Improving Health Professionals' Onboarding with AI and XAI for Trustworthy Human-AI Collaborative Decision Making

MIN HUN LEE, Singapore Management University, Singapore

SILVANA CHOO XINYI, Singapore General Hospital, Singapore

SHAMALA D/O THILARAJAH, Singapore General Hospital, Singapore

With advanced AI/ML, there has been growing research on explainable AI (XAI) and studies on how humans interact with AI and XAI for effective human-AI collaborative decision-making. However, we still have a lack of understanding of how AI systems and XAI should be first presented to users without technical backgrounds. In this paper, we present the findings of semi-structured interviews with health professionals (n=12) and students (n=4) majoring in medicine and health to study how to improve onboarding with AI and XAI. For the interviews, we built upon human-AI interaction guidelines to create onboarding materials of an AI system for stroke rehabilitation assessment and AI explanations and introduce them to the participants. Our findings reveal that beyond presenting traditional performance metrics on AI, participants desired benchmark information, the practical benefits of AI, and interaction trials to better contextualize AI performance, and refine the objectives and performance of AI. Based on these findings, we highlight directions for improving onboarding with AI and XAI and human-AI collaborative decision-making.

1 INTRODUCTION

Artificial intelligence (AI) has been increasingly being explored to provide data-driven insights for improving various decision-making tasks (e.g. health [5, 12, 15, 35, 67] and other social services [29, 74]). Even if recent research has demonstrated that these AI systems can have competent performance that can rival domain experts [12, 19, 33, 48, 59, 62], a fully autonomous approach of AI systems in high-stake contexts (e.g. health) is not desirable due to safety and ethical issues. A growing body of research has been conducted to investigate how humans and AI systems can complement each other's strengths [12, 35, 62] and integrate these AI systems in practice [5, 48, 59]. However, it is still challenging to integrate these systems in practice [27, 28, 61, 67] due to several factors, such as lack of user acceptance and trust [27, 28, 61] and difficulty with understanding the rationale of an AI output [12, 38, 55].

To address these challenges of integrating AI systems in practice, there is growing work that aims to make AI human-centered [5, 13, 34], explainable [1, 4, 31, 53, 68], and trustworthy [3, 20, 37]. Along this line, recent research works involved stakeholders to understand their practices and needs [34, 67, 72] and socio-environmental factors [5] to design explainable AI techniques to provide new insights on a decision making task and study how clinicians or health professionals can make use of AI outputs [11, 12, 34]. However, previous studies [10, 12, 34] assume that users without technical backgrounds can be onboarded with an AI system and AI explanations. Some research described the failures of effectively using AI explanations as they might be inadvertently the most understandable for users with technical backgrounds [60]. Additionally, there has been limited understanding of how AI systems should be introduced to users without technical backgrounds [13] and whether they can specify a desirable basic performance of AI to consider using it.

In this work, we focus on the context of physical stroke rehabilitation assessment and explore how AI and AI explanations should be introduced to users without technical backgrounds (e.g. health professionals and students majoring in medicine and health). To this end, we leveraged previous research of guidelines for human-AI interaction [2, 50], AI model card [43], onboarding recommendations [12], and tutorials of XAI techniques [31] to create onboarding tutorial materials of an AI-based decision support system and XAI techniques for the context of the study.

Our onboarding tutorial materials include 1) the description of the context (Figure 1a and 1c) and primary usage of AI (Figure 1b), 2) the introduction of AI (i.e. Figure 2a: inputs and outputs of AI & how it can be developed and operate, Figure 2b: dataset, and Figure 2c performance metrics and performance of AI), and 3) the descriptions of the motivation and meaning of an AI explanation (Figure 3a) and three widely used AI explanations (i.e. feature importance, counterfactuals, and prototype/example-based) for the context of the study (Figure 3b, 3c, and 3d).

Using our onboarding tutorial materials, we conducted a semi-structured interview with 12 health professionals and 4 students majoring in medicine and health. Throughout the interview, we learned their practices to build a trustworthy relationship with their colleagues. We also collected participants' feedback about the tutorial materials including their confusions to understand their information needs and areas for improvement. In addition, they were asked to describe a desirable performance AI to consider using it, rank the usefulness of three AI explanations for onboarding and decision support, and share suggestions on how to improve the onboarding and decision support with AI and AI explanations.

Our findings highlight the value of tutorials on AI and AI explanations along with the information needs of users without technical backgrounds (e.g. health professionals and students in medicine and health) on functional, developmental, and evaluation aspects of AI and how to make use of AI explanations. Specifically, participants suggested the context-specific required AI performance and evaluations to determine the usage of AI. Beyond presenting a numerical traditional performance metric, they also recommended communicating the benchmark information and the benefit of AI to contextualize AI capabilities and limitations. As they build a trustworthy relationship with their colleagues over time, they suggested providing iterative trials to refine AI objectives and tune it with feedback for trustworthy interactions with AI. Additionally, our study uncovered challenges of how AI explanations can be designed and used to improve onboarding with AI and support interactive communications with AI: creating a way to measure the level of understanding of AI and AI explanations, aligning goals between users and AI, and specifying the practices to audit AI.

Overall, our study provides insights into how AI and AI explanations can be presented to users without technical background and contributes to design considerations and challenges to improve their onboarding with AI. Our work advances ongoing discussions around onboarding and education of non-technical, domain users with AI [13, 29, 58] for effective human-AI collaborative decision-making in various high-stake domains.

2 RELATED WORK

2.1 Challenges of Deploying AI-based Decision Support Systems for Human-AI Collaboration

As AI achieves high performance to replicate expert's decision-making [19, 33, 48, 59], such as diagnosing prostate cancer [48] or assessing the quality of post-stroke rehabilitation exercises [33], AI has been investigated in the form of a decision support system [12, 35]. Specifically, ongoing research efforts explore to integrate AI-based decision support systems that provide data-driven insights (e.g. quickly retrieving similar cases from previously diagnosed patients [12] and identify important input features [35]) to enhance domain experts' accuracy and efficiency of decision making into practice [5, 12, 35, 67]. However, it remains challenging to integrate these systems into practice [27, 28, 61, 67]. One impediment to adopting these systems has been the lack of user acceptance and trust [27, 28, 61]. As these systems

utilize a complex AI algorithm and often operate as a black box [12, 55, 61], users have difficulty with understanding why the system provides a certain outcome [12, 38, 55]. If domain experts (e.g. clinicians, health professionals) do not understand the intended use, functionalities, or capability of an AI-based decision support system [40, 61], they may resist and abandon its usage [27].

2.2 Towards Human-Centered, Trustworthy, and Explainable AI

Researchers have emphasized the importance of making AI human-centered [5, 13, 34] , trustworthy [3, 20, 26, 37], explainable [1, 4, 31, 53, 68] to make it more deployable in practical settings. In the following subsections, we summarize the prior work on human-centered, trustworthy, explainable AI and describe how we build upon and differentiate with the prior work.

2.2.1 Designing Human-Centered AI. For human-centered designs and evaluation of AI, increasing recent research works [5, 34, 57, 72] highlight the importance of involving stakeholders to understand their challenges and needs [34, 67, 72] and socio-environmental factors [5]. For instance, Wang et al. [67] conducted interviews with clinicians in China and conducted observations to examine how AI-based decision support systems are used and discussed the issue of misalignment with local context and workflow and usability barriers. Lee et al. [34] interviewed and conducted a focus group session with therapists to understand their challenges and needs during rehabilitation assessment to design a human-centered decision support system. Beede et al. interviewed and observed the eye-screening workflows of clinics in Thailand, characterized the user expectations and post-deployment experiences of the AI-assisted screening process, and discussed the necessity of evaluating a system in socio-technical contexts [5].

In this work, we focus on the context of an AI-based decision support system for assessing physical stroke rehabilitation assessment. Building upon a growing body of research that highlights the value of human-centered approaches to invite feedback on designs of AI systems from the target users [5, 34, 67, 72], we engaged with stakeholders without technical backgrounds (e.g. health professionals) to explore how to improve their onboarding with an AI-based decision support system. As stakeholders without technical backgrounds may not provide sufficiently detailed suggestions on a narrow design aspect (e.g. the overall problem formulation) [29], our interdisciplinary team of a technical researcher and domain experts in stroke rehabilitation has worked together to create onboarding materials of AI and AI explanations and conducted semi-structured interviews with health professionals to collect their critiques and suggestions on how to improve onboarding with AI and AI explanations.

2.2.2 Efforts on Framework for Trustworthy AI. Trust is considered as a critical component of the successful deployment of AI and increasing research discusses about creating trustworthy AI [20, 63, 65, 70]. Although there has been little common understanding of what constitutes trust or trustworthy AI [22], researchers have discussed several definitions and frameworks of trustworthy AI. For instance, Vashney [65] builds upon the definition of trust and describes four attributes of trustworthy artificial intelligence: 1) technical competence that refers to the basic performance and accuracy of an AI model, 2) reliability and fairness that indicates maintaining good and correct performance across varying operating conditions, 3) understandability that describes whether users can comprehend the pipeline and lifecycles of an AI model, and 4) personal attachment/benevolence, which refers whether the purpose of an AI model can be aligned with a society's wants. In addition, Toreini et al. [63] based on the widely accepted principles of trust, ABI (Ability, Benevolence, Integrity) [41] and described a framework of trustworthy AI that includes a temporal dimension from initial trust to continuous trust [36, 63] and four technologies: fairness, explainability, auditability, and safety.

Although there are increasing efforts to make frameworks for trustworthy AI, there are still remaining questions on how these frameworks can be applied to create a new application of trustworthy AI (e.g. how we can effectively build initial trust with AI? what would be desirable basic performance of trustworthy AI and role of explainable AI techniques?). In addition, most prior work of designing human-centered, AI-based clinical decision support systems assumed that clinicians or health professionals onboard with AI and then studied how they can interact with these AI-based systems [5, 34, 57, 67, 72]. The problem of how users without technical backgrounds can be onboarded with AI [13] (i.e. understandability aspects of trustworthy AI [65]) is underexplored.

Building upon previous research of trustworthy AI [20, 63, 65, 70] and guidelines of human-AI interaction [2, 13, 14, 50] including AI model cards [43], we explored how onboarding tutorial materials of AI and XAI can be created and presented to users without technical backgrounds for trustworthy AI. Among various aspects and components of trustworthy AI, our work focuses on exploring how to build initial trust [63] and effectively onboard with AI while understanding the information needs of health profesionals on AI and XAI and exploring the possibility of defining the user's notion of basic performance of an AI model to start using it.

