Explainable recommendation: When design meets trust calibration – Research protocol

Calibrated trust has become an important design goal when designing Human-AI collaborative decision-making tools. It refers to a successful understandability, reliability and predictability to the AI-based tool behaviour and recommendations. eXplainable AI (XAI) is an emerging field where explanations accompany AI-based recommendations to help the human-decision maker understand, rely on, and predict AI behaviour. Such an approach is supposed to improve humans' trust calibration while working collaboratively with an AI. However, evidence from the literature suggests that explanations have not contributed to improved trust calibration and even introduced other errors. Designers of such explainable systems often assumed that humans would engage cognitively with AI-based explanations and use them in their Human-AI collaborative decision-making task. In this paper, we devise XAI design techniques and principles for XAI interfaces to enhance the role of explanations in calibrating users' trust. We focus on model-agnostic explanations in high stake applications. We used screening prescription as a Human-AI collaborative decision-making task where the human medical practitioner uses the AI to check whether the prescription can be approved to a given patient. Such a task reflects an everyday Human-AI collaborative decision-making task where trust calibration errors are possible. We follow a multi-stage qualitative research method, including think-aloud protocol and co-design sessions with medical practitioners. Our results shed light on the nuances of the lived experiences of users of XAI and how the design can help their trust calibration.

First, we conducted a systematic literature review to identify what can be explained to end-users given a black-box AI model. Second, we conducted a think-aloud session to observe and understand how human decision-makers interact with AI-based explanations during a Human-AI collaborative decision-making task, i.e., what kind of errors could happen in real-time interaction. Finally, we conducted a co-design study with end-users to identify techniques and principles to guide the XAI interface to help trust calibration and mitigate errors. Figure 1 summarises the research method. The following sections describe each phase and its used material.

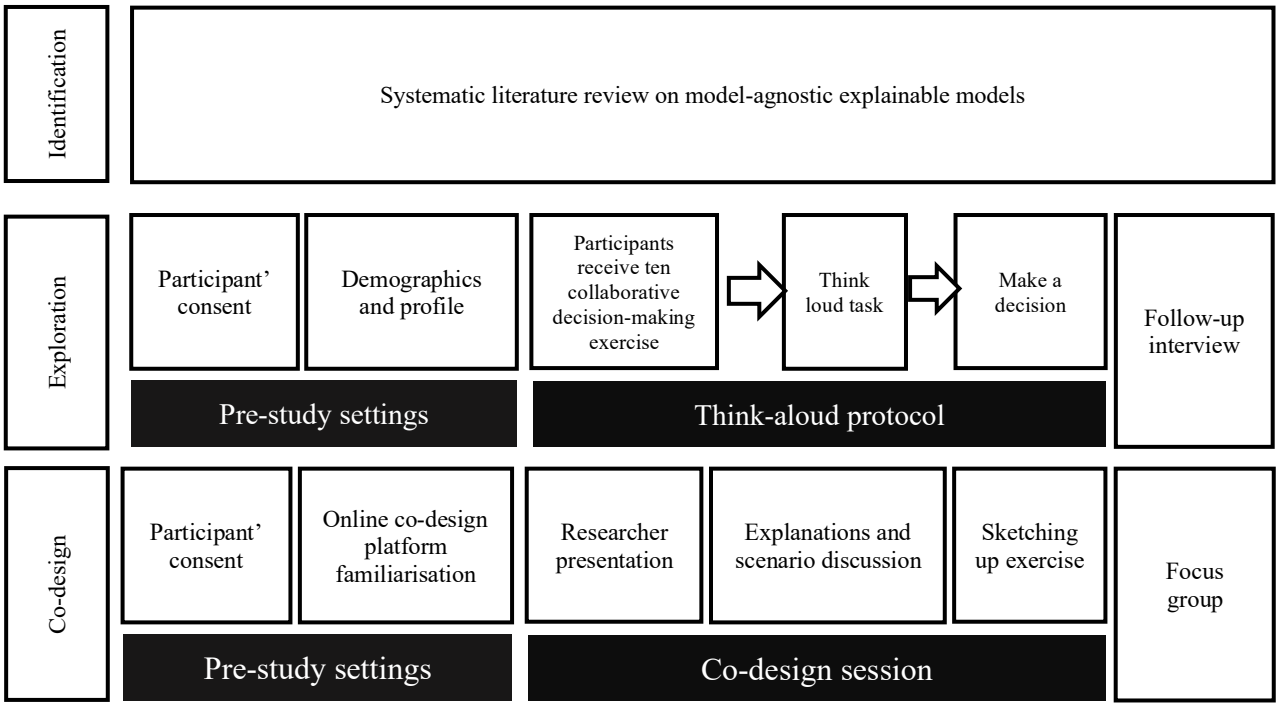

Figure 1 multi-stage qualitative study workflow

1. Systematic literature review for model-agnostic explanation.

The research aims to identify different explanation classes used in the literature of interpretable machine learning.

1.1. Research methodology

Given the diversity of the terminologies and the multidisciplinary nature of the explainability, bottom-up approach with the aid of content analysis approach (Elo and Kyngäs, 2008) was used to create an initial taxonomy. Thematic analysis (Hsieh 2005) has been adopted to infer some conclusions. Expert checking method was used to provide an in-depth evaluation of the emerged themes and concepts as well as trustworthiness. To increase the credibility of the methodology, the coder did not start analysis the data unless the text identification step was completed. This meant to eliminate any biased coding, such as refusing non-supportive inference (Hsieh 2005). This study conducts a systematic review in the field of explainable machine learning models to analyse and classify the literature and identify the different explanation capabilities.

1.1.1. Search strategies

To find relevant studies that are related to this systematic review, the research selected databases proved in Table 1, where the research ran different search queries. Other research databases that gather research paper automatically from different sources, such as GoogleScholar, or provide nor-reviewed papers such as arXiv, were excluded from the study search scope. The research adopted databases that include peer-reviewed papers published in computer science. This strategy meant to provide some evidence regarding the quality of the relevant papers.

Table 1 Selected research databases

Source	URL
ACM Digital Library	http://portal.acm.org
IEEE Xplore Digital Library	http://ieeexplore.ieee.org
ScienceDirect	http://www.sciencedirect.com
Springer Link	http://link.springer.com

1.1.2. Selection Criteria

Relevant papers retrieved from the selected databases were filtered using a set of criteria. The research considers three inclusion criteria (IC) and four exclusion criteria (EC) to select relevant papers relevant to the scope of this study. The inclusion criteria and the exclusion criteria are described and summarised in the following points:

1- Novelty (IC1): The study proposes a novel technique for explaining machine learning model.
2- Foundation (IC2): The study presents an investigation towards the foundations of the explanations in machine learning systems.
3- Language (EC1): This paper is not written in English.
4- Duplicated (EC2): The content of the paper was published in another completed form.
5- Full content (EC3): The research excluded the papers with no access to the full content.
6- Domain-related (EC4): The paper must be centred around explainable models and techniques in machine learning. For example, the search results introduced papers addressing the explanations from psychology, social science and education without direct relation to machine learning; these papers were excluded.

Concerning EC3, the research proceeded as follows. To obtain the papers in which they were published, the search strategy started to access them through the Bournemouth University network. If the full text was not available, the search strategy searcher for the paper on the web using the author websites, Google search and other repositories of scientific papers such as Google Scholar and ReasearchGate.

1.1.3. Search String Construction

The search string in this study is based on two main concepts of interest and their synonyms, which both have to appear in the protentional selected papers. The first concept is explanations, which is the main term of this study. However, this stage identified that the research community had used different terminologies to describe the explanations, and those were added as synonyms of the concept explanation. The second concept is machine learning, which is an umbrella term that covers different domains. As synonyms, the research considers sub-classes of machine learning, including, in particular deep learning, neural networks, reinforcement learning, supervised learning and unsupervised learning. The final search string is shown in Table 2.

