The Pennsylvania State University

The Graduate School

SELF-DIAGNOSIS THROUGH CHATBOT-BASED SYMPTOM CHECKERS: USER EXPERIENCES AND DESIGN CONSIDERATIONS

A Thesis in

Informatics

by

Yue You

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

May 2020

The thesis of Yue You was reviewed and approved* by the following:

Xinning Gui
Assistant Professor of Information Sciences and Technology
Thesis Advisor

Mary Beth Rosson
Professor of Information Sciences and Technology
Head of the Graduate Program of Information Sciences and Technology

Saeed Abdullah
Assistant Professor of Information Sciences and Technology

ABSTRACT

In recent years, there has been a growing interest in developing Artificial Intelligence (AI)-based chatbots in the healthcare market, which use human-like conversations to interact with users. One popular type of AI-based chatbots is an AI-based symptom checker (AISC) app, which provides potential diagnoses for users and assists them with self-triaging. Despite the popularity of such AISC apps and their high ratings in major app stores, little research has been undertaken to investigate users' perception, accountability, transparency, and data policies of AISC apps. To investigate AISC apps and explore how users evaluate and perceive the effectiveness of AISC apps, we conducted a feature review, a review analysis, and an interview study. We found that existing AISC apps lack credentials and transparency to verify their credibility and safety, which brings challenges to users' trust and privacy protection. We also found that users evaluated AISC apps by comparing their experiences of using AISC apps with offline consulting experiences. Users perceived existing AISC apps lacked support for diverse diseases and user groups, flexible symptom inputs, user-friendly conversations, and comprehensive health history. Based on the results, we derived discussions and implications for AI policymaking, AI algorithms design, and conversational design of healthcare chatbots.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I wish to show my gratitude to my advisor Dr. Gui and Dr. Rosson. Dr. Gui has provided for me helpful instructions and suggestions on literature review, methodology study, thesis organization, and whole thesis writing. Her advice was essential to the completion of this thesis and has taught me insights on the workings of academic research in general. I would also like to pay my special regards to Dr. Rosson for her help during my studies in the field of Human-Computer Interaction.

My thanks also go to Dr. Abdullah for reviewing this thesis, Dr. Hume, and Erica Mi for proofreading this thesis. This thesis would not have been possible without the assistance of them.

Last, but not least, I would like to thank my parents and my boyfriend for their understanding and love during the past few years. Their support and encouragement were in the end what made this thesis possible.

Chapter 1

Introduction

Recently, AI-based healthcare chatbots (AIHCs) have proliferated in the mobile application market. AIHCs are virtual conversational agents utilizing natural language processing techniques to mimic human interactions in the healthcare domain [43]. The AIHC with the function of a symptom checker (in this paper called "AI-based symptom checker" (AISC)) can assess users' symptoms and give diagnoses by using human-like conversations with users (see Figure **1-1**). The downloads of some AISC apps (e.g., Ada, K Health, and HealthTap) have reached more than 1,000,000 on the Google Play Store based on our calculation. Developers of AISC apps promise various benefits, such as providing accurate diagnosis [25], suggesting medication information [73], and giving triage decisions [83].

However, using AISC apps can pose risks for users. First, using AISC apps may jeopardize the privacy of users' health information. When interacting with AISC apps, users are required to input their personal information, such as age, gender, health history, and symptoms. As AISC apps lack guidelines and strict regulations, it poses a breach of confidentiality that would infringe on users' privacy. Therefore, it is essential to review the data policies of these AISC apps and ensure effective measures for privacy protection. Second, relying on AISC apps inappropriately may result in negative consequences. Outcomes of using symptom checkers can directly make an effect on the wellbeing of users [54]. For example, researchers have found that some symptom checkers have poor performance in identifying common illnesses [41]. Blindly trusting the inaccurate outcomes may put patients' lives at risk, especially for patients with high-risk diseases [106]. Thus, users' perceptions of the accuracy and effectiveness of AISC apps is essential, influencing their trust-building and further decision-making. Considering these risks, the information is important regarding AISC apps' accountability (whether the reliability and validity of these apps can be

guaranteed), transparency (transparency of apps' development process and developers' commercial interests), and data policies that have been disclosed to users. Users should know who is held accountable if such risks happen, emphasizing the importance to review these aspects of existing AISC apps. It is critical to analyze AISC apps using a user-centered approach [54], focusing on users' needs and perceptions to enlighten the further design of AISC apps to decrease potential risks.

Despite the fact that researchers have formulated strategies to evaluate AISC apps [73,91], few studies applied the user-centered approach to explore the effectiveness of AISC apps from users' perspectives. Also, little research examined accountability, transparency, and data policies of AISC apps. There is a crucial need for an empirical understanding of how users actually perceived the accuracy and effectiveness of these AISC apps. We also need to evaluate the accountability and transparency of AISC apps.

To fill these gaps, this work utilized methods of a feature analysis, an app review analysis, and ten semi-structured interviews aiming to inform the policymaking and examine the value of AISC apps from users' perspectives. We first conducted a review on AISC apps (mobile apps embedded with AISC apps) to analyze their history, context, data policies, and functionalities. We then examined how users evaluated these AISC apps by analyzing a set of online reviews in the U.S. Apple App Store and the Google Play Store. Finally, we conducted ten interviews with people who have used AISC apps to cross-validate the online review analysis results. We found doubtful the reliability of AISC apps' developers, the transparency of the source of the provided medical information, the transparency of developers' commercial interests, and the effectiveness of privacy protection. We also found that users noted four main deficiencies of AISC apps: the ignorance of diverse health conditions, rigid input requirement, failures in conversational design, and insufficient consideration of diagnostic information.

Our contributions are four-fold: First, we examined accountability, transparency, and data policies of AISC apps, shedding light on further policymaking in healthcare applications; Second, we identified the strategies users utilized to evaluate AI algorithms, suggesting further AI algorithms design should accommodate diverse contexts; Third, we identified a social-technical gap between the features of AISC apps and users' offline experiences, illustrating the importance of factoring into users' offline experience; Finally, our findings reveal new challenges in the conversational design of health chatbots regarding the human characteristics, input limitations, and the comprehensibility of language. Our work thus identified the factors influencing the services provided by AISC apps and facilitated our understanding of user's needs.

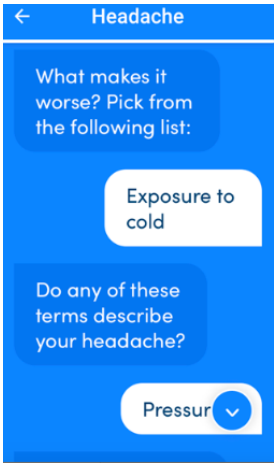

Figure **1-1**: A human-like conversation of K Health.

Chapter 2

Literature review

2.1 Accuracy evaluation for symptom checkers

A symptom checker, also known as a self-diagnosing tool, is a type of consumer-facing digital health tool that has emerged in past decades. Symptom checkers provide potential diagnoses and assist with triage (e.g., inform users whether they should seek medical care), by asking users to answer a series of questions regarding their symptoms [91].

The existing body of research focuses on evaluating the accuracy of symptom checkers. Several studies have evaluated symptom checkers' diagnostic accuracy by engaging multiple medical experts to assess checkers' features and performance through clinical vignettes [3,44]. While Babylon Health Inc. claimed their symptom checker, Babylon, has accuracy comparable to medical experts [62,83], a large amount of research has found that symptom checkers are less accurate than physicians. In 2015, the *British Medical Journal* evaluated twenty-three symptom checkers, disclosing that checkers have proposed the correct diagnoses only 34 percent of the time [91]. One study also found physicians outperform the symptom checkers through clinical vignettes [3]. In addition, the diagnostic capabilities of symptom checkers are inferior to the diagnostic capabilities of medical professionals for certain diseases and conditions, such as HIV or hepatitis C [10], some ophthalmic conditions [71], DCM symptoms [24], and inflammatory arthritis [81].

Very little research has examined the factors regarding to the accuracy of some symptom checkers. To the best of our knowledge, only one study stated that the low diagnostic accuracy of these symptom checkers may be due to an inaccurate database for diseases [64]. Furthermore, although evaluations for symptom checkers should assess the interaction between users and systems [106], few studies have been done from real-world users' perspectives [2,63] to determine

how users perceive accuracy of symptom checkers. Only one study found that an extensive proportion of users hesitated to accept AI-based symptom checkers partly due to their distrust for the accuracy [67]. Thus, it is still unclear how users perceive the accuracy of symptom checkers. Therefore, our study is targeted to explore the users' perceptions of symptom checkers' accuracy.

2.2 Social-cultural analysis for AI-based health applications

When evaluating AI-based health applications ("apps"), it is necessary to understand their integration and implementation in regard to their developers [80]. Several studies examined the socio-cultural background of health apps based on AI algorithms.

Recent research has studied the reliability and accountability of health applications, pointing out that there are no clear measures to assess and certify these applications [32] and guarantee their reliability and quality [34], which may lead to medical errors and unintended harms to users due to the smartphone nature [79]. Researchers also reviewed the application developers [55], regulation and service provision, consent and ownership [80], and the role of health applications in healthcare [52]. These factors can influence the credibility of health applications. These studies all focused on health applications market in general but did not specifically focus on AISC apps to explore AISC apps' reliability and accountability.

In addition, researchers have demonstrated the importance of transparency of mobile health applications. The transparency of the app development processes is desired by users, which is highly relevant to the reliability of mobile applications [89] because users make decisions according to the provided information [32]. Recent studies have found that there are a lack of professional medical involvement in the development process and design of mobile applications [104] and a lack of criteria to assess the professionality or authority of the medical content [79]. Further research is needed to examine users' requirements for the transparency of health applications [104].

Furthermore, previous studies have emphasized that health applications should ensure their security and safety to protect privacy due to data sensitivity and the great number of collected data [6]. Adding to the difficulty was that information is transferred through wireless networks frequently, which may lead to data breach [32]. Some researchers have found that users have concerns about the collection, transfer of their sensitive data [89] and the threats to privacy [36] when using health applications. However, there are few privacy guidelines for the existing mobile applications [6,59].

Even though accountability, transparency, and data policies have been explored by previous studies, few existing studies analyzed AISC apps. Although one study reviewed AISC apps, it did not analyze these apps from various viewpoints, only suggesting guidelines for AISC apps' implementation [80]. Another study gave an overview of chatbot-based health apps, but these apps were mainly used for behavioral changes and not for symptom checking [77]. Other studies examined the medical apps with detail, but these apps were not all in the form of chatbots [52,55].

Further research is essential to explore how developers of health applications consider data privacy, the authority [55], and governance mechanisms [58]. It is still unclear whether the developers of AISC apps have sufficient credentials, disclosed enough information to ensure the transparency, and guaranteed the data security. Thus, we aimed to conduct a social-cultural analysis on AISC apps, reviewing the accountability, the transparency, and the privacy policy of these applications in healthcare.

2.3 The effectiveness of AI-based chatbots

A chatbot is defined as a computer program using Artificial Intelligence techniques, which can be used to interact and communicate with humans [51]. The application of AIHCs spreads to dermatology, hospital, cardiology, nutrition, general practitioner, endocrinology, neurology, and

therapy. Researchers divide AIHCs into two categories: coaching chatbots and counseling chatbots. For the purpose of coaching, a chatbot can coach people to formulate healthy lifestyles and assist nutrition education [30]. For counseling, a chatbot provides health-related information and monitors the patients' conditions [22,42]. Based on this classification, AISC is a type of counseling chatbot.

Previous research has attempted to examine the effectiveness of the second type of AIHCs. They have explored how these AIHCs provide guidance when intervening people's health behaviors [60], assisting cognitive impairment [102], monitoring attention, and suggesting the decision processes for individuals [31].

Researchers also have provided design implications for AIHCs. Researchers have explored if users are satisfied with the interface design of AIHCs. They found users' requirements for the ability to choose voice and interaction styles [105], concise interfaces [30], and color [65]. In addition, several studies proposed design implications for the functions of AIHCs, such as providing a feedback feature [76] and contextualizing requests from users [98].

Aside from the analyses on the function design and interface design, several studies have explored the conversational design of AIHCs. Some studies illustrate enhancing the ability to recognize human inputs and symptoms [8] and factoring into the users' preferences [77,80] are important. However, none of these studies verified their statements and involved users' perspectives. Other studies conducted experiments with users, pointing out that chatbots should improve the tone and speed of the conversations [45] and show affinity as well as empathy [28,40,41]. Other issues influencing the acceptability of users for chatbots are as follows: the technological complexity of chatbots, the difficulty for users to describe symptoms accurately, and the shortage of human presence [67]. However, these studies simply focus on one particular chatbot, whether that is a chatbot for social needs screening [45], a chatbot for mental health interventions [66], or a general health chatbot [49,67]. None of them conduct user-involved testing specifically focusing on the AISC apps.

Beyond existing research, it is still unclear what information or features AISC apps should provide to effectively assess users' symptoms and make informed decisions. For example, the level of algorithms' transparency, the understanding of the context, and "white-box" explanations of algorithms (i.e., explain the inner mechanisms) may be taken into account when assessing a counseling chatbot [17,84]. Furthermore, users' perspectives and capabilities should be considered when evaluating healthcare chatbots' effectiveness [67]. Our study aims to fill these gaps.

Chapter 3

Methodology

To get an overview of the features of the existing AISC apps, we first selected eleven mobile AISC apps (mobile apps embedded with AISC apps) for a feature review through a screening process. We then analyzed the background, the implementation process, accountability, and transparency of selected AISC apps. We also did a critical analysis regarding how these AISC apps deal with users' privacy by analyzing listed data polices of these apps and the functions these AISC apps provide. With an overview in mind, we selected four apps based on pre-defined criteria and investigated their consumer reviews. We intended to understand users' needs and the relationship between users and AISC apps; thus, we can recognize the dynamic nature of the interactions between users and AISC apps and give insights for the features design of AISC apps. We downloaded the online reviews and acquired the ranking information with the help of two app analytics tools (i.e., ASO100 and App Annie). We then conducted thematic analysis for these reviews. To acquire more in-depth insights about how users perceive AISC apps, we also conducted ten semi-structured interviews and before that we utilized a screening survey to filter out unqualified participants. The interviews were then transcribed and analyzed using thematic analysis. Below we offer details regarding each method's data collection and analysis.