2.2.3 *Technically Oriented AI Explanations.*
To address the user's difficulty with understanding the rationales of an AI output/recommendation, researchers have explored techniques to make AI interpretable and explainable [4, 31, 53, 68]. These explainable AI techniques can be broadly categorized into 1) inherently interpretable models (e.g. rule-based models or linear regressions) whose internal mechanisms are directly interpreted and 2) post-hoc explainable AI (XAI) techniques that provide explanations of a complex algorithm (e.g. a deep learning model) [31]. Various post-hoc XAI techniques can be further classified into explaining the model's overall or instance-specific behavior [31]. Among various post-hoc XAI techniques, this work focuses on three widely used local XAI techniques: feature importance, counterfactual, and prototype/example-based explanations. A feature importance explanation describes how much input features contribute to a model output [23, 46, 56]. A counterfactual explanation describes how input features should be changed to update an AI output [25, 45, 66]. A prototype/example-based explanation aims to identify samples that are the most relevant and influential to an AI output [11, 23].

Explainable AI techniques that generate rationales of an AI output aim to serve a variety of users: technical AI/ML developers, who monitor and debug an AI model, or a non-technical, domain users, such as clinicians or health professionals, who review AI explanations as relevant evidence and outcomes on a decision-making task. However, prior research has shown that these AI explanations are not useful for people without technical background (e.g. clinicians or health professionals in clinical practice) [10, 52]. These failures of effectively using AI explanations might have occurred because these AI explanations are not designed for specific end-users or tasks [60]. These AI explanation methods might be inadvertently the most understandable to people with technical backgrounds who build and debug an AI model. As the end-users might have different needs, goals, and tasks when interpreting and reacting to AI model outputs and explanations, it is critical to engage with the end-user and make AI explanations user-centered.

In this work, we utilized three widely used XAI techniques and explored how these techniques can be used to improve users' onboarding with AI for their AI-assisted decision-making.

The most relevant research to our work is research by Cai et al. [13] that describes pathologists' information needs on an AI model (i.e. known strengths and limitations and its design objective). Although the previous work [13] provides several suggestions on onboarding with AI, it remains unclear how we can introduce users without technical backgrounds to the functionality, strengths and limitations, and design objectives of AI and AI explanations. Building upon this previous research [13], our research further investigates the usefulness of an AI model card [43] to communicate the competence

of an AI model [50] for user's onboarding with AI. Specifically, we studied whether users without technical backgrounds (i.e. health professionals) can leverage a traditional performance metric from an AI model card to understand the strengths and limitations of an AI model and determine whether an AI model can be used in practice. In addition, we conducted a deeper examination of aspects to faciliate users' onboarding with AI and AI explanations and how three widely used AI explanations can be used for onboarding and decision-making with AI. Our work further discusses considerations to improve onboarding with AI and AI explanations and human-AI collaborative decision-making.

3 STUDY DESIGN

This work aims to understand how an AI-based decision support system can be introduced to medical practitioners for its trustworthy usage. Specifically, we focused on studying (1) how well medical practitioners can understand the onboarding tutorial materials of an AI decision support system and (2) whether they can leverage a traditional evaluation metric which is commonly used by AI/ML researchers to indicate how well an AI/ML model can classify/predict ground truth scores (e.g. F1-scores) to determine the usage of the system, (3) the usefulness of three AI explanations for onboarding and decision support, and (4) informing the design of onboarding tutorial materials and considerations for trustworthy usage of an AI-based decision support system.

To address these research questions, we leveraged existing guidelines for human-AI interaction [2, 50] and onboarding with AI [13], an AI model card [43], tutorials of AI explanations [31] to create the onboarding tutorial materials of an AI-based decision support system for physical stroke rehabilitation assessment (Figure 1, 2, 3). In addition, we had iterative online synchronous and asynchronous discussions with domain experts in stroke rehabilitation to inform a set of semi-structured interview questions and refine onboarding tutorial materials of an AI-based decision support system for physical stroke rehabilitation assessment. During the online synchronous discussions, the leading researcher with a background of human-AI interaction and machine learning presented the draft of interview questions and onboarding materials and collected feedback on areas to be improved and revised questions, onboarding materials, and scripts for the follow-up discussions. After refining and finalizing the interview questions and onboarding materials, we conducted a pilot interview session with a student who majors in law and does not have technical backgrounds to check the length of a session and whether onboarding materials are understandable for people without technical backgrounds. Then, we conducted a semi-structured interview with healthcare professionals (i.e. therapists and a medical social worker) and students majoring in medicine and healthcare (e.g. nursing, therapy). This study including onboarding tutorial materials, protocol, and recruitment methods was approved by the Institutional Review Board.

3.1 Onboarding Tutorial Materials

Our onboarding tutorial materials are composed of three parts: introducing 1) the context and AI applications for physical stroke rehabilitation, 2) the development and evaluation of an AI model (e.g. how it is trained and operates on new data, dataset, evaluation metrics, and performance), 3) AI explanations (e.g. the motivation of AI explanations, feature importance, counterfactual, and example/prototype-based explanations).

First, building upon guidelines [43, 50], we described the context and challenges that an AI-based system aims to address (Figure 1a), the primary applications for the users of this study (i.e. therapists and post-stroke survivors) (Figure 1b), and envisioned use cases of quantitative stroke rehabilitation assessment (i.e. assessing the range of motion, smoothness, and the presence of compensatory motions) (Figure 1c). The descriptions of tutorial materials (Figure 1) to introduce the contexts of the study can be found below:
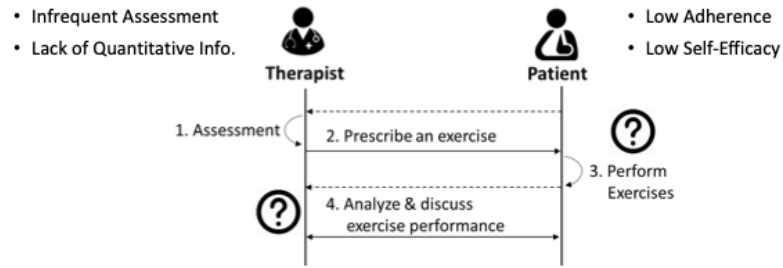
(a)
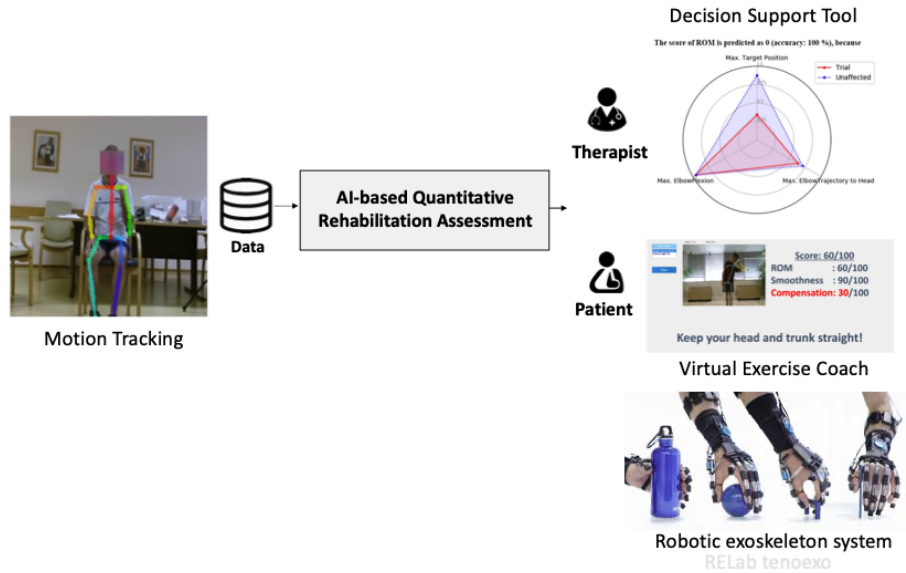
(b)

Performance Components	Score	Guidelines
Range of Motion (ROM)	0	Does not or barely involve any movement
	1	Less than half way aligned with an 'Target' position
	2	Movement achieves an 'Target' position
Smoothness	0	Excessive tremor or not smooth coordination
	1	Movement influenced by tremor
	2	Smoothly coordinated movement
Compensation	0	Noticeable compensation in more than two joints
	1	Noticeable compensation in a joint
	2	Does not involve any compensations

(c)

Fig. 1. Onboarding Tutorial Materials of an AI that introduce (a) an (c) the context of physical stroke rehabilitation assessment and (b) AI applications of physical stroke rehabilitation assessment and therapy.

Figure 1a: *"When stroke occurs, post-stroke survivors will have paralyzed and limited functional abilities. They typically involve therapy sessions to regain their functional and cognitive abilities. During therapy sessions, therapists assess the functional & cognitive status of a patient and prescribe a set of exercises to practice. In this work, we focus on the functional assessment of post-stroke survivors. During rehabilitation therapy, therapists often prescribe a set of exercises to a patient due to their limited availability. A therapist and a patient have regular follow-up meetings to discuss the patient's status and progress and adjust the rehabilitation program accordingly. During the follow-up meeting, there is limited quantitative information on the patient's status for therapists to make informed decision making."*

Figure 1b: *"To address this challenge, there have been increasing explorations on an AI-based system for rehabilitation. This system typically utilizes a vision-based or wearable sensor to estimate body joint positions and extract various kinematic features to quantify the patient's quality of motion. This quantitative assessment can be provided to therapists as a decision support system for improving their rehabilitation assessment or a virtual or exoskeleton system to improve patients' engagement in rehabilitation."*

Figure 1c: *"For the rehabilitation assessment, therapists assess patient's quality of motion in the following three aspects: Range of Motion, Smoothness, Compensation. ROM indicates whether a patients can achieve a specific target motion. Smoothness indicates whether a patients can coordinate their motion smoothly. Compensation checks whether a patient involves any unnecessary joint motion; here this patient compensates with his shoulder and trunk to move his arm that is affected by stroke."*
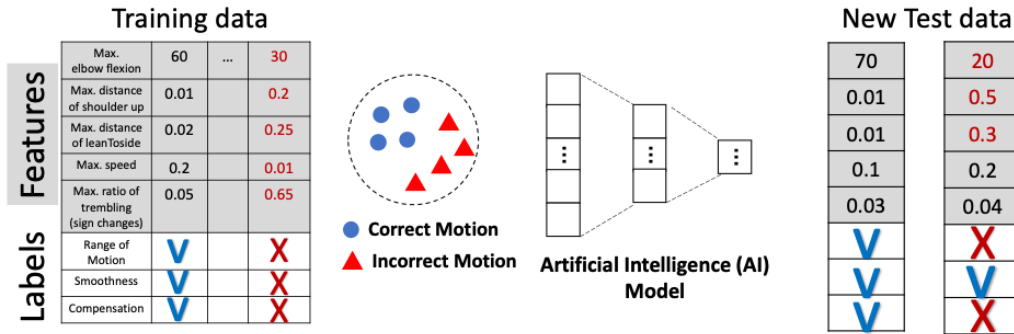
In the second part of the onboarding tutorial materials, as suggested by the guidelines [43, 50], we explained the inputs and outputs of an AI-based system and how a typical AI-based system for rehabilitation assessment can be developed and operated (Figure 2a) and described the dataset [33] (i.e. how it is collected and labeled) (Figure 2b). In addition, we elaborated on performance metrics and the performance of an AI model [43, 50] (Figure 2c). For reporting the AI performance, we utilized the dataset [33] and followed the previous research on quantitative stroke rehabilitation assessment [33] to implement a feed-forward neural network model from using Pytorch libraries [51]. The implementation details of the AI model can be found in the Appendix. A. We then reported how well an AI model can assess three common performance components of rehabilitation assessment (i.e. range of motion, smoothness, and compensation). The descriptions of tutorial materials (Figure 2a) to introduce the development, operation, and evaluation of AI can be found below:

Figure 2a: *"Here, we describe the pipeline of developing an AI model in more detail. When the system estimates body joints, it extracts kinematic features, such as elbow flexion, how much each joint moves in a certain direction, and computes the overall statistics during an exercise, such as the maximum value of elbow flexion and the corresponding labels of exercises, whether an exercise has the full range of motion or not; smooth or not; involves any compensations or not;*

For training an AI model, we collect samples of patients' exercises, extract features, and collect labels. Here, we have only 5 kinematic features, but in a real case of the development, we have a lot more features and samples. Given these paired features and labels, an AI model learns a function that maps features and corresponding labels as closely as possible.

