Journal of Medical Systems (2021) 45:64
https://doi.org/10.1007/s10916-021-01743-6

PATIENT FACING SYSTEMS

Effect of AI Explanations on Human Perceptions of Patient-Facing AI-Powered Healthcare Systems

Zhan Zhang[1] · Yegin Genc[1] · Dakuo Wang[2] · Mehmet Eren Ahsen[3] · Xiangmin Fan[4]

Received: 15 February 2021 / Accepted: 28 April 2021

Abstract

Ongoing research efforts have been examining how to utilize artificial intelligence technology to help healthcare consumers make sense of their clinical data, such as diagnostic radiology reports. How to promote the acceptance of such novel technology is a heated research topic. Recent studies highlight the importance of providing local explanations about AI prediction and model performance to help users determine whether to trust AI's predictions. Despite some efforts, limited empirical research has been conducted to quantitatively measure how AI explanations impact healthcare consumers' perceptions of using patient-facing, AI-powered healthcare systems. The aim of this study is to evaluate the effects of different AI explanations on people's perceptions of AI-powered healthcare system. In this work, we designed and deployed a large-scale experiment (N = 3,423) on Amazon Mechanical Turk (MTurk) to evaluate the effects of AI explanations on people's perceptions in the context of comprehending radiology reports. We created four groups based on two factors—the extent of explanations for the prediction (High vs. Low Transparency) and the model performance (Good vs. Weak AI Model)—and randomly assigned participants to one of the four conditions. Participants were instructed to classify a radiology report as describing a normal or abnormal finding, followed by completing a post-study survey to indicate their perceptions of the AI tool. We found that revealing model performance information can promote people's trust and perceived usefulness of system outputs, while providing local explanations for the rationale of a prediction can promote understandability but not necessarily trust. We also found that when model performance is low, the more information the AI system discloses, the less people would trust the system. Lastly, whether human agrees with AI predictions or not and whether the AI prediction is correct or not could also influence the effect of AI explanations. We conclude this paper by discussing implications for designing AI systems for healthcare consumers to interpret diagnostic report.

Keywords Artificial intelligence · Radiology report · Diagnostic results · Trust · Decision making · Healthcare

Introduction

Text-based diagnostic radiology reports are an important type of medical data, which are written by radiologists to describe their interpretation of the findings of imaging studies (e.g., computerized tomography (CT) scan or magnetic resonance imaging (MRI)). Typically, the reports are only shared with and used by the referring physician to augment their diagnosis. However, growing evidence has shown that patients are increasingly interested in getting access to this type of data to empower their self-management and participation in the decision-making process [1–3]. As such, healthcare organizations have started making the radiology reporting data available to patients through patient-facing technologies, such as online patient portals [4–7]. However, the literature points out that these radiology reports are too technical as they overly use medical jargons, making them not comprehensible to lay patients [3, 8, 9]. Despite many efforts [9–11], patients still have difficulty making sense of and acting upon their radiology data [12–14].

This article belongs to the Topical Collection: *Patient Facing Systems*

✉ Zhan Zhang
zzhang@pace.edu

[1] School of Computer Science and Information Systems, Pace University, New York, USA

[2] IBM Research, Cambridge, USA

[3] College of Business, University of Illinois At Urbana-Champaign, Champaign, USA

[4] The Institute of Software, Chinese Academy of Sciences, Beijing, China

Published online: 04 May 2021

In recent years, researchers have begun exploring the potential of using artificial intelligence (AI) technology to help patients interpret their clinical data and make informed decisions [15–18]. This novel technology is expected to help patients identify meaningful information out of their radiology report [19, 20]. Despite the high potential in enhancing patient-centered care, current AI-based healthcare systems are mostly working as a "black box"—the inner workings of the system are not visible to the user, making it hard for the patients to determine whether they can trust the suggestions and use that to make decisions [21]. To promote trust and acceptance of AI-based decision support systems, the literature highlighted the importance of explaining AI's predictions to help users understand the recommendations made by the system [21–26]. Prior work has investigated what information or explanations users would like to see [27–30]. For example, research has recommended providing *local explanations* for the rationale of a prediction (e.g., how each attribute of the case contributes to the model's prediction) to help people determine whether to trust AI's predictions [31, 32]. Also, the information about *model performance* may help users determine whether they should base their decisions on the system's predictions [33–35].

While these explanations are demanded, we know little about how these explanations are perceived by healthcare consumers and how they impact healthcare consumers' perceptions in AI-assisted decisions. That is, the effect of AI explanations has not been empirically tested in the context of patient-facing, AI-driven health systems. As healthcare is a high-stake domain, without rigorous, empirical and thorough evaluation, designers and developers of health AI systems are left with little guidance regarding how to present AI predictions in a format that is meaningful, understandable, and trustworthy to the patient [27, 36]. It is, therefore, vital to investigate what explanatory information has the greatest effect on establishing trust and promoting system acceptance [37].

