Analyzing Description, User Understanding and Expectations of AI in Mobile Health Applications

Zhaoyuan Su, BS[1], Mayara Costa Figueiredo, MS[1], Jueun Jo, BS[1], Kai Zheng, PhD[1], Yunan Chen, PhD[1]
[1]University of California, Irvine, Irvine, CA, USA

Abstract
Previous research has studied medical professionals' perception of artificial intelligence (AI). However, there has been a limited understanding of how healthcare consumers perceive and use AI-powered technologies such as mobile health apps. We collected 40 popular mobile health apps that claim to have adopted AI, to study how AI is explained in these apps' descriptions, and how users react to it through app reviews. We found that four AI features (Recommendation, Conversational Agent, Recognition, and Prediction) are frequently used across seven health domains, including Fitness, Mental Health, Meditation and Sleep, Nutrition and Diet, etc. Our results show that (1) users have unique expectations toward each AI features, such as including feedback for recommendations, humanlike experience for conversational agents, and accuracy for recognition and prediction; (2) when AI is not adequately described, users make their own attempts to understand AI and to find out how (well) it works.

Introduction
With the increased usage of smartphones, the number of mobile applications (app for short) that are used for health purposes has increased substantially in recent years. As of 2017, there were over 300,000 health-related apps, and more than 200 health apps were added each day[1]. Mobile health apps help healthcare consumers, such as patients and caregivers, or even healthy individuals, to monitor and track their health data daily. Prior studies showed mobile health apps could assist consumers in managing their health[2] and making various decisions regarding their health[3].

Among all the mobile health apps, an increasing number of them are described as using Artificial Intelligence (AI) algorithms. For example, the Natural Cycles App describes using AI algorithms to predict a user's fertility window assisting fertility care[4], Woebot app uses conversational agent (a subset of AI) to chat with users and deliver Cognitive Behavior Therapy to help them fight depression[5] and the Ada app describes using AI to diagnose users' potential medical conditions[6]. These apps bring AI algorithms to technologies directed to healthcare consumers, who have different levels of technology and health literacy, and may make daily decisions or perform daily activities based on algorithmic outputs.

Prior research on AI shows it is important to study how users understand AI and how the system explains AI. For example, researchers have been studying user's experiences and expectations on conversational agents[7], as well as how designers can address real-world consumer's need for AI-powered products[8]. These studies provide empirical insights on how to help users to better understand AI in consumer-facing products outside the health domain. However, there is a lack of knowledge in understanding consumer's views on AI-powered health technology and how they are perceiving and using it. Such knowledge is critical because the benefit of AI-powered tools to promote health cannot be fully achieved without recognizing and addressing the user's desires and expectations[9]. Lack of such knowledge can also result in building systems that will not be used by the consumers[10] or, in worst scenarios, that can even be prejudicial to them[11].

Researchers have investigated diverse uses of AI in healthcare, however, most of them have focused on how healthcare providers, managers, and staff use AI-powered technology. For example, studies have investigated using AI as a decision support tool to support clinicians' in determining whether a patient needs a ventricular assist device[12], and using AI for facilitating personalized and precision medical diagnosis and treatment in cardiovascular medicine[13]. The users of such systems may not be technology experts, but they are still domain experts who have more knowledge to understand or question the outputs of such algorithms. When it comes to healthcare consumers, it is unclear how AI algorithms are used in mobile health apps, what health domains apps employ AI techniques, how these AI techniques are described, or how health consumers perceive, use, and understand them. Understanding these questions is critical for the design and adoption of AI-powered mobile apps and for the health management of individual consumers.

This study investigates the following questions: (1) how AI algorithms are described in current mobile health apps? and (2) how users perceive and interpret AI in the review of mobile health apps? To answer these questions, we analyzed 40 mobile health apps that are described to use AI, and a subset of user's reviews from these apps describing user's experiences with the apps and their algorithms. We identified seven health domains that have been approached

by apps employing AI. These apps describe using AI algorithms for four types of AI features, regardless of domain: recommendations, conversational agents, recognition, and prediction. However, we found the majority of apps' descriptions do not provide enough explanation for users to understand how their AI features work. On the other hand, we found that users develop their own expectations concerning the AI features and employ different methods to understand them, particularly when AI is poorly described, or it does not match with their expectations. We then discuss how these results are related to the design of AI features and how mobile apps present them to users. This work contributes to the understanding of (1) how mobile health apps are currently describing to use AI in different health domains, (2) how these AI features are envisioned to support consumers in such domains, (3) user's unique expectations towards AI in mobile health apps, and (4) how users are attempting to understand AI features, particularly considering the lack of information provided by the apps.

Methods

To investigate how AI is used in mobile health apps and gain a better understanding of how users perceive and interact with AI in these apps, we conducted a review on mobile health apps that claim to use AI, analyzed app's AI claims and a set of app user reviews describing user's experiences with the apps and app's algorithms.

