Explaining health recommendations to lay users: The dos and don'ts

Maxwell Szymanski[1], Vero Vanden Abeele[1] and Katrien Verbert[1]

[1]Department of Computer Science, KU Leuven, Leuven, Belgium

Abstract

In recent years, mobile health recommendations are used in an increasing number of applications. Researchers have highlighted the importance of explaining these recommendations to lay users, with benefits such as increased trust and a higher tendency to follow up on these recommendations. However, a different explanation modality can impact the way users perceive the recommendation, either in a positive or negative way. This paper will explore and evaluate six different explanation designs through a qualitative user study, and give general design guidelines and considerations regarding explaining pain-related health recommendations to lay users.

Keywords

explainable AI, explainable recommender systems, explanation interpretation, lay users, health recommendations, HRS

1. Introduction & Related Work

Recommender systems are becoming more prevalent in health-related domains. However, several key aspects have to be taken into account when designing recommender systems, such as transparency through explanations and end user expertise.

1.1. RecSys in Health

Recommender systems (RS) have become prominent in health applications, where they help retrieve relevant information or recommend possible next actions tailored to the needs of the end user. These health recommender systems (HRS) are used both in clinical settings as well as in personal contexts where health applications aid users in their daily lives. A recent systematic review [1] of HRS for lay users shows that the majority of HRS that used a graphical user interface focus on mobile applications. These mobile HRS span several fields, such as sports, mental health and nutrition, and include applications that e.g. suggest the appropriate action to take for users with diabetes [2], recommend activities to promote healthier lifestyles [3] or help with anxiety by recommending external apps that will suit the user's needs [4]. These recommender systems all have a shared main goal of potentially steering the user towards a better and healthier lifestyle.

However, the increased use of HRS is also paralleled with certain barriers. One such issue is a mismatch in recommendations with the user's expectations. Such mismatch can not only lead to a decrease in system effectiveness [5], but a decrease in trust towards the system as well, potentially steering the user away from future use of such HRS. Early research mainly focused on increasing the accuracy of RS in order to mitigate this issue. However, Valdez et al. [6] explain that recent research has undergone a shift in focus from improving accuracy, to exploring the effects of human factors. This broader approach in reasoning about RS should allow researchers to improve RS effectiveness beyond quantitative algorithmic capability. The new approach includes the research on and addition of: explanations to increase transparency, human-in-the-loop feedback to correct misunderstandings, and using conversational RS to increase familiarity towards the system's interface.

In this paper, we will focus on the *explanation* aspect, more specifically, on designing and assessing different explanation types for a mobile health recommender system. The research is conducted in the context of a personal coaching app that guides users with chronic musculoskeletal pain through various informative and interactive topics, such as activity- and stress-management, pain-education, etc. Additionally, the app also includes a pain logbook, that can be used for logging pain flare-ups. Using this logged information, which consists of the context in which the pain occurred, as well as the thoughts and reactions users had, the app is able to give personalised recommendations to better cope with pain flare-ups in the future. In this study, we look into several designs that are deemed fit for explaining these pain related recommendations to end users. There remain, however, several research challenges that need to be addressed, such as explanation interpretability and end user expertise which are discussed in the next related work section.

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland

maxwell.szymanski@kuleuven.be (M. Szymanski);
vero.vandenabeele@kuleuven.be (V. V. Abeele);
katrien.verbert@kuleuven.be (K. Verbert)

1.2. Explaining health recommendations

As highlighted earlier, adding explanations to recommendations can improve the overall effectiveness. These make the system interpretable, which in turn can improve trust towards the system [7]. There exist HRS that explain their rationale to the end user, such as the food recommender system of Wayman et al. that explains why certain recipes are recommended based on the user's nutritional intake [8], or a visualisation for medical experts that is able to explain breast cancer similarities [9]. However, the systematic review of De Croon et al. states that only 10% of HRS that focus on lay users make use of explanations. This makes HRS explanations for lay users a novel, but under-explored topic. Additionally, a study of Bussone et al. points out that providing overly detailed explanations for health recommenders can create unforeseen effects, such as creating over-reliance on explanations [10], which points out that health recommender explanations should be designed with sufficient care. This makes designing explanations with non-expert users in mind, and evaluating them with end users, paramount.

