the MHRA in the UK if they are designed for medical purposes like diagnosis, treatment, or monitoring. As a result, deploying these devices requires navigating a complex regulatory and ethical framework to ensure they comply with safety, quality, and performance standards. As AI continues to evolve at a rapid rate, the risks associated with not fully utilising its capabilities are becoming more concerning [9]. A primary challenge for AI/ML as a medical device is there is ambiguity in applying clinical evidence requirements and evaluating the performance and effectiveness of these models[10].

A major area to consider when evaluating the effectiveness of the systems is understanding that the link between a specific belief in automation and its actual capabilities varies. These Trust-related biases are the main factor behind algorithm trust (algorithm appreciation) ,overtrust and distrust (algorithm aversion) [11, 12]. Trust calibration in AI refers to the process of appropriately adjusting the level of trust that human users have in AI systems based on their actual reliability[13]. It is important for human-AI collaboration to have trust that is properly calibrated to ensure safety and efficiency[14]. Poorly calibrated trust in CDSS can lead to serious issues with safety [15]. Trust calibration involves understanding the limitations and likely failures of AI systems and adjusting trust in their outputs accordingly[16].

Explanation of AI predictions is thought to be a solution for this problem and is considered a prerequisite for medical AI [17]. However, some argue that trust calibration errors can also occur when users interact with explanations provided by AI systems, leading to irrational or ill considered agreement or disagreement with AI recommendations[18]. As a result, practitioners should retain the authority to make final decisions to ensure that AI systems are effective at diagnosing and treating patients[19].

XAI allows practitioners to make more nuanced and informed decisions. However, to ensure effectiveness it is important to evaluate the human–AI interaction. Observing how end-users understand explanations and evaluating the explanations influence on user behaviour requires engaging end-users such as healthcare providers, and assessing their usage in decision-making contexts[20].

"Assurance" in AI-CDSS refers to the confidence that a system will behave as intended in its intended environment, with a focus on patient safety. Assurance involves both verification and validation: verification ensures that the system is built correctly according to the defined requirements, while validation ensures that the right system is built, meeting the intended purpose and goals. In situations where requirements are

implicit, XAI methods are used to provide explanations that allow for direct validation of the ML model. These explanations show that the predictions are based on reliable clinical variables and are consistent with clinical knowledge[21].

When evaluating the CDSS performance, robustness is considered an important factor, which refers to the model's ability to maintain its performance even when input features vary slightly. Since safety in an AI/ML application is dependent on explainability and performance, and there is no binary choice between them, safety requirements should be partially, even if not entirely, transformed into explainability requirements[20]. Therefore, if an interpretable model can achieve performance levels comparable to a black box model, the interpretable model should be preferred [21, 22].

1.3 Technical Perspective for XAI

There is a debate about the trade-off between a model's performance and its interpretability. AI/ML models with higher performance tend to be based on more complex algorithms, which can make them less interpretable which in turn makes it difficult to understand how they arrive at their predictions[23, 24] .Conversely, models with greater interpretability, commonly known as white-box models, may compromise some performance to deliver transparent and understandable outputs[25].

The challenge lies in finding a balance between the two, where the model is more accurate and understandable. XAI methods have been proposed to address this issue by producing human-interpretable representations of ML/AI models. These methods can contribute to safety assurance in healthcare by providing evidence to support the safety of complex AI/ML-based systems[26]. However, this trade-off is not always gradual and can vary depending on the specific application[27, 28].

XAI can be achieved using intrinsically interpretable models which are models that are transparent and explainable by design or through post-hoc XAI methods that

provide explanations without opening the complex black box model [29, 30]. Model-agnostic XAI methods refer to techniques that provide explanations for the output of AI systems without relying on the internal workings of the specific AI/ML models used [22]. These explanations can be provided at both local and global levels, highlighting the contribution of different features to the model's output [31]. Local explanations explain an individual decision based on one case, supporting trust in that individual decision, while global explanations, explain a model more generally across the entire training set, thus ensuring that the model behaves reasonably when deployed[32]. In clinical settings, both local and global explanations are highly relevant, as they align with the methods clinicians use to justify their diagnoses and treatments. Clinicians often explain how a disease or diagnostic process works in general (global explanation) and justify specific diagnoses based on individual patient data, such as symptoms, test results, and medical history (local explanation)[33]. This parallel enhances the potential for XAI to support clinical decision-making effectively.

XAI methods offers explanations in various forms such as natural text, parameter influence and data visualisations[34]. The use of visualizations by XAI systems enhances the transparency and comprehensibility of decisions, although clinicians' preferences for explanation methods and types vary significantly and often differ from those of developers [35]. The Local Interpretable Model-agnostic Explanations (LIME) is a well-known method for providing local explanations based on which individual features impact a decision [32, 36]. LIME uses a model-agnostic and application-agnostic approach to extract explanations from AI models regardless of domain. However, this comes with a drawback that model-agnostic approaches cannot always meet the specific user requirements and the domain-appropriate explanations[37, 38].
Counterfactual explanations are a particular type of explanation which relates what may have occurred if a model's input had been altered in a particular way. Users receive practical feedback that they can employ to modify their features and move

towards the desirable side of the decision boundary. Unlike other XAI techniques, they offer recommendations on how to get the desired result rather than directly addressing *why* the model made a particular prediction[39]. This approach is particularly valuable in clinical settings, as it can help clinicians and patients understand how to alter risk factors to potentially reverse adverse health probabilities. Additionally, counterfactual explanations can enhance trust, as they allow users to familiarise themselves with unknown processes by understanding the hypothetical input conditions under which the output changes [40].

2 Stage 1: AI/XAI methods evaluation in the workshops

The study involved a series of four structured workshops, organised around three distinct themes: regulatory, clinical, and data analytic considerations. The regulatory and clinical themed discussions focused on the necessary level of transparency from both regulatory and clinical perspectives, while considering which information and metrics were particularly valuable for regulators and clinicians. The data analytic considerations were informed by data science experiments generating decisions with explanations using various modeling approaches in response to regulatory requirements. In the initial phase of the experiment, the objective was to present performance metrics for a variety of representative white-box and black-box models, such as logistic regression, random forest, and artificial neural networks (ANN), to the expert group. Interpretations were provided for the white-box models, while global and local explanations were provided for a black-box model. Specifically, odds ratios, mean decrease GINI, and permutation importance as were presented as global explainers, LIME as a local explainer, and Explainable Matrix (ExMatrix) for identifying counterfactual cases. The findings of these techniques were provided to the expert group to to obtain their views on using these strategies in CDSSs.
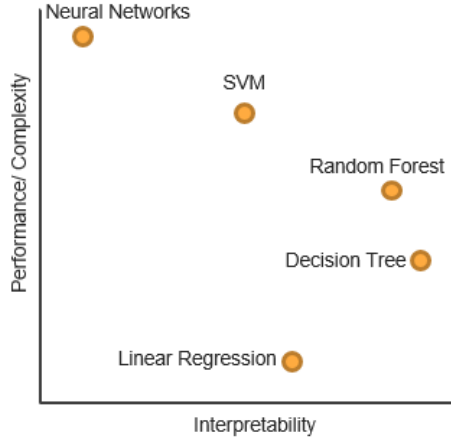
2.1 Data

The chosen task was to predict whether individuals were at a high or low risk of having a heart attack. A high-fidelity synthetic dataset based on anonymized CPRD primary care data was used for this purpose [41].This synthetic dataset focused on cardiovascular disease risk factors and included 22 variables, such as smoking behavior, age, and chronic conditions associated with cardiovascular disease, for 10,000 synthetic individuals randomly sampled from the dataset. The target variable was binary, confirming or denying the occurrence of a heart attack.