3.1 App feature review

3.1.1 Data collection

To obtain qualified AISC apps for the feature review, we searched the Apple app and Google Play stores for AISC apps (free, freemium, and paid) that are accessible within the U.S.

using the following keywords: “medical OR health OR healthcare AND chatbot OR symptom checker.” The initial search returned 70 apps on the Apple App Store and 289 apps on the Google Play Store. We finally selected 11 apps for the feature review through the following screening process (Figure 3-1).

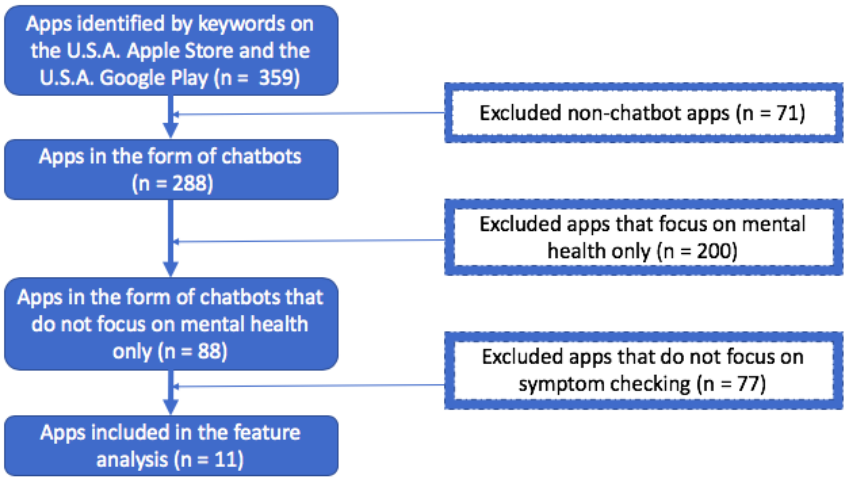

Figure **3-1**: App selection flow graph.

We selected apps for the feature analysis on the basis of the following six criteria: (1) the apps have the chatbot function; (2) the apps are in English; (3) medical diagnosis is the main focus of the app; (4) the apps target health consumers instead of medical professionals; (5) the apps focus on general health conditions (do not focus on mental health only); and (6) the apps can be downloaded and functional. We first excluded 71 apps that do not employ the forms of chatbots. After inspecting these apps, we found that a plenty of apps focus on mental health only. The chatbot of these apps coach users to release users’ mental pressure through conversations. Since in this study we are interested in apps designed for general symptom checking, these 200 apps specifically targeting mental health were excluded. We then filtered out 77 apps that do not focus on symptom checking. Finally, we found eleven AISC apps available for a feature analysis. Ten of these apps (Ada, K Health, Ask NHS, Your.MD, Mediktor, HealthTap, Apothēka Patient, Sensely, Babylon,

and NHS online: 111) are available on both the Apple App Store and the Google Play Store. One app called “Health Buddy” can only be found on the Apple App Store.

After we specified eleven AISC apps, we collected their listed app descriptions, the websites of developers, terms of use, and privacy policy for further analysis.

3.1.2 Data analysis

AISC apps may influence personal health decisions and behaviors. On the one hand, they help users conveniently acquire medical information, serving as an educational role for users. On the other hand, they may cause negative consequences due to unqualified developers and unreliable health information. They may also raise risks in breaching personal privacy data. Therefore, the features, the development processes, the app store descriptions, and the various policies listed by AISC apps can have implications for medical authority, delivery of health knowledge, and the relationship between medical professionals and patients.

Hence, we conducted an app feature review for these eleven AISC apps. A feature review is a common research approach in HCI (Human-Computer Interaction) and health informatics, involving reviewing app history (e.g., ratings, release time, and developers) [7,86], data policies [7,53], and app functions [13,57,85,86,110] to evaluate the validity of certain apps. Following the app feature analysis methodology, we carried out a critical discourse analysis, examining the following attributes for each app: app descriptions listed in the two stores, titles, app developers, logos, screenshots, and developer websites. To have an overview of these AISC apps, we investigated their past history, price, release time, and age limitations. To assess the credibility of these AISC apps, we dug into their credentials, the transparency of their development processes, and the accountability of their developers. The exploration for transparency and accountability helps us detect how and what information these apps should provide for users; this information can

facilitate users to understand potential risks and make more prudent decisions. To analyze how AISC apps use users' data and protect users' privacy, we studied data policies of AISC apps including how they store, transfer, collect, and use data. To shed light on the further feature design and interface design, we downloaded all apps, studied their main features, and then coded them based on the eight stages of a diagnostic process [39]: establishing a patient history, conducting physical exams, evaluating the symptoms, giving an initial diagnosis, ordering further diagnostic tests, performing and analyzing test results, providing a final diagnosis, and providing referrals or other follow-up treatments. We finally developed a coding sheet of feature categories with subcategories.

3.2 App review analysis and semi-structured interviews

3.2.1 Data collection

Review analysis is a process widely used in HCI and health informatics. A body of studies employed this approach to identify how users rated the apps [7] and how users evaluated the apps [70]. Following the app review analysis methodology, we selected apps based on the two following criteria to prevent bias from few individuals: (1) the apps have at least 700 reviews; (2) the apps have at least 1000 ratings. Apps were excluded from the study if they did not have enough reviews. Based upon these criteria, we narrowed our study to four AISC apps (Ada, K Health, Ask NHS, and Your.MD, details illustrated in Table 3-1). For each of the four apps, we analyzed the 500 most-recent reviews.

Table **3-1**: Description of the four apps for review analysis.

App Name	**Ada**	**K health**	**Your.MD**	**Ask NHS**
Category	Google Play Store, Apple App Store	Google Play Store, Apple App Store	Google Play Store, Apple App Store	Google Play Store, Apple App Store
Price	Free	Free/freemium (free three-month trial for chatting with a doctor)	Free	Free
Avg rating (out of 5.0)	4.7(Google)/4.8 (Apple)	4.4(Google)/4.8 (Apple)	4.3(Google)/3.6 (Apple)	4.3(Google)/4.6 (Apple)
Review count	232,373(Google)/ 1487(Apple)	4,674(Google)/ 685(Apple)	13,163(Google)/ 54(Apple)	1,811(Google)/ 2(Apple)
Functions	Users can use it to analyze symptoms and acquire diagnostic results.	Users can use it to analyze symptoms, acquire diagnostic results, and chat with expert doctors to discuss treatment options.	Users can use it to analyze symptoms, acquire diagnostic results, and find local healthcare services.	Users can use it to analyze symptoms, acquire diagnostic results, and find local healthcare services.

To delve into users' attitudes towards AISC apps and their positive and negative experiences, we also conducted ten semi-structured interviews. Participants were recruited through *Studyfinder*, a web-based participant recruitment tool provided by Penn State Clinical Translational Science Institute), and through social media. We first utilized a screening survey to filter out unqualified participants. We asked following questions: (1) how old are you? (2) what are your goals of using these symptom checkers? (3) what symptom check(s) did you use? (4) when was the last time you used a symptom checker? Eligibility criteria were as follows: (1) the user is over the age of 18; (2) the user has used AISC(s) to seek for possible diagnoses; and (3) the last time the user used AISC(s) is less than one year. We kept conducting interviews until the data from review analysis and interviews reached "theoretical saturation [20]." We finally recruited ten interviewees

and most participants were students and in their 20s. They used a couple of the AISC apps that we selected (see Table 3-2).

Table **3-2**: Participants' demographic information

#	**Age**	**Gender**	**AISC Apps**	**Recruiting Methods**
P1	23	Female	Ada, K Health, Your.MD, Ask NHS	Social media
P2	29	Female	Ada, K Health, Your.MD, Ask NHS	Social media
P3	24	Female	K Health, Ask NHS	StudyFinder
P4	24	Male	Ada, K Health, Ask NHS	Social media
P5	28	Female	Ada, K Health, Ask NHS	Social media
P6	27	Female	Ada, K Health	Social media
P7	26	Female	K Health	Social media
P8	27	Male	K Health	Social media
P9	25	Female	Ada, K Health	Social media
P10	40	Female	Your.MD	Social media

Then we acquired the participants' consent and conducted audio-recorded individual interview sessions with each participant. These interviews took about 30 minutes to 1 hour with questions regarding their personal experience of using one or more selected AISC apps. These interviews were semi-structured, entailing new topics for further questions. The interviews were started with the following open-ended questions: (1) which app do you prefer (if the user has used more than one AISC)? (2) how do you feel about having conversations with these apps? (3) what are your preferred forms of conversations when talking with the chatbot in these apps? Participants could decline to answer any questions, refuse to be audio-recorded, or withdraw from the study at any time. Together, a total of ten interviews were collected. We then manually transcribed these interviews for further analysis.

3.2.2 Data analysis

We then used thematic analysis to analyze the online reviews and interview transcripts in an inductive approach [12]. We first familiarized ourselves with the reviews and transcripts and then generated initial codes. The initial code list was comprised of over twenty codes. We then searched for themes and acquired a collection of candidate themes and sub-themes. Next, we reviewed and refined our themes to ensure internal homogeneity and external heterogeneity [56]. Finally, we acquired our final thematic map and defined five themes: useful functions when presenting diagnostic results, neglecting diverse health conditions, rigid input requirement, failures in conversational design, and failing to consider prerequisite information for a diagnosis. When reporting our findings, we used R1, R2, etc. to indicate the different users writing online reviews in our review analysis. When referencing interview quotes, we used P1, P2, etc. to denote each participant interviewed.

Chapter 4

Findings

4.1 App feature analysis

4.1.1 The overview for eleven AISC apps

We first examined the app names, overall ratings, numbers of downloads, price, release time, and age limitations for these eleven AISC apps (see details in Table 4-1). It is prominent that many apps leverage their titles to convey their medical authority and credibility. Terms such as "Health," "Medi," and "MD" indicate they can provide medical services to users. The term "NHS" suggests the authority of two apps – Ask NHS and NHS online: 111, given that NHS stands for National Health Service in the U.K.. NHS provides health services and gives medical assessment by telephones for problems that are urgent but not classified as emergencies [15]. These two apps are provided in partnership with the NHS. We also found that most apps appeared in recent years. Even the earliest one was released in 2014. This is in accordance with the development of AI algorithms that were applied in conversational chatbots. Between 2011 and 2014, a number of conversational chatbots that use natural language to answer users' questions appeared (e.g., Apple's Siri, Google's Google Now, and Microsoft's Cortana) [111]. In addition, most examined apps have relatively high ratings (above 4.0/5.0) and six of them have high number of downloads (above 10,000). This implies these apps are prevalent and users are generally satisfied with the functions provided by these apps. Most apps provide free services, while four of them (K Health, Mediktor, HealthTap, and Apothēka Patient) ask for fees when users choose to chat with medical professionals. All apps have various age benchmarks, showing these apps are not used for children. Some of them did not

elucidate the rationale as to why they set up such age limitations in their terms of use or developer's websites (e.g., Health Buddy).

Table **4-1**: Fundamental information of the eleven apps.

#	Name	Ratings (Google play Store/ Apple App Store)	Downloads in Google Play	Price	Release Time	Age Limitation
1	Ada	4.7/4.8	5,000,000+	Free	2016.07.15	16+
2	K Health	4.3/4.8	1,000,000+	Freemium	2017.11.05	18-85
3	Ask NHS	4.4/4.6	100,000+	Free	2017.04.11	18+
4	Your.MD	4.3/3.8	1,000,000+	Free	2014.8.18	17+
5	Mediktor	3.1/3	1,000+	Freemium	2018.05.28	17+
6	HealthTap	4.2/4.4	1,000,000+	Freemium	2014.09.16	17+
7	Apothēka Patient	NA	NA	Freemium	2019.03.20	17+
8	Sensely	3.3/5	1,000+	Free	2014.05.12	12+
9	Health Buddy	NA/3	NA	Free	2019.02.18	12+
10	Babylon	3.8/4.8	1,000+	Free	NA	17+
11	NHS online: 111	4.2/3.3	10,000+	Free	2017.01.31	17+

*NA: Not available

4.1.2 Accountability

The accuracy and effectiveness of these apps are essential for users, as it can influence users' further decisions and behaviors. Inaccurate diagnoses may lead to negative consequences to users. However, whether the accuracy and effectiveness can be guaranteed by these apps is unclear. Thus, we examined the app store descriptions and developer websites of these apps. We aimed to explore the extent of information reliability provided for users and accountability of the app

developers. From app store descriptions, we found most of them claim to cover certain illnesses such as cold and back pain. For instance, Ada and K Health claim that they can cover thousands of symptoms and conditions, e.g., common cold. Specifically, K Health promises trustworthy diagnostic results. While Ada and K Health do not specify whether they can diagnose all conditions in app store descriptions, Your.MD clearly notifies users that it cannot factor in all information as doctors can, identify all conditions, and satisfy all user groups. An interesting finding is that all apps except Health Buddy (which does not have the terms of use) state clearly in their terms of use that they make no warranties to the accuracy and cannot substitute for real doctors. For example, Ada posts in its terms of use, "*please note that the Ada app does not make any medical diagnoses*"; Ask NHS also states, "*the assessments, symptom checker, results and content provided on this App are provided as general information only and we cannot guarantee the accuracy of such assessments, results and/or any other content*" (Ask NHS). These claims reflect that the accuracy of the diagnoses provided by these apps is not guaranteed for users. It implies the apps are not responsible for potential resulting negative consequences if users entirely trust the diagnostic results from these apps.

We also found that all of these apps were developed by for-profit companies focused on customer-facing digital health products. None of them were developed by national health agencies, although Ask NHS and NHS online:111 have a partnership with NHS. These two apps can provide NHS primary care services, including making an NHS GP (General Practitioner) appointment and having face-to-face appointments with GPs. Ask NHS also integrates with a number of GP IT systems, such as 111 service providers (111 is a telephone service that can provide advice or medical treatment quickly). However, the module of symptom checker is operated by the developers and is irrelevant to NHS according to their developers' websites [112,123]. Thus, the reliability of these two apps is vague.