1.3. End user expertise

An increasing amount of research has pointed out that the expertise of end users should be taken into account when designing explanations. Ribera et al. [11] have proposed three main categories of end users: non-experts (lay users), domain experts (in our context medical professionals or health coaches) and software- and AI-experts. Each category of users comes with its own needs, goals and limitations. AI expert users, for example, use XAI to verify or improve the underlying AI system, whereas domain experts can leverage explanations to gain additional insights and learn from the system. Lay users have their own set of goals, but more interestingly their own array of limitations as well. Wang et al. have pointed out several shortcomings in non-expert users related to cognitive biases, such as confirmation and anchoring bias, due to a backward-oriented, hypothesis-driven reasoning process [12]. Tsai et al. also noticed a *reinforcing effect*, where users avoid interacting with content they are not familiar with [13]. Szymanski et al. additionally pointed out that non-expert users, despite having these biases and incorrectly interpreting certain complex explanations, can still have a preference for them over other, simpler explanation modalities [14].

Thus we see that interpretability through explanations has multiple benefits and can result in an increased trust towards the system. However, as previously mentioned, the adoption of explanations in HRS is still low. Furthermore, most health-related AI explanations are being researched with AI and domain expert users in mind [15], which leaves a big gap for explanations w.r.t. lay users. Keeping the aforementioned biases in mind that lay users are prone to, it is therefore tantamount to assess whether explanations are indeed interpretable to make sure no misalignment in trust is created.
With these considerations in mind, we investigate the following research questions:

RQ1 What explanation design do lay users prefer when explaining health recommendations and why?
RQ2 What design considerations are substantial when explaining health recommendations to lay users?