2.2 Methods

2.2.1 AI/ML Models

The expert group selected three ML/AI models to predict patients' risk of having a heart attack: logistic regression, random forest, and neural networks. This study is primarily concerned with assessing the performance of these models and investigating several XAI methods that can be applied to them. The selection of models aimed to cover varying degrees of interpretability and complexity—low, moderate, and high- as outlined in [27] and shown in Figure 1. We explored a number of parameterisations and train-test regimes to assess model performance. The results documented come from a 10-fold cross validation to reduce the risk of overfitting and improve robustness. The expert working group was provided with a mini-tutorial in the form of a presentation, designed to help them interpret the machine learning results. This tutorial introduced the machine learning models discussed in the workshop and aimed to demonstrate their complexity to the panel. It was presented to the group collectively, rather than individually. Although the experts came from varied backgrounds, they had sufficient familiarity with machine learning in healthcare to grasp the material. The tutorial also covered both local and global explanation methods, explaining how they are ranked and what key metrics, such as mean decrease Gini and odds ratios, represent.

9

Fig. 1: Tradeoff between Performance and Complexity of the model Vs Perceived Interpretabilty in Machine Learning

Logistic regression

Logistic regression is used for binary classification by modeling the probability of a given outcome based on one or more predictor variables. It estimates the relationship between the predictors and the log-odds of the outcome, enabling the prediction of categorical outcomes ans it considered the simplest model in the study [42].

Random forest

A random forest is an ensemble machine learning algorithm that enhances the robustness of decision trees. At its core, the random forest operates by constructing a multitude of decision trees during the training phase. These trees are grown using random subsets of both data and features, injecting an element of randomness into each tree's construction. During prediction, the random forest aggregates the outputs of all constituent trees, typically by taking a majority vote in classification tasks or averaging predictions in regression tasks. As a result, the ensemble approach improves the model's overall accuracy and robustness, while maintaining the interpretable nature of decision trees [42].

Neural Networks

ANN in its most general form consists of layered structures. These layers of interconnected nodes form a network that processes data in stages. The input layer receives the original data and then passes it through hidden layers before arriving at the final prediction in the output layer. Researchers have used ANN models as classifiers for risk prediction in the domain of medical diagnostics such as in [43, 44].

Simple ANNs have been shown to outperform recent specialised neural network architectures and even strong traditional ML methods. However, ANNs require careful pre- and post-processing to achieve good performance which can be challenging [45]. ANN classifiers are also known for their accurate results on imbalanced datasets where one class is more prevalent in the training data than the another [43].

ANNs with considerably more hidden layers (often known as Deep Neural Networks) are typically considered state of the art for many decision based problems in terms of performance, confront various challenges when applied to tabular data compared to white-box models. This includes a lack of localisation and lower performance due to the inner structure that cannot handle all feature types (numerical, ordinal, and categorical)[46]. ANNs often do not perform as well as some white-box models for tabular data[47]. An ANN is considered as a black box model, with little transparency or interpretability into how input data is used in the model's predictions. Thus, for our study, This model is considered the complexiest and we treat it as a black box model that required extra tools for explanation.

The performance metrics for these models, including sensitivity, specificity, precision, and AUC, the area under the receiver operating characteristic curve, were presented to the workshop attendees as in shown in Table 1. Sensitivity, which represents a test's ability to correctly identify individuals with a condition, and specificity, which indicates the ability to correctly identify those without the condition, are crucial metrics for evaluating diagnostic test performance. Precision, or positive predictive

11

value, reflects the proportion of true positive results among all positive results, indicating the reliability of a positive test outcome. AUC, is a comprehensive measure of a test's discriminative ability across all possible thresholds, providing a single value to assess overall test performance. These metrics are important for determining the effectiveness of diagnostic predictions in clinical settings [48].

Table 1: Performance Metrics for Machine Learning Models Developed to Predict Heart Attack Risk

Model	Sensitivity	Specificity	Precision	AUC
LR	0.78	0.85	0.46	0.90
RF	0.83	0.97	0.83	0.96
ANN	0.75	0.85	0.43	0.80

LR: Logistic Regression
RF: Random Forest

2.2.2 XAI Methods

- **Global Explanation**

For each selected ML/AI model we chose a proper method for the global explanation to understand the general structural characteristic. It can be either model-specific method that can only be applied on that specific model or it can be model agnostic method that can be build above any ML/AI model to provide information on its inner working. In global explanations it is common and desired by clinicians to use feature importance/ feature contribution approaches for assigning scores to input features in a predictive model that indicates the relative importance of each feature when making a prediction [49]. The relative scores can highlight which features may be most relevant to the diagnoses. This may be interpreted by a domain expert and can be used as the basis for gathering further data.

In the logistic regression model, feature importance was assessed using odd ratios, a fundamental measure of logistic regression interpretability. The odd ratio quantifies the change in odds of the outcome for a one-unit increase in a continuous predictor or for one category relative to a reference category in a categorical predictor, assuming other variables remain constant. This attribute makes the model's coefficients interpretable, as they directly indicate the influence of each predictor on the likelihood of the outcome[42].

The working group was presented with the odd ratios corresponding to the highest-ranked features in the logistic regression model utilised for predicting the risk of heart attack, as shown in Table 2. The global interpretation reveals that a history of angina heart attack emerges as a highly significant predictor in the model. Notably, the odd ratio indicates that patients with a history of angina heart attack exhibit around a 53-fold average increase in the odds of experiencing a heart attack compared to those without such a history. Additionally, features such as a history of atrial fibrillation, rheumatoid arthritis, and chronic kidney disease yielded odd ratios of 4, 2.2 and 2 respectively, highlighting their respective contributions to the predictive model. The confidence for these ranking are also appear in the table, showing for angina heart attack for example the true odd ratio lies between 39 and 75, indicating a statistically significant effect since the interval does not include 1. The relatively wide interval suggests some uncertainty around the estimate, but it confirms a strong positive relationship between the predictor and the outcome. and the same interpretation valid for all features.
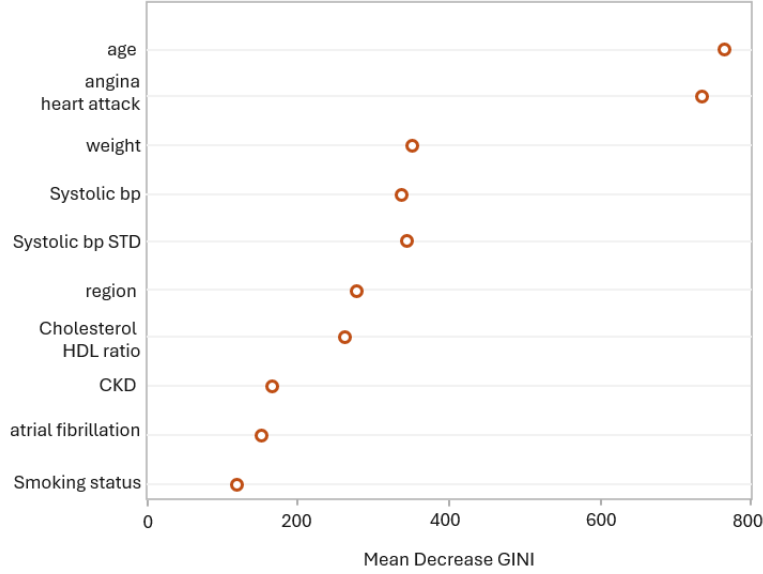
In random forest models, the Gini measure is employed to assess feature importance. Gini impurity quantifies the likelihood of incorrect classification within a dataset, where high impurity denotes a mix of classes and low impurity indicates homogeneity. During the construction of a random forest, data is iteratively split