Furthermore, only five apps (i.e., Ada, Your.MD, Mediktor, Babylon, and NHS online: 111) illustrate their credentials to demonstrate their quality (see Table 4-2). Only one app (Mediktor) mentions that it underwent a clinical trial with reliable results published in December 2017 in the journal *Emergencias*, volume 29, number 6. All of these five apps mention possessing the CE Mark, a certification mark signifying product sold in the European Economic Area are in conformity with health, safety, and environmental protection standards. However, having a CE mark does not ensure the safety of the product is certified by authority [118]. Four of these five apps also state they comply with other certifications. For example, Ada posts on its website that it complies with ISO 13485 and has acquired the BiM Badge. ISO 13485 is an international standard enacted for organizations of quality management systems, requiring organizations' medical devices and services to meet regulatory requirements [119]. BiM Badge is a certificate issued by a German agency BiM (i.e., Federal Association of Internet Medicine) who provides guidelines for digital health applications. To acquire BiM Badge, the application must meet the requirements of CE certification and FDA (Food and Drug Administration) [113]. Developed by the same company, Babylon Partners Limited, Babylon and NHS online: 111 claim that they follow guidance issued by the Food and Drug Administration (FDA), which is responsible for public health protection and safety of medical devices in the U.S. [114]. Babylon and NHS online: 111 are also regulated by the CQC (Care Quality Commission) in the U.K. and MHRA (Medicines and Healthcare Products Regulatory Agency) as they state on their websites. CQC is an independent regulator in England for health and social care [124]. It monitors, registers, and regulates care givers to meet fundamental standards, such as safety and effectiveness. MHRA regulates medicines and medical devices to ensure their safety and quality [115]. However, the credibility of these statements remains in doubt. It is unclear whether these AISC apps actually comply with the regulations of these credentials, given that MHRA stated they would require Babylon Health Inc.to change the statement that its app is "certified as a medical device with the MHRA" [61]. The CQC also has not verified the reliability of Babylon [61].

In summary, the developers do not guarantee the accuracy and effectiveness of these apps according to their disclaimer in their terms of use. In addition, few apps have credentials to prove their safety. Even though some apps stated they comply with regulations of CQC, FDA, and MHRA, it is difficult to evaluate the extent in which apps follow these regulations.

Table **4-2**: Credentials for app quality of the eleven apps.

#	Name	Developer	Credentials for App Quality
1	Ada	Ada Health GmbH	1. ISO 13485 Compliant 2. CE Mark (Class I medical devices for the European Economic Area) 3. BiM Badge (German agency for quality management)
2	K Health	K Health. Inc.	NA
3	Ask NHS	Sensely, Inc.	NA
4	Your.MD	Your.MD Ltd	CE Mark
5	Mediktor	Teckel Medical s.l.	1. CE Mark 2. Clinically validation: the results of the trial were published in December 2017 in the journal *Emergencias*, volume 29, number 6.
6	HealthTap	HealthTap, Inc.	NA
7	Apothēka Patient	Apotheka Systems Inc.	NA
8	Sensely	Sensely, Inc.	NA
9	Health Buddy	JOhn Lyons company	NA
10	Babylon	Babylon Partners Limited	1.Based on guidance issued by the Food and Drug Administration (FDA) 2. Regulated by the Care Quality Commission (CQC) 3. CE Mark 4. MHRA
11	NHS online: 111	Babylon Partners Limited	1.Based on guidance issued by FDA 2. Regulated by the Care Quality Commission (CQC) 3. CE Mark 4. MHRA

*NA: Not available

4.1.3 Transparency

The transparency of an app' development process and developer' commercial interests is important for users to perceive the effectiveness and credibility of these apps. By studying the apps' descriptions, terms of use, and developer websites, we attempted to analyze two types of transparency: (1) are these apps using reliable medical sources in their development processes, e.g., what databases they used and what professional technologies they employed; (2) how would these apps represent their ties to commercial interests.

For the first kind of transparency, six of all apps state that health professionals were involved in their development processes. For instance, Ada notified on its website, "*8 years' research & development, 60 in-house medical professionals*" (Ada); K Health points out on its website, "*Our medical sciences team includes doctors who are also data scientists and engineers*" (K Health); Mediktor illustrates on its website, "*Our medical chatbot incorporates knowledge of all medical specialties*" (Mediktor); Babylon states, "*Babylon's AI system has been created by experienced doctors*" (Babylon). These apps also employ technical terms from computer science to show that they were developed using up-and-rising technologies. For example, NHS online:111 emphasizes Babylon's deep neural network on its website. However, how the medical content was generated in detail, what medical databases they employed, and how these medical professionals played a role in their development teams have not yet surfaced.

For the second type of transparency, these apps fail to disclose commercial interests or conflicts of interests. Several of the apps (i.e., K Health, Mediktor, and HealthTap) provide the opportunities for users to consult with paid online experts. As such, these apps promoted their online expert services while these services are not apparently shown when users initially download these apps. The commercial interests between these apps and their affiliated experts are thrown into doubt. Furthermore, the commercial relations between the developers and their affiliated organizations are not apparent. For

example, Babylon claims that "*Who do we work with? Insurers, Health Systems, Retail & Pharmacy*" (Babylon). However, the details about how the developers collaborate with their affiliated organizations are not disclosed. It is also unclear how Ask NHS and NHS online: 111 cooperate with NHS in detail although these two apps state they can provide NHS services. Moreover, only three apps (i.e., Ada, K Health, and Sensely) disclose information about their investors on their websites.

In summary, these AISC apps lack transparency in two important regards. Though a few apps state they incorporated knowledge of medical professionals in their development processes, we still do not know how the knowledge influenced the development process in detail. In addition, most developers of these apps do not claim clearly how they collaborate with their affiliated organizations. Therefore, the commercial interests are unclear for users.

4.1.4 Data policies

When interacting with AISC apps, users need to input their personal health data, such as their demographic information and symptoms. Such data may be sensitive and potentially embarrassing. Therefore, keeping confidential, safe, and secure users' personal health information that they share with the apps plays a significant role in the privacy protection. To explore whether these apps can protect user privacy effectively, we examined their websites and privacy policies in two regards: (1) data storage and encryption and (2) data collection and usage.

Data storage and encryption

Nine apps (i.e., Ada, K health, Ask NHS, Your.MD, Mediktor, Healthtap, Apothēka Patient, Babylon, and NHS online: 111) illustrate their measures applied to store and encrypt data. For example, in regard to data storage, Ada and Your.MD claim that they storage health information in separate

servers to ensure the security. K Health promises that it will never sell users' data to others. In regard to data encryption, Your.MD states that it encrypts users' data with two keys and uses penetration testing to ensure security; Mediktor keeps data encrypted in the communication process. Other apps do not state clearly what measures they have employed. For instance, Apothēka Patient claims in its privacy policy, "*We employ commercially reasonable security methods to prevent unauthorized access, maintain data accuracy and ensure correct use of information*" (Apothēka Patient).

On the other hand, three apps (Apothēka Patient, HealthTap, and Babylon) warn users that they do not guarantee data security and information privacy. For example, Babylon posts in its privacy policy, "*we cannot and do not guarantee that information about you will not be accessed, viewed, disclosed, altered, or destroyed by breach of any of our physical, technical, or managerial safeguards.*" HealthTap claims that "*we can make no guarantees as to the security or privacy of Personal Information*" (HealthTap). K Health even claims, "*we may continue to retain your personal information even after you deactivate your user account or stop using K*" (K Health). This statement assigns K Health with the right to keep users' personal information even after users no longer use the app, leading to doubt about users' privacy protection. Specifically, Your.MD states that it stores information using AWS (Amazon Web Services) and Google Cloud Platform instead of its own infrastructures. While Your.MD describes the credibility of these two platforms in its privacy policy, it still remains in doubt whether the data is encrypted on these remote third-party servers and whether other third-party companies have access to the information.

Moreover, only five apps present their data security credentials on their websites, stating they comply with HIPAA (Health Insurance Portability and Accountability Act), GDPR (European Union General Data Protection Regulation), and ISO 27001 (see Table 4-3). These credentials certify their capability to protect personal data and privacy. Two apps (i.e., Ada and Mediktor) claim that they comply with GDPR. GDPR stipulates that organizations should collect and manage personal data legally [74]. Mediktor and Apothēka Patient state that they are comply to HIPAA regulations, which is

an act passed in 1996, aiming to protect the confidentiality of health information including data storage and data transmission [121]. Three apps (Ada, Babylon, and NHS online: 111) claim that they abide by ISO 27001 or ISO 27001 Stage 1. ISO 27001 is a standard specifying the requirements for the implementation of information security management systems, specifying the requirements for the management of information security risks within the context of an organization [116]. ISO 27001 stage 1 includes reviewing information security policies and inspecting risk treatment plans [109]. However, these apps do not clarify how they comply with these regulations in detail and it is still unclear whether following these regulations can ensure the data security.

Table **4-3**: Credentials for data security of the eleven apps.

#	**Name**	**Developer**	**Credentials for Data Security**
1	Ada	Ada Health GmbH	1. IST 27001 2. EU-GDPR compliant (European Union General Data Protection Regulation)
2	K Health	K Health. Inc.	NA
3	Ask NHS	Sensely, Inc.	NA
4	Your.MD	Your.MD Ltd	NA
5	Mediktor	Teckel Medical s.l.	1. HIPAA in the USA (Health Insurance Portability and Accountability Act) 2. GDPR compliant (GDPR is a regulation that requires businesses to protect the personal data and privacy of EU citizens for transactions that occur within EU member states)
6	HealthTap	HealthTap, Inc.	NA
7	Apothēka Patient	Apotheka Systems Inc.	1. HIPAA 2. Other digital security standards
8	Sensely	Sensely, Inc.	NA
9	Health Buddy	JOhn Lyons company	NA
10	Babylon	Babylon Partners Limited	ISO 27001 Stage 1 compliant
11	NHS online: 111	Babylon Partners Limited	ISO 27001 Stage 1 compliant

*NA: Not available

Data collection and usage

Most apps claim clearly that they collect three categories: personal data (i.e., registration information, profile Information, etc.), usage data (i.e., technical information about users' devices, details of visits, details of conditions and symptoms searched, etc.), and analytics data (i.e., IP address, length of visits to certain pages, etc.).

In addition to stating what data they collect, these apps also declare how they use these data. Most apps claim that they use the data to improve services. For example, HealthTap posts in its privacy policy, "*to operate, provide, maintain, improve and enhance our Services, to personalize your experience on our Services, to understand and analyze how you use our Services and to develop new products, services, features, and functionality*" (HealthTap).

Nevertheless, it is unclear how these apps share their data with third-party providers and ensure the data security during this process. For example, Sensely declares that it may disclose users' information to third-party vendors, service providers, or other third parties for marketing, advertising, and research. However, Sensely does not claim how it ensures the data security in the communication process. HealthTap claims that it discloses personal information to its affiliates, vendors, service providers, advertising partners, and analytics partners. Though it promises that it complies with applicable laws and regulations, it does not provide detailed information about what laws or regulations it has abided by. Moreover, Your.MD, which is embedded with external third-party apps, states clearly that it shares data with third-party technology providers and advertising providers. The data security is in doubt in this process. For example, Typeform, an external app to form surveys for Your.MD, hosts all data in Amazon Web Services (AWS); thus, it can acquire users' personal data and device data (e.g., IP address and browser type) and cannot ensure data security. It is also unclear how Your.MD protects users' data during the communication process with external apps.

4.1.5 Functionality

After coding features mapped by the eight stages of a diagnostic process (i.e., establishing a patient history, conducting physical exams, evaluating the symptoms, giving an initial diagnosis, ordering further diagnostic tests, performing and analyzing test results, providing a final diagnosis, and providing referrals or other follow-up treatments), we categorized their features (see Table 4-4). We found most apps can support the following five processes: establishing a patient history, evaluating the symptoms, giving an initial diagnosis, and providing referrals or other follow-up treatments. These apps do not support the following three processes: conducting physical exams, ordering further diagnostic tests, and performing and analyzing test results.

For establishing a patient history, six apps require users to input personal demographic information and health information from different levels. For example, HealthTap requires exhaustive amounts of information, including name, age, gender, ethnicity, current and past conditions, medications, allergies, vaccinations, lab test results, treatment, lifestyle (dietary restrictions, recreational drugs, alcohol, tobacco, sexually active), and pregnancy status. On the contrary, other apps require only basic information inputs. An example is Your.MD, which only involves information of name, date of birth, gender, and consultations history in the profile.

For giving an initial diagnosis, the presentations of assessments vary. Only four apps provide text-based expression to explain the diagnostic results; three apps reflect the likelihood of diagnostic results in comparison with similar users; four apps employ pictures or diagrams to explain conditions; only one app (Ask NHS) uses videos to explain diagnostic results.

For referrals or other follow-up treatments, these apps show a tendency of connecting with the offline medical services, including contacting with human doctors (K Health, Mediktor, HealthTap, and Apothēka Patient), getting prescriptions (Your.MD and K Health), finding pharmacy (Ask NHS and Your.MD), and finding hospitals (NHS online: 111). The function of

symptom tracking is also prevalent in these apps (Ada, K Health, and Your.MD). Users can utilize these apps to track severity of symptoms over the time.

Table **4-4**: Features of the eleven apps.

Name	**1.Establish A Patient History**	**3.Evaluate Symptoms**	**4.Give an Initial Diagnosis**	**5.Order Further Diagnostic Tests**	**8.Referrals or Other Follow-up Treatments**
Ada	Date of birth, gender, height, weight, medication, allergies, health background (e.g., diabetes)	√	Likelihood, cause-and-effect diagram	×	Symptom tracking
K Health	Date of birth, gender, height, weight, medications, allergies, ethnicity, smoking, surgeries, chronic conditions, family history	√	Likelihood, pictures for conditions	×	Chat with doctors, get prescriptions, symptom tracking
Ask NHS	×	√	Text / video	×	Find pharmacies
Your.MD	Date of birth, gender, consultations history	√	Text / pictures for conditions	Order tests using third-party apps	Get prescriptions, find pharmacies, symptom tracking
Mediktor	Age, gender, height, weight, medication, race, allergies, risk factors, past medical history, past surgical history	√	Text / pictures for conditions	×	Chat with doctors
HealthTap	Age, gender, medications, allergies, ethnicity, current and past conditions, vaccinations, lab test results, treatment, lifestyle, pregnancy	√	Likelihood, text	×	Chat with doctors
Apothēka Patient	Age, gender, weight, height, blood group	√	Text	×	Book an appointment with a doctor
Sensely	×	√	Text	×	×
Health Buddy	×	×	Text	×	×
Babylon	×	√	Text	×	×
NHS online: 111	×	√	Text	×	Find hospitals

In addition to different functionalities, these apps employ diverse forms of information expression (see Table 4-4). Most apps allow users to use text to input their symptoms. However, the flexibility of input is limited in certain apps (Ada, K Health, Ask NHS, Your.MD, Mediktor, and Sensely), which ask users to choose from a restrictive list. Only two apps (Ask NHS and

Sensely) allow users to input information via voice recording. Two apps (K Health and HealthTap) let users input information by clicking on separate parts of an image. Most apps use various forms of text, voice, and pictures to explain the questions presented during conversation. K Health uses cartoons and Ada uses pictures of real human body parts to explain the medical jargon that appears in the questions. For example, K Health uses a cartoon picture to illustrate a red joint and Ada uses a picture of a real human throat to elucidate a reddened throat (see Figure 4-1). Some AISC apps also use humanizing language in their questions, such as "*Thanks for telling me about your*…" (K Health). Two apps (Ask NHS and Sensely) employ avatars with a recorded human voice to ask questions (see Figure 4-2).