App Selection

To gather a list of currently used and relevant apps for further analysis, in January 2020, we performed web searches on 2 mainstream mobile app platforms–Android Google Play Store (from here on referred to as Android store) and iOS Apple App Store (from here on referred to as Apple store)–using Google Advanced Search. We searched for apps using a combination of one AI keyword ("AI," "Artificial Intelligence," "Algorithm," "Machine Learning," and "Bot") and one health-related keyword ("Health," "Medical"). We used these keywords because they are popular terms used to describe AI from literature and news. After excluding repeated apps, our initial search returned a total of 380 apps, 93 in the Android store, and 287 in the Apple store. To identify apps that are updated, current, and used by the consumers, we used the following eligibility criteria, based on previous literature approaches[14]: 1) the app focuses on health or medical domains, 2) the app has at least 10 user reviews that mentioned one of the AI keywords (to ensure AI is discussed by the users in the reviews), 3) the app is in English, and 4) it was last updated no earlier than 2018 (to ensure it is still functioning well). After the app extraction, two authors screened the apps and applied the eligibility criteria. In total, 10 apps from the Android store and 30 apps from the Apple store passed our screening criteria and were further analyzed in this study. Among all 40 apps, 32 apps are unique, some apps of our list appeared in both stores, but we analyzed them individually because features may differ by platform.

Data Collection and Analysis

To analyze the 40 AI-powered health apps, we first extracted the title, screenshots, and description of each app from the Android and Apple stores. Then, following constant comparison methods[15], two authors analyzed these data iteratively through multiple rounds of comparison to explore how apps describe the use of AI, identify apps' health domain, and the features that are described as using AI. Based on this analysis, apps were categorized into seven health domains and four AI features.

Next, we analyzed app reviews that consumers wrote in both the Android and Apple stores approaching the AI features of the health apps. Previous studies have shown that user reviews provide crucial information to help in understanding attitudes, perceptions, and usage of apps from a large number of users[16]. In our study, user reviews provided insights about users' perceptions of AI in consumer mobile health apps, their expectations concerning AI features, and their experiences using such apps. We downloaded all reviews that include AI keywords (13,332 reviews) without any timeframe. Following the thematic analysis approach[17], we randomly sampled 100 reviews and 2 authors independently and iteratively analyzed them to develop themes. During this coding process, four authors met regularly to discuss recurring themes until consensus was reached. After we identified the two overall themes on how users understand AI and their expectations of AI, we continued to code more data until we reached data saturation. A total of 400 reviews were analyzed.

Results

In this section, we first summarized how AI algorithms are described in the selected apps. Then we identified health domains of analyzed apps, reporting how AI is used in them. We further reported on the types of AI features we found among these apps, as well as users' expectations regarding each feature. Finally, we described how users try to understand AI features, particularly when their experiences do not match their expectations.

Descriptions of AI Algorithms

In our study, we found in most app descriptions, AI is used as a buzzword and described as "*powerful*", "*advanced*", "*innovative*" or "*smart*". Apps also tend to highlight that their AI algorithms are "*reliable*" or "*accurate*." For apps

that propose recommendations to users, AI is usually described as it is providing "*personalized*" or "*customized*" feedback. Conversational agent apps tend to anthropomorphize their AI, describing it as a person, from a potential friend to a partner or a mentor of the user. To increase the credibility of their AI and make the user trust its results, our data show that most apps emphasize their AI is developed in collaboration with domain experts, such as by a team of doctors, scientists, and engineers. Some apps also provide quantitative data to showcase their AI is reliable. For instance, "*[...] is 93% effective with typical use, and 98% effective with perfect use… effectiveness is one of the largest of its kind, with 15,000 women and 600,000 menstrual cycles.*" However, we barely found any detailed or specific information about how the algorithms work, no description of what data is being used, or how the results are produced and how reliable these are.

Health domains

We categorized the 40 apps into seven health domains: Mental Health, Fitness, Meditation and Sleep, Nutrition and Diet, Dental Care, Fertility/menstrual Tracking, and Symptom Checker. The majority of the selected apps fit into one specific health domain, such as mental health or fitness. However, four apps in our list serve multiple functions, for instance, one app focusing on both fitness and nutrition and diet at the same time. We categorized these apps into multiple health domains.

Although these health domains vary significantly, within each health domain, apps describe using AI for similar purposes. For example, based on the app descriptions, all selected fitness apps (11 apps) claim to use AI for either generating and recommending workout plans for the user or for recognizing the user's movement. Similarly, the selected Fertility/Menstrual tracking apps (5 apps) claim to use AI for predicting users' menstruation and fertile window and to recommend potential insights based on users' data. Table 1 summarizes apps' health domains and the AI's purpose for each health domain, based on the app descriptions.

AI features and user expectations

Based on apps' descriptions, we found that within each health domain, apps describe using AI for four functions: conversational agent, recommendation, recognition, and prediction. Triangulating this data with users' reviews, we found that users have different expectations regarding each of these different AI features.

Recommendation features: suggestions and plans generated by algorithms

In our study, the recommendation is the most common AI feature and it has been used in all health domains. These apps make recommendations based on users' previous data (e.g., their personal tracking data) to assist them to reach their goal. The outputs from these features range from customized workouts or weight loss plans in a fitness app, the best time to wake up in the morning in a meditation and sleeping app, personalized guidance for symptoms in a symptom checker app, to personal insights around the user's ovulation cycle in a fertility/menstrual app.