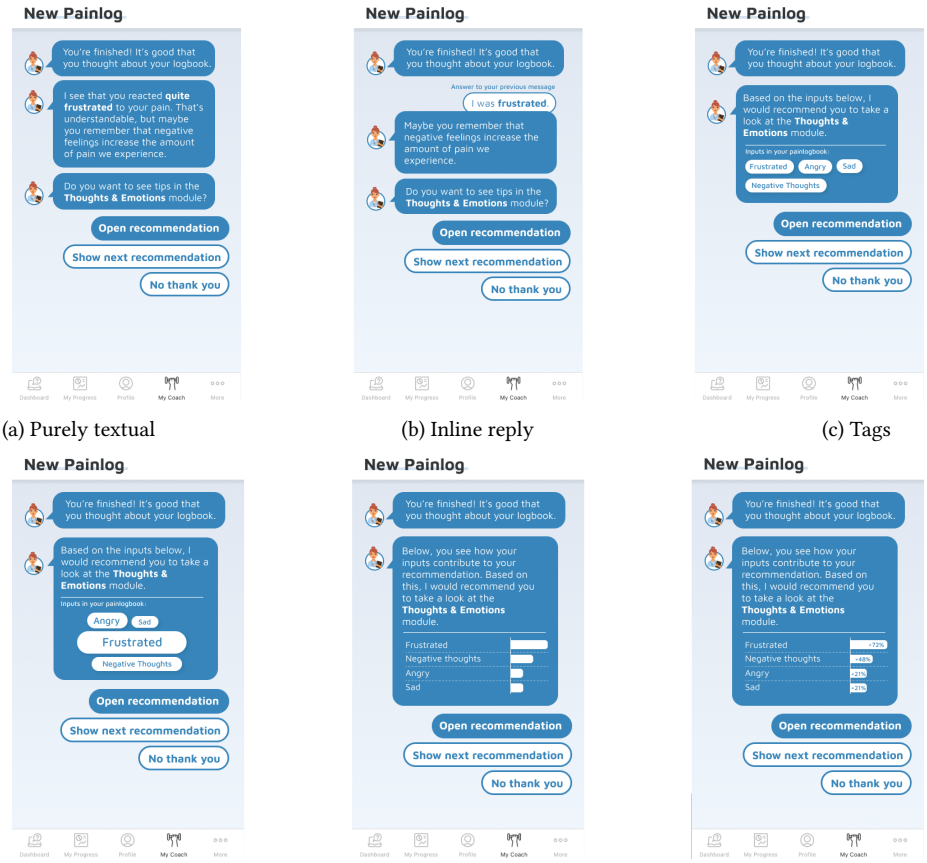
2. Explanation designs

As mentioned in section 1.1, we will focus on designing different explanations that will explain why users are receiving specific recommendations for their pain flare-ups. Keeping the context and type of end users in mind, the following design guidelines have to be kept in mind for all variants of explanations:

- **Mobile-friendly:** as the explanations will be offered within the context of the mobile health app, the explanations have to be well-suited for display on a small mobile screen.
- **Summative:** the explanations should possess the ability to summarise categorical data, as input consists of (semi-)unstructured user input.
- **Suited for non-experts:** as the end users are non-experts, the explanations should not use any advanced and statistical concepts to explain why the recommendation is suggested.

Keeping these criteria in mind, we came up with the following designs in Figure 1 based on well-known and widely used explanation types:

- **Text-based:** briefly explain why the recommendation is related to the most prevalent input. The wording is based on the "communicating health-related news to patients" guidelines described by [16] and these explanations were collaboratively designed for the purpose of this study by six ergo- and physiotherapists.
- **Text-based + inline reply:** an addition to the textual explanation, where the inline-reply shows which specific user message most contributed to the recommendation.
- **Tags:** tags are a common method of communicating all topics that are relevant to a recommendation (e.g. Bidargaddi et al. [17]).
- **Word clouds:** in addition to showing all relevant topics, word clouds are able to additionally communicate relative importance/relevance of these topics (e.g. [18, 19]).



(a) Purely textual (b) Inline reply (c) Tags

(d) Word cloud (e) Feature importance (f) Feature importance + %

Figure 1: Explanation designs for pain-related health recommendations used throughout the user study

- **Feature-importances (FI):** feature importance bars communicate contributing themes of the user input, as well as their input relevance, albeit in a more specific way compared word clouds.
- **Feature-importances (FI) + percentages:** adds percentages to the FI bars to communicate exact topic importances.

These explanation designs are sorted from least to most by the amount of information they convey regarding the inputs relevant to the recommendation. The textual explanation only focuses on one input, with the inline reply being able to also show which specific input triggered the recommendation, whereas the tags are able to display all relevant input categories that are related to the recommendation. The word-cloud further builds on this by also displaying the relative importance of each input related to the recommendation, and the FI shows the exact sorting of input according to importance. The added percentages give the most transparency regarding the inputs, by also displaying the exact values used by the underlying RS.

2.1. Participants

For the user study, we recruited 11 participants out of a pool of 286 people who were already using the mobile health coaching application without the pain logbook and its recommender system, as mentioned in section 1.1, and thus knew and have interacted with the content and different modules. The group consisted of nine women and two men, of which four finished graduate school, six college, and one high school. Age-wise, 2 participants were between 21-30, 5 between 31-40, 3 between 41-50 and 1 between 51-60. All 11 users noted to use the internet on the regular basis, with 6 participants stating to be average computer and IT users, and 5 participants stating to be advanced computer and IT users.

2.2. Protocol of the evaluation study

At the start of the study, users were briefed on the purpose and context of the think-aloud study, and gave their consent to having the audio recorded, after which they filled in the ResQue demographics questionnaire [20]. Afterwards, they were guided through the pain logbook, which they had to fill in with recent pain-episode they experienced in mind. Having done so, they received some information regarding the recommendations that are going to be given, along with the explanations. We briefly went over the six explanation designs in a fixed order, after which we asked the participant to "explain what they like or dislike about the explanation" separately for each design once they have seen them all. To conclude this *preference elicitation*, the users had to sort the explanations by preference, with 1 being their most preferred one, and 6 their least preferred. They also had to give (or repeat) a key reason as to why they are giving each explanation a certain ranking. The audio recordings of both the preference elicitation and ranking are used afterwards for a thematic analysis.