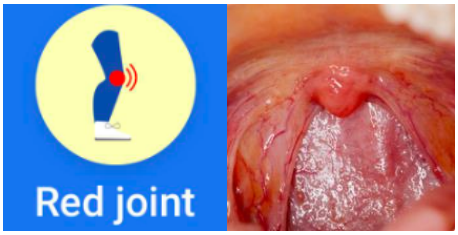

Figure **4-1**: A cartoon picture of red joint from K Health (left) and a picture of reddened throat from Ada (right).

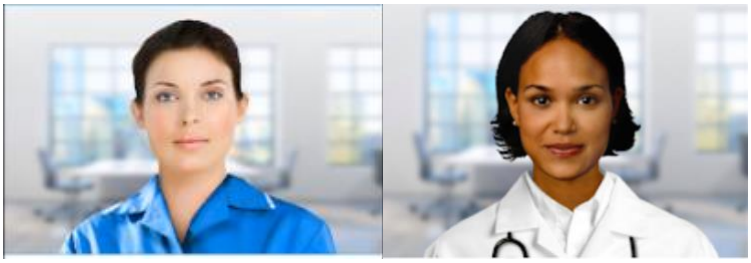

Figure **4-2**: The avatar of Ask NHS (left) and the avatar of Sensely (right).

4.1.6 Summary

Since the measures of data storage, data encryption, data transfer, data collection, and data usage are essential to ensure the data security, it is vital for AISC apps to disclose sufficient

information to users. Even though some apps claim they comply with credentials about data security, there are posed risks in users' privacy protection. In addition, how they communicate with third-party services is nontransparent. Furthermore, from our analysis of app functionality, these apps cannot support all diagnostic processes that may be required by off-line medical visits, which may influence users' attitudes towards these AISC apps.

4.2 App review analysis and interview analysis

After the feature analysis, we analyzed users' reviews of four AISC apps – Ada, K Health, Ask NHS, and Your.MD. Using the criteria that reviews should be written in English and longer than 15 characters, we then selected 2,000 latest reviews in total for analysis. We also conducted semi-interviews to study users' opinions for AISC apps.

From our review analysis, online users had different attitudes for AISC apps. Most users showed satisfaction in their online reviews, as portrayed in the high ratings, specifically regarding Ada. They appreciate the human-like conversations. For instance, two users stated "*I loved how it actually seemed like a real-life conversation. It also made me feel more relaxed and a little less stressed out…*" (R1; Ada) and "v*ery clear questions and to the point. Felt reassured as if I was having a face to face examination*…" (R2; Ask NHS). Some users also enjoy thorough explanations (e.g., "*I liked that the questions were pointed and easy to answer, additional simple explanations of what information was being asked for is provided…*" (R3; Ada)) and the capability to cross reference with other users (e.g., "*It also comes up with several diagnoses and compares my symptoms to what other people with those conditions experience. Very helpful!*" (R4; Ada)). They believe AISC apps provide more reasonable and realistic results in comparison to other diagnostic tools: "*Absolutely amazing! I truly love this app. WebMD just tells me I'm going to die of cancer lol*" (R5; K Health). In our interviews, some participants also expressed their approval for these AISC apps. For example, P3 stated, *"the K Health app, it tells*

you from your age, the number of people who are having the similar cases as you do, and it gives you a percentage of the possibility of each symptoms. Like even though it is not 100% accurate. You find it very reliable, because it's telling you the number of cases and you know the percentage" (P3; K Health). P1 enjoyed the human-like sentences for relief, as she stated, "*Also some feedbacks like 'you're so good', 'it is so great for you', you know, give you a warm heart*" (P1; K Health).

Nevertheless, some users hesitated to trust AISC apps. Some reviews show that users questioned the quality of Ada's questions, "*It asks a bunch of very shallow questions and then only asks for updates once a day on how bad your issues were and that is nowhere close to a guide*" (R6; Ada). Other reviews show complains from users that diagnostic results from AISC apps were not always consistent with professionals' diagnoses (e.g., "*…the first thing I typed in was the fact that I was having a headache and it immediately said that I was going to have a heart attack I got really scared and went to the hospital because you know what the app said come to find out I just had a migraine…*" (R7; K Health)). Some interviewees also queried the validity of AISC apps. For example, P5 stated, *"I don't think I fully trust this. Both of the apps they only gave me the percentage. And I always.. my case might be unique, might be different, it could be different from the other people. Even.. I might be the thirty two percent. I might just be the one percent. So I went to the doctor afterwards"* (P5; Ada, K Health).

In our study, we aimed to find out how users perceive the weaknesses of AISC apps. We found that some users in the review analysis perceived the existing AISC apps neglect diverse health conditions, such as chronic diseases and skin problems. Furthermore, some users both in the review analysis and interviews showed their dissatisfaction when interacting with these AISC apps. First, they complained they could not input the descriptions of their symptoms flexibly. Second, some users realized problems in the conversational design as they thought the conversational styles of these AISC apps were not user-friendly. Finally, some users noticed these AISC apps failed to consider important factors when making diagnoses, such as health history. In the following sections, we report on these findings in detail.

4.2.1 Neglecting diverse health conditions

We found that some users in the review analysis noted AISC apps lacked sufficient knowledge to support particular diseases (e.g., eye problems, skin problems, and certain chronic diseases) and certain special user groups.

Missing knowledge for particular diseases

According to reviews, we found that some users perceived AISC apps should have supported some particular diseases. For example, R8 found Your.MD could not solve eye problems though it claims that it can diagnose thousands of diseases. The following quotes show that users perceived eye problems should have been included.

"Can't solve eye problems." (R8; Your.MD)

Another user R9 noted that Ask NHS lacked support for skin problems. In the below review, R9 intended to acquire information requiring skin burns and throbs from Ask NHS. However, he/she simply received answers regarding menstruation, which were off track from R9's intended search. This implies Ask NHS cannot recognize certain medical terms and has difficulty in understanding users' input from the user's perspectives.

*"Absolutely useless app I wanted answers to why my skin burns and throbs for a long period of time her answers were about a f*****g period?!?! Absolute joke can't get to see me own useless doctor now this pointless app wow!"* (R9; Ask NHS)

Ada also could not provide well-rounded information for particular diseases perceived by users. In the below review, R10 expected a spider bite to be included in Ada's knowledge base. Nevertheless, the user could not find such information.

"Diagnosing a spider bite should be within its realm of capabilities but it's not... and half of the time the assessments are wrong." (R10; Ada)

Specifically, with regard to chronic diseases, some users noted that all of these AISC apps showed insufficient support for chronic diseases. First, some users thought AISC apps might give inaccurate results for chronic diseases. In the following review, R11 had chronic migraine, but K Health did not enquiry about the correct symptoms and gave wrong diagnostic results (Lupus, Lume Disease, or Mono).

"I have chronic migraine due to postnasal drip and allergies. It said i had Lupus, Lyme Disease or Mono. It didn't even list allergies as an option." (R11; K Health)

Second, some online users anticipated these AISC apps could not distinguish chronic diseases from acute diseases efficiently. The below review shows R12 perceived Ask NHS could not identify the difference between chronic kidney disease and acute kidney disease.

"There are major deficiencies in this app. For example, it is hazy about the terms chronic and acute. I asked about chronic kidney disease. And it responded about acute kidney infection, an entirely different condition. I shall remove the app as soon as I finish this comment." (R12; Ask NHS)

Third, some online users were aware AISC apps lacked knowledge of different types of chronic diseases and misunderstood different types of chronic diseases. R13 in the below instance complained Ada overlooked the difference between two types of diabetes. Type 1 diabetes and type 2 diabetes are two different types of diabetes. They are not mutually exclusive. Thus, these two types of diabetes should be asked in separate questions, implying some users perceived Ada ignored important factors which should have be offered for users.

"Oh and ask if a person has type 1 or type 2 diabetes as they are 2 completely different diseases with different origins, '1' being an autoimmune disease." (R13; Ada)

Finally, more comprehensive knowledge of AISC apps was desired by some online users because of the complexity of chronic diseases. For instance, R14 posted, "*I'd like to be able to tell her*

about chronic illnesses though. I have chronic bronchitis and heart problems due to birth defects, and I know the difference between those symptoms and when new symptoms arise. I only recommend adding a chronic illness feature, to get a more accurate report..." (R14; Ada). In this example, we can also find that the user realized that he/she needed to input more symptoms including new emerging symptoms about his/her chronic diseases, while Ada could not provide these functions due to its limited knowledge base for chronic diseases. Other AISC apps also encounter a similar situation. In the below reviews, some users of K Health and Your.MD also desired a more comprehensive knowledge base to provide information for their chronic diseases, such as OAB (overactive bladder) and IC (Interstitial cystitis).

"…this app doesn't know what OAB is nor IC which are two chronic illnesses I have and affect my health daily…" (R15; K Health).

"I needed a symptom checker after seeing multiple doctors with no legit reason behind my chronic symptoms and all it says was "seek help immediately" ive been seeking help for 13 weeks…" (R16; Your.MD)

Missing knowledge for diverse user groups

Aside from the lack of support for diverse diseases, some users in the review analysis also noticed the existing designs of AISC apps could not accommodate diverse user groups, such as caregivers of young children and those who identify as transgenders. The following review shows R17 wanted to use Your.MD to know sleep patterns for caregivers of young babies; however, Your.MD could not provide sufficient information.

"this app doesn't take into consideration women with young babies when asking about sleep patterns." (R17; Your.MD)

Furthermore, most online users wanted to use chatbots for children. These two reviews below imply that these users needed to use AISC apps for their children. They wanted the AISC can establish

profiles for children. In these cases, some users perceived Ada ignored the situation that caregivers would use Ada for their underage children.

"*Wanted to use it for my children but there is no option for that. Can only use it for one person and not children.*" (R18; Ask NHS)

"*I like it so far. As a mom, I just wish I could have a profile for my child.*" (R19; K Health)

Additionally, those identifying as transgender showed their dissatisfaction for chatbots. In the following example, R20 complained Ada could not support the terminology for transgenders. The user suggested Ada to accommodate with more variety of options in regard to gender setting. While the guidelines of inclusive design require taking into account the needs of all users, some online users perceived that Ada lacked inclusiveness for users with particular gender and ignored the diversity of user groups.

"This app is worse than useless unless you are cisgender; in fact, the poor handling of terminology that is needlessly binary and unaccommodating of trans health issues (like increased incidence of breast cancer after transitioning as an AMAB trans woman) are actively HARMFUL to trans people - not to mention the unpleasant and dysphoria-inducing experience that I MUST input my "sex at birth" with no other options, though they may be medically relevant..." (R20; Ada).

4.2.2 Rigid input requirement

The description of a symptom is complex. Some users both in the review analysis and interviews perceived the current algorithms of AISC apps did not have the capabilities to live up to the complexity of symptoms, especially when dealing with diseases with a large array of symptoms or user-defined symptoms. They found that K Health and Your.MD were restricted by the provided symptom list. Even for the AISC apps allowing freely input, they were unsure of appropriate input words.

Restricting descriptions for disease dimensions

In the review analysis, some online users expected to input more detailed information, similar to how they did in offline medical visits. These users wanted the algorithms of AISC apps to consider more factors (e.g., the frequency of symptoms) to cope with diseases with a plenty of symptoms. For example, some users wanted to describe the frequency of their symptoms.

"… but there has to be an option for "at times" as a response to a symptom question instead of a "yes, no or idk" response…" (R21; Ada)

"Only issue is when it asks questions, it doesn't have a response that says sometimes or it does occur but not all the time, like if appetite loss has occurred but not constantly..." (R22; K Health)

In the above examples, R21 and R22 desired to input further detailed information related to the frequency. These two users wanted the addition of the option such as "sometimes" in addition to the current "yes" or "no" in regard to the frequency of a symptom.

Some online users also showed their desire for providing the information about severity of their symptoms, which is required in the offline medical consultations, *"You could add in some rating how bad the symptoms are..."* (R23; Ada).

Moreover, locations of symptoms were also required by some online users, *"If you add a feature to capture part of body and mark the area of interest for more accurate investigation"* (R24; Your.MD). Some online users wanted to input multiple locations of their symptoms. Two users posted "…*Could do with a multiple-choice response on some answers, my pain is not located to just 1 area…*" (R25; Ask NHS) and "*The area of symptoms also only generally asks chest, legs, hands. There are other parts of the body also..."* (R26; Ada). In these two instances, these two users thought detailed location information was essential to reach a more accurate diagnostic result.

Furthermore, medical professionals are required to examine the appearance of some diseases for an accurate diagnosis. We found that some online users proposed their desire to involve images

when describing their symptoms for a better context and interpretation of the medical jargons. R27 posted, *"it would be awesome if the computer could see my photo and ask questions about it…"* (R27; Ada).

Limiting the quantity of symptoms input

One disease may have many symptoms [4]; however, some users in the review analysis could not input a sufficient number of symptoms using these AISC apps, which may affect the diagnosis and lead to inaccurate diagnostic results.

"Ada is a wonderful app however I found that after you put in "too many" symptoms it stops adding more. I think they need add more "slots" for more symptoms." (R28; Ada)

In the above review, R28 concerned the limited number of symptoms they could input due to the input limit of Ada.