In this paper, we report an experimental study during which we evaluated the effect of different explanations on healthcare consumers' perceptions of AI technology in the context of comprehending radiology reports. The first two research questions (RQs) included: RQ1) What is the effect of providing local explanation on healthcare consumers' perceptions of health AI systems? and RQ2) What is the effect of model performance information on the perceptions? Moreover, in theory people may rely on the AI in cases where it is correct or has high performance and use their own judgement in cases where it is inaccurate or has low performance. It is highly possible that there may exist misalignment between human judgement and AI recommendations that are in the form of predictions, which could result in trust issues. To that end, the last two research inquiries were RQ3) Whether patients' agreement with AI predictions could influence the effect of AI explanations on patients' perceptions of the system? and RQ4) Whether the correctness of AI prediction could also make an impact on patients' perceptions in AI-assisted decision making? This is part of a large research effort to design and develop AI-driven informatics tools to support patients making sense of their diagnostic reports. This study makes the following contributions: First, we present a large-scale experimental design for evaluating healthcare consumers' perceptions of AI systems, which can be replicated by other researchers. Second, we describe the effect of different explanations on healthcare consumers' perceptions of patient-facing, AI-driven health systems. Third, we provide suggestions for designing patient-centered AI-driven health systems.

Related Work

AI in Healthcare

Over the past decade, artificial intelligence technology has rapidly grown in popularity. Many modern AI systems are designed to utilize and learn from large datasets while imitating human cognitive functions to process input information to generate relevant outcomes for decision support. One field that has greatly benefited from the rise of AI is healthcare. Advanced AI techniques (e.g., deep neural networks and knowledge graph) have been used in medicine for various applications, such as cancer detection and diagnosis [38, 39], genetic diagnosis [40], and imaging study interpretation [41, 42]. In these cases, AI techniques serve as a "second set of eyes" to help clinicians diagnose or inspect a clinical case, with the goal of reducing diagnostic errors [30, 43].

However, prior work primarily focused on developing AI systems for clinician use, with little attention paid to how to design patient-facing AI systems to enhance patients' participation in the shared decision-making process [44]. A notable exception is healthcare chatbots, which allow patients to seek medical information and triage their conditions in a timely and cost-effective manner [45, 46]. However, they usually do not have the capability to interpret clinical data, such as radiology imaging reports. Our study contributes to bridging this knowledge gap by investigating how to design patient-facing, AI-powered systems to help patients interpret and act upon their radiology reports.

User Perception of AI-Driven Healthcare Systems

Beyond the technology development, user perception is critical to study as it influences acceptance and adoption of the technology. User perception of technology consists of many aspects. In established fields such as information systems, researchers have been using the Technology Acceptance Model (TAM) [47] or Unified Theory of Acceptance and Use of Technology Model (UTAUT) [48] to explain variables that

influence user perceptions. In contrast, as AI-driven technology is a rapidly evolving domain, a robust and well-accepted user perception model of AI systems has not been developed. However, it is worth noting a few exceptions. For example, Ehsan et al. [49] tested the effects of automated AI explanation on human perceptions from several perspectives, including confidence, human-likeness, adequate justification, and understandability. Another study [50] examined the usefulness and naturalness of AI-generated explanations.

In addition, human trust is critical to adoption of AI systems, especially in high-critical context, such as healthcare [51]. As prior studies have highlighted, the lack of patient and clinician trust in AI is a significant barrier to adoption of healthcare AI systems [52, 53]. In particular, lay individuals tend to have less trust towards AI systems by default compared to trust towards humans [54]. Also, trust in intelligent systems is slower to build up and faster to decrease than trust in humans [55]. The literature argued that one primary barrier of promoting users' trust and acceptance of AI systems is the "black box" problem—decision makers have difficulty understanding how AI systems produce certain outputs [56]. To foster trust in AI-assisted tools, researchers have suggested rendering them more transparent [22, 25, 26].

While a substantial body of literature is dedicated to evaluating the effects of AI explanation, these previous studies have focused on low-risk contexts in such settings as recommender system [57] and e-commerce system [25]. Little attention has been paid to the effect of AI explanations in domains with greater risk, such as healthcare. In addition, research has largely focused on developing or improving models to make AI systems more explainable, but there are few studies examining what explanations can improve users' acceptance and trust of AI systems. For example, *local explanations* that explain the rationale for a single prediction (in contrast to global explanations which describe information related to the overall logic of the AI model) are recommended to help users determine whether to trust AI predictions on a case-by-case basis [31, 32]. In addition, prior work has suggested presenting model performance information, i.e., the confidence score of each single prediction (which reflects the chances that the AI is correct), to help users determine whether they should rely on the AI [33, 34]. Despite some efforts [35, 58], we still know little about how local explanations and model performance are perceived by lay patients, or how they impact patients' perception in the context of AI-assisted radiology data interpretation.

Methods

To evaluate the effect of AI explanations on healthcare consumers' perceptions of using AI-powered healthcare systems to comprehend radiology reports, we designed and conducted a large-scale online experiment. Details of the experiment are described below.