2.3. Data analysis

The thematic analysis was done in two phases, with the first phase consisting of deriving granular themes from the thematic analysis with two researchers, and the second phase focusing on merging them to higher level themes with a third researcher. The resulting higher level themes are displayed in Figure 3, along with the frequencies in which they occur per explanation design. The agreement percentage of the first phase two-coder thematic analysis is 88.1%, with Cohen's kappa being $\kappa = 0.66$, resulting in a substantial inter-coder agreement [21].

3. Results

Taking the average raking scores of all explanation designs, we are now able to rank the 6 explanation modalities from best to worst ranked, along with the results from the thematic analysis to explain why each explanation type scored poorly or adequately. Figure 2 shows the frequencies of the rankings given to each explanation design.

3.1. Feature importance + percentage

Rank: 1 (best) · This explanation type was favored by most users, mainly due to the fact that it provided the most insight and transparency ($n = 10$). Only three out of 11 people found the addition of the percentages to feature importance bars to be inefficacious.

Insights through XAI (+)

Six users liked the fact that they were able to gain **more insight** through this explanation modality. Four users also stated that the percentages were a "**nice-to-know**", making the explanation more useful and informative.

Negative sentiment towards XAI (-)

On the flip-side, two users disliked the addition of displaying percentages, stating that when it comes to emotions and feelings, certain aspects are **not quantifiable**. U4 stated: "*Personally I think feelings are not quantifiable. The bars are good, but don't put an exact number on it. It's okay if you're communicating frequencies, like how often an emotion occurred for example.*".

Visual/information overload (-)

Two users also stated that the addition of percentages is **unnecessary**, mentioning that only using bars to communicate importances is sufficient.

3.2. Feature importance

Rank: 2 · The feature importance explanation was among the most preferred explanations, liked for the fact that is was able to give a summary of the user input ($n = 11$), as well as being able to give additional insights ($n = 2$).

Provides summary (+)

Six users found the feature importance bars to be a **clear** way of communicating input topics and their importance. Four users stated that it gives them a **nice overview** of their input.

Insights through XAI (+)

Two users specifically liked the **additional insights** that they were able to get from the feature importances. U4 mentioned: "There are of course no numbers given, but I can assume that I am really frustrated, and a bit less angry. I find it interesting to reflect on results that come out of a questionnaire."

Negative sentiment towards XAI (-)

Three users were **unsure of the ranking** of some topics, stating that they agreed with the general content, but not as to why one topic was deemed more important over others. This caused these users to slightly dislike and distrust the system, and give it a lower ranking.

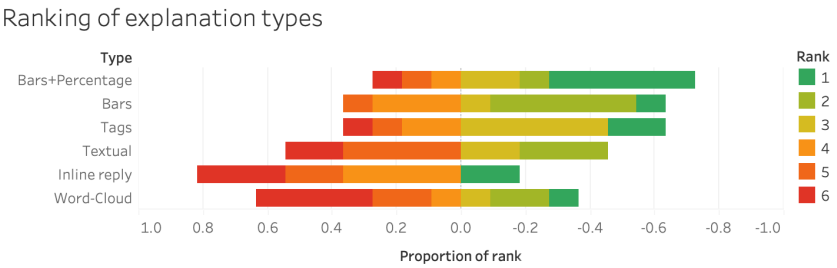


Figure 2: Frequencies of rankings per explanation type

Explanation Type	Category / General Theme						
	Positive				Negative		
	Summary	Insights through XAI	More control	Positive sentiment towards XAI	Problems with presentation / content	Visual / information overload	Negative sentiment towards XAI
Textual	8			3			5
Inline reply	7	3			3		
Tags	8	3	3			2	
Word Cloud		4			8	4	
FI	11	2	4			2	3
FI + %		10				2	2