Furthermore, some users complained the input rules of Ask NHS and K Health in reviews and interviews. For Ask NHS, some users perceived that they could not select multiple symptoms for one category, *"It won't let you select multiple options when symptoms are asked about…"* (R29; Ask NHS). For K Health, some users complained the limited list of symptoms for them to select, *"…It doesn't ask detailed questions and doesn't allow you to type questions or give detailed answers. Only let you choose from an inadequate predetermined list of symptoms…"* (R30; K Health). These two rules restricted users' ability to input sufficient symptoms according to these users' perceptions. P5 and P6 also complained that limited symptoms could be input to K Health. P5 stated, "*I think, I would prefer some text typing, cuz most of the conversations here are like multiple choice, so maybe sometimes if I can type in something, or give me more options, that would be better. I really don't like the "yes, no, maybe" kind of conversation. I think that would work in this scenario, although I do wish I could type in more personalized text*" (P5; K Health). In this case, since K Health provides the input form of multiple

choice, the participant desired K Health to provide a greater variety of response items for her to choose. She preferred that she could input personalized text so that she could provide enough information about her symptoms. P6 stated, "*It has limited options. Not many choices and not matching my situation*" (P6; K Health). She also perceived the response items listed in K Health were not enough for her to describe her symptoms.

Limiting the presentation of symptoms input

First, For AISC apps that users can type in symptoms using their own words, such as Your.MD, Ada, and Ask NHS, some users felt lost in which words or phrases to employ. For example, R31 posted, "*The app is so useful I like it, just that I don't know the correct word to use"* (R31; Your.MD). In this case, R31 needed more explanations or guidance when inputting his/her symptoms. P1 also showed her confusion regarding which words to use, stating, "*the thing is I don't know what they have or not have in their database. Maybe they should give similar words to users*" (P1; Ada). In this case, P1 did not know which words Ada had in its database, and she thought Ada should give hints about what words she could use to express her symptoms.

Some users also found that some AISC apps lacked support for medical terms. The below review reflects Your.MD lacked knowledge for a simple medical term, "pimple", according to the user' statement, *" … it just do not understand what a simple 'pimple' means, suggests me babies skin rash and many more irrelevant information over and over again…"* (R32; Your.MD). R32 wanted to search for information about a "pimple"; however, Your.MD did not comprehend it as a medical term and gave wrong answers. P2 perceived that she could not provide enough information when inputting symptom, stating, "W*hen I use the Your.MD, when I say I have trouble in like breathing, it doesn't understand. It just says, 'what is the problem?' It can't understand…So I have to find out the words I think the checker would understand. I need to use general words, not use sentences. First time I say I have trouble in*

like breathing, it doesn't understand. just say which part? What is the problem? It can't understand" (P2; Your.MD). From this statement, P2 thought Your.MD could not recognize some simple and short sentences, such as breathing. She realized that she must type in general words instead of sentences for her symptoms after several failures.

Second, for AISC apps that users have to select symptoms from a pre-structured list (e.g., K Health), some users perceived that they needed more efforts to determine ways to express symptoms. In the below review, R33 realized only words exactly in line with the symptoms pre-defined by K Health could be recognized, implying some online users perceived that K Health lacked the capability of approximate string matching for symptoms and the support for medical terms.

"only thing is that you have to get their wording exactly or you won't find your symptom with the search. (for example, if you were worried about cauliflower ear, you couldn't search up bumps on ears or enlarged ears) this can be really frustrating especially when you're sick." (R33; K Health)

P1 complained K Health could not recognize symptoms she input. The participant P1 recounted, "*when I searched the hyperthyroidism, when I typed 'hyperthyroidism', it didn't respond. I realized you have to input common symptoms, like tired. I also typed 'heart rate', but it didn't have this word. I know they can't recognize all words; the thing is I don't know what they have or not have in their database*" (P1; K Health). From her perspectives, K Health could not recognize particular medical terms, i.e., hyperthyroidism and heart rate. She was also unsure about what words she should use to describe her symptoms.

In addition, when answering questions, some online users hesitated to give definite answers when they were unsure if they had the symptom. In the below reviews, R34 and R35 hoped Ada could give them more choices for input (i.e., users can input "I don't know" option), requiring the algorithms to consider how to deal with vague answers.

"Sometimes we don't know some things in these questions. The questions need to have more I don't know answers or I'm not sure or even a skip button would be very useful. But it helps me get an understanding of what could be going on with me or someone I know. Very useful app" (R34; Ada)

"What I mean is that there's just a yes or no to choose from. Can they add a "I think so" and a "I don't think so" as choices? That's the only issue for me. Overall, I love the app!" (R35; Ada)

4.2.3 Failures in conversational design

The communication skills of medical professionals are important in the probing process, since various probing questions are associated with patients' feelings [69]. Good communication skills can promote users to describe their symptoms and feelings and a clear explanation for symptoms can help medical professionals reassure patients [33]. To be specific, these communication skills in real-world probing processes are consisted of showing empathy, building rapport, utilizing nonverbal techniques, and cross-referencing the information from different sources to gain reliability from patients [4]. Even though some AISC apps employ the human-like style and avatar in their conversations, some users still perceived that they could not provide similar communication skills as humans.

We found that some users both in the review analysis and the interviews desired user-friendly conversational styles with nuanced requirements. The AISC apps we studied use text, voice, or visual presentations to actively interact with users. For the AISC apps (Ada, K Health, and Your.MD) only with text input and output, some users yearned for human-like styles, better forms of explanations, and fast interaction. However, for the AISC employing the avatar in the conversation (i.e., Ask NHS in our study), some users did not show satisfaction with these human characteristics, but rather showing concerns for the odd voice, low speed, and the uncanny avatar.

Textual presentation

Excluding Ask NHS, the other three AISC apps only allow text input. Though some users in the review analysis appreciated the human-like conversations, many online users thought chatbots' conversational style as pushy, rigid, and dreary, as shown in the following reviews. These online users perceived the conversational styles of AISC apps were lack of communication skills of humankind.

"but it some cases it seems really pushy." (R36; Ada)

"Conversation stilted and unlike a real human at all." (R37; K health)

"The symptoms checker is tedious to fill in but thorough…" (R38; Ask NHS)

Some online users hoped these AISC apps could mimic human conversations. As posted in the following review, R39 desired Ask NHS could express emotion and have increased response flexibility beyond fixed answer.

"…need more interaction and give one the ability to express their feelings more than prepared answers…" (R39; Ask NHS)

However, P5 queried the necessity of some sentences in Ada representing human feelings. She stated, "*I think in Ada there are also some words telling you don't worry. I don't think that would be helpful. I wouldn't be relieved until doctors exam me, like hear my lung, unless the app hear my lung and check my throat. Because I study computer science, maybe there's just... with some probability the app just displays those words to me, maybe that doesn't mean anything*" (P5, Ada). P5 thought these words showing human feelings (e.g., "Don't worry") were meaningless because the app could not do physical tests, so P5 did not trust these words and was not relieved by these words.

Specifically, for the questions asked in the conversations, first, users in our interviews mentioned the difficulty of comprehending complex sentences. For example, P2 perceived the questions of Ask NHS were difficult to understand. She described, "*The questions of ASK NHS are vague. It's not that readable, like 'How does general behavior or thoughts seem at the moment'. So just use*

some simple words to ask. I don't need you to be that fancy words. But just at least, let me understand what you're asking. I need to read it like at least twice to understand what it is asking. For the other three apps, Ada, K Health, and Your.MD, their questions are closer to your feelings. The questions are simple to understand, not like this one (ASK NHS). I mean this question is okay 'so what does the chest pain feel like'" (P2; Ask NHS, Ada, K Health, and Your.MD). From these statements, we can know that P2 perceived that these AISC apps should use simple terminology for users to better comprehend what is being asked.

P1 also thought that it was better for the AISC to provide an explanation to its questions. She described, "*K Health, when it asks questions, it doesn't give you some information about what this question is about. But Ada, it tells you 'why ask this question'. I think in this aspect, Ada is better than k health, helping me understand the question better. When I used Ada to check the red eyes, it asked me a question "Whether the eyes are sticky". I thought I had this symptom, but after I click the explanation, I found it means there are some yellow things in eyes, which is inconsistent with my symptom*" (P1; K health, Ada). She thought Ada was better than K Health as Ada could give explanations for questions. When P1 misunderstood the question "whether the eyes are sticky," the explanation provided by Ada helped her understand the question correctly.

Second, some users disliked slow interactions with the AISC apps in their reviews and interviews. They complained the quantity of questions is overly large, leading to burdensome questions, especially for Ada and K Health. In the below reviews, R40 complained Ada and Ask NHS had too many questions, with no diagnostic results following the questionnaire.

"rubbish! at first was told call 999 for minor problem I described. then was not given a diagnosis at the end of her hundred questions..." (R40; Ask NHS)

Some online users even compared these AISC apps with other diagnostic tools, such as Google，complaining the slow conversation responses of AISC apps due to the great number of questions compared with the instant information required from Google, as shown in the following reviews.

"Too many questions. Google gave me the answer in seconds. This app asks like 100 questions some repeated." (R41; Ada)

"Too many questions... some of them only MD can answer. Worse then dr. Google." (R42; K Health)

P5 also disliked the large number of questions. She described, "*I think they are asking me too many questions, they are trying to cover every single possible aspect. The app was checking whether I have any allergy. Because I was coughing. And also like that high blood pressure or something. But the real doctors did not involve these symptoms or something. Because they can just look at me*" (P5; K Health), asking if P5 had allergies and high blood pressure. However, P5 thought these questions made the conversation much longer than consultations with offline doctors. Also, P5 doubted the necessity of some questions. She stated, "*So say I'm coughing, it's asking me 'do you have pain in the muscle', that just makes me feel worried. Umm, I was wondering why I need to tell them about blood pressure if I'm just coughing, sometimes I wonder why they ask these questions?*" (P5; K Health). P5 thought the questions regarding muscle pain and blood pressure were unrelated to her symptom (i.e., cough). She doubted the rationality of questions asked by K Health who failed to provide explanations for the appearance of these seemly irrelevant questions as the participant perceived. P6 also doubted the necessity of some questions. For example, she stated, "*But k health asked me 'does the pain radiate to any of these areas?'. That's really not a very useful question. It is not only hard to understand, but it's also hard to describe it with the options it's given. It's given several options to choose. but not including the temple or eye that's related to my symptoms*" (P6; K Health). She thought one question of K Health was useless, because this question was hard to understand and the options following this question were irrelevant to her symptoms. In line with their opinions, P10 perceived the long-winded questions bothered her because of their unreasonable asking order. She stated, "*It should ask highly relevant first and then ask lowly relevant questions. But Your.MD, its questions are not logical. When I say I'm vomiting, it asked you first, like 'have you had more alcohol then you think your body can cope with?'*

Then it asked me 'do you have any of these symptoms today?' Then it asked, 'Have you drunk any alcohol today?' So I think these questions are not logical. When I consult a doctor, the doctor will ask questions logically. Like he can get an overview, like understand seventy percent of my conditions only by asking five questions. But Your.MD, I can only provide fifty percent of my conditions after twenty questions. The questions are too many and inaccuracy" (P10; Your.MD). P10 thought the questions of Your.MD were asked in an unclear order. Due to this, she perceived these questions were irrational and lead to a large number of questions.

Aside from the questions, the number of options that users chose to answer questions also bothered some participants in our interviews. P2 stated, "*the number of the options will bother me. When I used Ask NHS, its answered, sometimes I don't know what it means, what is the difference. It gives a lot of options that you need to select and choose which one is most appropriate*" (P2; Ask NHS). She complained numerous options was overwhelming, especially when she felt confused about the meaning of options and difference among these options. This implies P2 wanted the app to explain the meaning of options and difference among options so she could answer questions correctly.

Visual presentation

1) The design of the Avatar

Ask NHS uses an avatar, Nurse Olivia, to communicate with users. The avatar is presented as a real nurse and can blink her eyes (Figure 4-1). While the avatar seems enhance friendliness of AISC apps, most online users showed their aversion for the avatar especially for the blinking action. From the below reviews, these users thought the avatar was patronizing and needless, emphasizing the bad design of an avatar did not enhance the interaction between users and AISC apps; in contrast, it impeded better interactions.

"Olivia (the virtual assistant) is slow, useless, and pretty patronizing." (R43; Ask NHS)

"Blinking awful." (R44; Ask NHS)

"put off by the blinking robot and menu style questions. should just be a natural spoken language interface where the va simply asks "what would you like to talk about?" (R45; Ask NHS)

P7 also perceived the avatar as useless since the text could already present enough information. She stated, "*I think the text-based is enough. I don't want human characteristics. The avatar or voice also provides same information for me as the text. They wouldn't give me more information*" (P7; K Health).

At first, some online users wanted the avatar to increase its conversational speed. As shown in the following example, the user R47 stated clearly that he/she needed to interact with the Avatar more efficiently. This is in accordance with our findings in the textual presentation. Combining with these findings, we notice clearly that users desired for faster conversations.

"User May reasonably expect more rapid interactions than are experienced." (R46; Ask NHS)

Second, some users perceived the avatar was creepy and frightening in their reviews and interviews. In the following reviews, all these online users thought the avatar was creepy.

"thorough symptom check although robot lady is creepier than necessary" (R47; Ask NHS)

"I didn't even like the virtual nurse. Spooky" (R48; Ask NHS)

"Downside is Olivia is scary, far too much for me I don't need a CGI person it's creepy." (R49; Ask NHS)

"Just wish the avatar wasn't quite so creepy." (R50; Ask NHS)

R51 in the following review noticed the avatar design of Ask NHS failed to follow the uncanny valley design principles—the avatar was designed uncanny for the user. In addition, the user thought the avatar design was out of the ark, leading to the similar feelings acquired from using old information systems, such as the ELIZA chatbot.

"The avatar fell down the uncanny valley for me I'm afraid, but even cutting-edge research isn't there yet, so what hopes have you for a commercial product? After 10 minutes of use trying to re-

diagnose an eye issue I had last month—it feels like I'm talking to the ELIZA chatbot from the 1960s crossed with a multiple choice expert system from the 1980s, just sexed-up with Android's built-in text↔speech. You've seen what can be done with Siri & Alexa & Google Now: Sensely is nowhere close. Very disappointing from a Valley company." (R51; Ask NHS)

In our interviews, all participants perceived the avatar in Ask NHS as "freaky". For example, one participant P3 stated, "*It's so freaking. I don't see too much meaning of it. I won't use, I won't listen to it, I won't use this one, just prefer to read. Even though it is a real person, a real person's picture I know it is fake. Just on is the same similar mechanics, machine learning*" (P3; Ask NHS). The participant disliked the avatar and preferred the text-based conversations. Because she knew this avatar was not a real human, she would still dislike it even if the avatar could look like a real person.