Dataset

In this experiment, we used a dataset created in a previous study during which participants were asked to annotate sentences retrieved from de-identified radiologist reports as describing *normal* or *abnormal* findings regarding a specified anatomical structure of interest [59]. For example, "the stapes is thickened" describes an abnormal finding, while "there is no evidence of bony erosion of the ossicles or the scutum" dictates a normal finding. The purpose of the original study was to investigate the feasibility and effectiveness of using crowd workers to label and curate biomedical datasets for training machine learning algorithms. The annotation dataset generated through this previous work consisted of 276 sentences that were annotated by the subject matter experts (e.g., radiologists), and 727 sentences that were annotated by crowd workers. We used the entire dataset in the present study.

Experimental Design

Since our study aims to examine the effect of local explanations and model performance on people' perceptions of AI-based healthcare systems, we created four groups based on two factors: the extent of explanations for the prediction (High versus Low Transparency) and the model performance (Weak versus Good Model). To that end, we employed a between-subject experimental design methodology and assigned participants randomly to one of the four conditions. We denoted these four conditions as Weak AI-Low Transparency, Weak AI-High Transparency, Good AI-Low Transparency, and Good AI-High Transparency.

To study the effects of AI model performance, we created two predictive models using Support Vector Machines (SVM) with linear kernels. SVM was chosen over other alternatives (e.g., deep learning, logistic regression) because this method offers a fair balance between performance and interpretability [60]. Prior to model training, we extracted features by creating vectoral representations of the textual data – commonly known as word embeddings. We follow a standard bag-of-word approach, where each textual entry was represented by its word frequency vector. Word frequencies were then transformed with tf-idf statistics to emphasize the important words in our dataset. For model building, we used the crowd-annotated labels reported in [59]. The accuracy of the crowd-annotated data was 93.49%. We then used the 276 expert-annotated labels as the "gold standard" to evaluate the performance of the models. Our reasoning for using crowd labelled data, as opposed to expert-annotated data, for training our models is twofold. First, expert-annotated data was not large enough to

train an AI model and the accuracy of the training data was reasonably high. Second, by using less accurate annotations, we were able to collect more data for the scenario that AI fails to provide correct recommendations during the experiment.

Two AI models were trained using the crowd-annotated labels. More specifically, the Weak AI model was trained using 20% of the available training data (143 rows and labels) while the Good AI model was trained using all the available data (727 rows and labels). The Area Under the Curve (AUC) statistics from Receiver Operating Characteristic (ROC) curves were 0.94 for the Good AI model and 0.78 for the Weak AI model, suggesting both models are reasonably accurate and different from one another.

Transparency conditions (High versus Low Transparency) differ in the extent of explanations provided. The differences can be shown in Fig. 1. In the Low Transparency condition, we provided the AI prediction (normal versus abnormal findings) for the presented radiology report with the associated posterior probability of the prediction (i.e., confidence score). In contrast, in the High Transparency condition, in addition to the confidence score, the SVM features with highest scores within the presented sentences were provided as the reasoning underlying the prediction. For example, as shown in Fig. 1a, for the given report, the high transparency AI system provided local explanation in addition to its prediction and confidence score: "*The suggestion is based on the following word(s): 'normal' and ' clear,' which AI associates with a normal diagnosis*."

Participants

We conducted the experiment on Amazon Mechanical Turk (MTurk), an online crowd-sourcing platform that facilitates worker recruiting and human intelligence task (HIT) completion. We invited MTurk workers who were based in U.S. and had completed at least 1000 previous HITs with at least a 95% approval rate to participate in our task. In total, we recruited 3,432 participants, who were randomly assigned into one of the four conditions. Table 1 summarizes the demographic characteristics of all participants. Each eligible worker earned $0.2 to $0.4 as compensation.

Procedure

On MTurk, we collected annotations in batches, which were deployed in sequence. Each batch contained tasks designed for one condition. For each batch, we populated the HITs by randomly selecting a sentence from the current collection of unlabeled sentences. We embedded two quality assurance mechanisms in each annotation task designed to minimize the impact of varying expertise of crowd workers. First, we required three unique workers to label each sentence. Second, workers could only complete one task to avoid any carry-over effects.

Participants completed the study on a survey-like interface. They first filled out the consent form acknowledging their understanding of the study procedure and purpose. Upon giving consent, participants completed a demographics questionnaire. Then, we provided instructions and examples to help the participants get familiar with the task. Following instructions, participants were asked to complete only one annotation task, that is, classifying the highlighted sentence of a radiology report as describing a normal or abnormal observation of the specified component (Fig. 1). An AI-generated prediction was provided to assist the annotation task; an explanation of the AI prediction was also provided for High-Transparency conditions (Fig. 1). At the end of each task, we asked participants to fill out a post-study survey to indicate their perceptions of the demonstrated AI system, all on a 5-point Likert Scale: 1 denotes strongly disagree and 5 denotes strongly agree. We developed the post-study survey based on prior work [49] and tested it with a small group of people (e.g., researchers and students) to ensure its validity. Some questions and what they measured are listed below:

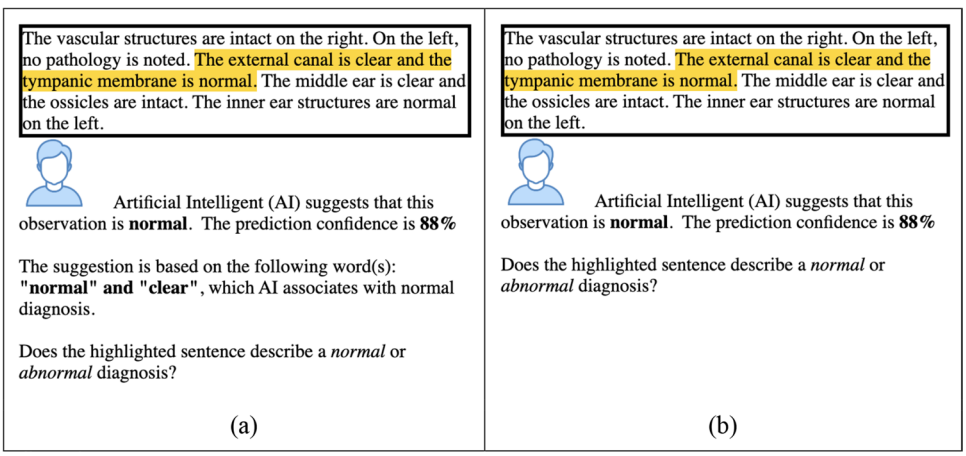

Fig. 1 An illustrative example of the annotation task under (a) with prediction explanations, (**b**) without prediction explanations

Table 1 Summary of participants' demographics

		Conditions			
		Weak Model		Good Model	
		with Explanation (N = 831)	without Explanation (N = 855)	with Explanation (N = 907)	without Explanation (N = 839)
Gender	Female	55% (452)	56% (476)	54% (486)	56% (471)
	Male	45% (375)	44% (376)	46% (416)	44% (365)
	Other	% (4)	% (3)	% (5)	% (3)
Age	18–25 years	16% (131)	17% (148)	21% (186)	19% (157)
	26–49 years	68% (563)	66% (568)	65% (588)	66% (553)
	50–64 years	13% (111)	13% (109)	12% (106)	13% (110)
	65 and older	3% (26)	4% (30)	3% (27)	2% (19)
Occupation Industry	Education	12% (101)	12% (105)	13% (120)	13% (112)
	Finance	8% (63)	8% (69)	10% (86)	10% (87)
	Government	4% (39)	6% (50)	5% (48)	6% (46)
	Health Care	18% (147)	17% (146)	18% (164)	21% (173)
	IT	18% (148)	21% (178)	18% (160)	18% (153)
	Other	40% (333)	36% (307)	36% (329)	32% (268)

Q1. I believe that AI predictions are useful (this is a measure from TAM [47] to evaluate perceived usefulness).
Q2. I believe that AI provides sufficient information to help me understand the report (this is a transformed measure from TAM to evaluate output quality).
Q3. I understand how the AI system generates its predictions (this is a transformed measure from TAM to signal perceived ease of use).
Q4. I trust the AI predictions (this is a measure from TAM).

Data Analysis

We evaluated average user responses to questions above with respect to the availability of local explanations (transparency: High vs. Low), and the performance of the AI model providing the recommendation (performance: Good vs. Weak.). First, we checked for possible interaction effects between the two independent variables: transparency and performance conditions, with a two-way ANOVA analysis for each question. We did not detect any significant interaction effects. Since each testing condition (transparency or performance) yielded two subgroups, we proceeded with comparing the average responses between each subgroup pairs and validated statistically significant differences between them with t-tests. The t-test results were in line with the main effects observed in our ANOVA analyses.

In addition, we split the responses into two groups based on whether a participant's final response was the same as the AI prediction they observed ("AI-Human Agree") or not ("AI-Human Disagree"). We repeated the statistical analysis described above for each group to determine whether the human-AI agreement could affect the observed effect of model performance and local explanation. We followed the same procedure to examine the relationship between the effect of AI explanations and the correctness of AI prediction.

Results

The Effect of Explanations on the Perceived Usefulness of AI Predictions

We found providing local explanation (high transparency) has no effect on the perceived usefulness of AI predictions (Table 2). However, we did observe that the usefulness of AI predictions was rated lower in a high transparency condition when human disagreed with AI predictions ($p = 0.047$) (Table 3).

With the same transparency level, we found that when participants saw a prediction generated by the good AI model, they tended to perceive the AI prediction as significantly more useful (low-transparency as a control variable, $p < 0.001$; high-transparency as a control variable, $p < 0.001$) (Table 4).

Table 2 When controlling model performance, the association between providing local explanations and people's perceptions of different aspects of the AI system

	Conditions					
	Weak Model			Good Model		
	With Explanation (N = 831)	Without Explanation (N = 855)	*P* value	With Explanation (N = 907)	Without Explanation (N = 839)	*P* value
Q1 (usefulness)	3.23	3.28	0.42	3.48	3.48	0.98
Q2 (sufficiency)	3.13	3.20	0.24	3.41	3.38	0.52
Q3 (understandability)	3.40	3.13	< 0.001***	3.61	3.23	< 0.001***
Q4 (trustworthiness)	3.09	3.20	0.04*	3.33	3.37	0.43

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

The Effect of Explanations on the Perceived Sufficiency of AI Information

Providing local explanation had no effect on the perceived sufficiency of AI information (Table 2). But it is worth noting that when people disagreed with AI or the AI prediction was incorrect, providing local explanations actually led to lower ratings, despite not significant (Table 3).