Figure 3: TA themes per explanation design and their frequencies

Visual/information overload (-)

Two users found the bars to be **unnecessary**, giving them information as to what contributed towards the recommendation, but not why, like the textual explanation did. U6 stated: "*There is not a lot of background given. It shows that these inputs contributed to my recommendation, but not why.*"

3.3. Tags

Rank: 3 · Tags scored relatively better than the previous three explanations in terms of average ranking, and were liked for their summative ability ($n = 8$). Only people who disliked having a lot of information, were less in favor of the tag explanation ($n = 2$).

Provides summary (+)

Four users found using tags to be a nice way of providing a summary of their input. Four users also stated that doing in such a way is a **clear and concise** method of explaining why the recommendation is given.

Insights through explanation (+)

Three users were fond of the **additional insights** they got from the tags and the general themes that were present in their input. U3 stated: "*When inputting my feelings I did not necessarily perceive them as negative or angry. But based on these tags, I'm able to see: okay, this is how the app interprets my feelings.*"

Visual/information overload (-)

Only two users stated that tags were **unnecessary** or provided too much information. U6 stated: "*Yes it's clear, but less practical. I tend to focus on one thing at a time.*"

3.4. Purely textual

Rank: 4 · Purely textual explanations received mixed reactions during the think-aloud study. When users liked or agreed with the recommendation, the textual explanation was a welcome addition helping them understand the recommendation process and the recommendation itself, and gave users a nice summary of why the recommendation matched their inputs ($n = 8$). However, when the recommendation wasn't in line with the user's expectations, the textual explanation highlighted the mismatch

even more and caused a poor reception of the recommender system in general ($n = 5$). Here is an overview of these topics:

Provides summary (+)

Six users found that the textual explanation was able to **summarize their input** quite well, albeit only focusing on one topic (the most relevant one) surrounding the recommendation.

Positive sentiment towards explanation (+)

Two users stated that the written explanation was **confirming and comforting**. One user also stated that the wording of the textual explanation felt **less confronting** regarding their negative input.

Negative sentiment towards explanation (-)

On the other hand, three users mentioned that they **cannot relate** to the recommendation, and that the textual explanation highlighted this fact. U4 also found the explanation to also be **provoking**, stating the following: "*I know that I'm frustrated and that it does not help. However, explaining that acts like waving a red flag in front of a bull.*"

3.5. Inline-reply

Rank: 5 · During the think-aloud study, the inline reply received relatively positive feedback and comments regarding the succinct summary it gave of the users input ($n = 7$), with only some minor remarks regarding the presentation of the explanation ($n = 3$). However, it scored quite low during the preference ranking itself due to other explanation modalities simply being preferred over the inline-reply.

Provides summary (+)

Six users found the explanation modality to be **clear and more concrete**, and one user additionally stated that showing which message triggered the recommendation requires **less analysis** from the user.

Insights through explanation (+)

Three users liked the fact that the inline-reply **raises awareness** of the fact that the recommendation is related to one of their own inputs. U3 stated: "*I find it better than the textual explanation. There, they state 'You seem to be frustrated', and here you really are made aware of the fact that it's your own input.*"

Problem with representation (-)

Only some minor and infrequent negative remarks were given surrounding inline replies. Three users disliked the fact that by highlighting or repeating their negative input, they are **more confronted** with it. One user additionally mentioned that this explanation feels like the recommendation is only tuned to one input instead of multiple user inputs, making it feel **too specific**.

3.6. Word cloud

Rank: 6 (last) · The word cloud received the lowest average score. In general, users like the addition of displaying keyword or topic importance, however using a word cloud to do so proves to be an inferior solution. The thematic analysis points out two main negative themes as to why this explanation is disliked: problems with representation and content ($n = 9$) and visual/information overload ($n = 4$) and one positive theme, insights through explanation ($n = 4$).