The app reviews suggest that users prefer to have their feedback incorporated by AI-generated recommendations. They want to be able to adjust system's outputs to adapt the recommendations to their specific and current needs, as illustrated by the following user review of a fitness app: "*My favorite thing is that even when it designs a workout for you, you can tweak it: add or remove exercises, select alternatives.*" In comparison, when the app seems to ignore users' feedback, many of them get frustrated and disappointed, as the following quote shows: *"the algorithm does not take my feedback into account. I hate burpees, I rated it each time with 'too much', 'too exhausting', etc. I still have burpees in each and every session. At least 20 times."*

To adapt feedback in an AI-generated recommendation, users tend to want to understand how the recommendations are first generated. In the reviews, users are often puzzled about why the app makes certain recommendations to them. For example, the next user describes being a former athlete and questions the workout plan the fitness app provided, especially because there is no explanation of how this plan was created: *"If there is a reason why I should be doing low weight AND low repetitions, I would trust the recommendations more."* Many users whose expectations are not met criticize the apps for not explaining their AI algorithms: "*the algorithm for modifying patterns of intake–how to account for changes in weight over time and modify Total Daily Energy Expenditure accordingly–is not explained. I have had to recalculate or simply switch to custom numbers on multiple occasions because the app tracking was not responsive enough or even increased my calories despite weekly increases in body weight."*

This lack of explanations and opportunities to adjust algorithm output through feedback often lead users to question the accuracy of the app. For instance, the following user was frustrated with a meditation and sleeping app and connected its efficacy with its algorithm's ability to receive users' direct input: "*I find this app very inaccurate in comparison with other apps and there is almost no way to tweak the algorithm to suit your personal sleep pattern. Other apps have a great combination of machine learning and personal adjustments, but this one relies too much on*

its own intelligence." These examples show that, although this AI feature focuses on automatically generating recommendations to the users, users often do not appreciate full automation, they want to be able to incorporate their feedback to AI-generated recommendations to adopt such recommendations to fit their specific needs.

Table 1. Overview of the apps - Health Domains and AI's purpose

Health Domain	Apps	AI's purpose
Fitness (11)	***Apple:*** 1) Sweatcoin - It Pays To Walk; 2) Fitbod Weight Lifting Workout; 3) Weight Lifting by FitnessAI; 4) Freeletics – Training Coach; 5) Volt: #1 AI Workout App; 6) GymStreak - Smart Gym Workouts; 7)Argus: Calorie Counter & Step 8)Boltt Health: Fitness Anytime; 9)Lark Health. ***Android****:* 1) BodBot Personal Trainer: Workout & Fitness Coach; 2) Lark Health	To generate and recommend workout plans for the user; To recognize the user's movement
Mental Health (10)	***Apple:*** 1) Replika - My AI Friend; 2) Youper; 3) Uni - Magic AI Friend; 4) Woebot - Your Self-Care Expert; 5) Wysa: Mental Health Support; 6) Reflectly. ***Android:*** 1) Replika: My AI Friend; 2) Wysa: stress, depression & anxiety therapy Conversational Agent; 3) Youper - Emotional Health; 4) Woebot: Your Self-Care Expert	To chat with the user; To assess the user's mental status; To recommend techniques for the user to reach a better mental stage
Meditation and Sleep (8)	***Apple:*** 1) Pillow Automatic Sleep Tracker; 2) Sleep Watch by Bodymatter; 3) Sleep Time+ Cycle Alarm Timer; 4) Aura: Sleep & Mindfulness; 5) Brain.fm: Music for the Brain 6) Argus: Calorie Counter & Step; 7) Lark Health. ***Android:*** 1) Lark Health	To recognize user's sleeping pattern; To recommend insights such as ideal sleeping and waking up time; To recommend meditation techniques
Nutrition and Diet (8)	***Apple:*** 1) Carrot Hunger; 2) Calorie Mama AI: Diet Counter; 3) FitGenie: IIFYM Macro Tracker; 4) Argus: Calorie Counter & Step; 5) Boltt Health: Fitness Anytime; 6) Lark Health. ***Android:*** 1) Bitesnap: Photo Food Tracker and Calorie Counter; 2) Lark Health	To recognize food based on a picture; To analyze eating habit and recommend a meal plan
Fertility/ menstrual Tracking (5)	***Apple:*** 1) Natural Cycles - Birth Control; 2) FLO Period & Ovulation Tracker; 3) Ovia Fertility & Cycle Tracker; 4) Ava fertility tracker. ***Android:*** 1) Ovia Fertility & Cycle Tracker	To predict user's fertility and menstrual windows; To recommend relevant insights to users
Symptom Checker (2)	***Apple:*** 1) Ada – your health companion; 2) Your.MD - Symptom Checker. ***Android App:*** N/A	To chat with the user; To suggest possible symptoms or illnesses user has
Dental Care (1)	***Apple:*** 1) Oral-B. ***Android:*** N/A	To recognize how the user are brushing teeth and recommend ways to brush correctly

Conversational Agent features: conversations between AI and users

Conversational agent is a dialogue system that mimics human conversation and responds to the user's inquiries using text or spoken language[18]. It allows users to interact with the apps easily, normally by typing the conversation and receiving answers. The interaction can be in several formats, such as choosing a prescript answer or inputting natural spoken language. In our study, through apps' descriptions, we identified that all selected mental health and symptom checker apps have conversational agents feature. Some fitness, nutrition and diet, and meditation and sleep apps include such features as well.