With the same level of information transparency, participants expressed that good AI model conditions provided more sufficient information for their interpretation (low-transparency as a control variable, $p < 0.001$; high-transparency as a control variable, $p < 0.001$) (Table 4).

The Effect of Explanations on the Perceived Understandability of AI Logic

With the same level of AI model performance, participants in the high transparency group believed they had a better understanding of how AI arrived at its prediction (weak AI model as a control variable, $p < 0.001$; good AI model as a control variable, $p < 0.001$) (Table 2). Even when people disagreed with AI or the AI prediction was incorrect, the rating of the perceived understandability of AI logic was higher when providing local explanation (Table 3).

When presented with the same amount of local explanations, participants saw a good AI model were more likely to agree that how AI generated its predictions was understandable ($p < 0.001$) (Table 4). However, presenting a good AI model didn't always result in high understandability of the AI logic. For example, if people disagreed with AI or when the AI prediction was incorrect, people would provide a lower rating ($p < 0.0001$ in both conditions) (Table 5).

The Effect of Explanations on the Perceived Trustworthiness of AI Predictions

Our results revealed that as long as participants saw a prediction generated by the good AI model, there is no significant difference of trust between low and high transparency conditions (Table 2). However, we observed that when the model performance was low (Table 2) or when human disagreed with the AI prediction (Table 3), the more explanations provided, the lower trust of AI predictions was expressed ($p = 0.04$ and $p < 0.001$, respectively). In addition, we found that participants reported a higher trust in AI when they were presented with a good AI model (low-transparency as a control variable, $p < 0.001$; high-transparency as a control variable, $p < 0.001$) (Table 4).

Table 3 Under two conditions ("Incorrect AI" and "AI-Human Disagree"), the association between providing local explanations and people's perceptions of different aspects of the AI system

	Conditions					
	Incorrect AI			AI-Human Disagree		
	With Explanation (N = 286)	Without Explanation (N = 291)	*P* value	With Explanation (N = 407)	Without Explanation (N = 428)	*P* value
Q1 (usefulness)	3.07	3.23	0.14	2.96	3.13	0.047*
Q2 (sufficiency)	3.05	3.12	0.48	2.94	3.01	0.35
Q3 (understandability)	3.31	3.08	0.003**	3.31	3.08	0.005**
Q4 (trustworthiness)	2.87	3.03	0.08	2.75	3.03	< 0.001***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 4 When controlling transparency level, the association between model performance and people's perceptions of different aspects of the AI system

	Conditions					
	High Transparency			Low Transparency		
	Present Weak Model (N = 831)	Present Good Model (N = 907)	*P* value	Present Weak Model (N = 855)	Present Good Model (N = 839)	*P* value
Q1 (usefulness)	3.23	3.48	< 0.001***	3.28	3.48	< 0.001***
Q2 (sufficiency)	3.13	3.41	< 0.001***	3.20	3.38	< 0.001***
Q3 (understandability)	3.40	3.61	< 0.001***	3.13	3.23	0.06
Q4 (trustworthiness)	3.09	3.33	< 0.001***	3.20	3.37	< 0.001***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Discussion

Our study revealed that model performance can influence people's trust toward health AI systems. When the model performance was higher, participants tended to rate AI predictions as highly trustworthy and useful. These findings are consistent with prior work [35, 61] showing that model performance information plays an important role in promoting users' trust and acceptance of AI systems. However, the challenge is that the performance and reliability of many today's patient-facing, AI-driven health systems are not good enough [62, 63]. As our study shows, when model performance was poor or when the local AI prediction was incorrect, the more information disclosed to users, the less people would trust AI. Prior work also suggest that showing a performance indicator (e.g., confidence score) might have drawbacks [35, 64]. For example, a numeric score may not be meaningful to lay people. Also, just presenting model performance with a confidence score may be insufficient for many people to develop a good understanding of how accurate or reliable the system is, especially when dealing with complex, high-critical scenarios (e.g., interpreting clinical data). Therefore, we suggest that the designers should scrutinize about presenting model performance information to patients. One possible design solution is combining the use of color and explicit description of low-confidence zones to deliver an intuitive view of the model performance to help users calibrate their trust. For example, when the model performance is not optimal, the system can use red color to display the predictions, accompanied by a clear description of why the confidence score is low. Future research could study the effect of this design technique on patients' trust while using the AI system to make decisions.