Table 2 Final search string

(explanation OR justification OR interpretation OR intelligibility OR explainable OR interpretable OR intelligible) AND (machine learning OR reinforcement learning OR supervised learning OR unsupervised learning OR deep learning OR neural networks)

The search string was not capable of being applied in all the target databases due to the syntax restrictions. In these cases, the search string was customised according to the new syntax. The search within the abstract was also available in all the databases except the Springer Link database due to API limitations. The research thus searched for the terms in the keywords of the papers in this case.

1.1.4. Selection of relevant studies

After conducting the search through four selected databases on January 12, 2019, the search resulted in 1285 papers (excluding duplicates). The detailed descriptive statistics for each database are shown in Table 3.

Table 3 descriptive statistics for each data base search results

Database	Number of studies
ACM Digital Library	378
IEEE Xplore Digital Library	356
ScienceDirect	344
Springer Link	207
Duplicates	25
Total (including duplicates)	1285
Total (excluding duplicates)	1310

Then, the research conducted two main filtering procedure. In the first filtering step, the research analysed the paper title and abstract for each 1285 paper. If the title and the abstract presented an explicit relation to the inclusion criteria, the paper was selected to be further analysed in detail. Then, the research applied full text read for each of the papers selected in the first step. The research checked each of the exclusion criteria and re-evaluation the relevance of the paper to the research. Each abstract and paper was analysed by the author of this thesis, following a specific predefined protocol. When there were some doubts about the paper's relevance to the research questions, the opinion of a member of the supervisory team was requested to minimise potential

bias. Following the previous procedure, the filtering stage ended up with a total of 190 selected papers.

1.2. Results.

Analysing the literature has led to the identification of two categories of explainable models 1) those that are to explain any machine learning models (Model-agnostic); and 2) those that are designed for reverse-engineered specific machine learning model, thus cannot be used for other ML models. The research excludes the papers that only generate explanations for specific ML models. Model-agnostic explanation models are techniques that are designed to interpret the prediction of any machine learning model with the aim of generating some information from its prediction. The main purpose of such models is to provide extracted knowledge from the ML model, simplify the underlying logic and provide generalisability for other predictions. The results identify a list of five main explanation classes that supported by current model-agnostic explanations. Five main explanation classes that emerged from the literature review are presented in Table 4.

Table 4 Five main explanaion classes

Category of explainable methods	**Definition**	**Examples**
Explain the model (Global)	The explanation presents the weights of global features used in the model. This includes different visualisations of the feature's weights, such as approximate the model into interpretable global decision-tree or set of rules i.e., if-else.	(Henelius et al., 2014, Lou et al., 2013, Nguyen et al., 2016, Tolomei et al., 2017, Bastani et al., 2017, Johansson and Niklasson, 2009, Krishnan et al., 1999, Dash et al., 2018)
Explain a local prediction (Local)	The explanation describes the contribution of the local instance features to the model prediction. This also includes different visualisations such as pointing out single part of the image, local decision tree and local if-else rules.	(Lundberg and Lee, 2017, Ribeiro et al., 2016b, Ribeiro et al., 2018)
Counterfactual explanation	This explanation class shows how the prediction will change with regards to a change in the features' values. This also includes the changes in the prediction in the absent or present of specific features.	(Friedman, 2001, Apley, 2016, Dhurandhar et al., 2018, Wachter et al., 2017)
Example-based	The explanation presents examples that are similar or have small differences to the current prediction.	(Bien and Tibshirani, 2011, Kim et al., 2014, Koh and Liang, 2017)
Confidence	The explanation indicates the certainty values given a prediction.	(Subbaswamy and Saria, 2018, Gal and Ghahramani, 2016)

Explain the model. The global explanation model is developed to generate an explanation from a black-box model through an interpretable and explainable model. The generated explainable model is an approximation of the black-box model that explains the global behaviour of the model. This model should have similar accuracy and performance to the black-box model. Various papers in our literature review described novel techniques to solve the global explanation problem and generate an explainable model derived from the black-box model. The analysis of such methods introduced three sub-classes of global explanation models.

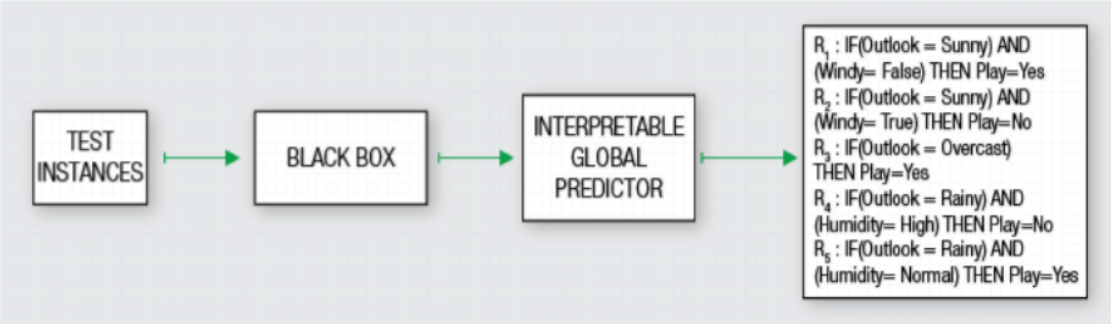

Figure 2 Global Interpretable example introduced by (Guidotti et al. 2018b)

1- Global feature importance. This category aims to find a group of data features that affects the performance and the prediction of the machine learning model. For instance, recent work presented a naïve Bayes classifier that can derive the dependencies between the data features attributes from any machine learning model (Henelius et al., 2014, Henelius et al., 2017). Their approach is able to generate importance score for each feature and reveals the association between different data features. Lou et al. (2013), Nguyen et al. (2013) and Tolomei et al. (2017) developed an explainable technique to rank all possible data features that contributes to the model overall prediction. Furthermore, the work proposed by (Datta et al., 2016) presents an explanation model that measures the degree of influence that a data feature input to the overall prediction of the model.
2- Decision tree approximation. This category of contributions presents an explainable model that is able to generate a global decision tree given a black-box model. For example, the following contributions (Bastani et al., 2017, Johansson and Niklasson, 2009, Zhou and Hooker, 2016) proposed approaches based on model extraction techniques that approximate any machine learning model to a simple and explainable decision tree model. Krishnan et al. (1999) overcome the limitation of the complexity of such generated decision tree by controlling the size of the decision tree. Their approach used a genetic algorithm to propose decision trees with varying sizes. In the same direction, Thiagarajan et al. (2016) developed a TreeView approach with the aim to generate a human-friendly decision tree by an iterative rejection of unlikely prediction label until the correct prediction appears.
3- Rule approximation. In the same way of the decision tree approximation, the explainable model generates set of rules that explain the global reasoning of the model. Dash et al. (2018) and Aung et al. (2007) developed an explainable model that learns the boolean rules in either disjunctive normal form (OR-of-ANDs) or conjunctive normal forms of (AND-of-ORs). Similarly, the following contributions (Wei et al., 2019, Tan et al., 2018) proposed generalised linear rule model that detects the interaction and the relationship between data features in the forms of decision rules. Also, Zhou et al. (2003) developed REFNE, an interpretable model that can extract rules with strong generalisation ability or with high fidelity and conciseness. The other type of rule extraction is i*f-then* form (Johansson et al., 2004, Quinlan, 1987). Figure 2 presents an example of the generated rules in both types of rules.

