## Research findings

Interview and Usability Study

# Analysis of AI-powered personal health assistant apps in order to determine ways to improve user satisfaction.

Angela Yung, Saransh Solanki, Sayena Majlesein

HCID 530 A - Usability and User Research

# Content

# Introduction

A chatbot, by definition, is a software system which can interact or "chat" with a human user in natural language. For our research study, we focused on AI-powered chatbots and their role as virtual health assistants. Ada, an AI-powered health assistant, claims to be an app that is able to bring personalized healthcare to its users. It is marketed as a "smart phone-doctor app". [1]

The doctor-patient relationship is a vital component of the healthcare system, and its effectiveness directly impacts the quality of care, achievement of a successful treatment, and patient satisfaction. [2] In order to replicate the real-life doctor experience, using these apps would need to feel as though the user is developing a similar rapport.

We aimed to look into the success of such health assistants to replicate the real-life doctor experience, the situations when usage of these apps are most effective, and the underlying reasons which make patients not trust them. In order to do this, we gathered information about users' experiences with both their General Practitioner (GP) and the health assistant app.

**References**

1 Brown, A. (2018). This Startup Wants To Replace Your Doctor With A Chatbot.
Retrieved from https://www.forbes.com/sites/forbesdigitalcovers/2018/06/29/this-startup-wants-to-replace-your-doctor-with-a-chatbot/#166349e32fe4

2 Kaba, R., & Sooriakumaran, P. (n.d.). The evolution of the doctor-patient relationship.
Retrieved from https://www.sciencedirect.com/science/article/pii/S1743919106000094

# Objective of the research study

1. Find out more about users of AI-powered personal health assistant apps and the situations these users find it most effective for

2. Identify the most impactful factors that contribute to user satisfaction when using an AI-powered personal health assistant app

3. Identify the current pain points of using an AI-powered personal health assistant app as a medical care companion

# Recruitment and participant profiles

All participants were recruited via a survey sent to several social media platforms. The study was comprised of 5 participants, all of which matched the following criteria:

- 18-30 years of age
- Has at least a moderate knowledge of technology
- Has access to Android or iOS devices
- Recently (in the past 3 months) visited their primary care doctor to seek medical advice
- Have used Ada in order to obtain additional medical information.

# Research methods

Both, the interview and usability study, were designed to be done in-person or remotely. Zoom application was used to conduct the interview and the usability study over video calls and to record participants' phone screens.

## Interview

The session began with the interview which allowed us to obtain a better understanding of our users' experiences with real-life doctors. We used elements of directed storytelling to get a narrative about their experiences with finding a doctor and their most recent appointment with their GP. Participants were asked questions to about what made them use an AI personal health assistant and their experience and impressions of the app.

Information was gathered about both their experiences with real-life doctors as well as AI personal health assistants to pinpoint the major discrepancies between the two experiences.

## Usability study

After the interview, participants were instructed to remember and input the symptom(s) of their most recent condition that prompted them to see a doctor. Participants were instructed to think out loud as they walked us through the application and explained how they used the features and services. The usability study was conducted to build an effective contextual communication between our problem space and participants' experiences and interactions with AI-powered personal health assistant apps, and thus we modified the usability study to better understand participants' conceptual models towards these applications.

The scenario was broken down into 8 sub-tasks and their success criteria.

1 **Open the mobile application**

The participant was able to open the app and log in to get started with the assessment.

2 **Locate symptom checker**

The participant was able to locate the input box and rightly understand its purpose to enter symptoms.

3 **Enter required personal information**

The participant was able to understand the need to enter personal details and correctly enter them.

4 **Enter symptoms**

The participant was able to enter all of her symptoms correctly and without any confusion. She was able to understand each question asked by Ada, its purpose and how it was going to influence the assessment.

5 **View and comprehend the received response**

The participant saw the assessment of the symptoms from Ada. She understood it correctly and it matched with the diagnosis the doctor had done for the same symptoms.

6 **Explore the detailed diagnosis page**

The participant, if needed, entered the detailed diagnosis page and saw more details of the assessment.

7 **Close the application**

After the interview and usability study users were asked following post-test questions:

- What do you think is the purpose of this app?
- How do you feel about this app?
- How much do you trust or distrust this app?
- What would make you trust this technology (in general) more?

The main objective of conducting a usability study was to understand how users were interacting with the app and whether there were any major usability issues that hindered users from having a successful interaction with the app.

## Data collection

An assistant was in charge of setting up the screen capture in order to record the user's screen during the usability test and making sure the sound recorder set up and turned on for the duration of the session. The facilitator conducted the interview and administered the usability testing exercise. During the interview and usability testing, the note taker documented quotes and notes from the participant's responses and actions.