Table 2: Feature Importance Ranking Using Odds Ratio in Logistic Regression Model for Heart Attack Risk Predictiont

Variables	Odds Ratio	minimum	maximum
Angina heart attack	53.4	39.01	75.09
Atrial fibrillation	3.95	3.19	4.92
Rheumatoid Arthritis	2.24	1.44	3.52
Kidney Disease	2.03	1.70	2.43
Region (North East)	1.98	1.51	2.68
Stroke	1.92	1.59	2.51
Hypertension treatment	1.9	0.95	3.58
Smoking status (Heavy Smoker)	1.8	1.19	2.65
Diabete (Type 2)	1.78	1.51	2.15
region (West Midlands)	1.59	1.21	1.95

into smaller subsets (via nodes) based on different features, aiming to reduce impurity with each split. The decrease in Gini impurity resulting from each split reflects the influence of the feature used. By averaging the reduction in impurity across all trees in the forest, an overall importance score for each feature is obtained. Features are then ranked according to their average reduction in impurity, with higher scores indicating greater importance [42]. The expert group was presented with Figure 2, illustrating the mean decrease GINI plot for the random forest risk prediction model. Within this model, age emerges as the most influential predictor, boasting a mean decrease in Gini of 780. This indicates that age is significantly contributing to impurity reduction. Similarly, angina heart attack demonstrates considerable importance in predicting the target variable, with a mean decrease Gini of 730. Weight, systolic blood pressure, and systolic blood pressure STD exhibit mean decreases in Gini around 300. While these values remain relatively high, they suggest that these features exert less impact on impurity reduction compared to age and angina heart attack.
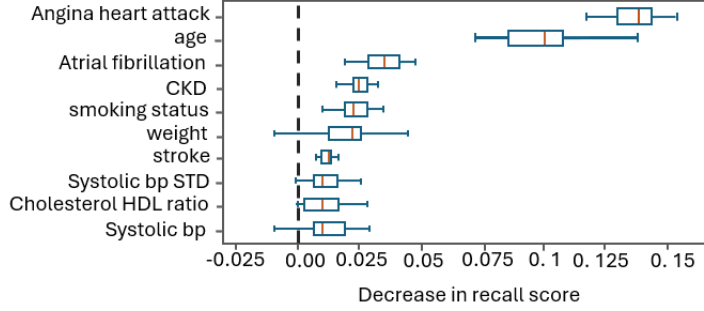
In this study, for the global explanation of the ANN model, a model-agnostic method known as permutation importance was presented in the workshops as a well known method for global explanation in healthcare [50]. Permutation feature importance assesses the impact of each feature on the model's performance by measuring the

Fig. 2: Feature Importance Ranking Using Mean Decrease GINI Metric in Random Forest for Predicting Heart Attack Risk

increase in prediction error when the feature values are permuted while keeping other variables constant [51]. Specifically, a feature is considered significant for the model if permuting its values noticeably increases the prediction error, indicating its importance in the model's predictive performance. Conversely, if the prediction error remains relatively unchanged after permutation, the feature is deemed less useful. Figure 3 presents the permutation importance plot for the ANN model, illustrating the decrease in recall (a metric related to sensitivity) following permutation of all features with a confidence bounds. Angina heart attack, with a permutation importance of 0.15, is the most influential feature, suggesting that shuffling its values resulted in the largest increase in prediction error. Age, with a permutation importance of 0.105, follows closely behind in importance. While not as influential as variable angina heart attack, it still significantly impacts the model's predictive performance. Atrial fibrillation, chronic kidney disease and smoking status with

permutation importance values of 0.035, 0.025, and 0.015 respectively, have lower importance scores.

Fig. 3: Permutation Importance Ranking of Features for the Neural Network Model in Heart Attack Risk Prediction, Assessing Reduction in Recall Error
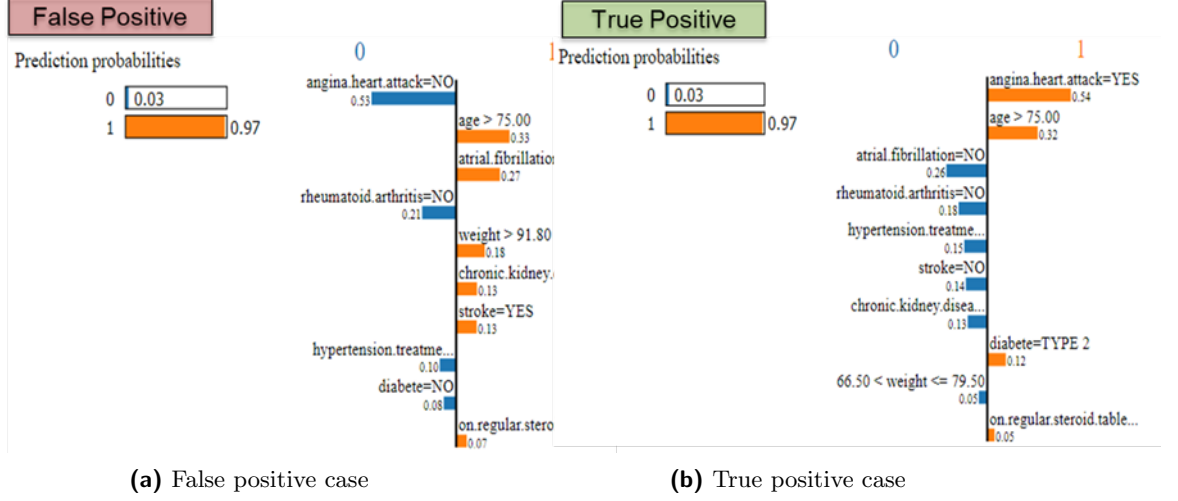
- **Local Explanations**

  LIME was chosen to obtain local explanations for the ANN model in this study. LIME can be considered as a model-agnostic post-hoc XAI method that provides explanations without opening the complex black box model. To explain a prediction for a specific instance, LIME generates a new dataset consisting of perturbed versions of the instance by slightly altering its feature values. The black-box model is then used to predict outcomes for these perturbed instances, creating a dataset of perturbations and their corresponding predictions. LIME assigns weights to these perturbations based on their proximity to the original instance, with closer perturbations receiving higher weights to emphasise local behavior. An interpretable model, such as linear regression, is then fitted to this weighted dataset, approximating the black-box model's behavior in the local region around the instance of interest. The coefficients of this interpretable model provide the local feature importance, indicating how each feature contributes to the prediction. The resulting explanation highlights the most influential features for the specific prediction [32]. However,

16

LIME is considered unstable due to the randomness in generating perturbed versions of the original instance which can result in different sets of perturbed data for different runs, leading to variability in the explanations. The choice of surrogate model and the specific data points used for fitting can also affect the resulting explanation, making it sensitive to the local sample. Additionally, LIME's weighting scheme, which assigns weights to perturbed instances based on their proximity to the original instance, can introduce instability.

In the workshops, two cases from the test set were used as case studies: a low-risk heart attack case misclassified by the model and a high-risk heart attack case correctly classified. To address stability and ensure consistent explanations, LIME was run 20 times for each example. Figures 4a and 4b show LIME local explanations for the two cases, respectively. Each figure provides the prediction confidence (prediction probability). In addition, it lists features on one axis, with bars indicating their importance; the direction of each bar shows whether the feature contributes positively or negatively to the prediction, while the length indicates the strength of this contribution. Feature values for each instance contextualise their importance. LIME provides local explanations specific to the examined instance, and feature importances can vary across predictions. Comparing the explanations for the two cases reveals the model's consistency in using the same features, with angina heart attack as the most important feature, followed by age, atrial fibrillation, and rheumatoid arthritis. In the misclassified case, the patient was older than 75 and had a history of atrial fibrillation, leading to a high-risk classification despite not being high-risk. In the correctly classified case, having an angina heart attack and being older than 75 led to correctly identify the patient as high-risk.