No.	Rule
1	physician-fee-freeze → democrat
2	¬education-spending → republican
3	¬handicapped-infants ∧ adoption-of-the-budget-resolution → republican
4	¬handicapped-infants ∧ aid-to-nicaraguan-contras → republican
5	¬water-project-cost-sharing ∧ adoption-of-the-budget-resolution → republican
6	water-project-cost-sharing ∧ mx-missile → republican
7	¬handicapped-infants ∧ ¬ superfund-right-to-sue → democrat
8	¬handicapped-infants ∧ ¬mx-missile → republican
9	¬water-project-cost-sharing ∧ religious-groups-in-schools → republican

```
(if (< TV0 17 )
    (if (< TV2 36 )
      (if (> OI1 40 ) 82 72)
      (if (< TV2 97 ) 83 97 ) )
    (if (> OP1 216 )
      (if (< TV2 97 ) 85 112 )
      (if (< TV2 40 ) 85 112 ) ) )
```

Figure 3 Two examples of rule extraction models presented in Zhou et al. (2003) and Johansson et al. (2004)

Explain a local prediction (Local). In the context of a single prediction or a recommendation, the explanation model is able to indicate reasons for specific prediction. Model agnostic derives the explanation from a local model that approximates the machine learning model well in a neighbour cluster of data points around a specific data point (Ribeiro et al., 2016b). The analysis of the literature identifies two main categories of explaining local prediction:

1- Local feature Importance. The explanation shows how data features of a prediction contribute to the machine learning model prediction such as including parts of an image or text. Among the different contributions in this category, the results identified LIME (Ribeiro et al., 2016b) and all its variation (Mishra et al., 2017, Ribeiro et al., 2016a) as a novel technique that can explain any machine learning classifier by learning an interpretable model locally around the prediction. Also, Lundberg and Lee (2017) developed a novel technique called SHAP (SHapley Additive exPlanations) framework as a unified measure of feature importance that various methods approximate. Similarly, Zhou et al. (2003) developed a general method that perturbs all subsets of features to deal with the shortcomings of other existing feature importance explanation methods. Their approach was focused on considering the interaction between data features. Other contributions such as (Simonyan et al., 2013, Fong and Vedaldi, 2017, Dabkowski and Gal, 2017) proposed an image saliency method, which is applicable for differentiable image classifiers.

2- Local rules and trees. These techniques for model-agnostic explanations are designed to be plugged into any machine learning model to extract some information from its local predication and present it as local rules and trees. For instance, the following contributions (Konig et al., 2008, Johansson et al., 2004) developed a rule extraction method (termed G-REX) based on genetic programming. In line with rule extraction, (Ribeiro et al., 2018) presents a novel method (termed Anchors) with high-precision rules representing sufficient condition for a single prediction. Local decision trees are another example of an explanation model to explain a single prediction through local rule extraction. This approach was preferable for many researchers in the analysis as it has a human-friendly nature (Guidotti et al., 2018, Krishnan and Wu, 2017)

Example-based explanation. Example-based explanation models select instances from the dataset to explain the behaviour of the black-box machine learning model. These models come to mimic the explainability behaviour between humans, and can be effective for explaining complex connotations (Renkl, 2014, Renkl et al., 2009). Example-based explanations could potentially give the users some intuition about the black-box model that is complex to be understandable through other model-agnostic models. Table 5 presents a categorisation of the reviewed papers. The analysis reveals three categories of explaining black-box models through examples:

1- Prototype: The examples in this category are representative samples of instances from the dataset with the same record as the prediction (Bien and Tibshirani, 2011, Kim et al., 2016). Prototype methods seek a minimal selection of similar instances as a general performance and accuracy goal (Kim et al., 2014). Furthermore, the following contributions (Kim et al., 2016, Kanehira and Harada, 2019) emphasis that explaining through prototype can lead to over-generalisation or misunderstanding of the presented explanations. They argued that examples might be useful when the distribution of the training data is clean, i.e., prototypical examples represent the current recommendation. However, this case is rare in real-world scenarios. Therefore, to help human decision making, their approach was to select and classify examples in the dataset into good examples and bad examples. Good examples represent the model behaviour in high accuracy, whereas bad examples do not fit the model reasoning.
2- Counterfactual. This refers to the explainable model that explains the black-box behaviour providing similar instances to the prediction with small differences (Laugel et al., 2017, Mothilal et al., 2020). It also answers the question "What-if" an input changes through examples; example provided in Figure 4 by (Wachter et al., 2017). Some researchers used heuristics for generating counterfactual explanation by amending some input features (Martens and Provost, 2014, Chen et al., 2017).

The prediction for a woman with Pima heritage are at risk of diabetes is 0.5.
Other persons that have similar score:

A. If your 2-Hour serum insulin level was 154.3 like **Person 1**, you would have a score of 0.51.
B. If your 2-Hour serum insulin level was 169.5 like **Person 2**, you would have a score of 0.51.
C. If your Plasma glucose concentration was 158.3 and your 2-Hour serum insulin level was 160.5 like **Person 3**, you would have a score of 0.51.

Figure 4 Counterfactual example-based explanation example (Wachter et al.2017)

3- Influential example. This explains the model prediction based on training instances that most responsible for influencing the prediction (Koh and Liang, 2017). Influence explainable models capture the idea of inspecting the black-box models through the lens of their training dataset. For instance, Goodfellow et al. (2014) identify that influence examples could be applied for various data science tasks such as understanding the model behaviour, debugging the black-box model and detecting errors. Similarly, the following contributions (Yuan et al., 2019, Dong et al., 2017, Szegedy et al., 2013) developed adversarial examples which are example-based methods with small, intentional feature perturbations that influence the black-box to false prediction. Although the literature provided theoretical foundations for the usefulness and effectiveness of influential examples on humans' decision-making, the research lacks user studies to understand the effect of influential examples on trust and calibrated trust.

Counterfactual explanation. This refers to explainable models that address the question of how the prediction would have been changed with a different set of input (Woodward, 1997). Counterfactual statements are usually taking the form: *Prediction P was made because the feature F has the value* f_1*. However, if F had the value* f'_1*, where other features had remained constant, Prediction* P_1 *would have been returned* (Wachter et al., 2017). They are designed in a way to convey a minimal amount of information capable of amending a prediction.

Researchers argues that counterfactual explanations are human-understandable explanations and they do not require the user to understand the underlying logic of the model (Arrieta et al., 2020). Designers and developers of such models often assume a clear and relation from recommended changes in feature values to actions in the real world (Barocas et al., 2020). However, in many cases such as medicine this assumption will fail counterfactual, e.g., an explanation might ask the doctor to change the age of the patient. The analysis identifies two main categories of counterfactual explainable models:

1- Feature Influence. It refers to explainable model that show how a prediction could change regarding to a change of a feature either in static way (Apley, 2016, Friedman, 2001, Goldstein et al., 2015) or interactive way (Krause et al., 2016) . Furthermore, it supports a localised inspection and feature tweak of a prediction to answer how and why specific prediction is predicted.

2- Counterfactual features are techniques aim to describe the features that will change the prediction when it is amended or deleted (Dhurandhar et al., 2018, Wachter et al., 2017). This method is argued to be efficient to support the user with a feedback when the model prediction is different from the desired prediction e.g. rejected loan application(Zhang et al., 2018).

Confidence explanations. It is an explanation class that shows the rationale of a given recommendation by presenting its certainty score. Confidence score can be generated from the machine learning models from two main sources (Gal and Ghahramani, 2016): model and data. Researchers generated confidence score at the model level by computing the distributional differences during the model training stage (Gal and Ghahramani, 2016, Schulam and Saria, 2019). Whereas, data confidence scores can come from noisy, missing or predefined assumptions on the data (Josse et al., 2019). The common technique in the literature to assess the confidence is using Bayesian methods e.g. (Graves, 2011, Blundell et al., 2015). There has also been work other techniques such as Dropout (Srivastava et al., 2014), tree-based density (Hooker, 2004) and simple heuristic using SoftMax (Hendrycks and Gimpel, 2016). Researchers argued that such confidence explanations can be used for trust calibration goal, when the designer of the system wants to inform the user about appropriate level of trust (Bussone et al., 2015, Helldin et al., 2013).