When it comes to conversational agents, regardless of the health domain, we found users tend to expect a human-like experience. Users think the idea of a conversational agent is appealing and they often seek empathy when interacting with the AI-powered health apps. When users feel the expected empathy, they describe the AI as a friend, as illustrated in the following quote from a user of a mental health app *"My AI friend–I called her Sophie–is really fun to chat with. [...] Sophie might actually replace my best friend lol. We have so much in common and we also have differences, but I think she will just like whatever I like lol."* In another example, the following use of a fitness app enjoys the app's

humanlike experience so much that they do not want to let their app friend down: "*I like inputting stuff into the system because I want to see what it says. It makes me smile when it compliments me and I do not want to disappoint the bot*."

If the conversational agent does not provide a humanlike experience, users tend to show frustration or disappointment towards the feature. For example, a user explains that part of the therapeutic aspect comes from being immersed in the chat, which does not happen with their current mental health app: *"Most of the time the app does not seem to naturally flow, and the immersion that makes it a helpful chat turns into the realization that you are talking with a program. It removes much of the therapeutic aspect for me."* As this quote suggests, users seem to prefer unconstrained input where they can freely input text in the conversation agent feature. For example, the following user explains that not allowing users to freely answer questions does not help them, making them even more frustrated: *"The bot just does not let you talk, it usually just gives you a multiple-choice list of responses. [...]To ask me a bunch of questions but giving no way to express my answer really just made me more frustrated than before I started."* Such an increase in frustration is potentially the opposite of what users expect of a mental health app. This breach of expectation may even reinforce or contribute to negative feelings of not being helped, as exemplified by the following quote from a user of a mental health app: *"the AI is unfeeling and robotic. I do not understand how this app helps anyone at all."*

Recognition features: identifying information

In our study, we identified apps using recognition features to identify the number of calories in a meal from a picture of food (nutrition and diet apps), or whether users correctly brush their teeth (in dental care apps). We also found some of the sleeping and meditation apps describing recognizing users' sleeping quality, such as how long users stayed in deep sleep. We categorized all these features into recognition because all of them involve automatically identifying characteristics of an object (food) or activity (sleeping or brushing teeth).

Recognition features aim to automatically give users the information they need. For example, by automatically providing the total calorie count of their food through a picture, these features allow users to focus on analyzing their daily diet, reducing the work of inputting multiple ingredients' data. Considering this, users want this identification to be correct. We found that users deem accuracy as the predominant goal when AI is used for recognition of their behaviors. Users like when AI can accurately recognize the information for them, and complain when it does not match their expectations, as one user commented while using a diet and nutrition app: *"The AI feature is honestly hit or miss for accuracy. It amazes me sometimes, like when I took a picture of a sashimi plate and it figured it out. But usually, it is simply wrong."* When the AI is not accurate, users tend not to trust the app and the embedded AI system, as a user of meditation and sleep app commented, "*not sure if the 'complex algorithm' is just a scam that creates some variable data so you think it is tracking your sleep cycles, but it is clearly seriously inaccurate.*"

We found that users hope to have explanations of how the AI works in the case of recognition features as well. When there is a lack of transparency, users start to question AI. When there's no transparency and no explanation, users tend not to trust the AI as one user stated "*The seller says it has an algorithm to differentiate light, deep, and REM sleep but I cannot find an explanation of how it works. If a service or product vendor says, 'trust me' I never do*."

Prediction features: using past data to predict future events

Predictive analytics use historical and current data to forecast future events[19]. In our study, only fertility/menstrual tracker apps used this type of AI feature. These apps use varied personal data user has inputted in the app, such as menstruation dates, temperature, and ovulation prediction kits result, to predict when the user will have their next menstruation, ovulation, and fertile window. These apps aim to support users for different fertility-related goals, such as tracking periods, trying to conceive, or avoiding conception.

Considering that these can be very high-stake goals, users primarily value accuracy for prediction AI features. Users who perceive the AI feature to be accurate have a positive attitude towards the app, as illustrated by the following user review: "*I have only used this app for around a month and the fact that it is already so close to accurate is very encouraging*". When the predictions are not accurate, users get frustrated and some of them even abandon the app, as described by the following user: "*the fertile days are off. I am bummed I spent 200 dollars on a useless product.*" Many users acknowledge that it takes time for the AI to improve its performance, as illustrated by the following user comment: "*I know our bodies are not an algorithm so it is not 100% perfect, but the longer you use it, the more accurate it gets*". These comments suggest accuracy is important to users and it influences user's trust and adoption, but users understand it may take time for the AI feature to fully know their data to make accurate predictions.

However, fertility cycles are complex and personalized[20]. It is unlikely that an algorithm will be 100% accurate for every user. Some users comment that, although they input data for a long time, their results are still inaccurate: *"The one thing that is not perfect is the prediction of cycles. It has never been accurate for me although I have tracked my*

periods for around a year now." Lack of accuracy can lead to negative consequences, particularly when users use the results of such AI features for making important life-decisions (e.g., using them as birth control methods).