We also found that providing local explanations for the rationale of a prediction led to a better understanding of how AI arrived at its prediction. However, it didn't help foster trust. These findings contributed to the on-going investigation of how AI explanations affect people's trust; some studies have found that interfaces designed to provide explanations are effective building users' trust in the AI systems [22, 25, 26], whereas other studies found contradicted results—providing explanations may not raise satisfaction, or even erode users' trust in a system [57, 65]. These previous studies conducted in settings such as recommender system [57] and e-commerce system [25]. Our findings revealed that in the context of comprehending radiology report, providing more explanations about AI predictions did not contribute to trust building. There are two possible explanations. One is that our study context has a high criticality, healthcare consumers are more concerned about whether the AI system

Table 5 Under two conditions ("Incorrect AI" and "AI-Human Disagree"), the association between model performance and people's perceptions of different aspects of the AI system

	Conditions					
	Incorrect AI			AI-Human Disagree		
	Present Weak Model (N = 381)	Present Good Model (N = 196)	*P* value	Present Weak Model (N = 481)	Present Good Model (N = 354)	*P* value
Q1 (usefulness)	3.09	3.27	< 0.001***	3.07	3.01	0.41
Q2 (sufficiency)	3.03	3.19	< 0.001***	3.00	2.94	0.47
Q3 (understandability)	3.22	3.17	< 0.001***	3.22	3.16	< 0.001***
Q4 (trustworthiness)	2.92	3.01	< 0.001***	2.93	2.84	0.23

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

is reliable and accurate, rather than its reasoning. Another explanation is that when participants disagreed with the AI prediction or when the AI prediction was inaccurate, providing more explanations of AI prediction actually had negative effect.

We found that people's agreement with AI predictions and the correctness of AI predictions could influence the effect of explanations on people's perceptions. For example, if human disagreed with the AI prediction, providing local explanation actually led to lower ratings of the perceived usefulness and trustworthiness of AI predictions. Similarly, in the condition of human-AI disagree or AI incorrect, presenting a good model didn't help at all (for example, it actually led to lower understandability of AI predictions). These observations indicated that people's perceptions of AI-driven health systems are not only affected by AI explanations or model performance, but also highly related to their knowledge level and personal judgement. These findings align with previous work showing that relatability (how relatable the explained rationale is to people's judgement) influences users' perceptions of how much the generated explanation helped them understand the rationale of the AI system [49].

Taken together, model performance information can promote people's trust and perceived usefulness of AI predictions, while providing local explanation can promote understandability. From this perspective, it might be useful to present both types of information to patients when they are using the tool to interpret clinical data. However, it is never easy to achieve that goal. There has been heated debate regarding the trade-offs between model accuracy and model interpretability—those state-of-the-art methods for generating more accurate models are less likely to be interpretable by end-users [60, 66]. For example, deep neural networks have achieved near-human accuracy levels in various types of tasks (e.g., classifying medical image [67]), but this approach remains operating as black boxes, offering little to no explanation as to why specific features are selected over others during training or whether and how users' inputs are taken and modeled by the algorithm [68]. Future work should look into how to achieve a good balance between model performance and model interpretability when developing patient-facing AI systems.

Our study has several limitations. For example, the context of our study only focused on the interpretation of radiology reports. As user needs for information and AI explanations may change in different contexts, future work should evaluate the generalizability of the results in other contexts and assess longitudinal behavioral outcomes. Second, our participants were recruited online through MTurk and they may not have experience with reviewing radiology reports. But as our goal is to develop AI-based systems for lay patients, especially those who do not have sufficient health literacy, using MTurkers for this study is appropriate. In addition, prior work has demonstrated that using MTurk to study healthcare consumers' perceptions is reliable and effective [69]. Third, it is well recognized that MTurkers could be more tech savvy than the wider population. This limitation of participant recruitment could affect the generalizability of our results because there is a lack of representation of marginalized population. In our future work, we will include more marginalized population (e.g., less literate people and older adults) to investigate their perceptions of using explainable AI systems to interpret clinical data. Lastly, other types of explanations that could feed into user trust, such as global explanations, were not included in our system prototype. Future work can compare the effects of global explanations with local explanations.

Conclusion

In this study, we conducted an experimental study to evaluate the effect of AI local explanations and model performance on establishing trust and promoting positive attitudes in the context of comprehending radiology report. We found that model performance can promote people's trust and perceived understandability, while local explanations can promote understandability of AI recommendations but not necessarily trust. We also observed that whether human agree or disagree with AI predictions or whether the AI prediction is correct or not could also influence the effect of AI explanations on healthcare consumer's perceptions. We discussed the implications of these findings, especially in relation to the design of health AI systems to help patients understand diagnostic results.

Declarations

Ethical Approval This study was approved by the first author's university Institution Review Board (IRB).

Conflict of Interest All authors declare that he/she has no conflict of interest.

References:

1. Ross, S.E., et al., *Expectations of patients and physicians regarding patient-accessible medical records.* J. Med. Internet. Res. 7(2):13, 2005.
2. Rubin, D.L., *Informatics methods to enable patient-centered radiology.* Acad. Radiol. 16(5):524-534, 2009.
3. Basu, P.A., et al., *Creating a patient-centered imaging service: determining what patients want.* Am. J. Roentgenol. 196(3): 605-610, 2011.