Problems with representation (-)

Three users pointed out having keyword size communicate importance was *unclear*, and would rather have something concrete like bars indicating exact relevance. Three users also pointed out that the inconsistent sizes inherent to the design of word clouds were **visually displeasing**. Two users additionally stated highlighting important keywords might be **too confronting** with respect to their own input, e.g. if a user inputs that they are feeling sad, having it displayed as a large word might confront the user too much with their state of mind.

Visual/information overload (-)

Three users found the addition of displaying relevance in such a way **unnecessary**, one of which additionally stated that adding the information in such way is **too distracting**.

Insights through explanation (+)

Four users stated however that adding this information of keyword relevance gives **more insight** due to not only showing the relevant topics, but their importance as well.

4. Discussion

We will now discuss some of the most prevalent observations that were present in several explanation designs, as well as suggest guidelines on how to design health explanations for lay users experiencing (chronic) pain.

4.1. Beware of confronting people with negative sentiments

People experiencing (chronic) pain or illness can feel distress when receiving negative information surrounding their state. In our study, we noticed that highlighting keywords that are potentially negative (e.g. negative emotions, reactions, etc.), can cause distress with users and therefore make them dislike the explanation. This was apparent with the inline reply and word cloud explanations, where visually highlighting negative sentiments that relate to the recommendation caused users to dislike the explanation.

4.2. Use tags or feature importance when control is needed

Due to the fact that tags and FI/FI+% are able to display multiple input categories, users positively expressed that this would provide them more control over the recommendation process, if the design or implementation allows for it. One user suggested that tapping certain topics could be useful to request recommendations in a more user-controlled way. Other users additionally suggested U9:"*It's nice if you can individually remove certain topics*", and U7: "*... especially of you notice something that wasn't interpreted the way you intended it*".

4.3. Design FI through a lay user's perspective

The FI and FI+% designs were favored by most users, giving most users the insight and summary they needed. However, as mentioned in section 3.2, U4 interpreted the FI bars as "*... I can assume that I am really frustrated, and a bit less angry*", indicating that they saw it as an overview of their input, and not how strongly their input relates to the recommendation. In total, 10 out of 11 lay users interpreted FI differently than intended. Only U4 was able to correctly interpret the bars (after reading the text above the FI bars - "This is how your inputs relate to the recommendation"), saying "*The frustrated bar is the biggest, okay, so that contributes most to my recommendation*". Having a wrong interpretation could lead to confusion towards the system when, for example, a next recommendation is shown, and the input keywords and their relevance change with respect to this new recommendation. However, overcoming biases and changing mental models of lay users often proves to be difficult. A possible design adaptations to the FI and FI+% design, may show a general overview/summary of the user input to be in line with what users were interpreting, and then highlight the keywords that are relevant to the recommendation that is being shown. This can be seen in Figure 4. Keeping the control aspect in mind from previous section, users are also able to tap on different topics to request recommendations regarding said topic.

Figure 4: Adapted feature importance explanation design

4.4. Insight vs. information overload

Users generally liked the holistic approach of the feature importances, and were more inclined to look into the recommendation itself. When asked why they liked the recommendations more when explained using FI compared to the purely textual explanation, they stated that the FI were able to show them a general overview of them as a person.

On the other hand, there were also some users who disagreed with the ordering of keyword importances that the feature importance bars were displaying, causing a slight increase in distrust towards the recommender system, ranking the explanation lower. This is to be expected, as increasing transparency of explanations can cause a higher drop in trust towards the system if the content of the explanation or recommendation does not align with the user's expectation. However, the effect of a misaligned textual explanation is still stronger, as users who did not agree with either the recommendation or the explanation expressed a more negative sentiment towards the recommendation, and gave the textual recommendation a lower ranking. This is in line with similar research by Balog et al. [5], in which they state that misaligned recommendations that focus on a single topic or item are more susceptible to a lower perceived quality of explanation compared to multi-item recommendations.

