Explainable recommendations and calibrated trust – Research protocol

Calibrated trust has become an important design goal when designing Human-AI collaborative decision-making tools. It refers to a successful understandability, reliability and predictability to the AI-based tool behaviour and recommendations. Explainable AI is an emerging field where explanations accompany AI-based recommendations to help the human-decision maker understand, rely on, and predict AI behaviour. Such an approach is supposed to improve humans' trust calibration while working collaboratively with an AI. However, evidence from the literature suggests that explanations have not contributed to improved trust calibration and even introduced other errors. Designers of such explainable systems often assumed that humans would engage cognitively with AI-based explanations and use them in their Human-AI collaborative decision-making task.

This research explores users' behaviour and interaction style with AI-based explanations during a Human-AI collaborative decision-making task. Such an investigation will help further studies address design solutions for AI explanations to enhance trust calibration and operationalize explainability during a Human-AI decision-making task. To achieve this goal, we conduct a multi-stage qualitative study. It includes think-aloud protocol, follow-up interviews and observations. The results of these studies will guide the research to develop an understanding of the main research question in the literature: "Why explanations do not improve trust calibration?". It will also help our future research to devise a design method for the XAI interface to enhance trust calibration. In the following subsection, we explained the procedures and provided the supplementary materials used in each study. The study workflow is summarised in Figure 1.

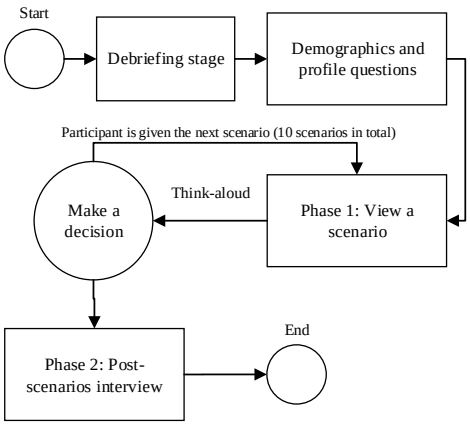
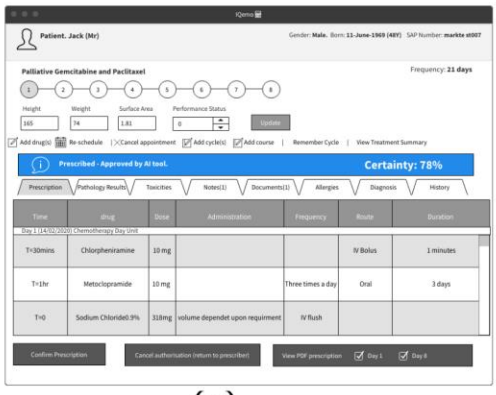


Figure 1 Study workflow for each participant

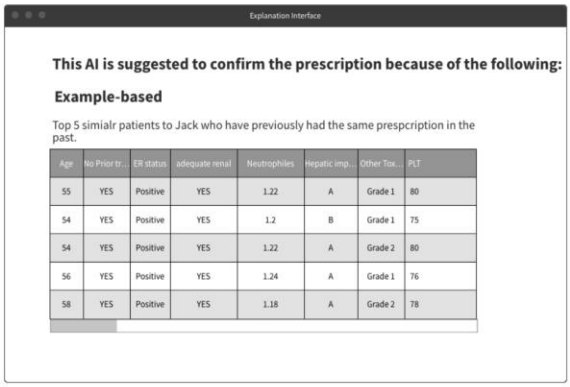
Phase 1: think-aloud protocol – first stage.

We aim to provide explanatory information that supports the medical practitioners in their trust calibration during Human-AI collaborative decision-making task. Our participant's inclusion criteria were based on their experience of using clinical decision support systems in their settings and experience in screening chemotherapy prescription (See Appendix A). We designed ten recommendations accompanied by ten different explanations. The adopted recommendations were generated to be non-trivial, which was based on a literature review on related work and medical expert judgment. We tested the

material and activities with two participants and refined them to optimise their fulfilment of these criteria (See Appendix B). Also, we validated the material with a medical oncologist with a focus on the border cases that need an investigation from the participants in the actual study. This ultimately helped put our participants, who were medical experts, in a realistic setting: exposing them to an imperfect AI-based recommendation and its explanations where trust calibration is needed and where errors in that process are possible. We consulted with one AI expert and one medical expert, presenting them with ten explainable interfaces, and asked them for their expert opinion regarding the relevance of the explanations and the validity of the recommendation. We used these opinions, as well as the results from our pilot study, to refine the interface design. Each scenario considered a hypothetical patient profile and AI-recommendations that suggests either rejecting or accepting a chemotherapy prescription for the patient. Patients have been initialised with fictional names and profiles to make it more realistic to our practitioners. Each scenario was accompanied by one different explanation class and was meant to be either correct recommendation or incorrect recommendation. We used our five main explanation classes revealed from our previous literature review (See Appendix C). We encouraged them to think aloud during their decision-making process. Then, they were asked to think freely and encouraged to make optimal decisions. Examples of explainable interfaces used in our study settings are shown in Figure 2.