4. Berlin, L., *Communicating results of all radiologic examinations directly to patients: has the time come?* Am. J. Roentgenol. 189(6):1275-1282, 2007.
5. Peacock, S., et al., *Patient portals and personal health information online: perception, access, and use by US adults.* J. Am. Med. Inform. Assoc. 24(e1):e173-e177, 2016.
6. Ma, X., et al., *Professional Medical Advice at your Fingertips: An empirical study of an online.* Proceedings of the ACM on Human-Computer Interaction. 2(CSCW):116, 2018.
7. Zhang, Z., et al., *Understanding Patient Information Needs about their Clinical Laboratory Results: A Study of Social Q&A Site.* Stud. Health.Technol. Inform. 264:1403, 2019.
8. Rosenkrantz, A.B. and E.R. Flagg, *Survey-based assessment of patients' understanding of their own imaging examinations.* J. Am. Coll. Radiol. 12(6):549-555, 2015.
9. Hong, M.K., et al. *Supporting families in reviewing and communicating about radiology imaging studies*. in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, 2017.
10. Arnold, C.W., et al., *Imaging informatics for consumer health: towards a radiology patient portal.* J Am Med Inform Assoc. 20(6):1028-1036, 2013.
11. Oh, S.C., T.S. Cook, and C.E. Kahn, *PORTER: a prototype system for patient-oriented radiology reporting.* J. Digit. Imaging. 29(4):450-454, 2016.
12. Alpert, J.M., et al., *Applying multiple methods to comprehensively evaluate a patient portal's effectiveness to convey information to patients.* J Med Internet Res. 18(5):e112, 2016.
13. Reynolds, T.L., et al. *Understanding Patient Questions about their Medical Records in an Online Health Forum: Opportunity for Patient Portal Design*. in *AMIA Annual Symposium Proceedings*. 2017. American Medical Informatics Association.
14. Zikmund-Fisher, B.J., et al., *Graphics help patients distinguish between urgent and non-urgent deviations in laboratory test results.* J. Am. Med. Inform. Assoc. 24(3):520-528, 2016.
15. Chen, H., S. Compton, and O. Hsiao. *DiabeticLink: a health big data system for patient empowerment and personalized healthcare*. in *International Conference on Smart Health*. Springer, 2013.
16. Long, J., M.J. Yuan, and R. Poonawala, *An Observational Study to Evaluate the Usability and Intent to Adopt an Artificial Intelligence–Powered Medication Reconciliation Tool.* Interact. J. Med. Res. 5(2):e14, 2016.
17. Palanica, A., et al., *Physicians' Perceptions of Chatbots in Health Care: Cross-Sectional Web-Based Survey.* J. Med. Internet Res. 21(4):e12887, 2019.
18. Zhang, Z., et al., *Lay individuals' perceptions of artificial intelligence (AI)-empowered healthcare systems.* Proc. Assoc. Inform. Sci. Technol. 57(1):e326, 2020.
19. Hoermann, S., et al., *Application of Synchronous Text-Based Dialogue Systems in Mental Health Interventions: Systematic Review.* J. Med. Internet Res. 19(8):e267, 2017.
20. Harwich, E. and K. Laycock, *Thinking on its own: AI in the NHS.* Reform Research Trust, 2018.
21. Johnson, H. and J. Peter. *Explanation facilities and interactive systems*. in *Proceedings of the 1st international conference on Intelligent user interfaces,* 1993.
22. Muramatsu, J. and W. Pratt, *Transparent Queries: investigation users' mental models of search engines*, in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery: New Orleans, Louisiana, USA 217–224, 2001
23. Sinha, R. and K. Swearingen, *The role of transparency in recommender systems*, in *CHI '02 Extended Abstracts on Human Factors in Computing Systems*. 2002, Association for Computing Machinery: Minneapolis, Minnesota, USA. 830–831, 2002.
24. Herlocker, J.L., J.A. Konstan, and J. Riedl, *Explaining collaborative filtering recommendations*, in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. Association for Computing Machinery: Philadelphia, Pennsylvania, USA. 241–250, 2000.
25. Pu, P. and L. Chen, *Trust building with explanation interfaces*, in *Proceedings of the 11th international conference on Intelligent user interfaces*. 2006, Association for Computing Machinery: Sydney, Australia. p. 93–100.
26. McGuinness, D.L., et al. *Explanation interfaces for the semantic web: Issues and models*. in *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.
27. Vorm, E.S. *Assessing Demand for Transparency in Intelligent Systems Using Machine Learning*. in *2018 Innovations in Intelligent Systems and Applications (INISTA)*. IEE, 2018.
28. Bussone, A., S. Stumpf, and D. O'Sullivan. *The role of explanations on trust and reliance in clinical decision support systems*. in *2015 International Conference on Healthcare Informatics*. IEEE, 2015.
29. Poursabzi-Sangdeh, F., et al., *Manipulating and measuring model interpretability.* arXiv preprint arXiv:1802.07810, 2018.
30. Cai, C.J., et al., *" 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):1–24, 2019.
31. Ribeiro, M.T., S. Singh, and C. Guestrin. *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*. ACM, 2016
32. Lundberg, S.M. and S.I. Lee. *A unified approach to interpreting model predictions*. in *Advances in neural information processing systems,* 2017.
33. Yin, M., J. Wortman Vaughan, and H. Wallach. *Understanding the effect of accuracy on trust in machine learning models*. in *Proceedings of the 2019 chi conference on human factors in computing systems,* 2019.
34. Lai, V. and C. Tan. *On human predictions with explanations and predictions of machine learning models: A case study on deception detection*. in *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 2019.
35. Zhang, Y., Q.V. Liao, and R.K. Bellamy, *Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making.* arXiv preprint arXiv:2001.02114, 2020.
36. Vorm, E.S. and D.M. Andrew. *Assessing the Value of Transparency in Recommender Systems: An End-User Perspective.* in *ACM Conference on Recommender Systems*. Vancouver, Canada, 2018.
37. Kizilcec, R.F., *How Much Information? Effects of Transparency on Trust in an Algorithmic Interface*, in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, Association for Computing Machinery: San Jose, California, USA. 2390–2395.
38. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature. 542(7639):115–118, 2017.
39. Sirinukunwattana, K., et al., *Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images.* IEEE Trans. Med. Imaging. 35(5):1196-1206, 2016
40. He, J., et al., *The practical implementation of artificial intelligence technologies in medicine.* Nat. Med. 25(1):30-36, 2019.
41. Arimura, H., et al., *Magnetic resonance image analysis for brain CAD systems with machine learning*, in *Machine learning in computer-aided diagnosis: medical imaging intelligence and analysis*. IGI Global. 258–296, 2012
42. Erickson, B.J., et al., *Machine learning for medical imaging.* Radiographics. 37(2):505-515, 2017.