5. Conclusion

This paper introduced several explanation designs for mobile pain related health recommendations, and compared them among lay users. Most users preferred the added transparency that was provided by the tags and FI / FI+% designs, stating that it gave them a brief and clear overview of their input which helped them understand why they received certain recommendations. Another interesting aspect is the fact that designs should be careful with visually highlighting negative sentiments of users. Designs that did so, i.e. the inline-reply and word cloud, were received poorly by users. Lastly, we confirmed that lay users might interpret certain visual explanations differently than intended, yet still prefer them over others. Given their feedback, we presented an adapted design of the favoured FI / FI+% explanation to be in line with what lay users expect.

6. Limitations & Future work

The qualitative aspect of this study was already able to point out several key aspects related to designing health explanations for patients experiencing chronic pain. However, a larger scale quantitative user study is needed to further investigate these results. One such aspect is the fact that some users preferred textual explanations over explanations that offered more information. Investigating whether this correlates to the user's need for cognition (NFC), and what its implications are, can prove to be an interesting research direction similar to the research of Millecamp et al. [22]. Another aspect is the fact that while most users disliked being confronted with their negative input, some did not mind. This could be related to the "warriors vs. worriers" research, in which some users experiencing chronic pain actually prefer being exposed to negative feedback so they could address it, and could prove useful for further research [23]. Future research should also consider other designs to explain health recommendations and elaborate design guidelines that can be used by researchers and practitioners in this exciting domain. In addition, an interesting further line of research is to personalise these explanations on-the-fly, based on interaction data of end-users. As in work of [24], clicks and hover interactions as well as eye gaze data can be considered for such personalisation.

Acknowledgments

This work is part of the research projects Personal Health Empowerment (PHE) with project number HBC.2018.2012, financed by Flanders Innovation & Entrepreneurship, and IMPERIUM with project number G0A3319N, financed by Research Foundation Flanders (FWO).

References

[1] R. De Croon, L. Van Houdt, N. N. Htun, G. Štiglic, V. Vanden Abeele, K. Verbert, Health recommender systems: Systematic review, J Med Internet Res 23 (2021) e18035. URL: https://www.jmir.org/2021/6/e18035. doi:10.2196/18035.

[2] F. Torrent-Fontbona, B. Lopez, Personalized adaptive cbr bolus recommender system for type 1 diabetes, IEEE Journal of Biomedical and Health Informatics 23 (2019) 387–394. doi:10.1109/JBHI.2018.2813424, robin's Paper: [93].

[3] R. Gouveia, E. Karapanos, M. Hassenzahl, How do we engage with activity trackers? a longitudinal study of habito, UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (2015) 1305–1316. doi:10.1145/2750858.2804290.

[4] K. Cheung, W. Ling, C. J. Karr, K. Weingardt, S. M. Schueller, D. C. Mohr, Evaluation of a recommender app for apps for the treatment of depression and anxiety: An analysis of longitudinal user engagement, Journal of the American Medical Informatics Association 25 (2018) 955–962. doi:10.1093/jamia/ocy023.

[5] K. Balog, F. Radlinski, Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 329–338. URL: https://doi.org/10.1145/3397271.3401032. doi:10.1145/3397271.3401032.

[6] A. Calero Valdez, M. Ziefle, K. Verbert, Hci for recommender systems: The past, the present and the future, in: Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 123–126. URL: https://doi.org/10.1145/2959100.2959158. doi:10.1145/2959100.2959158.

[7] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019). URL: https://www.mdpi.com/2079-9292/8/8/832. doi:10.3390/electronics8080832.

[8] E. Wayman, S. Madhvanath, Nudging Grocery Shoppers to Make Healthier Choices, in: Proceedings of the Ninth Conference on Recommender

Systems, ACM, 2015, pp. 289–292. doi:10.1145/2792838.2799669.

[9] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Séroussi, Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach, Artificial Intelligence in Medicine 94 (2019) 42–53. URL: https://www.sciencedirect.com/science/article/pii/S0933365718304846. doi:https://doi.org/10.1016/j.artmed.2019.01.001.