- **Counterfactuals**

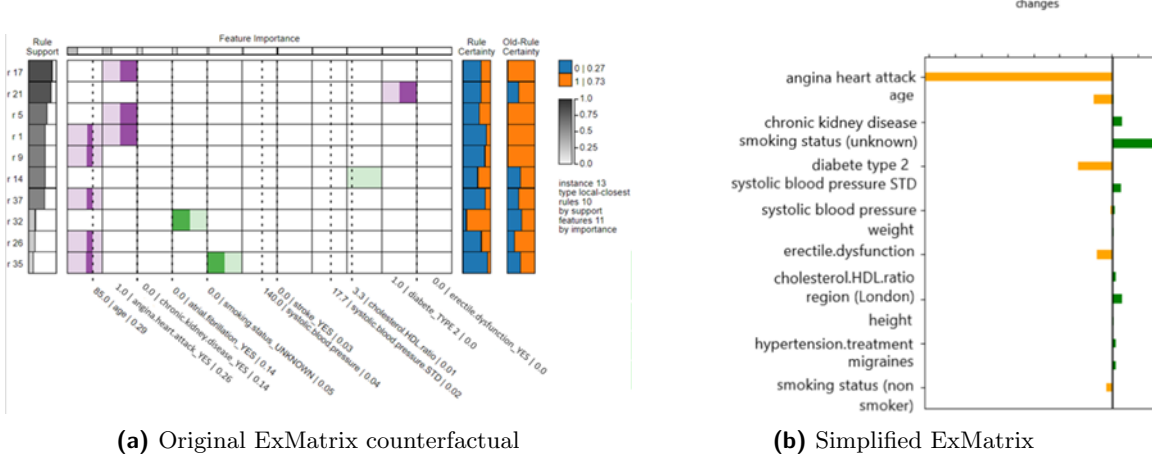**(a)** False positive case   **(b)** True positive case

Fig. 4: LIME local explanations for two selected cases predicted by the neural network model: (a) local explanation for a false positive output, (b) local explanation for a true positive output.

For counterfactual generation, we used the Explainable Matrix "ExMatrix" technique to extract counterfactual explanations from random forest. ExMatrix identifies decision paths within the random forest that lead to different predictions and searches for the nearest path resulting in an alternate decision, such as moving from "low risk" to "high risk." The method focuses on identifying the minimal feature changes required to switch decision paths, providing clear guidance on which features need to be modified and by how much [52].

Both the original version of ExMatrix, as introduced by [52], and a simplified version was presented to the expert group to assess their ability to interpret the complexity of the visualizations. This dual presentation allowed us to evaluate how the complexity of data presentation impacted interpretation. Simplification was achieved by aggregating the feature-wise changes required across all decision trees in the random forest, deriving a single value for each feature that represents the extent to which modifying that feature would alter the prediction.

Figure 5 shows the original and the simplified version of the ExMatrix counterfactual visualisation for a selected true positive case. In the modified version, the yellow

18

bars indicate negative alterations and the green bars positive ones needed to change a high-risk case to low-risk. For example, transitioning a high-risk patient with angina, type 2 diabetes, and erectile dysfunction to low-risk involves removing these conditions, particularly focusing on treating angina, type 2 diabetes and erectile dysfunction and theoretically reducing the patient's age.

(a) Original ExMatrix counterfactual

(b) Simplified ExMatrix

Fig. 5: Exmatrix counterfactual cases for a true positive case. Figure (a), Original matrix visualisation . Figure (b), simplified 1D plot of ExMatrix.

2.3 Results & Expert Working Group Insights

In this study, we presented the AI/ML models employed to predict heart attack risk to the expert group. Performance metrics for these algorithms were subsequently shared with them, as detailed in Table 1. All models demonstrated acceptable performance, but random forest outperformed the others. Sensitivity, essential for accurately identifying high-risk cases, was particularly stronger for random forest. However, regulators expressed concerns about the selection of predictive models for CDSS. While complex models like neural networks may initially seem attractive, regulators emphasise the importance of interpretability and ease of understanding to ensure patient safety. Simpler models, such as logistic regression or simple random forest, might be more

suitable if they offer comparable performance, as they are easier to understand and interpret. Comprehensive research is needed to evaluate the trade-offs between model complexity and interpretability in the intended clinical settings.

Clinicians on the other hand expressed their interest in the overall model performance and ensuring regulatory approval rather than delving into the technical aspect of the AI/ML models.

All the global explanation methods presented in the workshops utilised feature ranking to determine the most influential features on ML/AI models' predictions. Table 3 compared the top five features across the three chosen models. Commonly ranked features included angina, age, and atrial fibrillation, but there were clear differences between models. For regulators, this variation in influential features across different models highlighted the necessity of assessing multiple ML/AI models before selecting one for AI-CDSS. The assessment needs to consider not only performance metrics but also the features affecting model predictions. Regulators also highlighted the importance of evaluating how these features related to the specific use case and aligned with clinical knowledge. Which require comprehensive testing and active involvement of clinicians to identify the most suitable model. This also aligned with clinicians' preference for being informed about the underlying considerations and features that AI-CDSS relied on for their predictions, expressing that this transparency is essential for them to build trust in the CDSS system's outputs.

Table 3: The top five important features ranked from the global explanation methods

Rank	Logistic Regression	Random Forest	Neural Network
1	angina heart attack	age	angina heart attack
2	atrial fibrillation	angina heart attack	age
3	rheumatoid arthritis	weight	atrial fibrillation
4	chronic kidney disease	Systolic blood pressure	chronic kidney disease
5	region (North East)	Systolic blood pressure STD	smoking status

In the LIME local explanations presented in Figure 4, attributes were ranked by their influence on heart attack risk, along with the model's confidence score (prediction probability) for each class. The ANN model misclassified the first example (Figure 4a a false positive case) but correctly predicted the second (Figure 4b a true positive case). In both cases, the model was 97% confident in its predictions.

For the correctly classified case, this high confidence was seen by both clinicians and regulators as a way to enhance clinicians' confidence in their predictions. However, the discrepancy between model confidence and prediction error in the false positive case was highlighted as a significant issue. The explanation revealed that the false positive classification was due to the patient being older than 75, having a history of atrial fibrillation, kidney disease, and stroke, and a weight above 91.80 kg which are clinically valid explanations but still flagged an issue. While this specific false positive case may not pose an immediate issue, a high-confidence misclassification of a high-risk case as low risk would raise significant safety concerns for both clinicians and patients.

Another point raised by the clinicians was that the LIME visualizations were not easy for them to interpret, even though the data science team found them simple. Clinicians argued that while LIME outputs show how each feature contributes to the final prediction, in clinical settings, these visualizations would require significant time to interpret and evaluate how combinations of features diagnose a patient as high or low risk. This complexity could hinder their practical use in fast-paced clinical environments, where quick and accurate decision-making is crucial.

The original visualization for ExMAtrix counterfactual explanation, represented in Figure 5a, was initially considered informative as it presents the decision paths for all trees in the random forest. However, the expert group found it to be complex and overwhelming for use in clinical settings, especially with a significant number of trees in the model. In its simplified form (Figure 5b), the counterfactual cases were deemed easier for clinicians to interpret. Despite this improvement, the clinicians noted that,

similar to LIME, it would still require considerable time to utilize the output effectively in a clinical environment. There was a consensus among the experts that certain counterfactual features, such as age, gender, and some features of medical history, are difficult or impossible to change. Interpreting these features in a clinical setting would be overwhelming and would not necessarily guarantee a change in the model's output if those features were excluded from the required changes to alter the risk prediction.