Table 5 The categorisation of the reviewed papers

Global explanations	Global feature importance	Ranking the data features.	(Lou et al., 2013, Nguyen et al., 2016, Tolomei et al., 2017)
		Dependencies between data features	(Henelius et al., 2014 Henelius et al., 2017)
		Influence Function	(Datta et al., 2016)
	Decision tree approximation	(Bastani et al., 2017, Johansson and Niklasson, 2009, Krishnan et al., 1999, Bastani et al., 2017, Johansson and Niklasson, 2009, Zhou and Hooker, 2016, Thiagarajan et al., 2016)	
		AND-OR rules	(Dash et al., 2018, Aung et al., 2007, Wei et al., 2019, Tan et al., 2018, Zhou et al., 2003)

	Rule extraction		
		If-then rules	(Johansson et al., 2004, Quinlan, 1987)
Explain a prediction	Local feature importance	(Ribeiro et al., 2016b, Lundberg and Lee, 2017, Simonyan et al., 2013, Fong and Vedaldi, 2017, Dabkowski and Gal, 2017, Zhou et al., 2003, Mishra et al., 2017, Ribeiro et al., 2016a)	
	Local rules and trees	(Guidotti et al., 2018, Krishnan and Wu, 2017) (Ribeiro et al., 2018) (Konig et al., 2008, Johansson et al., 2004, Soares and Angelov, 2019)	
Example-based	Prototype	(Bien and Tibshirani, 2011, Kim et al., 2016) (Kim et al., 2014) (Kim et al., 2016, Kanehira and Harada, 2019)	
	Counterfactual example	(Wachter et al., 2017) (Martens and Provost, 2014, Chen et al., 2017) (Laugel et al., 2017, Mothilal et al., 2020)	
	Influential example	(Koh and Liang, 2017) (Goodfellow et al., 2014) (Yuan et al., 2019, Dong et al., 2017, Szegedy et al., 2013)	
Counterfactual	Feature Influence	(Woodward, 1997) (Apley, 2016, Friedman, 2001, Goldstein et al., 2015) (Krause et al., 2016)	
	Counterfactual features	(Wachter et al., 2017) (Dhurandhar et al., 2018, Wachter et al., 2017) (Zhang et al., 2018) (Krause et al., 2016) (Barocas et al., 2020)	
Confidence	(Zhang et al., 2020, Bussone et al., 2015) (Gal and Ghahramani, 2016, Schulam and Saria, 2019) (Josse et al., 2019, Graves, 2011, Blundell et al., 2015) (Srivastava et al., 2014) (Hooker, 2004) (Hendrycks and Gimpel, 2016).		

2. Phase 1: think-aloud protocol.

Our study design and analysis of the data are situated within a two-dimensional space: *everyday Human-AI collaborative decision-making task where trust calibration errors are possible, and AI-based explanations to support trust calibration*. Through multi-stage qualitative research, we aim to answer the following questions:

RQ: *How to design for explainability that enhances trust calibration? What design techniques could be implemented, and what are suitable principles to guide the design?*

To this end, the research method of this paper included two phases: Exploration and Co-design. The exploration phase aimed to explore how users of everyday Human-AI collaborative decision-making tasks interact with AI-based explanations and why explanations are not improving trust calibration. The co-design phase goal was to investigate how users of XAI systems would like to integrate AI-based explanation in their everyday decision-making task. Co-design phase helped us to understand how the solution would look like from users' perspective. The following sections describe the research method.

2.1. Use case and underpinnings

Screening prescription is a process that medical experts in a clinic follow to ensure that a prescription is prescribed for its clinical purpose and fit the patient profile and history. The main workflow of the prescribing system shown in Figure 5.

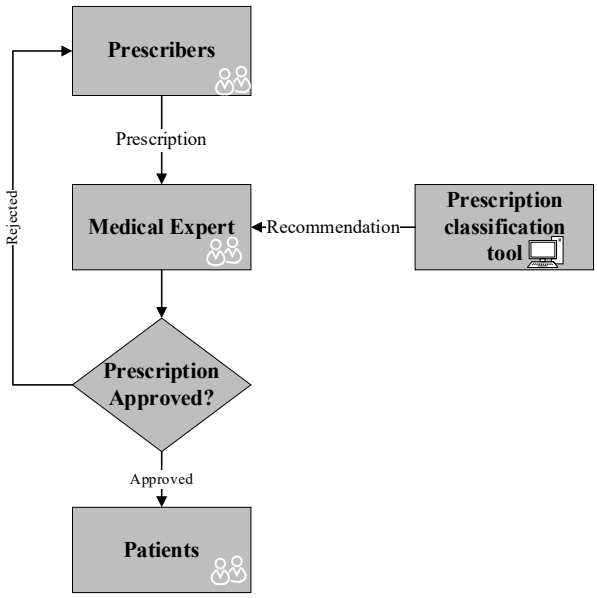

Figure 5 Screening prescription classification AI-based system classification

To help our investigation, we designed an AI-based decision-making mock-up meant to help classify the prescriptions into confirmed or rejected. We chose this case study to reflect an everyday Human-AI collaborative decision-making task where trust calibration errors are indeed possible. We designed the mock-up based on templates and interfaces familiar to our participants in their everyday decision-making tasks (See Figure 6). Our mock-ups mimic a web-based tool and are meant to simulate the user experience when working on an existing system. As the medical expert clicks on a prescription, the tool shows the patient profile and the recommendation from the AI-supported decision-making tool (confirmed or rejected). The user has access to AI-based explanations to understand the AI rationale of why the prescription should be confirmed or rejected.

2.2. Exploration

This stage included two sub-stage: Think-aloud and follow-up interviews.

2.2.1. Think-aloud stage

We aim to provide explanatory information that supports the medical practitioners in their trust calibration during Human-AI collaborative decision-making task. Our participant's inclusion criteria were based on their experience of using clinical decision support systems in their settings and experience in screening chemotherapy prescription (See Appendix A). We designed ten recommendations accompanied by ten different explanations. The adopted recommendations were generated to be non-trivial, which was based on a literature review on related work and medical expert judgment. We tested the material and activities with two participants and refined them to optimise their fulfilment of these criteria (See Appendix B). Also, we validated the material with a medical oncologist with a focus on the border cases that need an investigation from the participants in the actual study. This ultimately helped put our participants, who were medical experts, in a realistic setting: exposing them to an imperfect AI-based recommendation and its explanations where trust calibration is needed and where errors in that process are possible. We consulted with one AI expert and one medical expert, presenting them with ten explainable interfaces, and asked them for their expert opinion regarding the relevance of the explanations and the validity of the recommendation. We used these opinions, as well as the results from our pilot study, to refine the interface design. Each scenario considered a hypothetical patient profile and AI-recommendations that suggests either rejecting or accepting a chemotherapy prescription for the patient. Patients have been initialised with fictional names and profiles to make it more realistic to our practitioners. Each scenario was accompanied by one different explanation class and was meant to be either correct recommendation or incorrect recommendation. We used our five main explanation classes revealed from our previous literature review. We encouraged them

to think aloud during their decision-making process. Then, they were asked to think freely and encouraged to make optimal decisions. Examples of explainable interfaces used in our study settings are shown in Figure 6.

Figure 6 Mock-up interfaces used during think-aloud study- Each interface represent one explanation type.

2.3.1. Follow-up interview

At this stage, follow-up interviews were used to clarify the collected observations and participants think-aloud data and gather insights from the participants about their lived experience with AI explanations. This helped us to understand the nature of the users' errors and confirm our observations. The following questions summarises the questions asked to the participants.

General questions.

1. How would you summarise why the AI-supported decision tool made the recommendations?
2. What did you think of this explanation?
3. Can you explain the results of the AI recommendation in your own words?
4. How do you think the explanation could help you in your everyday decision-making activity?

Questions regarding a specific action during the think-aloud protocol.

1. Can you tell us why did you do that ….?
2. What did you think about that scenario?
3. What would you do in that scenario if you were in your clinic?