Given new test data and extracted features from a patient, the AI model generates an outcome whether an exercise has full ROM or not, smooth, or involves any compensations or not. Here, the elbow flexion angle is similar to the normal case. And the distance of shoulder up is also small and close to the normal and along with a similar ratio of trembling. Thus, we have AI outputs of normal ROM, smooth, and no compensation."

Figure 2b: *"For an explorative study, we collected 300 samples of exercises, in which 15 post-stroke survivors performed 10 trials on their unaffected and affected sides. We collected labels of these exercises from a therapist, who*

Training data

Max. elbow flexion	60	...	30
Max. distance of shoulder up	0.01		0.2
Max. distance of leanToside	0.02		0.25
Max. speed	0.2		0.01
Max. ratio of trembling (sign changes)	0.05		0.65
Range of Motion	V		X
Smoothness	V		X
Compensation	V		X

Features / Labels

● Correct Motion
▲ Incorrect Motion

Artificial Intelligence (AI) Model

New Test data

70	20
0.01	0.5
0.01	0.3
0.1	0.2
0.03	0.04
V	X
V	V
V	X

(a)

Dataset

- **300 Samples**: 15 post-stroke survivors performed 10 trials on their unaffected & affected sides

- **Labels**: Range of Motion, Smoothness, Compensation - a therapist, who have experience in stroke rehabilitation for 5 years & conducted Fugl Meyer Assessment on 15 post-stroke survivors

Patient ID	Total Fugl (0-66)	Age	Sex	Affected Side	Type
P01	65	69	M	Left	Not Specified
P02	65	60	M	Left	Hemorrhagic
P03	66	61	M	Left	Not Specified
P04	66	63	M	Right	Ischemic
P05	55	51	M	Left	Ischemic
P06	13	63	M	Left	Ischemic & Spastic
P07	42	86	F	Right	Ischemic
P08	15	71	M	Left	Ischemic
P09	35	78	M	Left	Hemorrhagic
P10	21	53	M	Right	Ischemic
P11	16	37	M	Right	Ischemic
P12	11	61	M	Left	Hemorrhagic
P13	46	59	M	Left	Ischemic
P14	11	67	M	Left	Ischemic
P15	34	66	F	Left	Ischemic

(b)

- If you want fewer false positives, consider optimizing precision - TP / (TP + FP)

- If you want fewer false negatives, consider optimizing recall - TP / (TP + FN)

- If you want fewer false positives & negatives, consider optimizing F1-Score - (Precision * Recall) / (Precision + Recall)

- **Performance of an AI Model**: Neural Network
  - ROM.   : 82% F1-Score | therapists' agreement: 95%
  - Smoothness   : 79% F1-Score | therapists' agreement: 55%
  - Compensation: 77% F1-Score | therapists' agreement: 72%

		Prediction	
		Positive (Incorrect)	**Negative (Correct)**
Reference	Positive (Incorrect)	True Positive (TP)	False Negative (FN)
	Negative (Correct)	False Positive (FP)	True Negative (TN)

(c)

Fig. 2. Onboarding Tutorial Materials of an AI: (a) a diagram of how AI is developed and operated, (b) dataset, and (c) evaluation metrics and performance

has 5 years of practicing stroke rehabilitation and conducted a fugl meyer assessment on 15 post-stroke survivors. Using this dataset, we developed an AI model to replicate therapist's assessment on ROM, Smoothness, Compensation of patient's exercises."

Figure 2c: *"Any AI model that we build is guided by a reward function, which the AI model uses to determine 'right' or 'wrong' outcomes. We should consider how we specify this reward function that the system will optimize for [50]. When an AI generates outcomes of whether an exercise is correctly conducted or not, there are four possible outcomes [50]: True Positive: indicates AI outputs an 'Incorrect' motion given an 'Incorrect' motion; True Negative: indicates AI outputs "correct" motion given a 'correct' motion; False Positive: indicates AI outputs an 'Incorrect' motion given a 'correct' motion; False Negative: indicates AI outputs 'correct' motion given an 'Incorrect' motion*
If you want fewer false positives, you can consider optimizing precision; If you want fewer false negatives, you can consider optimizing recall; If you want fewer false positives & negatives, you can consider optimizing the F1-score
Given our dataset, we optimize an AI model to have fewer false positives & negatives. An AI model achieves a 82% F1-score on ROM; 79% on Smoothness; 77% on compensation;
To understand the competence of an AI model to replicate a therapist's assessment, we computed how well a secondary therapist agrees with the therapist, who generated annotations. Overall, our AI model can achieve comparable performance with a secondary therapist."

In the third part of the tutorial materials, we first described the motivation and meaning of an AI explanation [31] using an image classification task [56]. In addition, we explained three commonly used local AI explanations (i.e. feature importance, counterfactuals, and prototypes/example-based) [18, 31] for the context of the study. The descriptions of tutorial materials (Figure 3) to introduce the AI explanations can be found below:
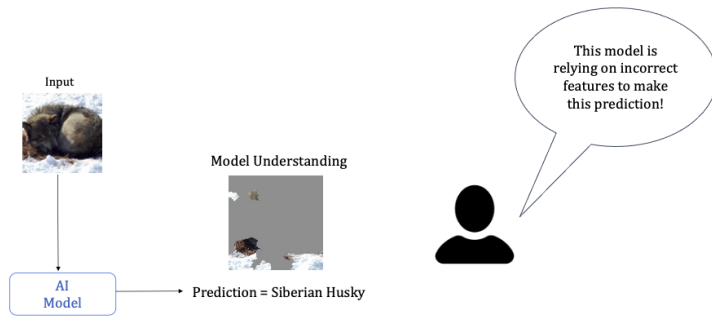
Figure 3a: *"There has been increasing research on AI explanations, which aims to improve users' understanding of how the AI-based system works and determine when to trust an AI output*
Here, we have an AI model that classifies the type of animals from an image; Given an input image, an AI model classifies it as Siberian Husky. Here I give you one example of an AI explanation; An AI explanation describes which parts of an image an AI model relied on for its output. We can find that an AI model focuses on snow parts and identify the limitation of an AI model. Thus, we need to be careful of using this AI model."

Figure 3b: *"Feature importance describes the overall importance of different features on AI model outcomes; Once we identify important features, we can pick the top most important features and show the comparison of these feature values on patients' unaffected and affected sides to check the threshold/boundary between unaffected/affected side makes sense or not and determine whether to trust AI or not."*

Figure 3c: *"Counterfactuals describe how input features need to be changed to generate an opposite outcome. By reviewing which inputs lead to different outcome, we can understand how an AI operates on specific inputs and determine whether we can trust AI or not. For instance, given an AI output of detecting compensation (1), the counterfactual explanations describe that an AI model will generate the output of no compensation if the feature value of max. leaningbackward-shouldervalue is decreased to 0.35."*

Figure 3d: *"Prototype/Example-based Explanation shows which sample data points are the most similar to input data. By reviewing relevant samples and the outputs of AI models on these samples, we can identify whether an AI model has the right outputs or not. Also, we can summarize the input feature into two key dimensions and use these representations to identify which samples are closely aligned and similar with each other. Specifically, we can project the features representation of each patient in the visualization. For instance, by reviewing this Figure, we can understand how each patient has been represented and which patients are represented/considered similar by an AI model and check whether an AI model utilizes correct feature representations or not."*

For a feature importance explanation, we utilized the SHAP library that support consistency and local accuracy [39] to compute importance scores (i.e. SHAP values) of each feature (Figure 3b). In addition, we utilized the top three most important features to show the difference between patient's unaffected and affected sides by stroke using a radar chart [2, 35], following the practices of therapists to compare patient's unaffected and affected side [34].

Input

Model Understanding

This model is relying on incorrect features to make this prediction!
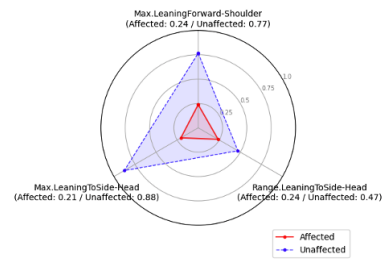
AI Model → Prediction = Siberian Husky

Lakkaraju, Himabindu, Julius Adebayo, and Sameer Singh. "Explaining machine learning predictions: State-of-the-art, challenges, and opportunities." *NeurIPS Tutorial* (2020).
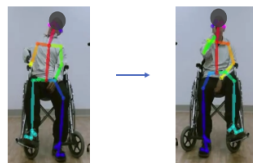
(a)

- Feature Importance
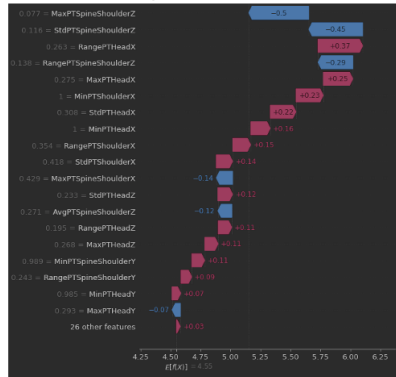
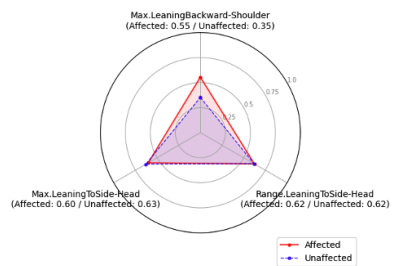The score of COMP is predicted as 2 (Confidence: 28%), because

(b)

- Counterfactual

What features need to be changed to flip a model prediction?
(i.e. compensation -> no compensation)

The score of COMP is predicted as 1 (Confidence: 32%), because

What If Explanations:
The AI prediction will be updated to the score of 2 if
Max. LeaningBackward-Shoulder value is decreased to 0.35

(c)

- Prototypes/Example Based

Which samples have the most relevance on the prediction?

AI Model → Limited ROM Compensations

Relevant samples

(d)

Fig. 3. Onboarding Tutorial Materials of an AI explanations: (a) motivation of an AI explanation, (b) a feature importance explanation, (c) a counterfactual explanation, and (d) prototype/example-based explanations.