3 Stage 2: Pilot Study

In this stage, we conducted a pilot study to evaluate the impact of AI/XAI methods and visualizations on clinicians' diagnostic processes. Clinicians, who were not part of the expert group, were presented with the same XAI outputs as outlined in section 2.2. This investigation sought to evaluate the influence of these outputs on their diagnostic processes and to investigate the key considerations and challenges entailed in clinicians' engagement with explanations in CDSS. The results of this pilot study were then discussed by the expert group. The aim was to investigate both the actual results of the pilot study and the expert group's reflections on these findings.

In accordance with the Health Research Authority (HRA) guidelines for clinical research within the UK's National Health Service (NHS), ethical approval was not required for this study. We used the HRA's decision tool to confirm this, and the tool's decision is provided in Appendix A.

3.1 Methods

This study involved several stages. Initially, the data science team selected fifty synthetic patient records from the same test dataset used in earlier phases of the research, ensuring they were representative of the AI/ML model's performance and covered a range of patient demographics. The clinical team then refined this selection to ten

patient records, focusing on cases that would provide diverse and clinically interesting scenarios. The cases chosen consisted of four true positives (correctly classified as high-risk), four true negatives (correctly classified as low-risk), one false negative (misclassified as low-risk), and one false positive (misclassified as high-risk). This selection allowed the expert group to examine a diverse range of decision scenarios.

These records were presented to eight practicing clinicians, ensuring that they were provided with exactly the same data that has been used by the AI/ML models and consists of patients medical history. These clinicians were shown each patient's full medical history individually, one patient at a time. For each patient case, clinicians were first asked to make their own diagnoses based solely on the patient's medical data and identify the top features that influenced their decision-making process.

Once they had completed their diagnosis for a given patient, they were shown the corresponding AI diagnosis, along with explanations from the XAI models and the confidence level of the CDSS predictions. This allowed the clinicians to assess whether and how the AI/XAI outputs might influence their diagnostic decisions and confidence levels. The process was repeated for all ten patients in sequence — clinicians reviewed the data, made their diagnosis, reviewed AI/XAI outputs, and then moved to the next patient.

Finally, the outcomes of this study, including both the clinicians' diagnoses and their responses to the AI/XAI outputs, were presented to an expert group for further discussion and analysis.

3.2 Pilot Study Results & Regulatory Reflection on the Results

To evaluate the pilot study results, we began by comparing the clinicians' diagnoses with the output from the ML models. Specifically, we assessed clinician confidence in their diagnoses prior to seeing the ML
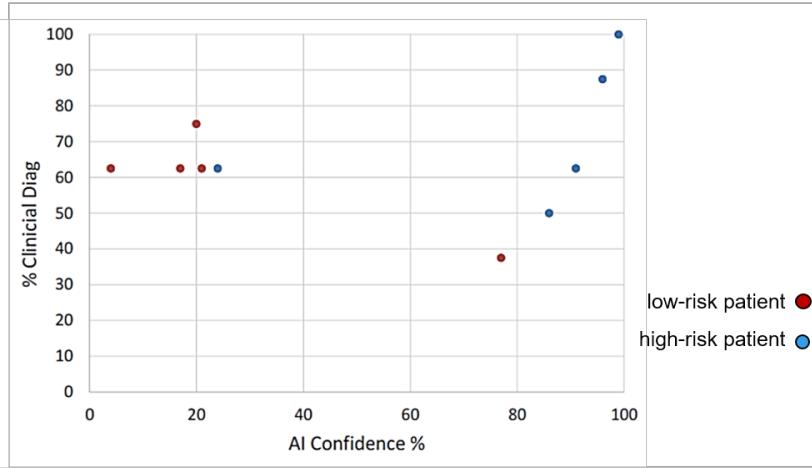
outputs and compared it to the confidence of the ANN model. Clinical confidence was measured by the percentage of clinicians diagnosing each case as high-risk. Confidence in high-risk diagnoses is complementary to that in low-risk diagnoses, meaning a lower confidence in high-risk implies a higher confidence in low-risk and vice versa. Confidence levels below 50% are considered low-risk diagnoses, while levels of 50% or higher are considered high-risk diagnoses.

Figure 6 represents this comparison, with the ten cases color-coded according to their actual risk status. Ideally, high-risk patients would cluster at the top right of the plot, and low-risk patients at the bottom left, reflecting high confidence in accurate diagnoses. The ANN model's confidence was either very high or very low, correlating with the actual risk status of the patients, while clinician confidence tended to cluster in the middle, showing a tendency for false positives as clinicians often rated more cases as high-risk.

Clinicians and the AI model agreed on 4 out of 5 high-risk cases, accurately diagnosing them as high-risk. Therefore, clinicians succeed in diagnosing all five high-risk cases being as high-risk. For the low-risk cases, the AI model correctly identified four, but only one was correctly diagnosed as low-risk by the majority of clinicians. This particular low-risk case was also the only negative case that the AI misdiagnosed.

For regulators, the disagreement between the actual risk status, the output of AI/ML models, and clinicians' diagnoses highlighted a potential area for consideration, suggesting that the disagreement may not

24

Fig. 6: Clinical Confidence Vs ANN model Confidence in Diagnosis Patients as High Risk to Have a Heart Attack

necessarily be problematic. Specifically, for actual low-risk patients who were diagnosed as high-risk by clinicians and/or models (false positive cases), this discrepancy could be attributed to incomplete data coverage of the patients' full medical histories. It is possible that these patients might have experienced a heart attack after the data collection period, indicating that the model's and clinicians' high-risk diagnoses could be justified despite the initial low-risk status.

In assessing influential features, there was generally strong agreement between clinicians and AI/ML models on high-ranking features such as age, previous angina, type 2 diabetes, smoking, and cholesterol. However, a notable divergence emerged for other features, with AI models incorporating a wider range of variables that did not consistently align with clinicians' considerations. Regulators suggested that this divergence

25

could make AI models a complementary tool, providing a new perspective during the decision-making process.

In assessing influential features, there was generally strong alignment between clinicians and the AI/ML models on key features such as age, history of angina, and type 2 diabetes. This assessment compared the top 3 features identified by clinicians with those highlighted by LIME explanations after running LIME 10 times. Features appearing at least once in the top 3 in LIME were included for comparison.

For the high-risk cases that were accurately diagnosed by both the ML model and the clinicians:

Patient 1: The ML model ranked diabetes, angina, and weight as the top features. Among clinicians, 5 out of 8 included diabetes in their top 3, 6 ranked angina, and 2 ranked weight.

Patient 2: The ML model identified angina, atrial fibrillation (AF), and weight as the most important factors. Here, 7 of 8 clinicians included angina in their top 3, 2 included AF, and 4 ranked weight.

Patient 3: The ML model ranked diabetes, angina, and AF as the key features. In this case, 4 of 8 clinicians included diabetes and angina in their top 3, 3 included AF, and 4 ranked weight.

Patient 4: The ML model highlighted angina, age, and weight, and 4 clinicians included these features as the most influential features in their diagnosis.

Other variables mentioned by fewer clinicians in these high-risk cases included smoking status and Chronic Kidney Disease (CKD), which the

ML model typically ranked lower, in the 4th and 5th positions. Additionally, the ML model sometimes included features like stroke or systolic blood pressure that clinicians did not prioritise.