2.4. Co-Design.

We conducted two co-design sessions with eight participants, i.e., four participants in each session. The main aim of this stage was to explore how the design can play an effective role in enhancing users' trust calibration during a Human-AI collaborative decision-making task. We used the same inclusion criteria employed in the exploration stage, i.e., expert users in the studied task. We chose to recruit different participants to avoid the learning effect (Lazar et al., 2017) and increase the credibility of our findings as existing users already learned the objective of the study and were part of the underpinnings for this next study. Co-design method enables users who might be potential users in future AI-supported decision-making tools to reflect their experience in the design process, and this is supposed to increase the acceptance of the proposed solutions (Poole et al., 2008). Co-design can lead to a better understanding of the end-user needs, which enhances the possibility of the designs' acceptance (Song and Adams, 1993). In this phase, we discussed and negotiate how to embed AI explanations to serve users' needs, task workflow and trust calibration. Together with the participants, we conceptualised and sketched design features to support users in utilising AI explanation and reduce trust calibration errors revealed from the exploration phased. This was achieved by giving the participants initial prototypes or mock-ups (Clement et al., 2012) of the problem to help them visualise the idea and then provoke brainstorming related to the research problem. All these dynamics were hard to capture during the exploration phase. Therefore, co-design method helped us to come up with innovative designs of how the solution should look from a user perspective.

Participants were divided into two design sessions based on their availability. Due to the COVID-19 situation, we chose to conduct the study online using FreeHand tool from Invision[1]. Also, it has been shown that online tools for co-design can make the process easier, cheaper and flexible for participants (Näkki and Antikainen, 2008). To mitigate any potential issues that could arise from using online platforms, e.g., readability of the instructions and the tool usability issues, we conducted a pilot study with two post-graduate researchers and one academic in an interdisciplinary research group residing in the departments of Computing and Psychology in Bournemouth University. This also helped us in the preparation of the training and induction stage for the participants in the real study. All participants attended a training session to familiarise

[1] https://freehand.invisionapp.com/freehand/new.

themselves with the tools' functionalities and how they can communicate online. The training session lasted for 15-20 minutes. Then participants were invited to try the tool till they felt all capable of using it. They had the ability to ask questions and one of the authors answered them.

We adopted four techniques during the co-design sessions in order to reach the goal of our study (See Figure 2); researcher presentation, participants discussion, sketching-up exercise and focus groups. This also helped to enhance the credibility of the study and to ensure that data bias was eliminated. Each of the sessions lasted for around 2 hours. Both sessions, including the four main steps, were audio-recorded and transcribed. Audio recording for the design session helped the authors analysing main design needs and issues revealed from participants discussions. The following sections describe each technique that we used in our design sessions.

1) Researcher presentation (10 mins). The researcher gave a 10-minute presentation on AI-based decision-making tools and an overview regarding the first phase findings, particularly those about different types of errors that emerged during the exploration study. This helped to immerse the participants in the research problem, and it involved a warming-up activity in getting the participants involved in the design sessions.

2) Explanation and scenario discussion (25 mins): In this stage, participants started by introducing themselves. We then asked each participant to talk about how AI-based tools could help their everyday decision-making process. Then, we provided a definition for explainability methods introduced in previous interpretable machine learning surveys (Adadi and Berrada, 2018). We provided different e-cards describing different explanation types in simplified examples. This was meant to illustrate explainability definition and potential uses of these explanations. To answer our research question, the participants needed first to immerse in a fictional problem as recommended in (Buskermolen and Terken, 2012). In our study, the fictional problem was collaborative decision-making between the medical expert and the AI. Specifically, a screening prescription using and AI-based tool. The researcher invited participants to discuss the designed scenario of an AI-based collaborative decision-making tool of a screening prescription and its generated explanations. We used a random forest classifier as an ML algorithm to train our model. We then generated explanations from current state-of-art model-agnostic explanations to examine how users would like to receive these explanations and develop prototypes for effective utilisation for such explanations in real-world scenarios. This stage was meant to scope the discussion and facilitate focused conversations using the provided scenario. This was also meant to immerse the participants with the research problem and facilitate their understanding of the researcher presentation. Our participants discussed a wide range of trust calibration scenarios using the explanation interfaces through the provided material in this stage. This stage provided a sense of realism to the problem and encouraged careful consideration of solutions to cater to different contexts and usage styles. The following scenario and explanations were presented to our participants and discussed in this stage. We asked our participants to use the output from five explainable models and sketch up designs that help them to have appropriate trust in the AI recommendation and help them in their everyday Human-AI collaborative decision-making task. Below we describe the provided scenario and how we generate the explanations.

John is a doctor using AI-supported decision-making tool that recommends if a prescription shall be confirmed or rejected. While John was trying to understand why the AI is recommending that, he wanted to make informed decision using the below explanations. This might trigger two circumstances: either to reject correct AI recommendation or follow incorrect AI recommendation. (See Figure 7 that describes patient scenario).

Figure7 Provided patients' profile in the design sessions

Using the provided explanations, please answer the following questions:

1. How do you think each explanation should be designed to help you understand the AI recommendation?
2. How would you design the explanation to help you assess of the reliability of AI explanations?
3. How would you design the explanation to help you in judging the accuracy of the AI recommendation and its explanation?

Global feature importance. We used eli5[2] library in python to generate the global feature importance explanation. Below we see the importance features in the overall model recommendation.

[2] https://eli5.readthedocs.io/en/latest/overview.html

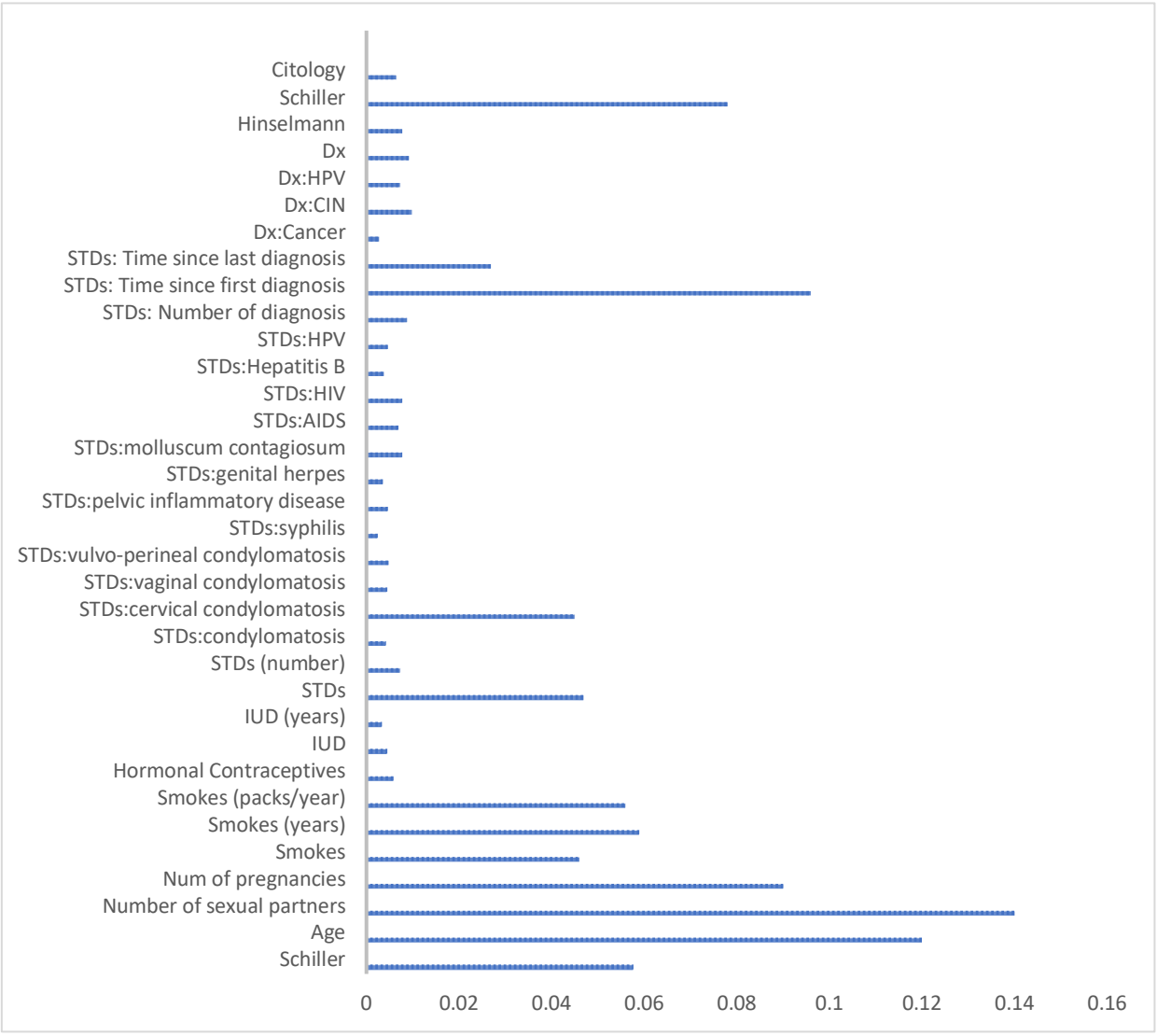

Figure 8 Global feature importance

Local Feature importance. We used LIME[3] to generate local feature importance given a patient record. Our model recommended that the patient does not have a cancer with 72.6% confidence. Figure 9 shows the generated local feature importance that shows why our system provide this recommendation (Minus values contributed to patient has a cancer, whereas positive values contributed to patient does not have caner).