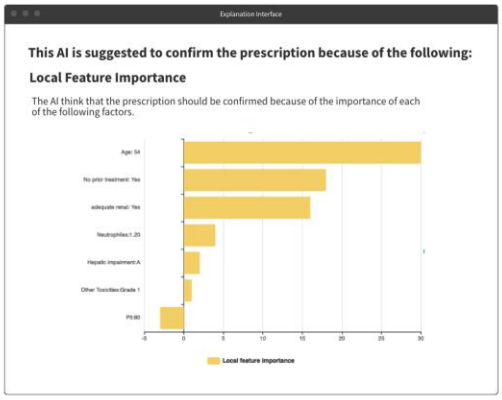


(a)

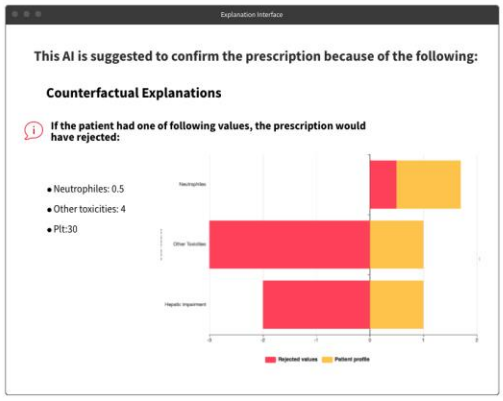


Age	No Prior tr...	ER status	adequate renal	Neutrophiles	Hepatic imp...	Other Tox...	PLT
55	YES	Positive	YES	1.22	A	Grade 1	80
54	YES	Positive	YES	1.2	B	Grade 1	75
54	YES	Positive	YES	1.22	A	Grade 2	80
56	YES	Positive	YES	1.24	A	Grade 1	76
58	YES	Positive	YES	1.18	A	Grade 2	78

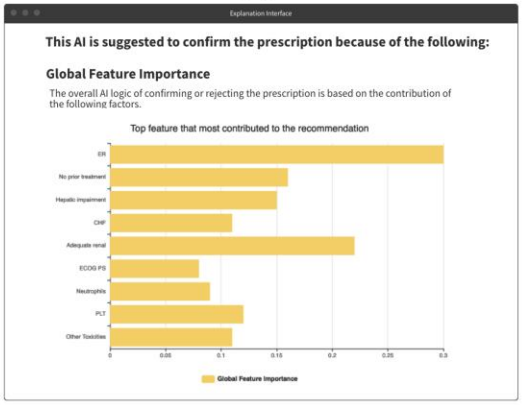
(b)



(c)



(d)



(e)

Fig 2. Five explanation classes mock-up interface presented to our participants. (a) confidence explanation. (b) Example-based explanation (c) Local feature importance (d) Counterfactual explanation (e) Global feature importance.

Phase 2. Post-interview questions.

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

General questions.

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

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

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

Appendix A. Scenarios characteristics

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

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

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

Please indicate your level of agreement with the following statements.

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

Appendix B. Scenarios characteristics

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

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

	SC1 Male:54 CHF	SC2 Male:47 CHF	SC3 Female:56 CHF	SC4 Male: 44 not CHF
ER	Positive	Positive	Positive	Negative
No prior treatment with CDK 4/6	Yes	Yes	Yes	Yes
Adequate renal and hepatic function	Yes	Yes	Yes	Yes
ECOG PS	2	0	1	2
Neutrophils	1.20	0.9	1.00	0.7

Plt	80	74	33	84
Hepatic impairment	A	B	A	C
Other Toxicities	Grade 1	Grate 2	Grade 1	Grade 4

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

Appendix C. Taxonomy for model-agnostic explainable models.

Global explanation	Global feature importance	Ranking the data features.	(Lou et al., 2013, Nguyen et al., 2016, Tolomei et al., 2017)
		Dependencies between data features	(Henelius et al., 2014 Henelius et al., 2017)
		Influence Function	(Datta et al., 2016)
	Decision tree approximation	(Bastani et al., 2017, Johansson and Niklasson, 2009, Krishnan et al., 1999, Bastani et al., 2017, Johansson and Niklasson, 2009, Zhou and Hooker, 2016, Thiagarajan et al., 2016)	
	Rule extraction	AND-OR rules	(Dash et al., 2018, Aung et al., 2007, Wei et al., 2019, Tan et al., 2018, Zhou et al., 2003)
		If-then rules	(Johansson et al., 2004, Quinlan, 1987)
Local explanation	Local feature importance	(Ribeiro et al., 2016b, Lundberg and Lee, 2017, Simonyan et al., 2013, Fong and Vedaldi, 2017, Dabkowski and Gal, 2017, Zhou et al., 2003, Mishra et al., 2017, Ribeiro et al., 2016a)	
	Local rules and trees	(Guidotti et al., 2018a, Krishnan and Wu, 2017) (Ribeiro et al., 2018) (Konig et al., 2008, Johansson et al., 2004, Soares and Angelov, 2019)	
Example-based	Prototype	(Bien and Tibshirani, 2011, Kim et al., 2016) (Kim et al., 2014) (Kim et al., 2016, Kanehira and Harada, 2019)	
	Counterfactual example	(Wachter et al., 2017) (Martens and Provost, 2014, Chen et al., 2017) (Laugel et al., 2017, Mothilal et al., 2020)	
	Influential example	(Koh and Liang, 2017) (Goodfellow et al., 2014) (Yuan et al., 2019, Dong et al., 2017, Szegedy et al., 2013)	
Counterfactual	Feature Influence	(Woodward, 1997) (Apley, 2016, Friedman, 2001, Goldstein et al., 2015) (Krause et al., 2016)	
	Counterfactual features	(Wachter et al., 2017) (Dhurandhar et al., 2018, Wachter et al., 2017) (Zhang et al., 2018) (Krause et al., 2016) (Barocas et al., 2020)	

Confidence	(Zhang et al., 2020, Bussone et al., 2015) (Gal and Ghahramani, 2016, Schulam and Saria, 2019) (Josse et al., 2019, Graves, 2011, Blundell et al., 2015) (Srivastava et al., 2014) (Hooker, 2004) (Hendrycks and Gimpel, 2016).