# Analysis of data

## Interview

In the process of analyzing the interview data, the session notes were coded. After coding, we created the following affinity map for a visual representation of the results.

**Ada was not comprehensive**

- Asked many strange and irrelevant questions **P3**
- Typing my symptoms... no results were found. **P5**
- I could not find my symptoms, had to chose something close. **P3**
- Unable to add to symptoms **P5**
- Not able to select my correct symtoms, i'm sure it would not give the correct results. **P5**

**No guidance on what to do next**

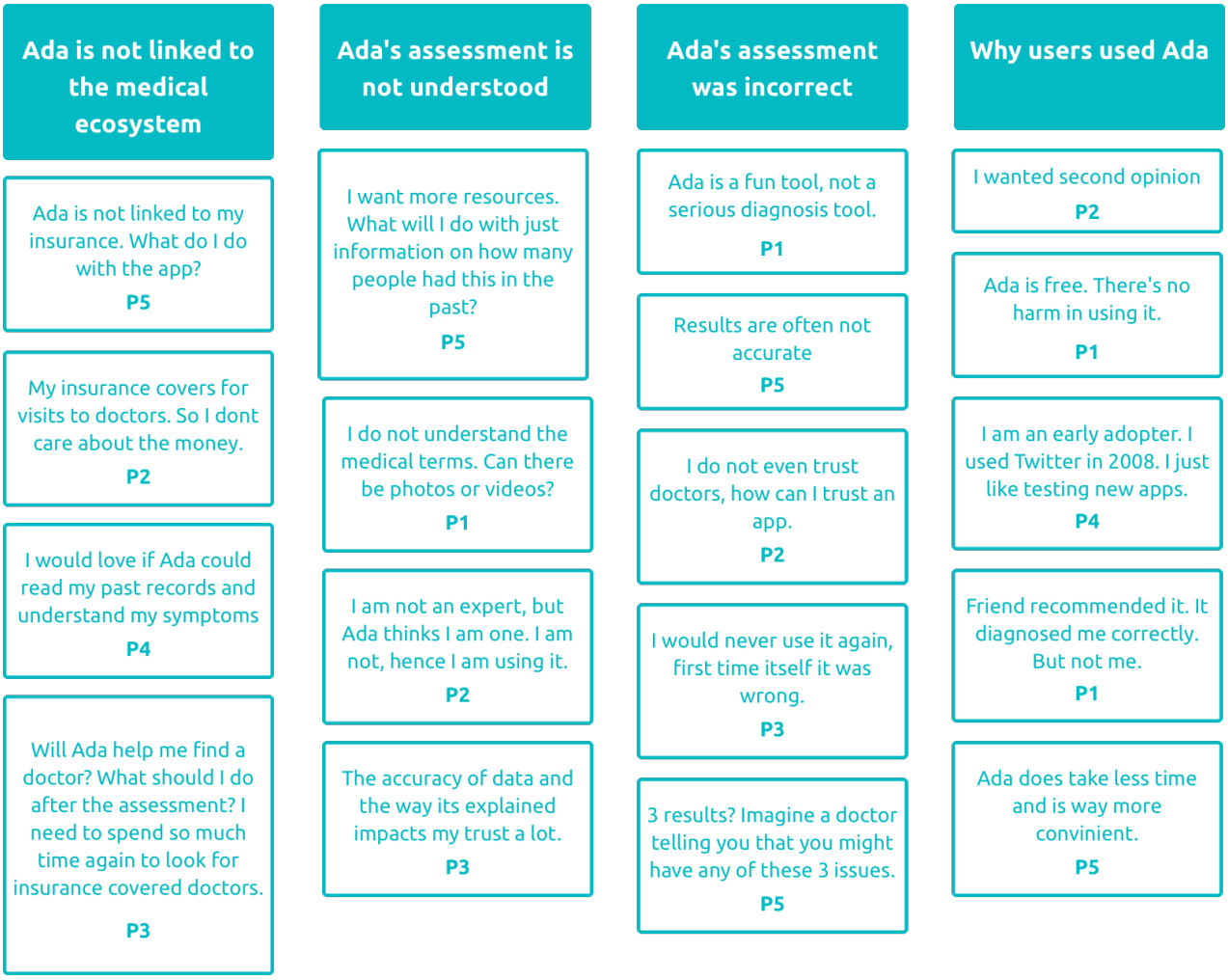
- I just wasted 15 minutes on Ada. What next? Go back and search a doctor. **P5**
- The app doesn't give me what to do next. I feel helpless. **P4**
- Doctor always tells me steps for the future. I can always ask them if I have doubts. **P1**