A counterfactual explanation indicates what changes in feature values will lead to updating an AI output in a certain way [31, 32, 45]. For a counterfactual explanation, we utilized the DiCE library [45] to apply a genetic algorithm [49] to find counterfactuals close to the query point. We specified the features to be changed in the DiCE library using the identified salient features by the SHAP library and their desired range using patients' held-out normal data to avoid generating varying and unfeasible explanations. After identifying counterfactual explanations, we generated textual descriptions of the changes in feature values and AI outputs (Figure 3c). For example, Figure 3c describes that the value of *'Max.LeaningBackward-Shoulder'* should be decreased to 0.35 to update the AI output from 1 (i.e. noticeable compensation) to 2 (i.e. no compensation).

A prototype/example-based explanation describes representative or similar samples of a current instance along with AI outputs and ground truths on those samples [11, 12, 31] (Figure 3d). We utilized the kinematic features of a patient's motion [33] and computed the cosine similarity score to identify similar samples, which assist a user in understanding and validating an AI model output. In addition, we utilized a Principal Component Analysis (PCA) [71] to reduce the dimension of a feature representation. PCA was utilized because it does not require hyperparameter tuning and is deterministic unlike another widely used dimensional reduction technique, t-Distributed Stochastic Neighbor Embedding (t-SNE) [64]. We then visualized this reduced feature embedding space [7] for a user to check whether the feature representation of an AI model is valid or not (Figure 3d).

Although we utilized a particular technique/library to identify important features, select counterfactual explanations, and reduce the dimension of a feature representation, this work does not intend to communicate the usage of these specific techniques to create tutorial materials without any considerations. Alternative techniques/libraries can be explored for different applications.

3.2 Recruitment and Demographics

We recruited sixteen participants for our study (Table 1). The detailed demographics can be found in Appendix. Table 2.

Table 1. Demographics of Participants: Therapists who have experience in stroke rehabilitation (P1 - P10) and other health professionals (P11 - P12) and students majoring in medicine or health (e.g. therapy, nursing) (P13 - P16).

PID	Occupation	# of yrs	Q. Tech Experience	Q. ML Outputs	PID	Occupation	# of yrs	Q. Tech Experience	Q. ML Outputs
P1	PhysioTherapist	7	4.8 out of 7	3 out of 3	P11	Speech Therapist	5	4.8 out of 7	2 out of 3
P2	PhysioTherapist	2	5.2 out of 7	1 out of 3	P12	Medical Social Worker	5	4.0 out of 7	1 out of 3
P3	PhysioTherapist	8	4.4 out of 7	2 out of 3	P13	Student in Occupational Therapy	n/a	3.8 out of 7	3 out of 3
P4	PhysioTherapist	11	5.8 out of 7	2 out of 3	P14	Student in Speech Therapy	n/a	3.8 out of 7	2 out of 3
P5	PhysioTherapist	9	5.4 out of 7	2 out of 3	P15	Student in Medicine	n/a	3.8 out of 7	2 out of 3
P6	PhysioTherapist	30	5.8 out of 7	2 out of 3	P16	Student in Nursing	n/a	4.0 out of 7	0 out of 3
P7	Occupational Therapist	14	5.4 out of 7	2 out of 3					
P8	Occupational Therapist	11	6.2 out of 7	0 out of 3					
P9	Occupational Therapist	6	4.4 out of 7	2 out of 3					
P10	Occupational Therapist	5	3.2 out of 7	3 out of 3					

Our participants were mostly therapists who had experience in stroke rehabilitation (P1 - P10). Among ten therapists in stroke rehabilitation, six of them are physiotherapists, who promote and maintain patient's physical impairments from bio-mechanical perspectives and four of them are occupational therapists, who assist patients to better engage in their daily activities. Therapists work in various settings: four participants are from outpatient clinics, three participants are from inpatient rehabilitation, two participants are from home care, and two participants are from skilled nursing facility

(Appendix. Table 2). As some outpatient clinics have an interdisciplinary team to support and manage a patient (e.g. physio/occupational therapists, speech therapists, nurses, doctors), we also included health professionals (P11 and P12) (e.g. speech therapist and a medical social worker) and students majoring in medicine and health (e.g. therapy and nursing) (P13 - P16). The student in occupational therapy (P13) and the student in speech therapy (P14) had an experience of working as an occupational/speech therapy assistant for stroke rehabilitation. Participants were recruited through advertisements sent to the hospital staff, the mailing lists, and the contacts of the research team.

We asked the participants to respond to a set of technical experience questions on recent technologies , which were based on questions designed by the Center for Research and Education on Aging and Technology Enhancement (CREATE) [17]. Specifically, they were asked to rate their experiences with recent technologies (i.e. computer/laptop, activity tracker, virtual voice assistant, unmanned convenient store, and autonomous vehicle) on a 7-point scale (1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = neutral, 5 = somewhat agree, 6 = agree, and 7 = strongly agree). A high score on technology experience (e.g. 7) indicates that a participant self-reported to be highly experienced with a recent technology. Overall, participants expressed a diverse level of experience with recent technologies, in which they had an average score of 4.31 out of 7.0. In addition, the column of 'Q.ML Outputs' in Table 1 describes how many times a participant correctly guess an AI output after introducing how an AI model is developed and operated (Figure 2a). In Section 4.2, we described overall results of participants and how well the scores of guessing AI outputs are correlated with the scores of technology experiences.

3.3 Protocol

We conducted semi-structured interviews with 16 participants: 12 healthcare professionals (11 therapists and 1 medical social worker) and 4 students majoring in medicine and healthcare (e.g. therapy, nursing). Our interview protocol is composed of five main parts. The list of interview questions for a semi-structured interview can be found in Appendix. Table 3

First, we asked participants to describe their work environment, how they build a trustworthy relationship with their colleagues, and how they discuss uncertain cases. We included this first question to understand any practices or aspects that should be considered for them to have trustworthy interaction with an AI system (Appendix. Table 3). P13 and P14 shared the responses based on their experience of working as a therapy assistant in a hospital. P15 elaborated on the experience of working with medical students and P16 described the experience of working as a part-time nurse.

Next, we introduced the context and primary application of this study (i.e. physical stroke rehabilitation and a decision support system) (Figure 1a, 1b, and 1c) and explained inputs and an output of an AI-based decision support system for rehabilitation assessment and how it can be developed and operated on a new case using onboarding tutorial materials (Figure 2a). Also, a participant was asked to review inputs of a new case to an AI model and guess expected AI outputs.

Third, we described the dataset (Figure 2b), evaluation metrics, and AI performance along with therapists' agreement levels (Figure 2c). The therapists' agreement levels indicate how well annotations of a secondary therapist are aligned with ground truth scores and could provide a reference on how well our AI model performs. After describing the dataset, evaluation metrics, and AI performance, we asked the participants without experience of using AI systems and AI explanations for their practices (1) if they have a particular baseline performance of an AI model that is required and (2) if they have any specific conditions or edge cases [13] that should be included in the dataset or which an AI model should be evaluated or good for considering using AI or for their trust usage (Appendix. Table 3). Note that we did not ask the second question to participants (P11 - P12 and P13 - P16), who do not have extensive experience in stroke rehabilitation as a therapist.

Fourth, after introducing the motivation of an AI explanation (Figure 3a) and three widely used AI explanations (i.e. feature importance (Figure 3b), counterfactual (Figure 3c), and prototype/example-based (Figure 3d) explanations), we asked participants to rank which AI explanations are useful to support onboarding (i.e. when a user initially starts reviewing and understands AI performance) and decision support (i.e. when a user starts reviewing AI outputs for their decision making) (Appendix. Table 3). Finally, we asked them to share any comments on how to improve onboarding with an AI system and validating an AI output during decision support (Appendix. Table 3).

Throughout the semi-structured interviews, after reviewing onboarding tutorial materials, we asked the participants whether they followed the tutorial materials or had any clarification questions and asked them to provide any feedback on materials. All interviews were conducted remotely on a video conference platform and recorded for data analysis. Each interview lasted between 60 to 80 minutes. The participants were compensated for their participation based on the rate recommended by the domain experts.

3.4 Data Analysis

We transcribed all 17.23 hours of interview recordings into text data for thematic analysis [9]. We utilized both deductive and inductive thematic analysis approaches [9, 21] to analyze our interview data. We first selected codes based on our interview questions for our research problem: practices of rehabilitation and building a trustworthy relationship, comments on AI outputs, a development pipeline, the dataset, and the minimum performance, and comments on the AI explanations, onboarding with AI, and AI-assistive decision-making. Three researchers including one who facilitated interviews then independently coded the transcript data to generate 519 codes. We refined the initial codes while discussing disagreements and ambiguities in the codes [9, 21, 42] through iterative sessions. Following the practice of a reflective analysis that collaboratively shapes codes through discussion for consensus of codes [9, 42], we did not calculate inter-rater reliability. After coding, we grouped similar codes to identify and conceptualize higher-level themes through affinity diagramming. Overall, this process yielded five high-level themes, twenty one second-level themes, and eighty five third-level themes (Appendix. Table 4). In Section 4, we summarize our findings that broadly corresponds to the higher-level themes that we identified from our interview data.

4 RESULTS

4.1 Clinical Practices of Therapy & Trustworthy Relationships

4.1.1 Clinical Practices. Our participant therapists were from various settings: outpatient clinic (4); inpatient rehabilitation (3); skilled nursing facility (2); and home care (2). For a holistic understanding of a patient (P5, P7, P14), these settings typically have different sizes of interdisciplinary teams ranging from 2 to 24 team members (e.g. doctors, nurses, occupational therapists, physiotherapists, therapist assistants, and speech therapists). In most cases, they work in pairs or with a team of colleagues to *"brainstorm how to assess patient's status"* (P9) and *"understand the needs of a patient to coordinate therapy sessions"* (P8).

When handling a case, participants might get referrals of experienced colleagues (P4) for discussions. Otherwise, they determine which colleagues will be adequate by checking the following aspects through their previous *"interactions with them"* (P8): (i) **experiences and knowledge** (all participants), (ii) **quality of work** (P10), and various **soft skills** (P5, P10, P12,P14,P15). Specifically, they will approach more senior and experienced colleagues with hard skills, such as a corresponding specialty (P1,P5,P7,P8,P10,P14,P15,P16) or experience with similar cases (P3,P6) for *"receiving task-oriented advice"* (P16). In addition, another hard skill that they can check is whether colleagues provide client-centered

quality work or not (P10,P11,P13). Lastly, they also highly consider soft skills of their colleagues, such as whether their colleagues are *"more approachable"* (P10) and comfortable to share and discuss (P14,P16) to determine appropriate peers or colleagues for discussion.
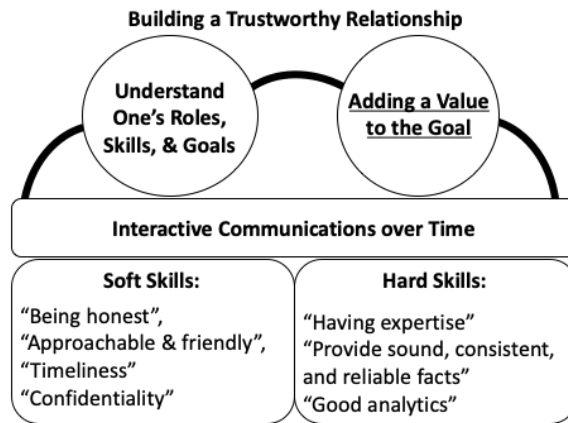

Fig. 4. Characteristics and skills of a trustworthy colleague and a process to build a trustworthy relationship with colleagues: a trustworthy relationship requires soft and hard skills to have interactive communications over time to understand colleagues' roles, skills, and goals and add a value on their goals.