Furthermore, though P5 disliked the avatar, she stated it would be better if the design of the avatar could reach a balance between the real human-like appearance and the bot-like appearance, as she described, "*If it's like a real doctor, like a headshot of a real doctor, I may think that's fake, because it's impossible there's real doctor talking to me on the app. And if it's a bot, like a cartoon picture of a bot, it's telling me I'm interacting with a bot. So maybe something in-between would be fine*" (P5; Ask NHS).

2) The design of pictures used for explanations

Some participants in our interviews showed their concerns for the pictures used for explanations in questions. First, some participants thought some pictures were unnecessary. For example, P4 stated, "*the pictures showed in questions are unnecessary. Some pictures are poor. Like this no smoking picture, it is easy to understand, so I think it's unnecessary. For this picture for running nose, I think this picture is hard to understand, I feel confused. But for some pictures that can show locations, like which body parts you have pain. This kind of pictures is good*" (P4; K Health). P4 perceived when the words in the questions were easy to understand, the apps did not

need the aid of a pictorial explanation, but when referencing parts of the body in pain, the pictures were useful.

Second, some users (P5, P8, and P9) preferred cartoons rather than pictures of real human body parts. For example, P5 expressed their preference for the abstract pictures of K Health but disliked real photos of human body parts showed on Ada. P5 stated, "*I remember I was coughing a lot, like two weeks ago. That's the symptom I told Ada, and when it tried to explain the words it was using, it showed me some photos... It should be the throat. It was disturbing, I don't like to see like a like a real body parts especially like body part with sickness. I remember seeing someone's throat and I didn't like it. But K health makes it very abstract, it uses cartoons, and just like very plain English words. And I could understand what K Health was talking about without the need of asking for explanations, so I think K health gave me a better impression*" (P5; Ada, K Health). When P5 input her symptom (i.e., cough), Ada showed a photo of a human throat, which annoyed her. She preferred abstract presentation, such as cartoons used in K Health to provide explanations. P8 stated, "*it's like a real people. It's a picture of real people getting sick, it will make me less comfortable. I guess it's just because like cartoon, cartoons make you feel like less stressful I think*" (P8; K Health). From this statement, we can see that P8 preferred cartoons because he would like to watch something to relax himself.

Finally, P10 perceived her preference for cartoons or pictures of real human body parts depended on different situations. In the below statement, P10 thought when users needed to point out the pain locations, both forms of pictures were reasonable, but when users needed to tell information about skin problems, a picture of real human body parts would be better. This is because to diagnose skin problems, patients need to examine the visual appearance of their skin accurately.

"*I think a picture of real human body parts can make you feel real. Like regarding to some skin problems, some real pictures can make it much easier to understand which type the skin*

problem is. But for some simple things, like if I only need to describe the location of my pain, I think cartoons are enough. So, like it depends on your symptoms." (P10; K Health)

4.2.4 Failing to consider prerequisite information for a diagnosis

A general diagnostic process is complex as it includes establishing a patient history, conducting physical exams, evaluating the symptoms, giving an initial diagnosis, ordering further diagnostic tests, performing and analyzing test results, providing a final diagnosis, and providing referrals or other follow-up treatments [39]. To reach an accurate diagnostic result, the probing process needs to guarantee sufficient information is collected, including health history and users' previous actives. The ignorance of medical records leads to wrong decisions of physicians [39]. However, some users in the review analysis and interviews perceived that AISC apps could not collect enough information in their probing processes. These users noted the existing AISC apps should inquire health history and sufficient information using their probing questions in order to reach an accurate diagnosis.

Failing to factor into health history

Common health history should consist of the history of present illness, past history (including past medical history, surgical history, childhood illness, obstetric/gynecologic history, health immunizations, screening tests), family history, and personal and social history (the use of tobacco, alcohol, and illicit substances, etc.) [95].

In our study, we found some of the existing AISC apps can record the basic health information of users. Ada can record users' height, weight, medication, allergies, high blood pressure, use of cigarettes, pregnancy, and diabetes. K Health can record the information of height, weight, use of cigarettes, surgeries, medications, chronic conditions, allergies, and family history. Your.MD can record

users' consultation history. Nevertheless, some users in the review analysis and interviews were unsatisfied with these profiles: "*Don't help, has no idea what it's patient's background is so how can it help at all? Keeps telling me I'm anorexic...I'm obviously not...*" (R52; Your.MD). They perceived none of these records were sufficient enough compared with the records required by offline medical consultations. They noted the lack of existing diagnoses, conditions, medications, personal history, and social history in some AISC apps.

1) Ignoring existing diagnoses, conditions, and medications

We found that in reviews such as the following ones, some users preferred AISC apps to add existing diseases and medications into the profile, including the past medication history and current condition of dose taking. In the below reviews, two users R53 and R54 showed their desire to keep existing diseases in the profile, e.g., anxiety, asthma, and narcolepsy. R54 also thought Ada should involve hereditary illnesses in the profile. In addition, R55 wanted Ada to involve medication information and R56 desired to link the medication condition with his/her symptoms so that he/she could acquire a more accurate assessment. P3 perceived that the medications and surgeries were important to reach a diagnosis for people who had relevant conditions.

"The only thing I would suggest is to have a way to include existing diagnosed conditions in your profile such as anxiety or asthma etc..." (R53; Ada)

"i would like to see an option to put already recorded diagnostics of illnesses or past illnesses in a chart i suffer from narcolepsy and ada doesn't know that so if i suffer from a pain or excessive tiredness over narcolepsy she wouldn't know my record also maybe a more in depth profile with hereditary illnesses..." (R54; Ada)

"The app is great, however it doesn't take into account whether you're taking any medications..." (R55, Ada)

"Good but I'd like to be able to add my conditions and what medication I'm taking, that way the app could see if my symptoms were linked to them or something new…" (R56; Your.MD)

"*if someone is like actually having a lot of medication and surgeries, the profile could be helpful. Because they're actually having surgeries and medications that might be relevant to their current symptoms.*" (P3; K Health)

2) Ignoring personal and social history

Aside from diagnosed diseases and medications conditions, some online users also noted these four AISC apps lacked consideration for personal and social history. First, they thought the prior activities failed to be taken into account in the current process.

"It diagnosed me with mono but I don't kiss people or share anything with people. I put my symptoms as throat pain and sharp pain on my ribs and it gave me this." (R57; Ada)

This above review shows that R57 was disappointed when the diagnostic result differed from what he/she expected based on his/her prior activities. Since the user thought he/she did not have any experience with mono before, this diagnostic result seemed unreasonable. This reflects the user thought that prior activities needed to be recorded in the health history.

Similarly, because some AISC apps do not keep information about drinking history, some online users perceived that AISC apps could give wrong diagnostic results. In the example below, Your.MD diagnosed R58 as an alcoholic while the user never had drinking history.

"wrong on the question saying I'm an alcoholic when i don't drink alcohol uninstalling the app." (R58; Your.MD)

Furthermore, some online users perceived that AISC apps overlooked some personal information, such as race and living conditions. In the below instance, the user thought the race information should be considered when making a diagnosis. Sickle cell disease (SCD) is a genetic disease [122]. This disease is specifically common among people whose ancestors are from sub-Saharan Africa, Spanish-speaking regions in the Western Hemisphere, and Mediterranean countries. Thus, this disease is strongly connected to racial information, which made R59 skeptical of the accuracy of K Health.

"Why wouldn't a health diagnosis app ask one's race? I was curious to see if you could diagnose my known disease Sickle Cell Anemia. Yet without asking my race it was almost impossible to do so with even with my symptoms of chronic pain, fatigue & jaundice." (R59; K health)

The living conditions also should be considered from users' perspectives. R60 in the below review thought Ada should ask information about living climate and weather conditions.

"Same with the duration ones. It should take into account the sort of climate you live in, the weather and the heat because I am Scottish and its warm today and it didn't take that into account when asking questions and for the symptoms as well." (R60; Ada).

Problematic probing questions

Asking probing questions is complex. These probing questions should be inclusive. Doctors may need to probe for a large amount of information, including symptoms, life stresses, and physical examination to provide an assessment [72].

In the probing process, some users both in the review analysis and interviews showed their desire for a comprehensive probing which is the prerequisite for an accurate diagnosis. Some online users disliked the too few or too generic questions given by all the four AISC apps as they thought they could not provide enough information. In the below reviews, these users thought the questions were too generic to let them describe enough symptoms and conditions.

"The questions were far too generic and not able to deal with the specifics of the situation…" (R61; Ask NHS)

"I couldn't discuss all of my symptoms and I didn't think that it asks enough questions." (R62; Ada)

"I feel like the chat bot shouldn't straight away diagnose you with a condition only after you answer a few easy questions, but overall I don't mind this app..." (R63; Your.MD)

In our interviews, P1 and P3 showed their concerns for the limited number of probing questions and they did not know whether they could provide enough information by answering the questions, as P3 stated, "*I always saw that I can provide enough information but who knows I'm not doctors, because you know, for example, if I went to hospital, and I thought I provided enough information but it turns out not, the doctor asked more symptoms I have never noticed. So I don't know if I can provide enough questions*" (P3; K Health). P3 was not sure about whether these questions were enough because she did not have enough medical knowledge. P1 also felt confused about the number of questions. She stated, "*I think it is necessary to ask enough questions, but I'm not sure if it misses any information, because I'm not a doctor*" (P1; Ada, K Health).

Aside from the number of questions, other users complained that the AISC apps just repeated same or similar questions without acquiring more information needed for making a diagnosis as shown in the following statements.

"… tried a couple other symptoms and it asked the same generic questions. I would suggest you fix it so it doesn't assume the worst and has more less critical issues." (R64; K Health)

"It's only asking questions for more than half an hour still no diagnosis. Same questions over and over again." (R65; Ada)

"*I feel like they ask repeating questions. They have several similar questions to ask me, like 'What kind of the stomachache? is that very strong stomachache? And is this just a lifestyle?' Okay. So I feel like it will be better by reducing the number of questions.*" (P9; K Health)

4.2.5 Summary

Based on our findings, we interpreted four limitations of existing AISC apps from users' perspectives. First, because human's diseases are very complicated, some online users perceived AISC apps could not support all kinds of diseases, user groups, and symptoms. Second, both online users and

interviewees complained the rigid input rules and lacked confidence in the words or phrases used in input. Third, some online users and interviewees desired humanization in the conversation, such as natural languages, but some of them doubted the necessary of human-like characteristics, such as photos of real human body parts that used for explanation and the avatar. Finally, both online users and interviewees perceived these AISC apps missed some prerequisite information for an accurate diagnosis. While AISC apps can establish profiles for users, some online users and interviewees found that more information was needed for a comprehensive diagnosis.

Chapter 5

Discussions

This work critically reviewed AISC apps aiming to explore their features and how users perceive them. First, our study shows that while several AISC apps claim they involve the medical professionals in their development processes, most of them lack accountability and transparency disclosed in their terms of use, descriptions, and privacy policies. Second, our study points out the social-technical gaps between the mechanisms of real-world medical consultations and that of these AISC apps. Users perceived that the complexity of human diseases, probing processes, and health history led to the insufficient support of AISC apps for users. Furthermore, our findings also provide insights into the conversational design of chatbots. We will discuss these findings in this section.

5.1 Governing the safety of AI in healthcare

We found the existing AISC apps do not disclose enough information to users regarding their accountability, transparency, and data protection measures. Though FDA deemed symptom checkers as low-risk [120], we found potential risks exist in using AISC apps.

First, few of these AISC apps claim that they comply with trustful credentials and do not disclose sufficient information about the qualification of their developers. Only one app (Mediktor) claims that its accuracy has been validated by clinical tests. Most apps were developed by for-profit companies, yet the reliability of these companies is not verified. Thus, the reliability of AISC apps is unclear. Additionally, these AISC apps claim that they do not guarantee the accuracy of diagnostic results. Researchers have pointed out that AI systems need regulations and strict testing before being released to the public [58,61]. Since developers disclose insufficient information about credentials

and qualifications, it is irresponsible for them to release these AISC apps without providing reliable credentials. The effectiveness and accuracy of AISC apps cannot be guaranteed and users have to make decisions by themselves regarding whether to trust diagnostic results or not. Our study stresses a need of improving laws and regulations for AI applications in healthcare. The accuracy and effectiveness of healthcare AI applications should be tested and verified by professional institutes before being released.

Second, few AISC apps provide sufficient information about their development processes and commercial interests, while explanations for the development processes of AI technologies are important. The European Union's latest General Data Protection Regulation (GDPR) emphasizes the "right to explanation" for citizens [103]. Researchers also have found that AI technologies can introduce risks to users due to hidden assumptions and opaque algorithms [58]. However, we found that these AI applications do not disclose enough information about the ground truth of algorithm mechanisms and details regarding the involvement of domain experts. For example, commercial relations between HealthTap and its online doctors are unclear. This indicates that more detailed information on algorithm mechanisms and commercial interests should be disclosed to users by healthcare AI applications. By improving the transparency of their algorithms, AI applications can allow users to understand their mechanisms better and make more rational decisions, decreasing the potential negative consequences caused by irrational decisions. If AISC apps strengthen their transparency for commercial interests, users can weigh the credibility of the developers.

Third, even though some AISC apps elucidate how they store, transfer, collect, and use users' data, they do not provide clear descriptions and justifications for these measures. In healthcare, privacy and security problems need to be more emphasized since the information of users is sensitive [11]. However, few suggestions have been provided regarding how to ensure safety and security for current mobile health apps [59]. Future researchers need to find out what information users need regarding health data collection, storage, and usage [46]. One study proposed that each step of the design and development of mobile health apps should consider privacy and security [11]. In our case, we found

that these AISC apps vaguely described the measures they employed to guarantee the safety, security, and privacy of users' data. It is important for AISC apps to disclose sufficient information to users. How they comply with credentials regarding data safety and security and how they communicate with third-party services are nontransparent, posing risks in users' privacy protection.

Existing studies found that AI can create ethical and social issues due to the lack of governance in transparency, safety, and privacy [68,99]. To enhance the safety of AI technologies, a body of studies has explored the security of AI and emphasized the importance of AI governance [14,50]. They have mainly proposed three categories of methods to enhance the safety of AI: (1) companies should involve experts or a committee to evaluate and oversee the AI issues [48]; (2) the society may create a global government model to govern AI systems [35]; and (3) governments and companies may strengthen relevant standards, policies and laws [19,38,78,99]. Nevertheless, these studies are still in the early stage while research for governance and policy is necessary [100]. The implications of our study resonate with the third category of AI governance. Our study reveals that AI governance in healthcare can be conducted by establishing comprehensive policies. Governments should refine their regulations and confine the release of unqualified healthcare apps. Companies in healthcare should disclose enough information to users, including more well-rounded information regarding accountability, transparency, and data protection policies.