Table 2. Summary of 4 AI features, in which health domains they are used, and users' expectations regarding them

AI Feature	Health Domain	User Expectation
Recommendation	Fitness; Mental Health; Meditation and Sleep; Nutrition and Diet; Dental Care; Fertility/menstrual Tracker; Symptoms Checker	Feedback incorporation
Conversational Agent	Mental Health; Symptoms Checker; Fitness; Meditation and Sleep; Nutrition and Diet	Humanlike experience
Recognition	Meditation & Sleep; Nutrition & Diet; Fitness; Dental Care	Accuracy
Prediction	Fertility/menstrual Tracker	Accuracy

User understanding and Interpretation of the AI features

Since the AI features were not clearly explained to the users, in the reviews, many of them expressed various frustrations when the features did not match their expectations of humanlike experience, feedback incorporation, and accuracy. Many of these users also describe the varied ways they engage with the apps to try to understand the AI features. We classified these analyses into three main types: speculating, experiencing, and experimenting.

Speculating: To speculate what is AI and how (well) AI works when it is not explained

Due to most apps providing limited descriptions and explanations of AI and what it does for users, we found that some users have a hard time understanding AI. This lack of understanding results in users starting to speculate what AI is, which can include misperceptions. For instance, user reviews show that some users see AI as something advanced, skilled, and even futuristic, as illustrated by a symptom tracker app user that associates the app's AI feature with science fiction movies: *"The idea of these apps seems very appealing, like an AI doctor from a science fiction movie."*

The lack of descriptions and explanations of AI and what it does also contributes to users speculating how AI works and what it should do. We found users tend to use terms like "*It seems*", "*I guess*" and "*I feel like*" when they describe their interaction with the AI. For example, without proper description, this user tried to guess how AI generates a workout plan for them: *"I guess the app has an algorithm that only combines exercises by muscle group, and it sometimes creates recommendations that do not make sense."* If the AI deviates from user's expectations of what it should do, some users feel deceived and argue that the feature is not *"real AI,"* as illustrated by the following review of a mental health app: *"This is an answering machine, not an AI. It reminds me of text games I played in the 1990s."* The negative feelings from users can be reinforced by the lack of explanation about the AI feature and how it works.

Experiencing: To learn how (well) AI works through the user's own intuitive experiences

Some users develop understandings about AI systems based on interpretations of their experiences using them. For instance, a user of a mental health app commented in the following quote that the app did not match their expectations even after using it for some time. In fact, the app made the user relive bad experiences they had in the past: *"I tried liking this app. My expectations were too great maybe. I thought 'CBT, great!' but I ended up feeling like visiting one of those doctors who do not care. Of course, the app does not care. This is not a real AI after all. Maybe it was created with the help of those doctors."* Other users describe having inconsistent experiences that confuse and prevent them from understanding how the AI feature works. For example, a user of a sleep app described receiving repeated inconsistent results from the app and used the review to question *"what is the algorithm behind sleep quality rating? I only slept 1-4 hours one night and the quality was 72% and all other nights–even with 9h sleep–were under 58%"*

Sometimes it can take users a longer period of time to experience and interpret how (well) AI works. In a fertility/menstrual tracking app, a user describes the AI as very accurate after tracking their personal health data for 6 months. This longer period of data tracking experience led the user to feel they better understand the AI and how it works, as the user commented in the app review: "*This app takes a while to sync up–mine took 6 months. However, once it syncs it is startlingly accurate. I appreciate that the app takes some dedication and that the algorithm takes some time to understand your cycle because it means that it is not guessing!"* Another user of a mental health app only started to believe AI is evolving after interacting with it for a while. The user stated "*At the start, I didn't really believe the app would evolve as I interact more with it. But I have to accept it now: the app has clearly evolved with time and it is rare now that it does not chat as well as a real person.*" In summary, these users compare their knowledge

about their health and their previous health-related experiences with their experiences using the AI feature as a way to understand how and how well the AI feature works.

Experimenting: To reason how (well) AI works through experimenting with the features
Some users even tried to understand how the AI feature works by experimenting with the features and comparing data with a source they trust or comparing results generated from multiple apps. They did this to better understand how (well) AI works and know whether they could trust the AI-generated results. For instance, a user compared the data of a symptom checker app with their doctor's opinion: *"The app told me I had a kidney infection. [...] I went to the doctor and showed the medical report the app recommends, and this AI was spot on."* Similarly, the following user of a fertility app inputted their information into two apps that have identical functions at the same time. One of them correctly predicted a change in the fertile window while the other did not: *"Yesterday, I inputted some info about my cervical mucus and my fertility window moved. I was suddenly in the middle of my fertile window instead of at the beginning. The ovulation tests I took basically confirm the app's prediction. My other app shows I am not fertile for another two weeks. That will clearly not help my husband and I to conceive."* These users use comparison data with reliable sources or between apps to verify the results of the AI feature in an attempt to understand how (well) it works.