4.1.2 Process to Build a Trustworthy Relationship. Participants commented that building a trustworthy relationship (Figure 4) requires soft and hard skills to have **interactive communications over time**, in which one **understands colleagues' roles, skills, and goals**, but also checks whether a colleague **adds a value on common goals** (P7,P9,P11,P12). For having a common goal, participants mentioned the importance of interacting with each other to know educational backgrounds and strategies (P4,P9) and *"understand each other's needs and roles"* (P12). As they *"learn from each other"* (P3) and *"show that they can help and do the work well"* (P5) over multiple interactions or sessions (P1,P2,P6), they build *"a deeper trustworthy relationship with colleagues"* (P1).

4.1.3 Skills & Characteristics of Trustworthy Colleagues. Participants shared several skills and characteristics of trustworthy colleagues for interactive communication. First, soft skills, such as being honest, approachable & friendly (P1,P7,P5,P10), and timely (P8,P15) are important to have iterative communications and identify common goals for building a trustworthy relationship. Given common goals, trustworthy colleagues also require hard skills to demonstrate one's expertise through *"provid[ing] sound, consistent, reliable facts"* (P7). In addition, having good reputation with others (P5,P15) and keeping the confidentiality of information (P12) are important to build a relationship over time.

4.2 Understanding and Information Needs of AI

Overall, participants understood the high-level ideas of an AI system for rehabilitation after presenting the tutorial materials. However, participants *"are not clear about how the statistical methods work in detail"* (P12).

When participants were asked to review inputs of an AI-based system and guess its outputs, participants had diverse ranges of correct guesses: among 16 participants, three participants (P1,P10,P13) correctly guessed all three outputs, nine participants (P3,P4,P5,P6,P7,P9,P11,P14,P15) correctly guessed two out of three outputs, two participants (P2,P12)

correctly guessed one out of three outputs, and two participants (P8,P16) incorrectly guessed all outputs. When we analyzed the correlation between participants' normalized scores on technology experiences and their normalized scores on guessing AI outputs by computing Pearson's correlation coefficient (r). We found that the coefficient value (r) of -0.31 and the p-value of 0.23, which indicates **a weak correlation between normalized scores on technology experiences and guessing AI outputs**.

While going through the tutorial materials, participants asked **questions about functional, operational, development, and evaluation aspects of AI**.

4.2.1 Functional & Operational Aspects. For functional aspects, participants (P1,P2,P4,P5,P6,P9) questioned how AI processes data and they could interact with AI. Specifically, participants wondered if AI can automatically identify incorrect data without therapists' input (P1,P5). In addition, participants questioned how AI will operate if input data is slightly off from normal ones (P1,P3,P9,P10). Participants also asked *"whether AI can be evolved with new data"* (P7) and *"inputs from therapists"* (P10) to have *"an adaptive goal/normality"* (P10).

In addition, participants inquired about the easiness of setting a device/system (P1,P4), *"how fast the system can process data to provide an assessment"* (P5), and *"how it might be repaired and maintained"*(P7).

4.2.2 Development & Evaluation Aspects. Participants questioned about how to interpret a confusion matrix (P11,P16), *"any benchmark to determine whether AI is trustworthy or not"* (P7), *"whether AI can become more accurate as it has a larger sample size"* (P15), and an evaluation setting (e.g. *"whether AI is being piloted in a hospital"* - P14).

When it comes to the dataset, participants suggested additional factors that can be considered to expand the dataset. Overall, participants considered that the usage of the clinically validated functional assessment scores (e.g. fugl meyer assessment [24]) is good to characterize and recruit post-stroke survivors for data collection. In addition to this functional assessment score, participants mentioned about distinguishing post-stroke survivors by other clinically relevant factors: the stage and severity of stroke (P2,P4,P5,P7,P10,P14), spasticity and muscle tone (P3,P7,P8), *"the status of finer motor functions"* (P10), *"cognitive status"* (P9) that a lot of post-stroke survivors struggle with.

In addition, participants suggested expanding the dataset to make a more balanced distribution of sex (P2,P4,P6,P11,P12,P12), age (P2,P5,P8,P9,P11,P13), race (P4) as different sex, ages, and races might have *"different characteristics and factors that affect their recovery"* (P2) and consider how to address *"possible bias on labels"* (P1).

4.3 Context-specific Required AI Performance & Evaluations

Most participants had **difficulty with enumerating how much percentage is sufficient** as they are not familiar with this metric (P4,P5,P10,P15) and *"wondered if there is any industry standard to determine a good score"* (P3). *"When AI has a lower performance, we[they] won't likely to use it"* (P5). Thus, participants still described that it would be good to achieve high accuracy (P3,P15) ranging from 80% (P2,P8,P12), 85% (P3,P9), to 90% (P6,P7,P10,P13) even if they *"do not know whether it is even possible"* (P10).

Participants mentioned the **importance of having context-specific required performance and a way to interpret the meaning of numbers**. P11 suggested a context-specific, desirable performance of AI: *"when the AI is used as a reference, 70 - 80% might be enough as therapists would make final assessment"* and *"when the AI is being used independently, I expect a much higher score at least 95%"*.

Participants **recommended (i) presenting a benchmark performance and (ii) having contextualized and iterative evaluations**. *"Even if we[they] are given this numerical performance value, we[they] have difficulty to interpret an actual meaning about what it implies"* (P10). Participants considered presenting a benchmark performance score (e.g.

therapists' agreement) would be useful to *"build an initial guide for comparison"* (P15) and check whether AI is reliable or not (P8,P14). Also, they suggested contextualizing evaluations by reporting how well AI will perform on the following aspects: common symptoms (P5,P6) (e.g. *"assessing the range of motion"* - P1) and difficult tasks (P4,P7,P8,P9,P10,P13) (e.g. muscle tone, pelvis, scapular shoulder compensation, spasticity, finer finger motions, knee replacement, and gait patterns), *"borderline cases"* (P3), and uncontrolled situations and data (P3,P5), such as presentation of other people than a patient and different setups of a system and a camera.

In addition, participants mentioned that *"having trials to see how AI works would be necessary"* (P10) and *"investigate how to improve it"* (P16) instead of presenting a number once as they *"need time to build trust with AI as we build trust with our colleagues"* (P10).

4.4   Understanding and Information Needs of AI Explanations

Overall, participants **desired descriptions on how they can leverage AI explanations**. Also, a few participants **inquired underlying processes of identifying AI explanations** and technical terms in visualizations.

After asking a few clarification questions, participants understood the high-level ideas of three AI explanations (i.e. feature importance, counterfactual, prototype/example-based explanations). For a feature importance explanation, most participants followed the high-level concept in the first place. However, when some participants (P2,P3,P10) first encountered a local bar plot of SHAP values for each feature [39] (Figure 3b), they asked how to read and interpret the graph (i.e. the meaning of blue and red plots, presented values).

For a counterfactual explanation, some participants can follow the concept of a counterfactual explanation by correlating their similar practice (P2,P4), *"'familiarization', in which we determine why a client cannot get a correct position"* (P4). However, even if some participants understood the high-level ideas of a counterfactual explanation, they were confused about how these explanations can be used to validate an AI output (P3,P10,P14,P15).

For an example-based explanation, most participants considered that it is *"easier to understand and review"* (P1). Some participants asked questions on the procedures of generating an example-based explanation: *"how similar items are identified"* (P2) and *"if the clothing color of a patient affects which similar cases will be identified"* (P2); *"whether an AI model generates a similar sample"* (P9); how many samples are needed to define a prototype or find relevant cases (P2,P8). In addition, after reviewing the visualization of embedding spaces of samples (Figure 3d), P3 asked about the meaning of the axes to project samples.

4.5   Usefulness Ranking of AI Explanations for Onboarding and Decision Support

Figure 5 summarizes the overall ratio of rankings on three AI explanations for onboarding and decision support using data from all participants, therapists with experience in stroke rehabilitation, and other participants.

For onboarding, both therapists and other health professionals and students considered an example-based explanation as the most useful. Therapists considered a counterfactual explanation (35.0%) and a feature importance explanation (28.3%) as the second and the third most useful. Other health professionals and students ranked a feature importance explanation (34.2%) and a counterfactual explanation (26.3%) as the second and the third most useful.

For decision support, therapists considered both an example-based explanation and a feature importance explanation as equally useful (33.9%) the most and a counterfactual explanation (32.2%) as the third most useful. **Other health professionals and students ranked an example-based explanation as the most useful** (44.4%) and both a feature importance explanation and a counterfactual explanation as the second most useful ones (27.8%).
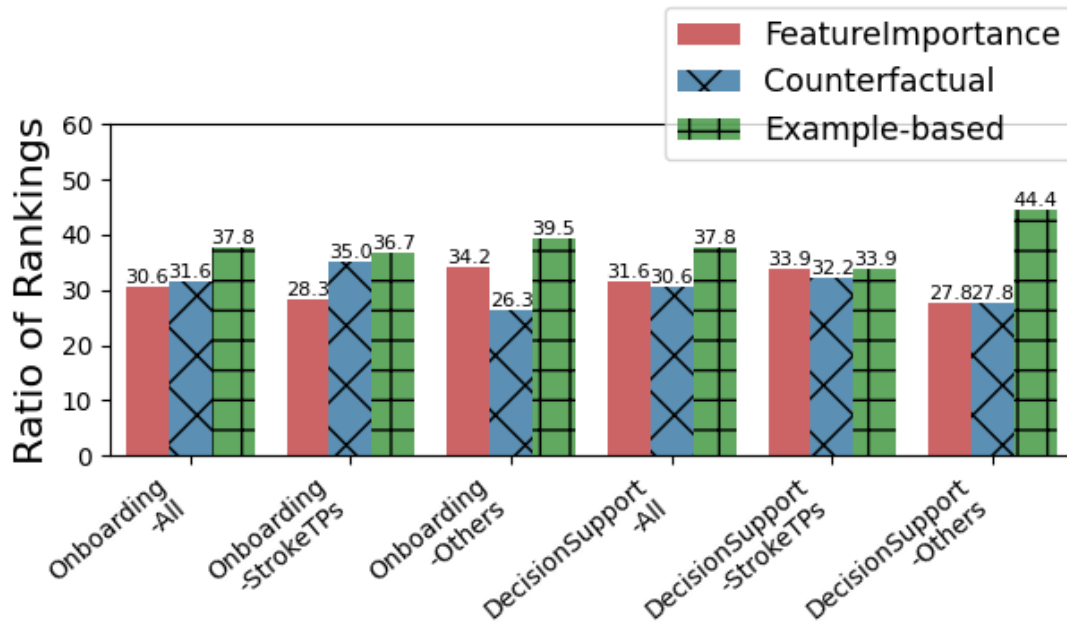
Fig. 5. Ratio of Rankings on the Perceived Usefulness of AI Explanations (Feature Importance, Counterfactuals, and Example-based) for Onboarding and Decision Support with AI from all participants, therapists with experience of stroke rehabilitation, and others (other health professionals and students majoring in medicine and health).