**Doctor is trusted more than Ada**

- Ada is not a replacement for doctor. How come then there is no link to a doctor on the app. **P5**
- Analysis of peptic ulcer was only possible after doing tests and running examinations. **P2**
- Ada does not get all my vital signs. **P4**
- No way I'm going to trust it. Can it tell me I'm pregnant? No. **P1**
- What I feel and what I say is always different. **P4**
- But Ada does not know I had fever last week. **P3**

**Ada doesn't explain like a doctor**

- A doctor listens carefully to what I explain. **P2**
- The app is not confident enough to give one answer, how would the user trust it? **P2**
- I would use this only as a second opinion app **P4**
- Feels like I'm talking to a machine **P1**




Ada's AI service and logical flow are among the highest rated in the field, however, Ada is a misunderstood product. The recent tech-blog coverage refers to Ada as a "smartphone doctor app", which users do not relate to. There are many facets and intricacies of a doctor-patient relationship which the AI-bot is not able to replicate. The experience felt mechanical and tech-driven rather than patient centered to the users. Ada's role in the current medical journey of a user is very ambiguous. Although Ada is currently being used by more than 3 million users across the globe, most of them turn out to be technology-savvy and the app has not yet reached mainstream adoption.

# Usability study

Following the usability study, each sub-task was analyzed based on the following success criteria and usability metrics.

### Success criteria for each sub-task

The sub-task was completed successfully when the participant indicated the task's goal had been obtained. If the participant was not able to complete the task or requests and received sufficient guidance from the interviewer to move forward, the sub-task was marked as unsuccessful.

### Usability metrics

We defined the following usability metrics and classification to better address usability issues and errors that we found in our study.

#### Critical errors

Critical errors were unresolved errors during the process of completing the task or errors that produced an incorrect outcome.

#### Non-critical errors

Non-critical errors were errors that the participant was able to recover from, or if not detected, did not result in incompletion of the task. These errors may have been procedural, in which the participant did not complete the scenario in the most optimal means.

### Sub-tasks analysis

The scenario was broken down into 8 sub-tasks.

**1 Open the mobile application**

Successfully completed by all 5 participants

**2 Locate symptom checker**

Successfully completed by all 5 participants

**3 Enter required personal information**

Successfully completed by all 5 participants

**4 Enter symptoms**

Successfully completed by 1 participant

4 of the 5 participants had a query regarding how to proceed. The interviewer hence gave hints and guidance to them so that they could continue with the scenario task.

**Issues found**

| | |
|---|---|
| **Non-critical error** | Only allows users to input the closest matching symptoms [4/5 participants] |
| **Critical error** | Unable to edit old symptoms that have been previously input [3/5 participants] |
| **Critical error** | Unable to go back to an old question without closing the app [1/5 participant] |

## 5 View and comprehend the received response

Successfully completed by 2 participants

3 of 5 participants had different issues regarding the assessment.

**Issues found**

| | |
|---|---|
| **Critical error** | Ada diagnosis different from what their doctor had stated [2/5 participants] |
| **Critical error** | Did not understand the assessment and were looking for more insights or information [2/5 participants] |
| **Non-critical error** | Needed information on how Ada come up with this assessment. Wanted to see if anyone of their entered symptoms was wrongly mentioned. [1/5 participants] |

## 6 Explore the detailed diagnosis pages

Successfully completed by 2 participants

As it was needed by the participants, they explored Ada for detailed information about the assessment. This included them entering the detailed diagnosis results page. 3 of 5 participants wanted to see more information.

**Issues found**

| | |
|---|---|
| **Critical error** | The explanation by Ada was around demography and very jargonish. [1/3 participants] |

## 7 Close the application (Ada)

Successfully completed by all 5 participants

# Insights and recommendations

## Recommendation Summary

- The service should support natural language conversations using voice-based conversational interface
- Enable photo and video input methods to impute symptoms and other vital signs easily and comprehensively
- Allow users to edit their previously entered symptoms easily.
- Generate easy and comprehensible demonstration of results and diagnosis using photo and video descriptions
- Provide users with tips and guides on further actions
- Store and contextually use patient's information and medical history to provide a better diagnosis
- Link the process of assessment with a real-life doctor (could be done remotely as well) to build trust in the system and the diagnosis
- The service should work alongside the current healthcare ecosystem. Link it to insurance providers, healthcare providers, and pharmacies to make the process of finding care more streamlined.