Design implications: Our study reveals that existing health AI applications should make more efforts in making transparent their qualifications, algorithm mechanisms, and commercial interests. Our work also reinforces the call for studies in health data protection. Relevant policies should be enacted regarding what responsibilities these AI applications must have and what data protection methods these AI applications should follow. To be specific, policies should regulate what data safety credentials these AI applications must comply with and what measures of the data storage, data encryption, data usage, and data collection should be stated in the privacy policies and terms of use. By doing this, users can acquire sufficient information to decrease the possibilities of blindly trusting.

5.2 Evaluating AI algorithms from a user's standpoint

Our study found that when users were interacting with AISC apps, they developed various strategies to evaluate the accuracy of the AI algorithms behind AISC apps. When interacting with AISC apps, users tried to compare their experience of using AISC apps with real-world consulting experience. Users also compared AISC apps with other diagnostic tools. For example, they compared the response time of Ada with that of Google. While evaluating AI algorithms is essential as AI algorithms can lead to inaccurate predictions and unintended consequences [47,68], few studies have considered users' standpoints for AI algorithms. Most previous studies focused on utilizing technical approaches [16,23,27,29,37,75,107,108], such as developing a platform or model to evaluate AI algorithms and models. Little research has explored the evaluation of AI algorithms from users' points of view [2,63]. Our study, however, demonstrates that we should focus on users' perspectives, as the interaction between users and systems should be emphasized [106].

Our findings reported that users evaluated AI algorithms not only through the diagnostic results but also through details in interactions. For example, when users input their symptoms, the restricting dimensions of symptoms made users doubt the accuracy of AISC apps. Users also tempered their expectations for AISC apps when they found some AISC apps lacked support for certain diseases, such as chronic diseases. Thus, AI algorithms design and research should be more comprehensive, considering diverse factors and enhancing the knowledge base. In our case, the AI algorithms of AISC apps should accommodate the complexity of describing symptoms and the diversity of human diseases.

Our study reveals how users evaluate the accuracy of AI algorithms. The way users evaluate the AI algorithms can affect users' further use and trust-building [40,88,93,101]. However, to our best knowledge, only one study has explored users' attitudes towards the AI system (i.e., face verification

systems), but this study only investigated users' satisfaction and acceptance and did not explore the approaches users employed to evaluate these systems and algorithms [28]. Our study, however, shows the strategies users utilized to evaluate AI systems, helping us understand user's needs that users desired more flexible AI algorithms that could be mapped into reality.

Design implications: Future design of AI algorithms or systems should accommodate diverse contexts and consider different users' needs. For example, in the case of AISC apps, enhancing their knowledge base for diverse diseases can improve users' trust.

5.3 Bridging social-technical gaps between AI systems and offline medical consultations

Our study identifies that there is a social-technical gap between the features of AISC apps and offline medical consultations during each diagnostic process that the AISC apps support.

When establishing a patient history, a clinical probing process usually collects a complete medical history and reviews a patient's previous activities [39]. However, the profiles of the AISC apps missed critical information in their medical history, which disappointed users. As such, the current profile of AISC apps needs to be more comprehensive.

We also found the AISC apps' knowledge base ill-suited to disease complexity during the evaluation of symptoms. For example, when inputting symptoms, users found it difficult to input the sufficient dimensions of their conditions. Users could not input multiple locations, the frequency, and the severity of their symptoms. However, in the offline setting, describing symptoms is complicated, requiring details regarding frequency, severity level, and locations**.**

Additionally, the real-world probing process requires communication skills, including empathy and the building of rapport [95]. Users complained about the stilted language of the apps'

probing questions. Although some AISC apps employ humanizing language, work is still needed to more closely mimic human conversation.

Finally, users perceived that the AISC apps lacked support for diverse users, echoing the guidelines of inclusive design, which requires product design to take into account the needs of all users without requirements for adaption [1]. The inclusiveness of offline healthcare services has already been emphasized [9]. Our study reveals that we should also consider inclusiveness in AISC apps design, reporting that users wanted AISC apps to accommodate diverse user groups.

These findings offered a new angle to consider in future consumer-facing diagnostic tools design and research: users' experiences can be improved if these tools can provide similar functions and experiences to clinical visits in the real world. Researchers should strive to take into consideration users' offline experiences and follow the guidelines of inclusive design to accommodate diverse user groups.

Design implications: Future design of AI algorithms or systems should follow the guidelines of inclusive design to take into consideration users' offline experience. The functions of an AI system should mimic the processes that already exist in the real world. In the case of the AISC apps, establishing a comprehensive health profile, allowing users to input different levels of symptoms, and improving response speed could improve the user experience.

5.4 Addressing users' needs in conversational design of chatbots

Our findings reported users had concerns regarding human characteristics and input limitations, and the comprehensibility of language in the conversational design of AISC apps.

First, from feature analysis, we found most AISC apps involve two types of human characteristics in their conversational design. The first type is an avatar. Ask NHS and Sensely use avatars to interact with users. These avatars have the human-like appearance (looking like a nurse), actions (blinking eyes), and a real human voice. An avatar is "an image that represents a user in a

multi-user virtual reality space" [92]. It is one kind of anthropomorphic design, involving human characteristics. The other type is a picture of real human body parts used to explain medical terms that appeared in probing questions. For example, Ada uses pictures of real human body parts to explain medical jargons in its probing questions.

A body of studies has explored if we should involve human characteristics into the conversational design of chatbots. On the one hand, plenty of studies believe the usage of human-like characteristics has a positive effect on the interaction between users and chatbots [5,49,90,92,97]. This series of studies resonate with the theory of Computers are Social Actors (CASA) [49], which states visual anthropomorphic features may promote the social presence of these health chatbots, helping users treat chatbots socially and naturally [49]. On the other hand, some studies pointed out users would experience more uncanny effects when using complex avatars [18,82]. Considering users' preferences is essential to improve the users' perceptions of the effectiveness of chatbots [77,80]. However, there is no final conclusion regarding whether human characteristics should be embedded in the conversational design of chatbots. Our study reported that users did not have a uniform attitude towards human characteristics. Some users desired that chatbots could show human feelings, such as a more human-like conversational style. Some users thought these human characteristics were unnecessary. For example, some users doubted the necessity of the sentences showing sympathy to users. Other users disliked human characteristics. For example, one user was frightened by a real photo of a human body part that was used to explain medical terms. Another instance is that most users perceived the avatar of Ask NHS as freaky. Among them, one user thought neither too human-like nor too bot-like would be an acceptable design for an avatar. This finding resonates with the Uncanny Valley of Mind (UVM) theory that people are frightened sometimes by an almost realistic human appearance [82]. Thus, it is critical for reaching a balance between the real human-like appearance and the bot-like appearance when involving human characteristics in the conversational design.

Furthermore, our study reveals users' needs regarding human characteristics in the conversational design of chatbots. Few studies have explored deeply the reasons why users like or dislike the human characteristics and most existing studies utilized quantitative methods or surveys to investigate users' attitudes. The exception is one study that explored the reasons why users dislike the appearance of avatars using in-depth interviews [94]. However, this study is conducted through the lens of experts instead of users. Further research, especially qualitative and phycological research should be conducted to explore the ways and effects of human characteristics embedded in chatbots from the users' side to facilitate better interactions [18,26]. By employing the qualitative methodology, our study explored deeply users' attitudes towards human characteristics through the case of AISC apps, providing another new stand of point to consider in the future conversational design of chatbots. Our findings reported that when introducing human characteristics into the design of chatbots, we should consider what human characteristics are unnecessary and what level of them we should involve in different situations. For example, according to P10, when explaining a medical term appeared in a question, cartoon pictures were enough for pain locations while a real picture of human skin was needed for terms regarding skin problems. We need to reach a balance between real human characteristics and virtual characteristics. This suggests that chatbot designers and researchers should pay attention to the nuanced requirements of users. Different strategies should be adopted in diverse situations in the conversational design of chatbots.

Second, our study reported users had difficulties when inputting symptoms. Users found that some AISC apps could not recognize their input, such as medical terms and common words they used in daily lives. Previous research has found that the difficulty of describing symptoms accurately can affect users' acceptability for chatbots [67] and recognizing human inputs accurately plays an important role in chatbot design [8]. Our study confirms these arguments by finding that our participants desired flexible input in the conversations with chatbots. They wanted to input

symptoms more easily, using words they perceived as simple and familiar. This calls for research on approximate string matching and character recognition.

Third, our findings reported that users perceived some language difficult to understand. Our study found that users had difficulty understanding the medical jargon that appeared in questions, which may lead to inappropriate input from users. Previous studies have emphasized the importance of decreasing the language complexity used in healthcare systems, such as the medication management systems of older patients [96] and the Electronic Health Records (EHR) [87] used in offline medical visits. Few studies have examined the users' perceptions of the language complexity of online healthcare chatbots. Our study, however, reveals that health chatbots also need to increase the comprehensibility of their probing questions based on users' perceptions, as a great number of people do not have a sufficient health literacy to interact with these healthcare systems [96]. In addition, users felt confused about the sequence of questions and the relationship among these questions; they did not know how these questions were related to the diagnostic results. Thus, information regarding why AISC apps ask certain questions, how these questions are related, and how these questions relate to the diagnostic results should be explained to users. This indicates that we should provide an explanation of the AISC apps' AI model to users. Previous studies drew on the social sciences and HCI knowledge to explain AI models [4,21] to users. While these studies provide insight as to what can constitute a good explanation, limited attention has been paid to users' needs for the explanations themselves. Since users are the one to evaluate explanations, in contrast to the studies that propose researcher-driven explanations, our study highlights the users' requirements for explanations in conversation with healthcare chatbots.

Drawing from our findings and discussion, the future conversational design of healthcare chatbots should consider how to improve input flexibility and the presentation of probing questions. First, the design should improve the functions of approximate string matching and character recognition

and assist users to input their symptoms. Second, healthcare chatbots should use comprehensible language and provide explanations during conversations.

Design implications: Drawing from our discussions, the future conversational design of chatbots should consider how to involve human characteristics and how to improve input flexibility as well as the presentation of probing questions. First, when designing avatars or pictures, the human characteristics and virtual characteristics should be balanced. Second, different situations should involve different extents of human characteristics. In our case, presentations for skin problems need to involve more human characteristics. Third, chatbots should enhance their functions regarding input recognition and approximate string matching, making the conversations more user-friendly. Finally, healthcare chatbots should use comprehensible language and provide explanations during conversations.

Chapter 6

Conclusions & limitations

Our study aims to explore how users evaluate the accuracy and effectiveness of AISC apps and understand the algorithms of AISC apps. By reviewing the development background and features of AISC apps, we found that existing AISC apps need to disclose more information regarding their accountability, transparency, and data protection measures. By studying user reviews and conducting interviews, we analyzed how users evaluated and understood AISC algorithms. We identified the information that should be disclosed in the privacy policies and terms of use, the approaches users utilized to evaluate algorithms, and methods that should be employed to the AI algorithms design. We also shed light on the future conversational design of chatbots with regard to human characteristics and input flexibility.

Our study also has some limitations. First, because the reviews from the Apple App Store and the Google Play Store do not include the demographic information or distribution of users, we did not factor in this information. Second, not all users posted their reviews online, resulting in a selection bias in the users' perspectives we acquired. Third, we only conducted ten interviews, and thus the interview sample is rather small. Although the combined interview and app review data did reach the theoretical saturation, larger scale studies need to be conducted in the future to acquire more comprehensive information regarding users' perceptions.

References

1. Julio Abascal and Colette Nicolle. 2005. Moving towards inclusive design guidelines for socially and ethically aware HCI. *Interacting with computers* 17: 484–505. https://doi.org/10.1016/j.intcom.2005.03.002
2. Stephanie Aboueid, Rebecca H Liu, Binyam Negussie Desta, and Ashok Chaurasia. 2019. The Use of Artificially Intelligent Self-Diagnosing Digital Platforms by the General Public: Scoping Review. *JMIR medical informatics* 7, 2: e13445. https://doi.org/10.2196/13445
3. Diagnostic Accuracy. 2019. Comparison of Physician and Computer Diagnostic Accuracy. *JAMA internal medicine* 176, 12: 2015–2016. https://doi.org/10.1001/jamainternmed.2016.6001
4. Arjun Akula, Changsong Liu, Sinisa Todorovic, Joyce Chai, and Song-chun Zhu. 2019. Explainable AI as collaborative task solving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 91–94.
5. Theo Araujo. 2018. Computers in Human Behavior Living up to the chatbot hype: The in fluence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85: 183–189. https://doi.org/10.1016/j.chb.2018.03.051
6. Shifali Arora, Jennifer Yttri, D Ph, Wendy Nilsen, and D Ph. 2014. Privacy and Security in Mobile Health (mHealth) Research. *Alcohol research: current reviews* 36, 1: 143.
7. Veronica Barassi. 2017. BabyVeillance? expecting parents, online surveillance and the cultural specificity of pregnancy apps. *Social Media+ Society* 3, 2: 2056305117707188.
8. Mary Bates. 2019. Health care chatbots are here to help. *IEEE Pulse* 10, May: 12–14. https://doi.org/10.1109/MPULS.2019.2911816
9. Carolyn Beniuk, James Ward, and P John Clarkson. 2011. Applying inclusive design principles in

the assessment of healthcare services. In *Proceedings of conference on design*, 22–36.

10. A C Berry, B D Cash, B Wang, M S Mulekar, A B Van Haneghan, K Yuquimpo, A Swaney, M C Marshall, and W K Green. 2019. Online symptom checker diagnostic and triage accuracy for HIV and hepatitis C. *Epidemiology & Infection* 147: 9–12. https://doi.org/10.1038/ajg.2017.305.Cite
11. Soumitra S Bhuyan, Hyunmin Kim, Oluwaseyi O Isehunwa, Naveen Kumar, Jay Bhatt, David K Wyant, Satish Kedia, Cyril F Chang, and Dipankar Dasgupta. 2017. Privacy and security issues in mobile health: Current research and future directions. *Health Policy and Technology* 6, 2: 188–191. https://doi.org/10.1016/j.hlpt.2017.01.004
12. Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2: 77–101.
13. Clara Caldeira, Yu Chen, Lesley Chan, Vivian Pham, Yunan Chen, and Kai Zheng. 2017. Mobile apps for mood tracking: an analysis of features and user reviews. In *AMIA Annual Symposium Proceedings*, 495.
14. Ryan Calo. 2017. Artificial Intelligence Policy: A Primer and Roadmap. *UCDL Rev.* 51: 399.
15. Duncan Chambers, Anna J Cantrell, Maxine Johnson, Louise Preston, Susan K Baxter, Andrew Booth, and Janette Turner. 2019. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ open* 9, 8: e027743. https://doi.org/10.1136/bmjopen-2018-027743
16. X L Chang, X M Mi, and J K Muppala. 2013. Performance evaluation of artificial intelligence algorithms for virtual network embedding. *Engineering Applications of Artificial Intelligence* 26, 10: 2540–2550. https://doi.org/10.1016/j.engappai.2013.07.007
17. Hao-fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. 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*, 559.
18. Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the Shades of the Uncanny Valley: An Experimental Study of Human–Chatbot Interaction. *Future*

Generation Computer Systems 92: 539--548.