For the patients whom clinicians diagnosed as high risk, but the ML model correctly classified as low risk:

Patient 6: Five clinicians ranked weight as a top 3 feature, consistent with the ML model. 5 clinicians cited diabetes, and 4 considered the patient's status as an ex-smoker to be key. However, the ML model aligned with only 2 clinicians, focusing on age and systolic blood pressure as the main factors for classifying this patient as low risk.

Patients 7 and 9: In both cases, 3 and 6 clinicians, respectively, agreed with the ML model that diabetes was a top feature. However, 6 clinicians considered slightly elevated cholesterol and erectile dysfunction to be significant risk factors, leading them to classify the patients as high risk. whereas the ML model did not prioritize these factors, instead focusing on the absence of a history of angina as key to assessing low risk.

Patient 8: The ML model correctly predicted the patient as low risk due to the lack of a history of heart disease (AF, angina) or diabetes, despite the patient's advanced age (84). However, 5 of 8 clinicians saw the history of stroke as a significant risk factor, and 4 considered slightly elevated hypertension important in their diagnosis.

Cases Misclassified by the ML Model:

For the patient incorrectly predicted as high risk by the ML model but correctly identified by clinicians, the model ranked age (over 75) and

CKD as important features. Three clinicians who also misdiagnosed the patient as high risk cited these same factors. The remaining clinicians correctly diagnosed the patient, citing a clear medical history and stable metrics (hypertension, blood pressure, and cholesterol) as indicators of low risk.

In the case of a patient falsely predicted as low risk by the ML model but correctly diagnosed by most clinicians, 4 clinicians cited erectile dysfunction, diabetes, and slight overweight as risk factors. While the ML model agreed on the importance of diabetes, it downplayed these factors, instead focusing on the absence of angina, AF, or CKD, leading to the low-risk classification.

From a regulatory perspective, the observed alignment between clinicians and the ML model explanations regarding major risk factors, such as diabetes, angina, and weight, indicates that the model effectively identifies critical clinical indicators associated with high-risk cases. This consistency suggests that the model has the potential to meet safety requirements for identifying high-risk patients, provided that these factors are rigorously validated against clinical standards and that the model's performance aligns with its intended purpose. Furthermore, the model's emphasis on the absence of certain conditions (e.g., angina or atrial fibrillation) as protective factors has resulted in several accurate low-risk predictions where human clinicians may have misjudged the risk. While this could imply enhanced performance in specific scenarios, it

underscores the necessity for careful evaluation of how the model balances the absence of conditions with existing risk factors. Clinicians also often considered traditional markers, such as cholesterol levels and erectile dysfunction, when assessing risk, even though these factors are not consistently prioritized by the ML model. This disparity indicates that clinicians may give greater weight to patient history, while the model is primarily driven by data patterns. In borderline cases (e.g., Patients 6, 7, and 9), notable discrepancies between the model's reasoning and the clinicians' decisions became apparent, highlighting potential gaps in the model's training. Nevertheless, these differences may also underscore the value of models in providing clinicians with new perspectives, albeit requiring rigorous testing to ensure their reliability and effectiveness.

After revealing AI diagnoses, explanations, and decision confidence, all clinicians agreed that these explanations boosted their confidence in their diagnoses, particularly when the explanations and AI predictions aligned with their clinical knowledge. In cases where discrepancies existed between AI and clinician diagnoses, the introduction of XAI outputs; including influential features and AI confidence, prompted a notable shift in clinicians' diagnostic decisions. Specifically, in five out of six instances, clinicians adjusted their diagnoses to align with AI prediction, resulting in an overall improvement in diagnostic accuracy, especially concerning low-risk cases. However, a notable exception occurred when the AI inaccurately diagnosed a high-risk case, prompting clinicians to adopt a wrong

diagnosis in alignment with the AI output. Regulators viewed this as an indicator of automation bias or trust calibration issues.

4 Discussion

The goal of this study was to evaluate and understand multidisciplinary perspectives on the use of AI/XAI technologies in clinical settings and ensure their safe introduction. Initially, we presented ML/AI models with different levels of complexity to regulators and clinicians to observe their reactions. We also introduced global and local explanations for these models to assess how utilising these methods could help clinicians and identify potential challenges for both regulatory and clinical applications. While there are a few studies that examine clinicians' perspectives on using AI [4, 5, 49] and others that consider regulating AI systems in healthcare[10, 19], our approach is unique in facilitating comprehensive discussions among various stakeholders through workshops. This approach allowed for a better understanding of the needs and concerns of each group.

In the second part of the work, we conducted a pilot study to evaluate human-clinician interaction with AI/XAI systems. Similar to other studies, we assessed the performance of clinicians before and after introducing the XAI outputs. Consistent with previous findings, we observed an overall increase in diagnostic accuracy after incorporating XAI explanations [17, 50]. However, incorporating AI in the decision-making process raised

potential issues previously noted in similar experiments, such as trust calibration problems and over-reliance on AI system outputs [5, 11, 17, 49]. Unlike prior studies, our research included regulators to assess these issues from a regulatory standpoint.

In this section, we discuss the lessons learned from this study and propose regulatory recommendations to consider for safely utilising AI/XAI in clinical settings. These recommendations aim to address the identified challenges and ensure that AI technologies improve clinical decision-making without compromising safety.

Lesson 1: Assessing scientific and analytical validity is essential for AI-CDSS adopting

Before adopting AI/ML based CDSS, it is essential from a regulatory perspective to ensure thorough validation. This is particularly important when using black-box AI systems that lack straightforward scientific and analytical validation methods, especially in high-risk clinical scenarios. Key performance metrics such as sensitivity, precision, and specificity must be rigorously assessed. Additionally, it is important to guide analytics teams to build models that are as simple as possible while still achieving the required tasks [21, 22]. Engaging clinicians for expert review, conducting clinical trials, and implementing XAI methods are crucial to ensure clinical relevance and transparency. Ensuring regulatory compliance, continuously updating model training, and establishing ongoing monitoring and feedback mechanisms are essential steps to

maintain safety in clinical settings.

Lesson 2: The divergent importance of global explanations for regulators and clinicians

When approving or adopting AI models, global explanations hold differing significance for regulators and clinicians. For regulators, global explanations of AI/ML models were crucial as they provided insights into how the model functions and what considerations are being made. This understanding is essential for ensuring the model's safety and transparency. Regulators also found value in using feature importance rankings to identify key features that influence the models' predictions. Notably, the variation in feature rankings among different models can be insightful, suggesting that models may learn differently and could be relevant for different scenarios and tasks. This variation should be considered in future research, despite some authors [12, 38] suggesting that inconsistent explanations across models might be an issue and invalidate their use.

Clinicians on the other hand, were primarily interested in understanding the clinical knowledge that has been incorporated into predicting patient outcomes. However, they were less concerned with the in-depth technical details of the model's inner workings. Clinicians expressed that they prefer a straightforward explanation of these considerations, ideally provided in a concise briefing before they begin using the AI/ML-based CDSS.

Lesson 3: Local explanations are essential for clinicians but demand careful consideration in both method selection and usage:

Clinicians in the workshop and pilot study found local explanations beneficial for enhancing their trust in AI model decisions, particularly when the explanations and key features aligned with their clinical knowledge.

Regulators, did not show a strong interest in understanding how a decision for a specific case is made. Their primary concern was ensuring that human clinicians can interact with these explanations in a safe and effective manner. Consequently, a brief overview on how to interpret XAI methods' outputs was provided to clinicians before the experiment in the pilot study. This preparation was reflected in their satisfaction and their ability to understand the outputs. In our study we used LIME as local explainer because its output is similar to how humans clinicians visualise explanations [3]. However, in a clinical environment, this might not be the best method. Clinicians in the workshops expressed that the visualizations can be confusing and time-consuming in a fast-paced clinical setting, especially when practitioners need to assess from the visualisation how a combination of features contributed to a specific diagnosis.