[10] A. Bussone, S. Stumpf, D. M. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, 2015 International Conference on Healthcare Informatics (2015) 160–169.

[11] M. Ribera, A. Lapedriza, Can we do better explanations? a proposal of user-centered explainable ai, CEUR Workshop Proceedings 2327 (2019).

[12] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing Theory-Driven User-Centric Explainable AI, Association for Computing Machinery, New York, NY, USA, 2019, p. 1–15. URL: https://doi.org/10.1145/3290605.3300831.

[13] C.-H. Tsai, P. Brusilovsky, Beyond the ranked list: User-driven exploration and diversification of social recommendation, in: 23rd International Conference on Intelligent User Interfaces, IUI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 239–250. URL: https://doi.org/10.1145/3172944.3172959. doi:10.1145/3172944.3172959.

[14] M. Szymanski, M. Millecamp, K. Verbert, Visual, textual or hybrid: The effect of user expertise on different explanations, in: 26th International Conference on Intelligent User Interfaces, IUI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 109–119. URL: https://doi.org/10.1145/3397481.3450662. doi:10.1145/3397481.3450662.

[15] J. Ooge, G. Stiglic, K. Verbert, Explaining artificial intelligence with visual analytics in healthcare, WIREs Data Mining and Knowledge Discovery 12 (2021). URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1427. doi:https://doi.org/10.1002/widm.1427.

[16] M. Schmid Mast, A. Kindlimann, W. Langewitz, Recipients' perspective on breaking bad news: How you put it really makes a difference, Patient Education and Counseling 58 (2005) 244–251. URL: https://www.sciencedirect.com/science/article/pii/S0738399105001473. doi:https://doi.org/10.1016/j.pec.2005.05.005, medical Education and Training in Communication.

[17] N. Bidargaddi, P. Musiat, M. Winsall, G. Vogl, V. Blake, S. Quinn, S. Orlowski, G. Antezana, G. Schrader, Efficacy of a web-based guided recommendation service for a curated list of readily available mental health and well-being mobile apps for young people: Randomized controlled trial, Journal of Medical Internet Research 19 (2017). doi:10.2196/jmir.6775, robin's Paper: [55].

[18] Y. Wu, M. Ester, Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 199–208. URL: https://doi.org/10.1145/2684822.2685291. doi:10.1145/2684822.2685291.

[19] C.-H. Tsai, P. Brusilovsky, Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 22–30. URL: https://doi.org/10.1145/3320435.3320465. doi:10.1145/3320435.3320465.

[20] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 157–164. URL: https://doi.org/10.1145/2043932.2043962. doi:10.1145/2043932.2043962.

[21] N. J.-M. Blackman, J. J. Koval, Interval estimation for cohen's kappa as a measure of agreement, Statistics in Medicine 19 (2000) 723–741. doi:https://doi.org/10.1002/(SICI)1097-0258(20000315)19:5<723::AID-SIM379>3.0.CO;2-A.

[22] M. Millecamp, N. N. Htun, C. Conati, K. Verbert, To explain or not to explain: The effects of personal characteristics when explaining music recommendations, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 397–407. URL: https://doi.org/10.1145/3301275.3302313. doi:10.1145/3301275.3302313.

[23] J. Geuens, T. Swinnen, L. Geurts, R. Westhovens, R. De Croon, V. Vanden Abeele, Worriers versus warriors: Tailoring mhealth to address differences in patients with chronic arthritis, in: 2020 IEEE International Conference on Healthcare Informatics (ICHI), 2020, pp. 1–12. doi:10.1109/ICHI48887.2020.9374322.

[24] M. Millecamp, T. Willemot, K. Verbert, Your eyes explain everything: exploring the use of eye tracking to provide explanations on-the-fly, in: Proceedings of the 8th Joint Workshop on Interfaces and Hu-

man Decision Making for Recommender Systems co-located with 15th ACM Conference on Recommender Systems (RecSys 2021), volume 2948, CEUR Workshop Proceedings, 2021, pp. 89–100.