43. Wang, D., et al., *" Brilliant AI Doctor" in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment.* arXiv preprint arXiv:2101.0152, 2021.
44. Stiggelbout, A.M., et al., *Shared decision making: really putting patients at the centre of healthcare.* Bmj. 344, 2012
45. Fan, X., et al., *Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case Study.* J. Med. Internet Res. 23(1):e19928, 2021.
46. Nadarzynski, T., et al., *Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study.* Digit . Health. 5:2055207619871808, 2019.
47. Davis, F.D., *Perceived usefulness, perceived ease of use, and user acceptance of information technology.* MIS quarterly. 319–340, 1989.
48. Venkatesh, V., et al., *User acceptance of information technology: Toward a unified view.* MIS quarterly. 425–478, 2003.
49. Ehsan, U., et al. *Automated rationale generation: a technique for explainable AI and its effects on human perceptions*. in *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 2019.
50. Broekens, J., et al. *Do you get it? User-evaluated explainable BDI agents*. in *German Conference on Multiagent System Technologies*. Springer. 2010
51. Larasati, R. and A. DeLiddo, *Building a trustworthy explainable AI in healthcare*. Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from. 209, 2009.
52. *Overcoming Barriers in AI Adoption in Healthcare*. 2018; Available from: https://newsroom.intel.com/wp-content/uploads/sites/11/2018/07/healthcare-iot-infographic.pdf.
53. Esmaeilzadeh, P., *Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives.* BMC Med. Inform. Decis. Mak. 20(1):1-19, 2020.
54. Dietvorst, B.J., J. Simmons, and C. Massey. *Understanding algorithm aversion: forecasters erroneously avoid algorithms after seeing them err*. in *Academy of Management Proceedings*. 2014. Academy of Management Briarcliff Manor, NY 10510.
55. Dzindolet, M.T., et al., *The role of trust in automation reliance.* Int. J Hum. Comput. Stud. 58(6):697-718, 2003.
56. Adadi, A. and M. Berrada, *Peeking inside the black-box: a survey on explainable artificial intelligence (XAI).* IEEE Access. 6:52138-52160, 2018.
57. Cramer, H., et al., *The effects of transparency on trust in and acceptance of a content-based art recommender.* User Modeling and User-Adapted Interactio. 18(5), 2008.
58. Kaltenbach1, E. and I. Dolgov. *On the dual nature of transparency and reliability: Rethinking factors that shape trust in automation*. in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2017. SAGE Publications Sage CA: Los Angeles, CA.
59. Cocos, A., et al., *Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation.* J. Biomed. Inform. 69:86-92, 2017.
60. Johansson, U., et al., *Trade-off between accuracy and interpretability for predictive in silico modeling*. Future Med. Chem. 3(6):647-663, 2011.
61. McGuirl, J.M. and N.B. Sarter, *Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information.* Hum. Factors. 48(4):656-665, 2006.
62. Strickland, E., *IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care.* IEEE Spectrum. 56(4):24-31, 2019.
63. Fan, X., et al., *Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case Study.* J. Med. Internet Res. 22(12):e19928, 2020.
64. Nguyen, A., J. Yosinski, and J. Clune. *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
65. Kizilcec, R.F. *How much information? Effects of transparency on trust in an algorithmic interface*. in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016.
66. Kamwa, I., S. Samantaray, and G. Joós, *On the accuracy versus transparency trade-off of data-mining models for fast-response PMU-based catastrophe predictors.* IEEE T. Smart Grid. 3(1):152-161, 2011.
67. Tajbakhsh, N., et al., *Convolutional neural networks for medical image analysis: Full training or fine tuning?* IEEE Trans. Med. Imaging. 35(5):1299-1312, 2016.
68. Chakraborty, S., et al. *Interpretability of deep learning models: a survey of results*. in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. 2017. IEEE.
69. Shapiro, D.N., J. Chandler, and P.A. Mueller, *Using Mechanical Turk to study clinical populations.* Clinic. Psychol. Sci. 1(2):213-220, 2013.