Lesson 4: Model confidence is a key for trust and safety but can cause issues in some scenarios

While confidence is considered as a reliable indicator of the degree that clinicians can trust the output of a machine learning model, it is vital

to devote careful attention to the basis of this measure as well as any inherent limitations (missingness and bias) in the training data. This includes factors like the severity of specific medical conditions, extra symptoms and conditions that the treated patient may be suffering and that were not included in the training data. As a result, practitioners must be aware of the data used for the development of AI systems to be able to better understand the decisions being recommended to them and accept or reject these based on this understanding. Special focus should be placed on the correlation between AI model confidence in its predictions and the accuracy of these predictions as was seen in the false positive case in the lime explanation output. There might be trust implications if the model confidence does not match the likelihood of the decision being correct. In such circumstances, an investigation into the potential causes of being *incorrect but confident* should be conducted.

Lesson 5: Transparency requirements are different for different stakeholders:

Regulators in the workshops highlighted the importance of focusing not just on assessing the overall performance of the model, but also, its fairness and any potential biases. This evaluation involves identifying situations where the model demonstrates sub-optimal performance and determining which subgroups are disproportionately affected by these

shortcomings. When applicable, documentation should explicitly highlight the affected subgroups and outline situations where the model fails to operate as intended.

Meanwhile, clinicians prioritised understanding the data and clinical considerations during the AI/ML based CDSS development over the actual working logic of the models deployed. This is to ensure that the model assumptions and parameters align with clinical knowledge. However, all workshop attendees including clinicians assured the importance of educating healthcare providers on when to utilise these models effectively and when it may be prudent to avoid their use due to concerns about fairness and safety.

Lesson 7: Counterfactual explanations are highly useful if they were introduced correctly

Counterfactual analysis is regarded as one of the most important tools in clinical healthcare by experts [39]. However, in a clinical setting, it is critical to communicate these counterfactual scenarios in a manner that is easily comprehensible for end users, highlighting the main features to change and their related values clearly. Furthermore, considering the nature of the clinical field, experts should be provided with a variety of feasible options. Options, for example, should not recommend modifications to fixed patient characteristics such as age, gender, or medical history as it was introduced in the ExMatrix output.

4.1 Recommendation for safely introducing AI/XAI tools to clinical settings

Before adopting AI/ML based CDSS, multiple factors must be considered to ensure their safety and efficacy in clinical workflows. It is generally preferable to use simple yet efficient models as has been suggested by [51], balancing complexity with interpretability, as simple models are often easier to understand and trust, which is crucial in clinical settings. During model evaluation, it is important to recognize the limitations of the test data, particularly the time span and scope of the patients' medical histories it covers. Assessing multiple models is essential in the selection process, and global explanations should be examined to identify the most influential features for each model. Clinical knowledge should be engaged to evaluate the relevance of the top-ranking features for each model to the specific use case, ensuring the chosen model aligns well with clinical needs and expectations.

Clinicians' preferences for explanation methods and types vary significantly and often differ from developers' preferences [35]. Therefore, involving a diverse group of clinicians from different backgrounds and experience levels in the development process is crucial. For example, while developers might find LIME to be straightforward, some clinicians may find these explanations confusing. Incorporating insights from psychologists and cognitive specialists can also enhance the design process by

ensuring that explanations are tailored to the cognitive needs of end-users [17]. Local and counterfactual XAI methods, are encouraged to be used as valuable educational tools, particularly for junior doctors, helping bridge the gap between theoretical knowledge and practical application, especially if their performance has been validated against actual clinical outcomes. Involving experienced clinicians in validating these tools can further enhance their educational value.

AI and XAI tools should be regarded as support systems rather than standalone solutions [19]. They are most effective when complementing human practitioners, highlighting information that might otherwise go unnoticed. To facilitate this complementary relationship, professional training and education for all stakeholders are essential. This training ensures that medical professionals are prepared for AI integration, understand how to calibrate their trust in AI outputs, and know when to rely on or discard CDSS recommendations.

In our workshop, the inclusion of perspectives from psychologists and cognitive specialists could provide deeper insights into how clinicians prefer to be provided with AI/XAI outputs. Additionally, employing other well-known methods for local explanations, such as SHAP [36], and counterfactual methods like DiCE, which allow restrictions on counterfactual cases, could be beneficial. This is particularly relevant in clinical settings where factors such as demographics and medical history cannot be changed. Future work will include a pilot study with a broader scope to assess how clinicians' experience levels and background knowledge of

AI affect their interaction with AI-CDSS. Observing practitioners' interactions with these systems over a longer period will also be beneficial to address potential challenges and opportunities in AI/XAI medical devices.

5 Conclusion

This study evaluated the perspectives of data scientists, clinicians, and regulators regarding the safe integration of ML/AI-based CDSS into clinical settings. Introducing XAI methods as a crucial tool for ensuring the safe deployment of these technologies. We found that performance metrics, in conjunction with global explanations and clinical knowledge, serve as valuable guides for selecting suitable AI models for specific tasks. Local explanations are essential for improving clinician trust. However, Regulators underscored the significance of viewing AI/XAI CDSS as support systems only and emphasized the need for proper trust calibration to mitigate potential risks such as over-reliance on AI outputs.

Abbreviations. AI: Artifcial Intelligence; ML: Machine Learning; ANN: Artificial Neural Network; CDSS: Clinical Decision Support System; XAI: eXplainable AI; MHRA: Medicines and Healthcare products Regulatory Agency; AI-CDSS: AI based Clinical Decision Support System; LIME: Local Interpretable Model-agnostic Explanations; ExMatrix: Explainable Matrix; SHAP: Shapley Additive explanations; DiCE: Diverse Counterfactual Explanations

Declarations

Johan Ordish (Roche), Maria Wilder (BEIS), Melissa Tucker (Health Education England), Michael Nix (Leeds NHS Trust), Paul Campbell (MHRA), Puja Myles (MHRA), Richard Branson (MHRA), Russell Pearson (MHRA), Salah Hammouche (Health Education England), Susan Hodgson (MHRA).

References

[1] Preim, B.: HCI in Medical Visualization. In: Hagen, H. (ed.) Scientific Visualization: Interactions, Features, Metaphors vol. 2, pp. 292–310 (2011). https://doi.org/10.4230/DFU.Vol2.SciViz.2011.292

[2] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., Consortium, P.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making **20**, 1–9 (2020)

[3] Alam, L., Mueller, S.: Examining the effect of explanation on satisfaction and trust in ai diagnostic systems. BMC medical informatics and decision making **21**(1), 178 (2021)

[4] Teach, R.L., Shortliffe, E.H.: An analysis of physician attitudes regarding computer-based clinical consultation systems. Computers and Biomedical Research **14**(6), 542–558 (1981)

[5] Rosenbacke, R.: Errors in physician-ai collaboration: Insights from a mixed-methods study of explainable ai and trust in clinical decision-making. Available at SSRN 4773350 (2024)

[6] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)

[7] Yang, G., Ye, Q., Xia, J.: Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion **77**, 29–52 (2022)

[8] Hasan, Z., Serajuddin, M., Alvi, S.A.M., Khan, A., Ayub, R.: Explainable deep learning models for healthcare decision support

[9] Pagallo, U., O'Sullivan, S., Nevejans, N., Holzinger, A., Friebe, M., Jeanquartier, F., Jean-Quartier, C., Miernik, A.: The underuse of ai in the health sector: Opportunity costs, success stories, risks and recommendations. Health and Technology **14**(1), 1–14 (2024)