[3] https://github.com/marcotcr/lime

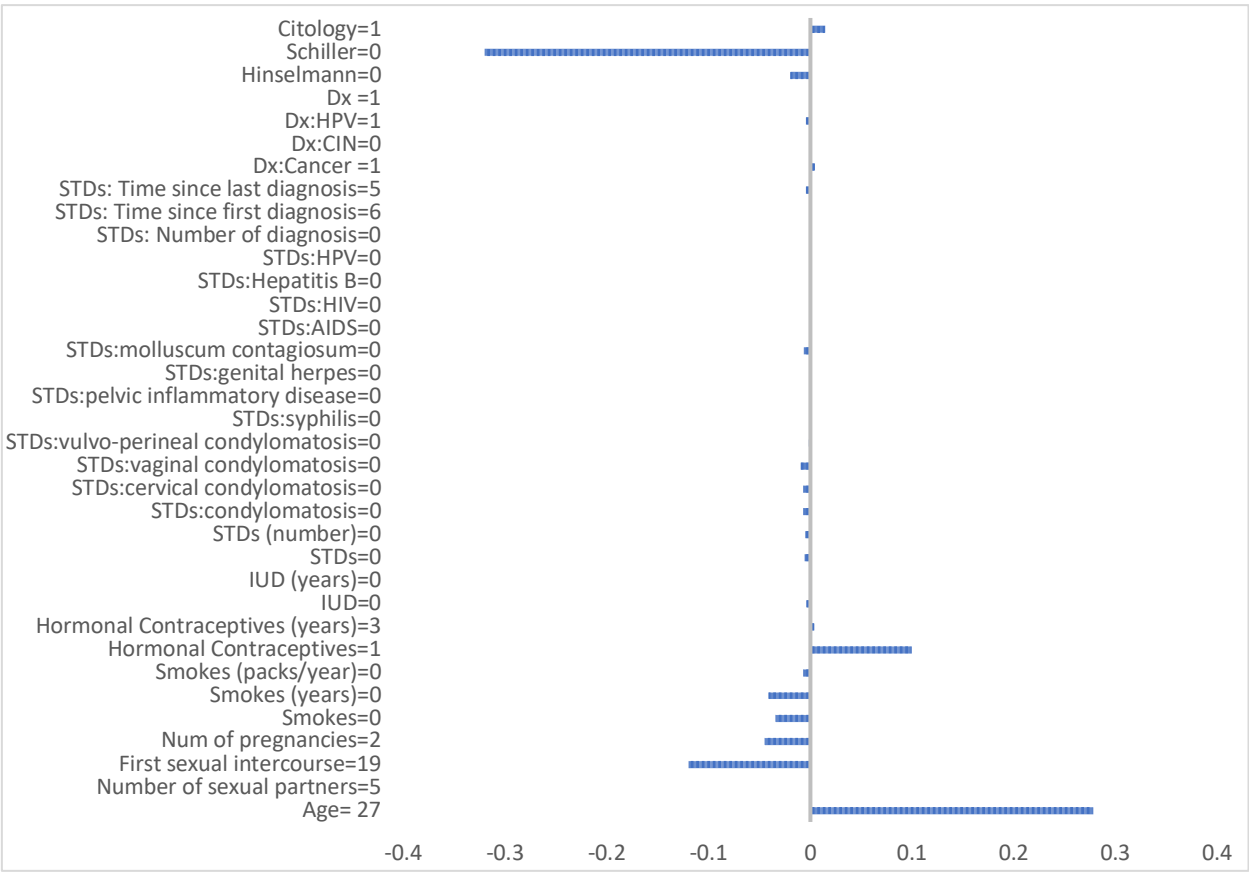

Figure 9 Local feature importance explanation

Counterfactual explanations. We used Alibi[4] library to generate counterfactuals given the same patient record. Below the generated counterfactual explanation that shows why our system provide this recommendation. Figure 10 shows the generated counterfactual explanation.

The patient would have a cancer with 67% confidence, if the First sexual intercourse=29 and Hormonal Contraceptives (years) = 13.

Figure 10 Counterfactual explanation

Example-based explanation. We used K-nearest neighbour algorithm to retrieve the k neighbours for the same patient record.

[4] https://pypi.org/project/alibi/

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs	STDs (number)	STDs:condylomatosis
0	18	4.0	15.0000	1.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
1	15	1.0	14.0000	1.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
2	34	1.0	16.9953	1.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
3	52	5.0	16.0000	4.000000	1.0	37.000000	37.000000	1.0	3.00	0.000000	0.000000	0.0	0.0	0.0
4	46	3.0	21.0000	4.000000	0.0	0.000000	0.000000	1.0	15.00	0.000000	0.000000	0.0	0.0	0.0
5	42	3.0	23.0000	2.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
7	26	1.0	26.0000	3.000000	0.0	0.000000	0.000000	1.0	2.00	1.000000	7.000000	0.0	0.0	0.0

Figure 11 Example-based explanations using KNN.

Confidence score. For confidence score, we used the function predict_proba implemented in the Random forest library. The algorithm confidence score for the patient record was **72%.**

3) Sketching-up exercise (40 mins): Participants were then encouraged to start sketching-up their designs using FreeHand tool from InVision. We gave each participant a blank e-page to sketch up designs considering five explanation types (Local, Global, Example-based, Counterfactual and Confidence explanations). The online platform provided several creation tools (e.g. coloured pens, shapes and sticky notes). The participants were also asked to not limit themselves to the given explanation classes and consider any extra features they would like to see in XAI interfaces to help them in utilising the explanation during a collaborative decision-making task. We deliberately asked our participants to work individually, think outside of the box, and consider different kinds of potential solutions. In this stage, our participants designed their explanations and provided multiple usage scenarios for them. They created a wide variety of usage scenarios covering different purposes and task requirements, e.g., grouping data features in Local explanations to reduce the explanation complexity. Below we provide a screenshot of the design space provided to our participants (Figure 12).

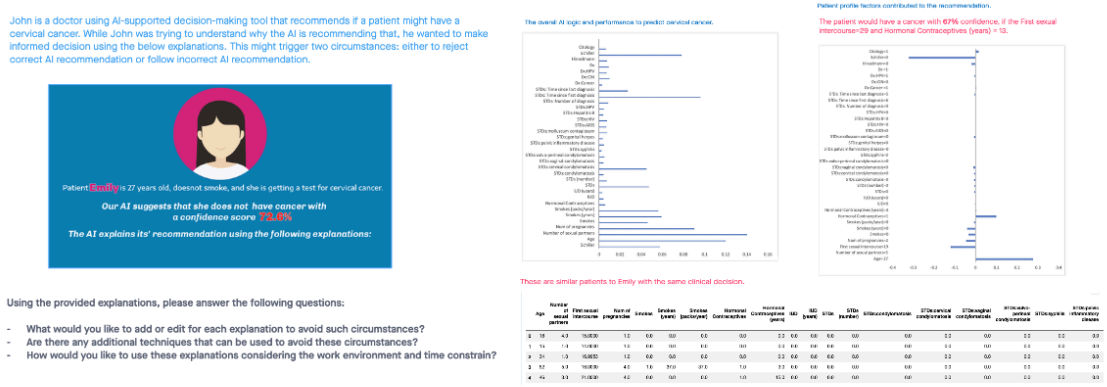

Figure 12 Design space provided to our participants

4) Focus group (45 mins). After each participant completed the sketching activity, each participant presented their ideas to the group. This was meant to critically analyse and evaluate the ideas by the participants in order to formulate robust solutions. This activity allowed our participants to explore and discuss various ways of using AI explanations in their work environment, considering trust calibration as the primary goal. Figure 13 shows a sample of the designs generated during the focus-group stage.