# Priority rating key

Priority ratings were assigned to our recommendations in response to the insights that we found from the interviews and usability studies. This rating scale provides information on the urgency of our recommendations. The ratings are based on the following criteria:

- **Impact on the user's satisfaction**
  Does the solution tackle a problem that left them feeling unsatisfied with the app?

- **Frequency of users mentioning the issue**
  Is this a solution to a problem for many participants?

### High priority

This rating is an indication that the solution targets an issue which has a severe impact on the user's satisfaction and a lot of users voices similar complaints. This rating means that attention and remediation of this issue, using this recommendation, is critical.

### Medium Priority

This rating is an indication that the solution targets an issue which has moderate impact on a user's satisfaction and that 3-4 of our participants have voiced concerns related to this issue.

### Low priority

This rating is an indication that the solution targets an issue which has low impact on a user's satisfaction but was an issue mentioned by 1-2 participants. Low priority items are minor.

# Users do not feel that symptom input on Ada is a comprehensive process

Ada's AI logical structure does not match users' mental models. Often times it asks questions which feel unnatural to the user. Participants also mentioned that Ada's symptom selection was not comprehensive and only allowed users to input closest matching ones. When this was the case, participants became doubtful of the accuracy of the results. Participants believed that trusting only what patients say without analyzing the other plethora of vital signs and medical information could lead to an incorrect diagnosis. As a result, participants were more likely to trust their doctor over the app when there is a discrepancy between the two opinions.

> **"I could not find my symptom so I had to choose something similar" P4**
>
> **"It asked many strange and irrelevant questions. Only one question seemed useful" P3**
>
> **"When my symptoms were not found I thought this won't work" P5**
>
> **"Ada is not able to get all my vital signs" P4**
>
> **"There is an app where you can take a picture of your skin and it tells you about your condition, I like that" P2**

**Design Recommendation**

**High priority** Allow users to input symptoms without a great deal of cognitive load by providing understandable cues and hints. The app should also allow users to input vital signs and symptoms using photos or videos. The ability to provide the service with additional information will allow them to feel more confident about the diagnosis.

# The information that Ada provides about the diagnosis is too technical.

The language used in the app was very technical which many users felt was unpalatable. They found it to be full of jargon, and as a result, it was not effective in providing useful information. In general, participants said they wanted resources and education about their medical conditions. While many doctors do not provide sufficient information post-diagnosis, participants felt that Ada provided less in the way of resources and feedback. Due to these limitations, participants believed the app to be insufficient even as a supplement to their doctor's opinion. Participants would feel more comfortable with the app if they were provided with photos or videos to aid diagnosis and condition explanations.

> **"The user is not an expert, but the app explains things as if the user is expected to know these things." P2**
>
> **"I wanted more resources from my doctor but he just diagnosed me, gave me a prescription, and left" P3**
>
> **"I don't really understand the terms so having photos and videos would help" P2**

**Design Recommendations**

| | |
|---|---|
| **High priority** | Provide the diagnosis and results of the assessment in easy and comprehensible terms. |
| **Medium priority** | Along with the current information (which is more around the demography of the past patients with the diagnosis), Ada needs to provide users with supplemental photo and video descriptions which help them understand the assessment better. |

# The interaction with Ada is not natural and does not feel like they are talking to a real doctor.

Participants mentioned that the aspects of their interactions with real doctors that foster trust do not occur when they are using Ada. For example, typing to a bot felt like an interaction with a machine. At a real doctor's appointment, you would able to ask the doctor further questions or request additional feedback. Ada currently does not use a holistic approach when gathering information to suggest possible diagnoses or to recreate the doctor-patient relationship. Relying on what patients say without analyzing the other plethora of vital signs and medical history could potentially lead to an incorrect diagnosis. In addition, Ada currently lists several possible diagnoses. In a written format, it was very noticeable to users and negatively impacted their trust.

> **"It feels like I'm talking to a machine. If you could speak to it, it might feel more trustworthy." P2**
>
> **"I don't know which one is my condition. I would still need to visit a doctor after using the app," P1**
>
> **"3 results? Which one is correct? Which one should I trust? Imagine a doctor telling you 'you have 3 possible issues'" P5**

**Design Recommendations**

**High priority** We recommend switching the Ada format to a voice interface so users can sustain a conversation with the AI assistant and ask questions. Being able to talk about their symptoms is a more natural interaction, wherein Ada can recreate a lot of the intricacies of the doctor-patient communication.

**Medium priority** During their assessment, doctors often come up with multiple possible diagnoses. However, they construct the next steps in a manner that doesn't make the patient realize the ambiguity of the process. Ada should replicate this recommendation process during their assessment.