[10] Li, P., Williams, R., Gilbert, S., Anderson, S.: Regulating ai/ml-enabled medical devices in the uk. In: Proceedings of the First International Symposium on Trustworthy Autonomous Systems. TAS '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3597512.3599704 . https://doi.org/10.1145/3597512.3599704

[11] Rosenbacke, R.: Heuristics and errors in xai-augmented clinical decision-making: Moving beyond algorithmic appreciation and aversion (2024)

[12] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602 (2022)

[13] Naiseh, M., Al-Thani, D., Jiang, N., Ali, R.: Explainable recommendation: when design meets trust calibration. World Wide Web **24**(5), 1857–1884 (2021)

[14] Okamura, K., Yamada, S.: Empirical evaluations of framework for adaptive trust calibration in human-ai cooperation. IEEE Access **8**, 220335–220351 (2020)

[15] Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., Kaplan, L.: Rapid trust calibration through interpretable and uncertainty-aware ai. Patterns **1**(4) (2020)

[16] Okamura, K., Yamada, S.: Adaptive trust calibration for human-ai collaboration. Plos one **15**(2), 0229132 (2020)

[17] Naiseh, M., Al-Thani, D., Jiang, N., Ali, R.: How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human-Computer Studies **169**, 102941 (2023)

[18] Schrammel, J., Fröhlich, P., Mirnig, A.G., Dinica, O., Lindley, A., Woitsch, R., Falconi, D., Baldauf, M.: Investigating communication techniques to support trust calibration for automated systems. In: Workshop" Automation Experience Across Domains" in Conjunction with CHI2020 (2020)

[19] Oppermann, I.: Regulating ai for health. BMJ Health & Care Informatics **30**(1) (2023)

[20] Panigutti, C., Beretta, A., Giannotti, F., Pedreschi, D.: Understanding the impact of explanations on advice-taking: A user study for ai-based clinical decision support systems. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3491102.3502104

[21] Jia, Y., McDermid, J., Lawton, T., Habli, I.: The role of explainability in assuring safety of machine learning in healthcare. IEEE Transactions on Emerging Topics in Computing **10**(4), 1746–1760 (2022)

[22] Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386 (2016)

[23] Herm, L.-V., Heinrich, K., Wanner, J., Janiesch, C.: Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. International Journal of Information Management **69**, 102538 (2023)

[24] Vellido, A.: The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications **32**(24), 18069–18083 (2020)

[25] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., Cilar, L.: Interpretability of machine learning-based prediction models in healthcare. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(5), 1379 (2020)

[26] Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., Celi, L.A.: The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health **2**(9), 489–492 (2020)

[27] Herm, L.-V., Heinrich, K., Wanner, J., Janiesch, C.: Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. International Journal of Information Management **69**, 102538 (2023)

[28] Singla, K., Biswas, S.: Machine learning explanability method for the multi-label classification model. In: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), pp. 337–340 (2021). IEEE

[29] Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)

[30] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences.

43

Artificial intelligence **267**, 1–38 (2019)

[31] Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., Mooney, C.: Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Applied Sciences **11**(11), 5088 (2021)

[32] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)

[33] Alam, L.: Investigating the impact of explanation on repairing trust in ai diagnostic systems for re-diagnosis. PhD thesis, Michigan Technological University (2020)

[34] Manresa-Yee, C., Roig-Maimó, M.F., Ramis, S., Mas-Sansó, R.: Advances in xai: Explanation interfaces in healthcare. In: Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects, pp. 357–369. Springer, ??? (2021)

[35] Bienefeld, N., Boss, J.M., Lüthy, R., Brodbeck, D., Azzati, J., Blaser, M., Willms, J., Keller, E.: Solving the explainable ai conundrum by bridging clinicians' needs and developers' goals. npj Digital Medicine **6**(1), 94 (2023)

[36] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[37] Panigutti, C., Perotti, A., Pedreschi, D.: Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 629–639 (2020)

[38] Brankovic, A., Cook, D., Rahman, J., Huang, W., Khanna, S.: Evaluation of popular xai applied to clinical prediction models: Can they be trusted? arXiv preprint arXiv:2306.11985 (2023)

[39] Verma, S., Boonsanong, V., Hoang, M., Hines, K.E., Dickerson, J.P., Shah, C.: Counterfactual explanations and algorithmic recourses for machine learning: A review. arXiv preprint arXiv:2010.10596 (2020)

[40] Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., Holzinger, A.: On generating trustworthy counterfactual explanations. Information Sciences **655**, 119898 (2024)

[41] Draghi, B., Wang, Z., Myles, P., Tucker, A.: Bayesboost: identifying and handling bias using synthetic data generators. In: Third International Workshop on Learning with Imbalanced Domains: Theory and Applications, pp. 49–62 (2021). PMLR

[42] Molnar, C.: A guide for making black box models explainable. URL: https://christophm. github. io/interpretable-ml-book **2**(3), 10 (2018)

[43] Yildirim, P.: Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), vol. 2, pp. 193–198 (2017). IEEE

[44] Yan, H., Jiang, Y., Zheng, J., Peng, C., Li, Q.: A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Systems with Applications **30**(2), 272–281 (2006)

[45] Kadra, A., Lindauer, M., Hutter, F., Grabocka, J.: Well-tuned simple nets excel on

tabular datasets. Advances in neural information processing systems **34**, 23928–23941 (2021)

[46] Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information Fusion **81**, 84–90 (2022)

[47] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)

[48] Mavrogiorgou, A., Kiourtis, A., Kleftakis, S., Mavrogiorgos, K., Zafeiropoulos, N., Kyriazis, D.: A catalogue of machine learning algorithms for healthcare risk predictions. Sensors **22**(22), 8615 (2022)

[49] 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. Applied Sciences **12**(20), 10323 (2022)

[50] Wu, H., Ruan, W., Wang, J., Zheng, D., Liu, B., Geng, Y., Chai, X., Chen, J., Li, K., Li, S., et al.: Interpretable machine learning for covid-19: An empirical study on severity prediction task. IEEE Transactions on Artificial Intelligence (2021)

[51] Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**(177), 1–81 (2019)

[52] Neto, M.P., Paulovich, F.V.: Explainable matrix-visualization for global and local interpretability of random forest classification ensembles. IEEE Transactions on Visualization and Computer Graphics **27**(2), 1427–1437 (2020)

Appendix A: Ethical Approval Requirements for the Pilot Study)

Is my study research?

ℹ️ **To print your result with title and IRAS Project ID please enter your details below:**

Title of your research:

> Integrating Explainable AI in Medical Devices: Technical, Clinical and Regulatory Insights and Recommendations

IRAS Project ID (if available): [   ]

You selected:

- **'No'** - Are the participants in your study randomised to different groups?
- **'No'** - Does your study protocol demand changing treatment/ patient care from accepted standards for any of the patients involved?
- **'No'** - Are your findings going to be generalisable?

> **Your study would NOT be considered Research by the NHS.**
>
> You may still need other approvals.
>
> Researchers requiring further advice (e.g. those not confident with the outcome of this tool) should contact their R&D office or sponsor in the first instance, or the **HRA** to discuss your study. If contacting the HRA for advice, do this by sending an outline of the project (maximum one page), summarising its purpose, methodology, type of participant and planned location as well as a copy of this results page and a summary of the aspects of the decision(s) that you need further advice on to the HRA Queries Line at **Queries@hra.nhs.uk**.

For more information please visit the **Defining Research** table.

Follow this link to start again.

[Print This Page]

NOTE: If using Internet Explorer please use browser print function.