When participants ranked the usefulness of AI explanations, they considered whether *"it provides clinically relevant and useful information"* (P9) and whether it is *"easy to understand and interact"* (P10).

Among 16 participants, eight participants (P1,P4,P7,P9,P11,P12,P14,P15) had the same rankings on three AI explanations for onboarding and decision support. They considered that *"both onboarding and decision support processes require the same process of reviewing and validating AI outputs"* (P7), in which they check how well *"AI provides clinically relevant information"* (P9). Thus, they mentioned that *"the usage of AI explanations would be similar"* (P11) for both onboarding and decision support.

In contrast, eight participants (P2,P3,P5,P6,P8,P10,P13,P16) had different rankings on three AI explanations for onboarding and decision support. They differentiated the processes of validating AI outputs for onboarding and decision support phases. For instance, they described that they want to *"validate how well AI outputs are useful to support a patient-specific assessment"* (P8) for decision support while they aim to *"briefly validate the correctness of AI outputs to develop a trust with AI"* (P10) for onboarding.

4.5.1 Feature Importance Explanation. For a feature importance explanation, participants considered that it is useful to review only important features as *"reviewing all features can be time-consuming"* (P7). However, participants also described the limitation of a feature importance explanation that it might identify features that are not important (P3,P4,P14) and less correlated with an outcome (P2,P16), which *"might not be applicable and useful for the practice"* (P9).

Some participants mentioned that *"reviewing which features AI identifies as important"* (P6) is useful to understand how the AI works (P12,P13,P16) and *"check the strength and limitations of AI"* (P15) for onboarding. In contrast, some participants elaborated that a feature importance explanation is more useful to validate an individual assessment for decision support than onboarding (P5,P6,P16).

4.5.2 Counterfactual Explanation. For a counterfactual explanation, participants described that it is useful to review how to update patient's movement features to flip an AI output (P7,P11), which is *"what we want to achieve for our patient (e.g. how to make a patient with compensation to not have compensation)"* (P9). However, participants found it is difficult to understand (P1,P10) and time-consuming to get used to and validate whether counterfactual explanations make sense or not (P1,P5,P12,P16).

Some participants mentioned that reviewing counterfactual explanations *"how features need to be changed to flip an AI output"* (P3) is helpful to understand how AI defines the medical conditions (e.g. compensation) (P6,P13) and understand the accuracy and performance of an AI model (P3,P8) for onboarding. Other participants described that counterfactual explanations *"bring richer insights on both normal and abnormal conditions than other explanations that provide information on a single condition"* (P2) and are *"useful to validate an AI output"* (P2) and *"applicable in practice"* (P3) for decision support.

4.5.3 Prototype/Example-based Explanation. For an example-based explanation, participants valued that it is useful to review relevant samples to a client/patient's condition instead of searching a whole dataset (P5,P13,P14) or *"relying on our memory to remember all past cases"* (P9) and easy to understand to validate whether AI is correct or not (P1,P10,P12,P16). However, some participants mentioned that as individual post-stroke survivors are very different and specific, reviewing relevant examples might not be very useful (P7,P8,P16). They also wondered whether AI might have enough samples to provide relevant samples (P3,P8).

Participants described the usefulness of an example-based explanation for onboarding as it shows how an AI model defines a clinical concept/symptom (P5,P6), and it is *"easier to interpret than others"* (P6). In addition, reviewing a pool of samples is useful to draw relevant conclusions (P2,P3) and *"confirm the validity of an AI output"* (P13) for decision support.

4.6 Suggestions to Improve Onboarding and Decision-Making with AI

For improving the understanding of the strengths and limitations of AI during onboarding and human-AI collaborative decision-making, participants recommended **(i) communicating benchmark information and the benefits of AI and (ii) a trial period to interact with AI for calibrating user trust, refining an objective of AI, and tuning AI with user feedback**.

4.6.1 Communicating Benchmark Information & Benefits of AI. When we asked the participants to define a desired performance value of an AI model, they had difficulty with enumerating what it is desirable performance. Participants considered that *"presenting benchmarkable information"* (P7), such as how much AI matches with therapists' agreement (P2,P3,P5,P6,P14) or characterizing the performance on different medical conditions (P4,P6,P14) is necessary to better estimate a desirable performance for trustworthy AI (P3,P5).

Participants also suggested that describing *"the benefits of using AI"* (P4). As most participants had difficulty with contextualizing what a specific score of an evaluation metric means, they considered that it would be more effective to communicate the benefits that they can easily understand. The example of these benefits of AI include *"how much time*

can be saved" (P5), *"how well they can improve their decision than a therapist alone"* (P5), or whether it can support a better health outcome for patients (P2,P4).

4.6.2 Interaction Trials to Calibrate User Trust, Refine AI Objective, and Tune AI with Feedback. Participants considered that demonstrating the strengths of AI through presenting a numerical value on an evaluation metric is still important. However, as a numerical value might not be sufficient to show the whole picture of AI performance, they also suggested providing a trial period with multiple interactions (P1,P5,P13,P14,P16) similar to how they build a trustworthy relationship with their colleagues over time (Section 4.1.2). By directly interacting with AI and observing how AI performs (P1,P12,P14,P16), participants considered that they can *"check whether AI outputs are similar to what therapists consider"* (P7), understand the strengths and limitations of AI, and determine *"whether AI is really helpful or not"* (P5).

For a trustworthy relationship with colleagues, participants described the necessity of interactive communications over time to align a common goal (Table 4.1.2). Along this line, participants wondered *"if AI can have an adaptive goal to set the notion of a correct movement based on patient's status"* (P10). In addition, participants desired a way to provide feedback to AI and update it accordingly (P5,P7) when any limitations of AI are identified and periodically refine AI with relevant, up-to-date data (P7,P9,P15) for more trustworthy interactions with AI.

4.6.3 Other Considerations: Periodic Audits, Multi-sites Validations, and Easy Setups & Usage. In addition to communicating the competence and benefits of AI and interaction trials, participants described other important factors to consider using AI in practice and make AI more trustworthy. First, participants considered that it is necessary to have periodic, internal & external audits (P2,P14,P16) *"even after passing the onboarding or a trial period"* (P12). These audits refer to whether *"AI can effectively provide relevant information"* (P7), *"therapists correctly use the system"* (P12), and refining AI (Section 4.6.2), but are not limited to these [44, 47, 54].

Participants also considered that having validations with multiple colleagues in multiple sites (P1,P12) would be helpful to consider using AI. As *"each therapist has different educational backgrounds"* (P16) and *"each hospital has different cultures"* (P14), participants wondered how an AI-based system can be deployed and applicable in various settings.

Finally, as *"time-consuming logistics and setups would be a big barrier"* (P8) to consider using an AI-based system, these systems should be easy to set up and use as health professionals do not have much time in a clinical setting (P1,P5,P7,P10).

5 DISCUSSION

In this section, we highlight key takeaways on the usefulness and value of tutorials on AI and AI explanations to improve communicating the strengths and limitations of AI for onboarding and human-AI collaborative decision-making. In addition, we discuss design recommendations and future research for more effective onboarding with AI and AI explanations and human-AI collaborative decision-making: 1) improving tutorials on AI and AI explanations, 2) AI explanations for onboarding and interactive communications with AI, 3) measuring the level of understanding of AI and AI explanations, 4) beyond presenting a numerical, traditional performance metric, 5) goal alignment between users and AI and refining AI, and 6) audits to build a reputation of AI.

5.1 Values & Limitations of Tutorials on AI and AI Explanations

An AI model card [43] aims to provide a useful way to provide the essential facts of AI models in a structured way. However, our results showed that an AI model card is not sufficient to support onboard health professionals with AI. Even before presenting an AI model card, as most of our participants haven't used much AI, they are clueless about AI (e.g.

how it works and helps or what the limitations are) (P1,P5,P6). Without any tutorials, *"it will be difficult for the first-time user without a technical background to use it"* (P4). Thus, they described the importance of having tutorials in simple languages (P6,P16) so that they can *"make use of AI in a short time"* (P1).

Much in the way that participants (i.e. health professionals and students majoring in medicine and health) desired the characteristics of being "friendly" and "approachable" as a trustworthy colleague (Figure 4), our findings echo the needs to contextualize technical terminologies and make them "friendly" and "approachable" to improve the understanding of AI for AI-assisted decision-making [13, 29, 58].

Our participants mentioned that they learned a lot about what's behind AI and AI explanations (P8,P10) through our onboarding tutorial materials. They highlighted the importance of educating and *"introduce[ing] about AI and AI explanations so that we[they] can have a better understanding on these to make the effective usage"* (P10). However, we found that some participants still asked clarification questions on AI or/and AI explanations after introducing our onboarding materials. In addition, even if participants developed understanding of AI and AI explanations, some participants were not clear how they can determine when AI is 'ready' to be used. *"We[Users without technical backgrounds] do not necessarily think about inspecting whether AI is trustworthy or not"* (P2) and they *"do not have any experiences to check the validity of AI"* (P1). Thus, it is necessary to educate and present a way (e.g. AI explanations) for them to effectively onboard with AI but also inspect AI outputs (e.g. *"communicating why AI generates a certain output and how to validate it"* - P3).

We expect participants might need additional interactive tutorials on the development pipeline of an AI model (e.g. how it processes data and metrics) and how to interact with and make use of AI and an AI explanation. Along this line, further research could explore how onboarding materials can be effectively delivered through interactivity, the choice of visual cues, and data visualizations [16].

5.2 Design Recommendations

5.2.1 AI Explanations for Onboarding and Interactive Communications with AI.
AI explanations that aim to describe the behavior of the entire AI model or a specific AI output [1, 18, 31, 68] have been increasingly explored to allow a user to understand when an AI model is right or wrong and can be trusted for AI-assisted decision making [10, 30, 69]. However, these AI explanations have been mainly explored to provide insights on an AI output during the inference phase of an AI model for user's decision support [12, 52, 75]. There have been limited explorations on how AI explanations can be used to support user's onboarding phase before moving to AI-assisted decision making phase.