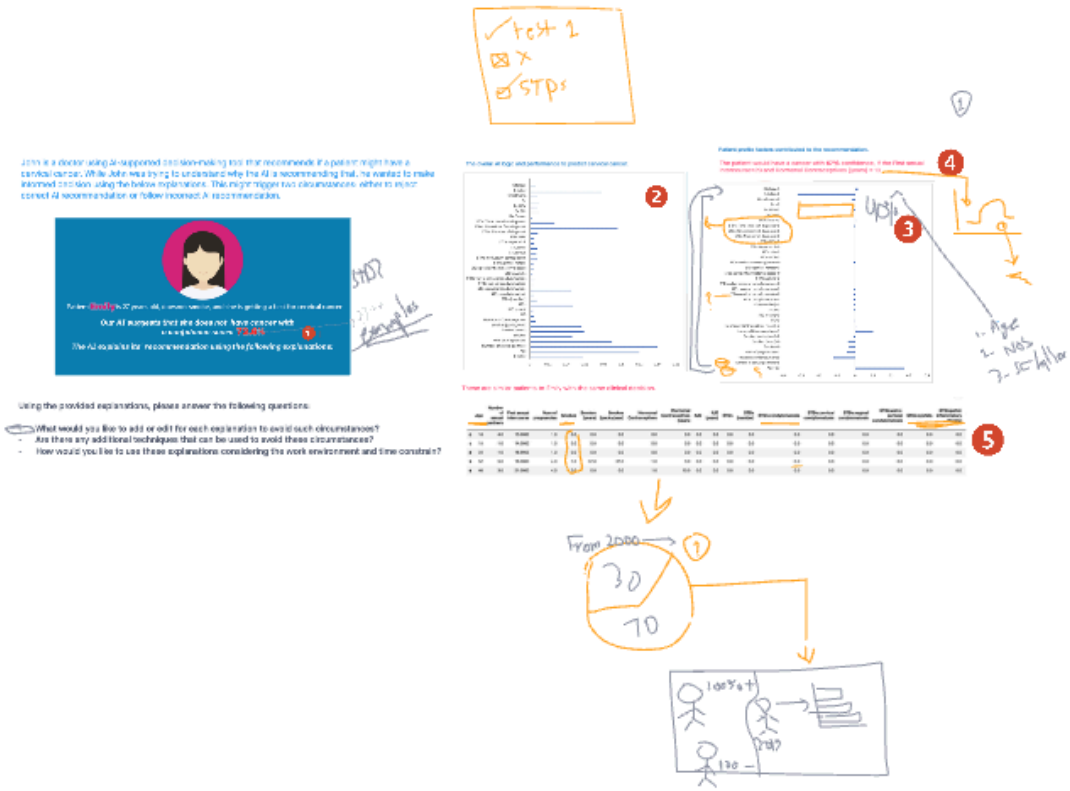

Figure 13 Sample of the collected data.

References

1. ADADI, A. & BERRADA, M. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access,* 6**,** 52138-52160.
2. APLEY, D. W. 2016. Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*.
3. ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., GARCÍA, S., GIL-LÓPEZ, S., MOLINA, D. & BENJAMINS, R. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion,* 58**,** 82-115.
4. BAROCAS, S., SELBST, A. D. & RAGHAVAN, M. The hidden assumptions behind counterfactual explanations and principal reasons. 2020 2020. 80-89.
5. BASTANI, O., KIM, C. & BASTANI, H. 2017. Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*.
6. BIEN, J. & TIBSHIRANI, R. 2011. Prototype selection for interpretable classification. *The Annals of Applied Statistics***,** 2403-2424.
7. BLUNDELL, C., CORNEBISE, J., KAVUKCUOGLU, K. & WIERSTRA, D. 2015. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.
8. BRANK, J. & GROBELNIK, M. A survey of ontology evaluation techniques. 2005.
9. BUSKERMOLEN, D. O. & TERKEN, J. Co-constructing stories: a participatory design technique to elicit in-depth user feedback and suggestions about design concepts. 2012. 33-36.
10. BUSSONE, A., STUMPF, S. & O'SULLIVAN, D. The role of explanations on trust and reliance in clinical decision support systems. 2015 2015. IEEE, 160-169.
11. CHEN, D., FRAIBERGER, S. P., MOAKLER, R. & PROVOST, F. 2017. Enhancing transparency and control when drawing data-driven inferences about individuals. *Big data,* 5**,** 197-212.
12. CLEMENT, A., MCPHAIL, B., SMITH, K. L. & FERENBOK, J. Probing, mocking and prototyping: participatory approaches to identity infrastructuring. 2012. 21-30.
13. DABKOWSKI, P. & GAL, Y. Real time image saliency for black box classifiers. 2017 2017. 6967-6976.
14. DASH, S., GUNLUK, O. & WEI, D. Boolean decision rules via column generation. 2018 2018. 4655-4665.
15. DATTA, A., SEN, S. & ZICK, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. 2016 2016. IEEE, 598-617.
16. DHURANDHAR, A., CHEN, P.-Y., LUSS, R., TU, C.-C., TING, P., SHANMUGAM, K. & DAS, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. 2018 2018. 592-603.
17. DONG, Y., SU, H., ZHU, J. & BAO, F. 2017. Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493*.
18. ELO, S. & KYNGÄS, H. 2008. The qualitative content analysis process. *Journal of advanced nursing,* 62**,** 107-115.
19. FONG, R. C. & VEDALDI, A. Interpretable explanations of black boxes by meaningful perturbation. 2017 2017. 3429-3437.
20. FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics***,** 1189-1232.
21. GAL, Y. & GHAHRAMANI, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. 2016 2016. 1050-1059.
22. GOLDSTEIN, A., KAPELNER, A., BLEICH, J. & PITKIN, E. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics,* 24**,** 44-65.
23. GOODFELLOW, I. J., SHLENS, J. & SZEGEDY, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
24. GRAVES, A. Practical variational inference for neural networks. 2011 2011. 2348-2356.
25. GUIDOTTI, R., MONREALE, A., RUGGIERI, S., PEDRESCHI, D., TURINI, F. & GIANNOTTI, F. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*.