# Users do not trust Ada because they are unsure of its role in their healthcare.

Many of our participants used Ada as a second opinion. During the usability tests, a few users got a diagnosis that conflicted with their doctor's opinion. They were quick to judge the app and were not likely to use it again due to the mistake. Another notable complaint was the app was not able to provide "next steps" or any additional feedback. Users simply did not know what to do with the information obtained from the app. At the same time, Ada does not keep a log of your assessments and diagnoses. It does not correlate with 2 adjacent diagnoses to find possible patterns. While Ada is marketed as a smart-phone doctor, users are finding that this is not an effective use of the current technology. Due to this major limitation, many participants were confused on what Ada's purpose is and how it fits into their already established medical system. Participants mentioned that if the app was somehow linked to their doctor, they would be more likely to trust it.

> **"I just wasted 15 mins on ADA. What next? Go back and find a doctor." P4**
>
> **"This is not the diagnosis my doctor gave. I don't trust Ada anymore" P3**
>
> **"I would trust it more if it was related to my doctor's appointment or if it was endorsed by my doctor" P3**
>
> **"I used Ada a week back to find out I had viral flu. Now I have headaches. Ada can't tell that they were due to the same issue." P4**

**Design Recommendations**

| | |
|---|---|
| **High priority** | Give users next steps and link the app with a platform for them to connect to a doctor and validate these results, similar to a triage doctor. |
| **Low priority** | Store patient's health records and previous assessments to give a better and more accurate diagnosis. |
| **Low priority** | Link Ada with their insurance provider to simplify finding care process. |

# Usability Issues

Participants pointed out that you are unable to go back to an old question without closing out of the app and you are unable to edit old symptoms that you had previously input. There were no other major usability issues that hindered participants from being able to successfully use the app. Most complains were about the output of the app rather than the usability.

> **"I can't edit/add to old symptoms. What if I wanted to add to an existing condition?" P3**
>
> **"When you press the back button, the app closes" P5**

**Design Recommendation**

| | |
|---|---|
| **Low priority** | Allow users to go back to change answers when inputting symptoms/answering questions |
| **Medium priority** | Allow participants to go back to old symptoms and add to it (for chronic or long-lasting conditions. |

# Study Limitations

## Only Ada users

We initially planned to test multiple health assistants but all of our participants happened to use Ada. We were able to gain many insights from just Ada so we focused our research on this particular app.

## Remote Testing Issues

Given the time constraints of the project and the low usage rate of Ada in the US market, a few of our users had to be tested remotely. Remote testing required additional commitment from the participants to download applications and complete steps online. This proved to be a hassle and time consuming for our participants. We also had some issues viewing the screens while they were completing user testings. Zoom user had a lag so the screens were not in real time.

## Lack of diversity in the participation group

All 5 participants identified themselves as early adopters of technology. Some of them tried Ada because of it getting featured in various tech blogs like Wired, which got their attention. We don't think that the issues identified are exclusive to only these set of users, however, a more diverse group (different nationalities, cultures, demography) could have given us some more insights. The economy impacts the problems and needs around healthcare so future research could explore how Ada's purpose changes according to these factors.

## Voice interface requires additional research

While switching to a voice interface can help Ada improve its interactions with users, we were not able to gain information in this study about what form the voice interface should take. Future research should include focusing on both secondary research and user's preferences in relation to voice interfaces.

# Conclusion

In attempting to determine ways to improve user satisfaction with AI-powered personal health assistant apps, we studied the quality of the user's experience when interacting with health assistant applications.

Users unanimously mentioned the gap between the experience offered by Ada and a real-life doctor experience. Ada is not patient-centric and does very little to replicate the intricacies of the current healthcare journey. Participants have yet to adapt it into their daily lives because it adds additional effort with little benefit. In addition, Ada still feels like speaking to a machine rather than a doctor. It lacks the personal connection with users and as a result fails to fosters the doctor-patient rapport. Due to the impersonal process, users are hesitant to trust Ada. The usability study results showed that it was not usability, but rather the perception of the app that led to low adoption rate.

These issues can be mended by designing a process which keeps the users in mind. To address this issue, we recommended providing a more effective and robust means to communicate with the app. We propose moving Ada towards a voice-based conversational user interface to establish a more natural interaction between users and the app. In addition, we believe that Ada can explicitly be bridged to the current healthcare system by assisting the user with the steps to do next. By linking the app to a doctor would further increase credibility and trust towards Ada and its diagnoses.

In conclusion, we hope that by proposing these key recommendations, we can reduce the gap between the current healthcare system and Ada. Furthermore, resolving the above-mentioned issues and addressing these problems can lead to a more trusted experience with AI-powered personal health assistant apps and can also effectively enhance user satisfaction.