When we asked the participants to rank the usefulness of three AI explanations for onboarding and decision support phases, some participants differentiated the characteristics of onboarding and decision support tasks. Our results (Section 4.5 and Figure 5) show that the usefulness rankings of AI explanations are different depending on the tasks (e.g. onboarding vs decision support) and also participants had different strategies of using AI explanations on a task. Aligned with previous research that describes the necessity of characterizing tasks and stakeholders [60], our study discusses a research problem of how AI explanations can be designed and used for onboarding phases. In addition, as interactive communications are critical to build a trustworthy relationship with colleagues Figure 4), a further study is required to explore how AI explanations can serve as a tool/medium for health professionals to have interactive communications with AI.

5.2.2 Measuring Understanding of AI and AI Explanations.
Our study also uncovered challenges and needs in measuring the level of understanding of AI and AI explanations for

more effective onboarding with AI/ML systems. When we recruited participants, we asked participants to respond with their experience of recent technologies (i.e. computer/laptop, activity tracker, virtual voice assistant, unmanned convenient store, and autonomous vehicle) based on the questions designed by the Center for Research and Education on Aging and Technology Enhancement (CREATE) [17] that aim to measure and profile older adults' experience with technology [6]. However, one's experiences and higher exposures to recent technologies including AI applications do not necessarily mean that he or she will also have a good understanding of AI and AI explanations. For instance, P8 rated own technology experience as 6.2 out of 7 and P16 had 4.0 out of 7 but both P8 and P16 incorrectly guessed AI/ML outputs (Table 1). In contrast, P10 and P13 had technology experience scores of 3.2 and 3.8 respectively, which is lower than those of P8 and P16. Both P10 and P13 correctly guessed all AI/ML outputs (Table 1). Along this line, we found that normalized scores of technology experiences are not correlated with the normalized scores of guessing AI outputs (Section 4.2).

In addition, even if our study explores the number of correctly guessing AI/ML outputs to measure participants' understanding of AI, this measurement is still limited. For instance, we observed that participants including ones who guessed all AL/ML outputs correctly asked clarification questions on AI and AI explanations. (Section 4.2 and 4.4).

Overall, our study results show that participants' technology experiences or their number of correctly guessing AI/ML outputs are not necessarily linearly correlated with their capabilities to understand how an AI operates nor how they appreciate the AI explanations. Thus, it is important to further explore how we can measure user's understanding of AI and AI explanations in a metric similar to how the previous studies assess the computer proficiency [8] and investigate what level of a metric indicates when a user has sufficient understanding to interact with AI and AI explanation.

5.2.3 Beyond a Numerical Traditional Performance Metric.

Our study results show the limitation of presenting the value of a traditional evaluation metric to either communicate the overall performance of an AI model or take a specific action (e.g. determining whether to deploy an AI-based system or not). Unlike participants' practice which they have interactive communications with colleagues over time to align a common goal and check the abilities of colleagues to add value to the common goal for their trustworthy relationships (Figure 4), one-directional presentation of a numerical performance value does not provide the whole picture and understanding of AI for trustworthy onboarding and usage.

For onboarding with an AI-based system, Cai et al. [13] recommended informing the overall performance of an AI-based system including its particular strengths and limitations. The Google's People + AI Guidebook [50] recommends specifying a threshold value of a performance metric to take a specific action in the 'User Needs + Defining Success' section. Given these recommended guidelines, we hypothesized that participants without technical background could discuss with AI/ML developers to review the reference performance of therapists' agreement level and specify a threshold value of a performance metric to determine whether an AI-based decision support system can be considered being used in the practice. However, even if we presented the performance of an AI model to assess three common performance components of rehabilitation assessment, our participants described the difficulty with understanding the overall performance of an AI model by reviewing a numerical value on a performance metric. *"If we are told that an AI model has 90% accuracy, it might give a wrong mental model on AI performance as its performance might be changed in new cases"* (P11). Thus, participants desire a better way to *"understand how well AI could perform on new data"* (P7), such as describing the benefits of using AI. However, the benefits of using AI cannot be communicated with a traditional evaluation metric. Thus, it is worthwhile to explore how to quantify the benefits of an AI-based system (e.g. clinical utilities) and describe its values to the end users in a more understandable way.

5.2.4   Goal Alignment between Users and AI and Refining AI.

In addition, our study results suggest a gap between the objective of AI and the goal of a user. Although health professionals typically have a goal of improving patient's status, AI systems are trained to maximize the probability of replicating a therapist's assessment. For a trustworthy relationship with colleagues, participants described the necessity of interactive communications over time to align a common goal and add a value to it (Figure 4). Along this line, participants wondered *"if AI can have an adaptive goal to set the notion of a correct movement based on patient's status"* (P10). In addition, participants desired a way to provide feedback to AI and update it accordingly (P5,P7). When any limitations of AI are identified, they desire to periodically refine AI with relevant, up-to-date data (P7,P9,P15) for more trustworthy interactions with AI. For enabling trustworthy, human-AI collaborative decision-making, it would be critical to explore how to align an AI's goal with a user's goal and refine AI with the user's feedback [35].

5.2.5   Audits to Build a Reputation of AI.

As health professionals considered colleagues' reputation with others is an important factor in building a trustworthy relationship with their colleagues (Figure 4), they also had a similar opinion that AI becomes more trustworthy if they hear more positive testimonials of colleagues from multiple sites (P4,P13,P14) (e.g. *"colleagues can interpret AI outputs and reliably make use of them"* - P13). To this end, our study also highlights the values of audits on AI [54]. As mentioned in Section 4.6.3, these audits can range from simply checking whether AI provides necessary information, refining AI with feedback [35]. Also, the audit process can be monitoring and anticipating the potential negative impact of a system, designing mitigation or informing when to abandon the development and usage of an AI technology [54]. For onboarding with an AI-based system and integrating it in practice, our study also unveiled challenges and needs on how these AI-based systems for high-stake contexts can be audited and how we can support to educate people without technical backgrounds to participate in this process of deploying, onboarding, using, and auditing an AI-based system.

5.3   Limitations

Our study provides insights on information needs for onboarding with AI and AI explanations and discusses several existing gaps and areas to improve for more effective onboarding with AI and trustworthy human-AI collaborative clinical decision-making. However, our study is limited to introducing and presenting the onboarding tutorial materials to participants (i.e. health professionals and students majoring in medicine and health) by an online session. Our study does not observe how health professionals would initially interact with an AI system and AI explanations in a practical setting, which would bring richer and more useful insights on information needs to effectively onboard and interact with AI.

In addition, our study has a limitation of generalizing the results as we mainly explore the research questions in the context of a single clinical decision-making tasks along with brief descriptions of an AI/ML model training and three AI explanations (i.e. feature importance, counterfactual, prototype/example-based). As the primary health professionals from the context of this study (i.e. rehabilitation therapy) are mostly females (e.g. around 62.7% of therapists are female [73]), the participants of this study are mostly females (87.5%). Also, our study does not involve a large number of participants even if such a small sample size is not unusual in similar previous works [10, 13]. A further in-situ study with other decision-making tasks and types of ML models and explanations is necessary for further generalizable insights on improving onboarding with AI and AI explanations.

6 CONCLUSION

In this work, we contributed to an empirical study that explored how AI and AI explanations can be first introduced to health professionals and students majoring in medicine and health and identified information needs on AI and AI explanations for effective onboarding and trustworthy usage of AI. Our study suggested the value of onboarding tutorial materials on AI and AI explanations and the necessity of designing AI explanations for improving onboarding and communications with AI. Also, our study highlighted the importance of exploring metrics to characterize the user's understanding of AI and AI explanations. In addition, our study discussed other considerations for effective onboarding and trustworthy human AI collaborative decision-making moving beyond describing a numerical traditional performance metric: presenting user-understandable benchmark information, interactive trials to communicate the practical benefits of AI, calibrate user trust, refine an objective of AI and AI with user feedback, and AI audits. Future research should explore how various types of AI/ML models and AI explanations on different contexts/tasks can be introduced to people without technical backgrounds.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society* 35 (2020), 611–623.

[4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[5] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[6] Jenay M Beer, Cory-Ann Smarr, Tiffany L Chen, Akanksha Prakash, Tracy L Mitzner, Charles C Kemp, and Wendy A Rogers. 2012. The domesticated robot: design guidelines for assisting older adults to age in place. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 335–342.

[7] Angie Boggust, Brandon Carter, and Arvind Satyanarayan. 2022. Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples. In *27th international conference on intelligent user interfaces*. 746–766.

[8] Walter R Boot, Neil Charness, Sara J Czaja, Joseph Sharit, Wendy A Rogers, Arthur D Fisk, Tracy Mitzner, Chin Chin Lee, and Sankaran Nair. 2015. Computer proficiency questionnaire: assessing low and high computer proficient seniors. *The Gerontologist* 55, 3 (2015), 404–411.

[9] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[10] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[11] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.

[12] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[13] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[14] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[16] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 427–439.

[17] Sara J Czaja, Neil Charness, Arthur D Fisk, Christopher Hertzog, Sankaran N Nair, Wendy A Rogers, and Joseph Sharit. 2006. Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE). *Psychology and aging* 21, 2 (2006), 333.

[18] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[19] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[20] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (2019), 261–262.

[21] Nicola K Gale, Gemma Heath, Elaine Cameron, Sabina Rashid, and Sabi Redwood. 2013. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC medical research methodology* 13, 1 (2013), 1–8.

[22] Felix Gille, Anna Jobin, and Marcello Ienca. 2020. What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine* 1 (2020), 100001.

[23] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[24] David J Gladstone, Cynthia J Danells, and Sandra E Black. 2002. The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. *Neurorehabilitation and neural repair* 16, 3 (2002), 232–240.

[25] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.

[26] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[27] Saif Khairat, David Marc, William Crosby, Ali Al Sanousi, et al. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e8912.

[28] Ajay Kohli and Saurabh Jha. 2018. Why CAD failed in mammography. *Journal of the American College of Radiology* 15, 3 (2018), 535–537.

[29] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[30] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[31] Himabindu Lakkaraju, Julius Adebayo, and Sameer Singh. 2020. Explaining machine learning predictions: State-of-the-art, challenges, and opportunities. *NeurIPS Tutorial* (2020).

[32] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–22.

[33] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2019. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 218–228.

[34] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[35] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[36] Xin Li, Traci J Hess, and Joseph S Valacich. 2008. Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems* 17, 1 (2008), 39–71.

[37] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 8 (2022), 669–677.

[38] Alex John London. 2019. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report* 49, 1 (2019), 15–21.

[39] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[40] Thomas M Maddox, John S Rumsfeld, and Philip RO Payne. 2019. Questions for artificial intelligence in health care. *Jama* 321, 1 (2019), 31–32.

[41] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[42] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[43] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[44] Jakob Mökander and Luciano Floridi. 2021. Ethics-based auditing to develop trustworthy AI. *Minds and Machines* 31, 2 (2021), 323–327.

[45] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[46] T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. 2019. Efficient saliency maps for explainable AI. *arXiv preprint arXiv:1911.11293* (2019).

[47] Ivy Munoko, Helen L Brown-Liburd, and Miklos Vasarhelyi. 2020. The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics* 167 (2020), 209–234.