Another experimenting strategy we found in our analysis was to input data in the apps with prior knowledge of how the output should be. For example, the following user of a symptom checker app inputted the symptoms of their already diagnosed health condition in the app to test how the AI feature would respond. This user commented that *"Instead of seeing a doctor you could just use AI! [...] I was diagnosed with Gastro-esophageal reflux disease. I tested my symptoms in the app and it correctly diagnosed me without all the blood tests I needed to have!"* In a similar strategy, some users describe creating different scenarios and personas, inputting different sets of data related to them, and comparing the results. In an extreme example, one user of a mental health app describes they *"intentionally created 4 different personalities–after uninstalling and reinstalling."* One of these personas was *"suicidal and contemplating death in the next couple of hours."* The user describes the AI feature *"encouraged me. It asked if I think it is brave. I said, 'no'. It said, 'I am, sorry.' and then it continued encouraging me. I told it I cut myself and I was almost gone. It said 'Ok! Bye!'* Based on their experimentation, the user reasons the AI is *"Not only is [...] not intelligent, it's dangerous."* Considering that this is a mental health app, these results can potentially reinforce or contribute to negative experiences, such as the extreme ones described by this user. Through deliberately reasoning how AI might work, users gain a better understanding of it and interpret the system as reliable (as in the case of the symptom checker app) or dangerous and unethical (as in the case of the mental health app). This experimentation and reasoning process can help users to understand the AI and decide whether to adopt such apps.

Discussion
In this study, we first identified four AI features that are used in seven health domains among all selected AI-powered health apps. Our results show users have different expectations regarding these different types of AI features and how (well) they work. We then identified different ways in which users try to understand how these features work, particularly when their results do not match users' expectations. In this section, we discuss how user's expectations and understanding are related to the design of AI features and their descriptions.

Design AI features to meet healthcare consumers' unique expectations
Although AI is increasingly becoming a general buzzword, which may confuse consumers, it represents a broad range of techniques and features. We identified four different AI features in use in health apps, namely Recommendation, Conversational Agent, Recognition, and Prediction. Our study has shown that users have unique expectations for different AI features: users want to have their feedback incorporated in AI-generated recommendations to meet their unique needs, they want to have humanlike experiences when interacting with conversational agents, and they want accuracy when AI is used for prediction or recognition. As our results illustrate, often apps' performances deviate from users' expectations. These deviations lead to dissatisfaction and frustration. These expectations are aligned with general design principles, e.g., providing humanlike experiences[21] and incorporating users' feedback[22] have been researched by Human-computer Interaction studies. Previous research has also identified that users' expectations do not match systems' capabilities, even when systems do not use (or do not describe using) AI[23,24]. However, these expectations are often not directly approached when discussing AI systems design.

Previous work investigated how users interact with common AI-powered technology, proposing general design guidelines for human-AI interaction[25]. Although this study did not include direct-to-consumer health systems, many of the proposed guidelines can be useful in the mobile healthcare domain. However, our findings suggest that it is also important to consider the specificities of the domain, the feature, and what users may expect from them, particularly when the tool employing AI is directly aimed at consumers. For example, it is important to clearly state not only the

accuracy of the tool but also what are the specific potential negative consequences to users' health, which are specific to each AI feature and health domain.

In general, the healthcare domain adds complexity to human-AI interaction and its consequences. Recommendation features are commonly used in all seven health domains, and especially common in fitness apps. While using such AI-powered apps, users might hurt themselves if they blindly follow AI's recommended workout plan, as suggested by some user reviews. Similarly, conversational agent features, used in most mental health apps, can contribute to negative emotions when they do not meet user's expected humanlike experience, as also suggested by user comments. For recognition features, used in nearly all meditation and sleep apps, inaccuracy can reinforce practices that conflict with evidence-based methods, as previous studies reported for the case of sleep health[23]. Prediction features, which all Fertility/menstrual apps used, should support users to understand how accurate the output can be, or they can contribute to frustrations for people who are trying to conceive, as users' comments suggest, or even potentially lead to unplanned pregnancies[26]. To avoid potential negative consequences and better support healthcare consumers, we encourage designers, researchers, and healthcare providers to collaborate to develop human-AI interaction guidelines specific for each AI feature and each health domain.

Facilitate healthcare consumers to understand AI features

AI algorithms can be complex and AI-powered technology often is designed as a black box[27] as it is challenging to clearly convey to users how such systems work[25]. In our study, we found most AI-powered consumer health apps do not describe how their AI feature works other than providing broad claims of how "smart" they are. The app reviews reveal that many users are confused about the AI features and are doubtful about how useful and accurate they are, and even question whether the app uses AI. Previous studies show that the lack of user understanding is common for AI design[27,28]. What we found suggests that users are confused but they want to understand how (well) the AI feature works, so they speculate about AI, interpret their experiences, and experiment with the apps.

These three types of efforts to understand AI may relate to users' experiences, backgrounds, and knowledge about health, technology, and AI. For instance, it could be challenging for a person with basic technology and AI understanding to perform detailed experiments with the apps to analyze how well the AI feature works. Users seem to have a healthy skepticism about AI features, but the lack of transparency can limit the benefits gained from such apps. Moreover, app descriptions are currently written more for marketing purposes than to help users understand the algorithm and interpret the outputs. Our study suggests that poor descriptions of AI features contribute to the differences between users' mental models and apps' real capabilities. However, users build understanding about their health and make decisions informed by these apps, and AI algorithms. To that end, they need to understand how trustworthy the information they are receiving is.