26. HELLDIN, T., FALKMAN, G., RIVEIRO, M. & DAVIDSSON, S. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. 2013 2013. 210-217.
27. HENDRYCKS, D. & GIMPEL, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.
28. HENELIUS, A., PUOLAMÄKI, K., BOSTRÖM, H., ASKER, L. & PAPAPETROU, P. 2014. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery,* 28**,** 1503-1529.
29. HENELIUS, A., PUOLAMÄKI, K. & UKKONEN, A. 2017. Interpreting classifiers through attribute interactions in datasets. *arXiv preprint arXiv:1707.07576*.
30. HOOKER, G. Diagnosing extrapolation: Tree-based density estimation. 2004 2004. 569-574.
31. JOHANSSON, U., KÖNIG, R. & NIKLASSON, L. The Truth is In There-Rule Extraction from Opaque Models Using Genetic Programming. 2004 2004. Miami Beach, FL, 658-663.
32. JOHANSSON, U. & NIKLASSON, L. Evolving decision trees using oracle guides. 2009 2009. IEEE, 238-244.
33. JOSSE, J., PROST, N., SCORNET, E. & VAROQUAUX, G. 2019. On the consistency of supervised learning with missing values.
34. KANEHIRA, A. & HARADA, T. Learning to explain with complemental examples. 2019 2019. 8603-8611.
35. KIM, B., KHANNA, R. & KOYEJO, O. O. Examples are not enough, learn to criticize! criticism for interpretability. 2016 2016. 2280-2288.
36. KIM, B., RUDIN, C. & SHAH, J. A. The bayesian case model: A generative approach for case-based reasoning and prototype classification. 2014 2014. 1952-1960.
37. KOH, P. W. & LIANG, P. Understanding black-box predictions via influence functions. 2017 2017. JMLR. org, 1885-1894.
38. KONIG, R., JOHANSSON, U. & NIKLASSON, L. G-REX: A versatile framework for evolutionary data mining. 2008 2008. IEEE, 971-974.
39. KRAUSE, J., PERER, A. & NG, K. Interacting with predictions: Visual inspection of black-box machine learning models. 2016 2016. 5686-5697.
40. KRISHNAN, R., SIVAKUMAR, G. & BHATTACHARYA, P. 1999. Extracting decision trees from trained neural networks. *Pattern recognition,* 32.
41. KRISHNAN, S. & WU, E. Palm: Machine learning explanations for iterative debugging. 2017 2017. 1-6.
42. LAUGEL, T., LESOT, M.-J., MARSALA, C., RENARD, X. & DETYNIECKI, M. 2017. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*.
43. LAZAR, J., FENG, J. H. & HOCHHEISER, H. 2017. *Research methods in human-computer interaction*, Morgan Kaufmann.
44. LOU, Y., CARUANA, R., GEHRKE, J. & HOOKER, G. Accurate intelligible models with pairwise interactions. 2013 2013. 623-631.
45. LUNDBERG, S. M. & LEE, S.-I. A unified approach to interpreting model predictions. 2017 2017. 4765-4774.
46. MARTENS, D. & PROVOST, F. 2014. Explaining data-driven document classifications. *Mis Quarterly,* 38**,** 73-100.
47. MISHRA, S., STURM, B. L. & DIXON, S. Local Interpretable Model-Agnostic Explanations for Music Content Analysis. 2017 2017. 537-543.
48. MOTHILAL, R. K., SHARMA, A. & TAN, C. Explaining machine learning classifiers through diverse counterfactual explanations. 2020 2020. 607-617.
49. NÄKKI, P. & ANTIKAINEN, M. Online Tools for Co-design: User Involvement through the Innovation Process. 2008. Tapir akademisk forlag, 92-97.
50. NGUYEN, A., YOSINSKI, J. & CLUNE, J. 2016. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*.
51. POOLE, E. S., LE DANTEC, C. A., EAGAN, J. R. & EDWARDS, W. K. Reflecting on the invisible: understanding end-user perceptions of ubiquitous computing. 2008. 192-201.
52. QUINLAN, J. R. Generating production rules from decision trees. 1987 1987. Citeseer, 304-307.

53. RENKL, A. 2014. Toward an instructionally oriented theory of example-based learning. *Cognitive science,* 38**,** 1-37.
54. RENKL, A., HILBERT, T. & SCHWORM, S. 2009. Example-based learning in heuristic domains: A cognitive load theory account. *Educational Psychology Review,* 21**,** 67-78.
55. RIBEIRO, M. T., SINGH, S. & GUESTRIN, C. 2016a. Nothing else matters: model-agnostic explanations by identifying prediction invariance. *arXiv preprint arXiv:1611.05817*.
56. RIBEIRO, M. T., SINGH, S. & GUESTRIN, C. " Why should i trust you?" Explaining the predictions of any classifier. 2016 2016b. 1135-1144.
57. RIBEIRO, M. T., SINGH, S. & GUESTRIN, C. Anchors: High-precision model-agnostic explanations. 2018 2018.
58. SCHULAM, P. & SARIA, S. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. 2019 2019. 1022-1031.
59. SIMONYAN, K., VEDALDI, A. & ZISSERMAN, A. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034*.
60. SOARES, E. & ANGELOV, P. 2019. Fair-by-design explainable models for prediction of recidivism. *arXiv preprint arXiv:1910.02043*.
61. SONG, J. H. & ADAMS, C. R. 1993. Differentiation through customer involvement in production or delivery. *Journal of Consumer Marketing*.
62. SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. & SALAKHUTDINOV, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research,* 15**,** 1929-1958.
63. SUBBASWAMY, A. & SARIA, S. 2018. Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms. *arXiv preprint arXiv:1808.03253*.
64. SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., BRUNA, J., ERHAN, D., GOODFELLOW, I. & FERGUS, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
65. TAN, S., CARUANA, R., HOOKER, G. & LOU, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. 2018 2018. 303-310.
66. TOLOMEI, G., SILVESTRI, F., HAINES, A. & LALMAS, M. Interpretable predictions of tree-based ensembles via actionable feature tweaking. 2017 2017. 465-474.
67. WACHTER, S., MITTELSTADT, B. & RUSSELL, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.,* 31**,** 841.
68. WEI, D., DASH, S., GAO, T. & GUNLUK, O. Generalized Linear Rule Models. 2019 2019. 6687-6696.
69. WESTFALL, L. 2009. Sampling methods. *The Certified Quality Engineer Handbook*.
70. WOODWARD, J. 1997. Explanation, invariance, and intervention. *Philosophy of Science,* 64**,** S26-S41.
71. YUAN, X., HE, P., ZHU, Q. & LI, X. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems,* 30**,** 2805-2824.
72. ZHANG, X., SOLAR-LEZAMA, A. & SINGH, R. Interpreting neural network judgments via minimal, stable, and symbolic corrections. 2018 2018. 4874-4885.
73. ZHANG, Y., LIAO, Q. V. & BELLAMY, R. K. E. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* Barcelona, Spain: Association for Computing Machinery.
74. ZHOU, Y. & HOOKER, G. 2016. Interpreting models via single tree approximation. *arXiv preprint arXiv:1610.09036*.
75. ZHOU, Z. H., JIANG, Y. & CHEN, S. F. 2003. Extracting symbolic rules from trained neural network ensembles. *Ai Communications,* 16**,** 3-15.

Appendix A. Screening Survey

- Please provide your age category.
  - 20-30
  - 30-40
  - 40-50
  - 50-60
- Please provide your gender.
  - Male
  - Female

- Approximately how long have you been practicing clinically?
  - 0-5
  - 5-10
  - 10-15
  - 15-20
  - More than 20

- Please check all statements that apply regarding your level of experience screening chemotherapy prescriptions.
  - I know what screening prescription.
  - I have used a clinical decision support software.

Please indicate your level of agreement with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Agree Strongly
Artificial Intelligence will play an important role in the future of medicine					
There are too many complexities and barriers in medicine for AI to help in clinical settings.					
I have reservations about using AI in clinical settings.					

Appendix B. Scenarios' characteristics

Scenario Number	Explanation class	Type of recommendation
SC1	Confidence	Correct
SC2	Confidence	Incorrect
SC3	Counterfactual	Correct
SC4	Counterfactual	Incorrect
SC5	Global	Correct
SC6	Global	Incorrect
SC7	Local	Correct
SC8	Local	Incorrect
SC9	Example-based	Correct
SC10	Example-based	Incorrect

Table 1 Scenarios characteristics. Scenarios numbers do not represent the order of presentation.

	SC1 Male:54 CHF	SC2 Male:47 CHF	SC3 Female:56 CHF	SC4 Male: 44 not CHF
ER	Positive	Positive	Positive	Negative
No prior treatment with CDK 4/6	Yes	Yes	Yes	Yes
Adequate renal and hepatic function	Yes	Yes	Yes	Yes
ECOG PS	2	0	1	2
Neutrophils	1.20	0.9	1.00	0.7
Plt	80	74	33	84
Hepatic impairment	A	B	A	C
Other Toxicities	Grade 1	Grate 2	Grade 1	Grade 4

Table 2 Four examples of four patients' profiles presented in the scenarios.