[48] Ju Gang Nam, Sunggyun Park, Eui Jin Hwang, Jong Hyuk Lee, Kwang-Nam Jin, Kun Young Lim, Thienkai Huy Vu, Jae Ho Sohn, Sangheum Hwang, Jin Mo Goo, et al. 2019. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 290, 1 (2019), 218–228.

[49] J Arturo Olvera-López, J Ariel Carrasco-Ochoa, J Martínez-Trinidad, and Josef Kittler. 2010. A review of instance selection methods. *Artificial Intelligence Review* 34, 2 (2010), 133–143.

[50] Google PAIR. 2019. People + AI Guidebook. https://pair.withgoogle.com/guidebook/

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[52] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[53] Alun Preece. 2018. Asking 'Why' in AI: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* 25, 2 (2018), 63–72.

[54] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[55] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature medicine* 28, 1 (2022), 31–38.

[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[57] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. " The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 99–109.

[58] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I Hong. 2020. Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[59] Ramandeep Singh, Mannudeep K Kalra, Chayanin Nitiwarangkul, John A Patti, Fatemeh Homayounieh, Atul Padole, Pooja Rao, Preetham Putha, Victorine V Muse, Amita Sharma, et al. 2018. Deep learning in chest radiography: detection of findings and presence of change. *PloS one* 13, 10 (2018), e0204155.

[60] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[61] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 3, 1 (2020), 17.

[62] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.

[63] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283.

[64] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[65] Kush R Vashney. 2022. *Trustworthy machine learning*. Independently published.

[66] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. 2020. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[67] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI doctor" in rural clinics: Challenges in AI-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–18.

[68] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[69] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*. 318–328.

[70] Jeannette M Wing. 2021. Trustworthy ai. *Commun. ACM* 64, 10 (2021), 64–71.

[71] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[72] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

[73] Steven Zauderer. 2023. Statistics, Facts & Demographics of Physical Therapy.   https://www.crossrivertherapy.com/research/physical-therapy-statistics#:~:text=67%25%20of%20physical%20therapists%20are,common%20gender%20in%20the%20occupation.

[74] Aleš Završnik. 2020. Criminal justice, artificial intelligence systems, and human rights. In *ERA forum*, Vol. 20. Springer, 567–583.

[75] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

A   IMPLEMENTATIONS OF AN AI MODEL

We followed the previous research [33] to learn an AI model for rehabilitation assessment. Specifically, we processed the estimated joint positions of post-stroke survivors' exercises to extract various kinematic features. The kinematic features of the 'Range of Motion' (ROM) include joint angles, such as elbow flexion, shoulder flexion, and elbow extension, and normalized relative trajectory (i.e. the Euclidean distance between two joints - head and wrist; head and elbow), and the normalized trajectory distance (i.e. the absolute distance between two joints - head and wrist, shoulder and wrist) in the x, y, and z coordinates [33]. The features of the 'Compensation' include the normalized trajectories, which indicate the distances between joint positions of the head, spine, and shoulder in the x, y, and z coordinates from the initial to the current frame over the entire exercise motion [33].

As previous research demonstrated the outperformance of a feed-forward Neural Network (NN) model to classify the quality of post-stroke survivors' motion [33], we utilized the extracted kinematic features and labels of post-stroke survivors' exercises to implement a feed-forward NN model using Pytorch libraries [51]. For the labels, we utilized the labels by the expert therapist, who conducted the clinically validated assessment test. We grid-searched various architectures (i.e. one to three layers with 32, 64, 128, 256, and 512 hidden units) and different learning rates (i.e. 0.0001, 0.0005, 0.0001, 0.005, 0.001) while training a feed-forward NN model with cross-entropy loss and the mini-batch size of 1 and epoch of 4. For training and evaluating the model, we utilized the leave-one-subject-out cross-validation, where we trained the model with data from all post-stroke survivors except one post-stroke survivor and tested the model with data from the held-out post-stroke survivor. The final model architectures and learning rates are three layers with 256 hidden units and 0.005 of the learning rate for the ROM, one layer with 16 hidden units and 0.0001 of the learning rate for the Smoothness, and three layers with 64 hidden units and 0.005 of the learning rate for the Compensation. The models achieved 82% F1-score, 79% F1-score and 77% F1-score to replicate therapists' assessment on 'ROM', 'Smoothness', and 'Compensation' components respectively.

B   DETAILS OF THE STUDY: PARTICIPANTS, INTERVIEWS, AND DATA ANALYSIS

Table 2. Detailed Demographics of Participants: Therapists who have experience in stroke rehabilitation (P1 - P10) and other health professionals (P11 - P12) and students majoring in medicine or health (e.g. therapy, nursing) (P13 - P16).

PID	Sex	Age	Occupation	Setting	# of yrs	Q. Tech Experience	Q. ML Outputs
P1	Female	25 - 34 years	PhysioTherapist (PT)	Outpatient Clinic	7	4.8 out of 7	3 out of 3
P2	Male	25 - 34 years	PhysioTherapist (PT)	Inpatient Rehabilitation	2	5.2 out of 7	1 out of 3
P3	Male	25 - 34 years	PhysioTherapist (PT)	Home Care	8	4.4 out of 7	2 out of 3
P4	Female	35 - 44 years	PhysioTherapist (PT)	Outpatient Clinic	11	5.8 out of 7	2 out of 3
P5	Female	25 - 34 years	PhysioTherapist (PT)	Inpatient Rehabilitation	9	5.4 out of 7	2 out of 3
P6	Female	45 - 54 years	PhysioTherapist (PT)	Skilled Nursing Facility	30	5.8 out of 7	2 out of 3
P7	Female	35 - 44 years	Occupational Therapist (OT)	Outpatient Clinic	14	5.4 out of 7	2 out of 3
P8	Female	35 - 44 years	Occupational Therapist (OT)	Homecare	11	6.2 out of 7	3 out of 3
P9	Female	25 - 34 years	Occupational Therapist (OT)	Skilled Nursing Facility	6	4.4 out of 7	2 out of 3
P10	Female	25 - 34 years	Occupational Therapist (OT)	Inpatient Rehabilitation	5	3.2 out of 7	3 out of 3
P11	Female	25 - 34 years	Speech Therapist	Community outpatient	5	4.8 out of 7	2 out of 3
P12	Female	25 - 34 years	Medical Social Worker	n/a	5	4.0 out of 7	1 out of 3
P13	Female	25 - 34 years	Student in Occupational Therapy	n/a	n/a	3.8 out of 7	3 out of 3
P14	Female	25 - 34 years	Student in Speech Therapy	n/a	n/a	3.8 out of 7	2 out of 3
P15	Female	18 - 24 years	Student in Medicine	n/a	n/a	3.8 out of 7	2 out of 3
P16	Female	18 - 24 years	Student in Nursing	n/a	n/a	4.0 out of 7	0 out of 3

Table 3. List of questions for a semi-strucutured interview

Parts of a Semi-Structured Interview	Prompt Questions
Rehabilitation practices & trustworthy relationships	How many colleagues you have & how frequently you interact with them for rehabilitation assessment? How do you build a trustworthy relationship with your colleagues? When you have an uncertain case, how do you know a particular colleague will be good for discussion? What aspects of your colleagues make them trustworthy?
Intro to AI	Any questions/information needs about how the system operates and the development and evaluation pipeline of an AI system We tried to collect data from post-stroke survivors with diverse FMA scores. Do you have any particular post-stroke survivors that should be included in the dataset? AI system cannot be perfect. Do you have a particular performance that is required for your trustful usage? At least XX F1-score an AI model needs to achieve to consider using AI and your trustful usage if being used? Do you have any specific conditions (e.g. Full ROM, shoulder/trunk compensation) or edge cases that AI should be evaluated and good at for consider using AI and your trustful usage if being used?
AI Explanations	Any questions or information needs about AI explanations (e.g. Feature Importance, Counterfactual, Example-based)? Rank which AI explanations are the most useful to onboard with AI and understand overall performance & strengths/limitations of AI? Why do you consider that a particular explanation is useful or not to onboard with AI and understand overall performance and strengths/limitations of AI? Rank which AI explanations are the most useful to make a decision with AI and determine whether to trust an AI outcome or not? Why do you consider that a particular explanation is useful or not to make a decision with AI and determine whether to trust an AI prediction or not?
Wrap-up	Any comments/suggestions on how we can address the following issues: >how to determine whether AI has sufficient performance >how to improve onboarding with AI and determine strengths/limitations of AI >how to improve decision-making with AI and inspect whether AI outputs are trustful or not

Table 4. The five high-level themes, twenty one second-level themes, and eighty five third-level themes of qualitative data analysis

High-level Themes	Second-level Themes	Third-level Themes
Practices of Rehabilitation	Practices on Therapy	Teams, Assessment & Therapy, Meetings, Communications
	Practices on Uncertain Decision-Making	Meetings, Other Info, Factors to identify colleagues (Referral, Expertises, Soft Skills)
	Process to Build a Trustworthy Relationship	Period of time, Understand role & needs, Share common goals, Add values, Honest, Friendly, Listen & Interact
	Characteristics of Trustworthy Colleagues	Personalities, Knowledge & Background, Consistent & Reliable, Work Ethics & Timeliness, Confidentiality
Intro to AI	Guesses on AI outputs	-
	Comments on the dataset of AI	Recruitment criterion - Demographics of participants (age, sex, race, etc.) - Stroke conditions (impact, spasticity, tone, finer motions, etc.) Number of samples Labels
	Required performance of AI	Context-specific thresholds, specific numbers, Unclear meaning of numbers, Reference
	Evaluation and edge cases of AI	Common symptoms, Difficult tasks, Borderline cases, Realistic, uncontrolled settings
	Questions about AI	Definitions, Operation speed, Meaning of a confusion matrix and a confidence score, How to perform and fail on a certain case
	Strengths of AI	Process data quickly, Reduce workload, Objective data
	Limitations of AI	Troublesome and time-consuming, Requires a specific setup, Can't show me exact symptoms
	Suggestions on AI	Not familiar with how to onboard & make a decision with AI Need to be usable Periodic audits, Studies from multiple sites Other rehabilitation-specific features Others
AI Explanations	Comments/questions on AI explanations Participants with the same/different rankings on onboarding & decision-support Ranking of AI explanations for onboarding Ranking of AI explanations for decision-support	Clarifications on the concept
	Example-based explanations	Questions, Comments (General, Pros, Cons), Onboarding, Decision-support
	Feature importance explanations	Questions, Comments (General, Pros, Cons), Onboarding, Decision-support
	Counterfactual explanations	Questions, Comments (General, Pros, Cons), Onboarding, Decision-support
Onboarding with AI	Suggestions on onboarding	Values of tutorials, Beyond numbers, Benchmark references, Demonstrate benefits, Interactions, Referrals and testimonials, Other factors
AI-assisted decision-support	Suggestions on decision-support	Trial periods, Update AI with feedback, Other factors (e.g. setups, user-friendly)