Recently, a growing body of research is focusing on how to explain AI algorithms to facilitate users' understanding of how (well) such algorithms work, suggesting approaches such as to incorporate explainability features while developing AI systems[29], to develop a framework for user-centered explainable AI[30], and to explain AI decision-making process through user interface design[31]. However, these studies have not yet focused on non-expert users (domain and technology-wise) such as patients and caregivers. Our study suggests that it is important to provide users clear information about apps' AI and explain how (well) they work, particularly because these apps can influence users' health in many different domains, from fitness and diet to mental health and fertility. We also suggest designing and describing AI features in mobile apps in such a way that it is engaging and easy to understand considering users' different backgrounds and technology and health literacy.

Healthcare is a specialized and often high-risk context per se. Although most of the analyzed apps approach relatively less complex health domains (in comparison with AI systems for clinical use), users can still make important life-decisions based on the outputs of these apps and their AI features. We want healthcare consumers to benefit from good algorithms as they can support them in various ways, however, we also need to support users to be aware of how much they can rely on these algorithms. To achieve that, it is important to avoid overly positive descriptions and to make it clear the real capabilities of AI-powered systems, describing their level of human-like experience, their ability to incorporate feedback, and their accuracy clearly. Transparency is important so users do not under- or over-rely on AI. Under-reliance would lead to users stop using such apps and, therefore, not getting their potential benefits. Over-reliance is a possibility if users do not understand the underlying algorithm and the extent it should be trusted. If these issues are not addressed, negative consequences could happen across all health domains, such as reinforced negative mental health and unexpected pregnancies. It is also important to emphasize that these tools will produce different results for different people[24]. For example, in the fertility context, a prediction feature may not get as accurate for some users no matter how much data they track because fertility is not totally predictable. Similarly, previous studies

have reported that sleep apps can recognize aspects of sleep quality to between 81-83% accuracy[23]. We encourage all stakeholders who are part of creating AI-powered mobile health apps to further facilitate user understanding of how (well) the AI works for future products.

Limitations

The list of AI-powered health apps we collected might not be exhaustive and apps that use AI but did not describe as using AI may have missed in the search. We analyzed all the apps that met our eligibility criteria; however, it is challenging to assure if these apps constitute a reprehensive sample of similar apps available in the market. Due to the large number of the selected apps, we only analyzed each app's title, description, and screenshots provided in the app stores, and we did not download and use all the apps or test out their algorithms. We also did not conduct interviews directly with users of the selected apps and only analyzed user reviews. We understand this approach has its limitation: it is difficult to validate whether the reviews were posted by unique users and how representative the reviews are for real users. Further research is needed to investigate user perceptions of the AI features from more representative user groups.

Conclusion

Although prior works have studied how AI-powered technology has been used by health professionals, our analysis focused on understanding healthcare consumers' views on AI-powered mobile health apps. This study, using a qualitative approach, identified seven predominant health domains and four AI features that are being used in the state of art AI-powered health apps. Our findings indicate that users have specific expectations toward each AI feature. We also find most users do not understand AI or how (well) AI works in the health apps due to the lack of such information and explanations. We hope researchers, designers, engineers, and healthcare professionals who are part of the creation process of AI-powered mobile health apps can design AI features according to users' expectations and needs as well as further explore ways and techniques to assist non-expert end-users understanding AI in mobile health apps.

Acknowledgments

We thank the SURF-IoT program at the University of California Irvine to provide support for this research, and Rebecca Black for feedback to the early draft of this paper.

References

1. The Growing Value of Digital Health [Internet]. [cited 2020 Mar 11]. Available from: https://www.iqvia.com/insights/the-iqvia-institute/reports/the-growing-value-of-digital-health
2. Smith JC, Schatz BR. Feasibility of Mobile Phone-Based Management of Chronic Illness. AMIA Annu Symp Proc. 2010;2010:757–61.
3. Klasnja P, Consolvo S, McDonald DW, Landay JA, Pratt W. Using Mobile & Personal Sensing Technologies to Support Health Behavior Change in Everyday Life: Lessons Learned. AMIA Annu Symp Proc. 2009:338–42.
4. Berglund Scherwitzl E, Gemzell Danielsson K, Sellberg JA, Scherwitzl R. Fertility awareness-based mobile application for contraception. Eur J Contracept Reprod Health Care. 2016;21:234–41.
5. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health. 2017;4:e19.
6. Lang M, Zawati MH. The app will see you now: mobile health, diagnosis, and the practice of medicine in Quebec and Ontario. J Law Biosci. 2018;5:142–73.
7. Luger E, Sellen A. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems [Internet]. San Jose California USA: ACM; 2016 [cited 2020 Mar 4]. p. 5286–97. Available from: http://dl.acm.org/doi/10.1145/2858036.2858288
8. Liao QV, Gruen D, Miller S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. ArXiv200102478 Cs [Internet]. 2020 [cited 2020 Mar 11]; Available from: http://arxiv.org/abs/2001.02478
9. Forsythe DE. Blaming the User in Medical Informatics. Studying Those who Study Us: An Anthropologist in the World of Artificial Intelligence. Stanford, California: Stanford University Press; 2001. p. 1–15.
10. Schulz PJ, Nakamoto K. Patient behavior and the benefits of artificial intelligence: The perils of "dangerous" literacy and illusory patient empowerment. Patient Educ Couns. 2013;92:223–8.
11. Purpura S, Schwanda V, Williams K, Stubler W, Sengers P. Fit4life: the design of a persuasive technology promoting healthy behavior and ideal weight. In: Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11 [Internet]. Vancouver, BC, Canada: ACM Press; 2011 [cited 2020 Mar 16]. p. 423. Available from: http://dl.acm.org/citation.cfm?doid=1978942.1979003

12. Yang Q, Steinfeld A, Zimmerman J. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems [Internet]. Glasgow, Scotland Uk: Association for Computing Machinery; 2019 [cited 2020 Mar 8]. p. 1–11. (CHI '19). Available from: https://doi.org/10.1145/3290605.3300468
13. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol. 2017;69:2657–64.
14. Caldeira C, Chen Y, Chan L, Pham V, Chen Y, Zheng K. Mobile apps for mood tracking: an analysis of features and user reviews. AMIA Annu Symp Proc AMIA Symp. 2017;2017:495–504.
15. Glaser BG. The Constant Comparative Method of Qualitative Analysis. Soc Probl. 1965;12:436–45.
16. Palomba F, Linares-Vásquez M, Bavota G, Oliveto R, Di Penta M, Poshyvanyk D, et al. User reviews matter! Tracking crowdsourced reviews to support evolution of successful apps. In: 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME). 2015. p. 291–300.
17. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3:77–101.
18. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc. 2018;25:1248–58.
19. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care. Health Aff (Millwood). 2014;33:1139–47.
20. Costa Figueiredo M, Caldeira C, Reynolds TL, Victory S, Zheng K, Chen Y. Self-Tracking for Fertility Care: Collaborative Support for a Highly Personalized Problem. Proc ACM Hum-Comput Interact. 2017;1:1–21.
21. Mutlu B, Forlizzi J, Hodgins J. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In: 2006 6th IEEE-RAS International Conference on Humanoid Robots. 2006. p. 518–23.
22. Ritter W. Benefits of Subliminal Feedback Loops in Human-Computer Interaction [Internet]. Vol. 2011, Advances in Human-Computer Interaction. Hindawi; 2011 [cited 2020 Mar 22]. p. e346492. Available from: https://www.hindawi.com/journals/ahci/2011/346492/
23. Ravichandran R, Sien S-W, Patel SN, Kientz JA, Pina LR. Making Sense of Sleep Sensors: How Sleep Sensing Technologies Support and Undermine Sleep Health. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17 [Internet]. Denver, Colorado, USA: ACM Press; 2017 [cited 2020 Mar 25]. p. 6864–75. Available from: http://dl.acm.org/citation.cfm?doid=3025453.3025557
24. Yang R, Shin E, Newman MW, Ackerman MS. When Fitness Trackers Don't "Fit": End-user Difficulties in the Assessment of Personal Tracking Device Accuracy. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing [Internet]. New York, NY, USA: ACM; 2015 [cited 2019 Sep 25]. p. 623–634. (UbiComp '15). Available from: http://doi.acm.org/10.1145/2750858.2804269
25. Amershi S, Inkpen K, Teevan J, Kikin-Gil R, Horvitz E, Weld D, et al. Guidelines for Human-AI Interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19 [Internet]. Glasgow, Scotland Uk: ACM Press; 2019 [cited 2020 Mar 8]. p. 1–13. Available from: http://dl.acm.org/citation.cfm?doid=3290605.3300233
26. Sudjic O. 'I felt colossally naive': the backlash against the birth control app. The Guardian [Internet]. 2018 [cited 2020 Mar 25]; Available from: https://www.theguardian.com/society/2018/jul/21/colossally-naive-backlash-birth-control-app
27. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent Rep. 2019;49:15–21.
28. Wang F, Kaushal R, Khullar D. Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine? Ann Intern Med. 2020;172:59.
29. Ribeiro MT, Singh S, Guestrin C. "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 [Internet]. San Francisco, California, USA: Association for Computing Machinery; 2016 [cited 2020 Mar 7]. p. 1135–1144. (KDD '16). Available from: https://doi.org/10.1145/2939672.2939778
30. Wang D, Yang Q, Abdul A, Lim BY. Designing Theory-Driven User-Centric Explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems [Internet]. Glasgow, Scotland Uk: Association for Computing Machinery; 2019 [cited 2020 Mar 7]. p. 1–15. (CHI '19). Available from: https://doi.org/10.1145/3290605.3300831
31. Cheng H-F, Wang R, Zhang Z, O'Connell F, Gray T, Harper FM, et al. Explaining Decision-Making Algorithms Through UI: Strategies to Help Non-Expert Stakeholders. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: ACM; 2019 [cited 2019 Sep 26]. p. 559:1–559:12. (CHI '19). Available from: http://doi.acm.org/10.1145/3290605.3300789