19. Peter Cihon. 2019. Standards for AI Governance : International Standards to Enable Global Coordination in AI Research & Development. *Future of Humanity Institute*: 1–41.
20. By Juliet Corbin and Anselm Strauss. 2010. *Basics of qualitative research techniques and procedures for developing grounded theory*. Sage publications.
21. Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, Milton Rosenberg, and Marina Rey. 2006. Building explainable Artificial Intelligence systems. *AAAI*: 1766–1773.
22. Rik Crutzen, D Ph, Gjalt-jorn Y Peters, D Ph, M Sc, Erwin M Fisser, B A, Jorne J Grolleman, and M Sc. 2011. An Artificially Intelligent Chat Agent That Answers Adolescents ' Questions Related to Sex , Drugs , and Alcohol : An Exploratory Study. *Journal of Adolescent Health* 48, 5: 514--519. https://doi.org/10.1016/j.jadohealth.2010.09.002
23. Abdul Dakkak, Cheng Li, Abhishek Srivastava, Jinjun Xiong, and Wen-Mei Hwu. 2018. MLModelScope: Evaluate and Measure ML Models within AI Pipelines. *arXiv preprint arXiv:1811.09737*.
24. Benjamin Marshall Davies, Mbchb Hons, Colin Fraser Munro, B A Hons, Mark Rn, and Robinson Way. 2019. A Novel Insight Into the Challenges of Diagnosing Degenerative Cervical Myelopathy Using Web-Based Symptom Checkers. *Journal of medical Internet research* 21, 1: e10868. https://doi.org/10.2196/10868
25. Erin Dietsche. 2018. Babylon Health says its AI's advice is "on par with practicing clinicians," but one medical organization is skeptical. *MedCityNews*. Retrieved from https://medcitynews.com/2018/07/babylon-health/
26. Laury Donkelaar. 2018. How human should a chatbot be?: The influence of avatar appearance and anthropomorphic characteristics in the conversational tone regarding chatbots in customer service field. University of Twente.
27. T H E Effectiveness and O F Retrieval Algorithms. 1991. Determing the effectiveness of retrieval algorithms. *Information Processing & Management* 27, 2–3: 153–164.
28. Mohamad El-abed, Romain Giot, Baptiste Hemery, Christophe Rosenberger, Mohamad El-abed,

Romain Giot, Baptiste Hemery, Christophe Rosenberger, Mohamad El-abed, Romain Giot, Baptiste Hemery, and Christophe Rosenberger. 2010. A study of users' acceptance and satisfaction of biometric systems. In *44th Annual 2010 IEEE International Carnahan Conference on Security Technology*, 170–178.

29. Ziv Epstein, Blakeley H Payne, Judy Hanwen Shen, Casey Jisoo Hong, Bjarke Felbo, Abhimanyu Dubey, Matthew Groh, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan. 2018. Turingbox: an experimental platform for the evaluation of AI systems. In *IJCAI 2018*, 5826–5828.
30. Ahmed Fadhil and Silvia Gabrielli. 2017. Addressing Challenges in Promoting Healthy Lifestyles: The AI-Chatbot Approach. In *Proceedings of the 11th EAI international conference on pervasive computing Technologies for Healthcare*, 261--265.
31. Mitchell J Feldman, Edward P Hoffer, G Octo Barnett, Richard J Kim, Kathleen T Famiglietti, and Henry C Chueh. 2012. Impact of a computer-based diagnostic decision support tool on the differential diagnoses of medicine residents. *Journal of Graduate Medical Education* 4, 2: 227–231.
32. Pierre-antoine Fougerouse, Mobin Yasini, Guillaume Marchand, Oliver O Aalami, Palo Alto, United States, and D M D Santé. 2017. A Cross-Sectional Study of Prominent US Mobile Health Applications: Evaluating the Current Landscape. In *AMIA Annual Symposium Proceedings*, 715.
33. By Jennifer Frank. 2017. Patient Concerns Are as Important as Symptoms. Retrieved from https://www.physicianspractice.com/blog/patient-concerns-are-important-symptoms
34. Quinn H Grundy, Zhicheng Wang, and Lisa A Bero. 2019. Challenges in Assessing Mobile Health App Quality. *American Journal of Preventive Medicine* 51, 6: 1051–1059. https://doi.org/10.1016/j.amepre.2016.07.009
35. The Harvard. 2017. A Layered Model for AI Governance. https://doi.org/10.1109/MIC.2017.4180835
36. Sarah C Haynes and Katherine K Kim. 2017. A mobile system for the improvement of heart failure management: evaluation of a prototype. In *AMIA Annual Symposium Proceedings*, 839.
37. Perfecto Herrera. 2008. A New Approach to Evaluating Novel Recommendations. In *Proceedings*

of the 2008 ACM conference on Recommender systems, 179–186.

38. C W L Ho, D Soon, K Caals, and J Kapur. 2020. Governance of automated image analysis and artificial intelligence analytics in healthcare. 74, 2019: 329–337. https://doi.org/10.1016/j.crad.2019.02.005

39. J Hoffman. 2014. Annual benchmarking report: malpractice risks in the diagnostic process. *Cambridge: CRICO Strategies*: 1–20.

40. Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain. *arXiv preprint arXiv:1712.09923*: 1–28.

41. Jeremy Hsu. 2019. Medical advice from a bot: The unproven promise of Babylon Health. *Salon*. Retrieved from https://www.salon.com/2019/12/14/medical-advice-from-a-bot-the-unproven-promise-of-babylon-health_partner/

42. David Ireland, Christina Atay, Jacki Liddle, Dana Bradford, Helen Lee, Olivia Rushin, Thomas Mullins, Dan Angus, Janet Wiles, and Simon McBride. 2016. Hello Harlie : Enabling Speech Monitoring Through Chat-Bot Conversations. In *Digital Health Innovation for Consumers, Clinicians, Connectivity and Community-Selected Papers from the 24th Australian National Health Informatics Conference, HIC 2016, Melbourne, Australia, July 2016.*, 55–60. https://doi.org/10.3233/978-1-61499-666-8-55

43. Mirjana Ivanovic and Marija Semnic. 2018. The Role of Agent Technologies in Personalized Medicine. In *2018 5th International Conference on Systems and Informatics (ICSAI)*, 299–304.

44. Sabin Kafle, Penny Pan, Ali Torkamani, Stevi Halley, and John Powers. 2018. Personalized symptom checker using medical claims. In *HealthRecSys@ RecSys*, 13–17.

45. Rafal Kocielnik, Elena Agapie, Alexander Argyle, Dennis T Hsieh, Kabir Yadav, Breena Taira, and Gary Hsieh. HarborBot: A Chatbot for Social Needs Screening.

46. David Kotz, Carl A Gunter, Santosh Kumar, and Jonathan P Weiner. 2016. Privacy and Security in Mobile Health: A Research Agenda. *Computer* 49: 22–30. https://doi.org/10.1109/MC.2016.185

47. Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia

Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Review conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9: 1248–1258. https://doi.org/10.1093/jamia/ocy072

48. By Anastassia Lauterbach and Andrea Bonime-blanc. 2016. Artificial Intelligence: A Strategic Business And Governance Imperative. *NACD Directorship, September/October*: 54–57.

49. Bingjie Liu and S Shyam Sundar. 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10: 625–636. https://doi.org/10.1089/cyber.2018.0110

50. Michael Lockwood, Julie Davidson, Allan Curtis, Elaine Stratford, Michael Lockwood, Julie Davidson, Allan Curtis, and Elaine Stratford. 2010. Governance Principles for Natural Resource Management. *Society and natural resources* 23, 10: 986–1001. https://doi.org/10.1080/08941920802178214

51. Joao Luis, Zeni Montenegro, Cristiano André, and Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems With Applications* 129: 56–67. https://doi.org/10.1016/j.eswa.2019.03.054

52. Deborah Lupton. 2014. Apps as Artefacts: Towards a Critical Perspective on Mobile Health and Medical Apps. *Societies* 4, 4: 606–622. https://doi.org/10.3390/soc4040606

53. Deborah Lupton. 2014. Quantified sex: a critical analysis of sexual and reproductive self-tracking apps. *Culture, health & sexuality* 17, 4: 440–453. https://doi.org/10.1080/13691058.2014.920528

54. Deborah Lupton and Annemarie Jutel. 2015. 'It' s Like Having a Physician in Your Pocket!' A Critical Analysis of Self-Diagnosis Smartphone Apps. *Social Science & Medicine*, June. https://doi.org/10.1016/j.socscimed.2015.04.004

55. Deborah Lupton and Annemarie Jutel. 2015. Social Science & Medicine ' It ' s like having a physician in your pocket ! ' A critical analysis of self- diagnosis smartphone apps. *Social Science & Medicine* 133, January 2014: 128–135. https://doi.org/10.1016/j.socscimed.2015.04.004

56. Courtney Rees Lyles, D Ph, Lynne T Harris, Tung Le, Jan Flowers, James Tufano, D Ph, Diane Britt, James Hoath, Irl B Hirsch, Harold I Goldberg, and James D Ralston. 2011. Qualitative

evaluation of a mobile phone and web-based collaborative care intervention for patients with type 2 diabetes. *Diabetes technology & therapeutics* 13, 5: 563–569. https://doi.org/10.1089/dia.2010.0200

57. Ulrik Lyngs, Kai Lukoff, Max Van Kleek, Nigel Shadbolt, and Reuben Binns. 2019. Self-control in cyberspace: applying dual systems theory to a review of digital self-control tools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–18.

58. Carl Macrae. 2019. Governing the safety of artificial intelligence in healthcare. *BMJ quality & safety* 28, 6: 495–498. https://doi.org/10.1136/bmjqs-2019-009484

59. Borja Martínez-pérez, Isabel De Torre-díez, and Miguel López-coronado. 2015. Privacy and Security in Mobile Health Apps : A Review and Recommendations. *Journal of medical systems* 39, 1: 181. https://doi.org/10.1007/s10916-014-0181-3

60. Gemma Flores Mateo, Esther Granado-font, Carme Ferr, and Xavier Monta. 2015. Mobile phone apps to promote weight loss and increase physical activity: a systematic review and meta-analysis. *Journal of medical Internet research* 17, 11: e253. https://doi.org/10.2196/jmir.4836

61. Margaret Mccartney. 2018. Margaret McCartney : AI in medicine must be rigorously. *Bmj* 361: k1752. https://doi.org/10.1136/bmj.k1752

62. Katherine Middleton, B M Bch, B A Hons, Mobasher Butt, Hons Mb, Nils Hammerla Ph D, Steven Hamblin Ph D, Karan Mehta, Bmedsci Bm, and Ali Parsa Ph D. 2016. Sorting out symptoms: design and evaluation of the'babylon check'automated triage system. *arXiv preprint arXiv:1606.02041*.

63. Michael L Millenson, Lorri Zipperer, Zippperer Project Management, and Hardeep Singh. 2018. Beyond Dr . Google : the evidence on consumer- facing digital tools for diagnosis. *Diagnosis* 5, 3: 95–105. https://doi.org/10.1515/dx-2018-0009

64. Tomohiro Morita, Abidur Rahman, Takanori Hasegawa, Akihiko Ozaki, and Tetsuya Tanimoto. 2017. The Potential Possibility of Symptom Checker. *International journal of health policy and management* 6, 10: 615. https://doi.org/10.15171/ijhpm.2017.41

65. Patricia Morreale. 2018. m-Health application Interface design for symptom checking. In

Proceedings of the 10th International Conference on e-Health 2018 (EH 2018), 17–19.

66. Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, Stephen M Schueller, and Robert R Morris. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research* 20, 6: e10148. https://doi.org/10.2196/10148

67. Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health* 5: 2055207619871808. https://doi.org/10.1177/2055207619871808

68. Kee Yuan Ngiam and Ing Wei Khor. 2019. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology* 20, 5: e262--e273. https://doi.org/10.1016/S1470-2045(19)30149-4

69. Mary K O'Brien, Keith Petrie, and John Raeburn. 1992. Adherence to medication regimens: updating a complex medical issue. *Medical care review* 49, 4: 435--454.

70. Jeungmin Oh, Service Engineering, Daehoon Kim, Service Engineering, Uichin Lee, Service Engineering, Jae-gil Lee, Service Engineering, and Junehwa Song. 2013. Facilitating developer-user interactions with mobile app review digests. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 1809–1814.

71. Main Outcomes. 2019. Accuracy of a Popular Online Symptom Checker for Ophthalmic Diagnoses. *JAMA ophthalmology* 137, 6: 690–692. https://doi.org/10.1001/jamaophthalmol.2019.0571

72. L A Page and S Wessely. 2003. Medically unexplained symptoms: exacerbating factors in the doctor-patient encounter. *Journal of the Royal Society of Medicine* 96, 5: 223–227.

73. Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 2019. Physicians' perceptions of chatbots in health care. *Journal of medical Internet research* 21, 4: e12887. https://doi.org/10.2196/12887

74. Danny Palmer. 2019. What is GDPR? Everything you need to know about the new general data protection regulations. *ZDNet Academy*. Retrieved from https://www.zdnet.com/article/gdpr-an-

executive-guide-to-what-you-need-to-know/

75. Manos Papagelis and Dimitris Plexousakis. 2005. Qualitative Analysis of User-based and Item-based prediction Algorithms for Recommendation Agents. *Engineering Applications of Artificial Intelligence* 18, 7: 781--789.

76. Hannah E Payne, Cameron Lister, Joshua H West, and Jay M Bernhardt. 2015. Behavioral Functionality of Mobile Apps in Health Interventions : A Systematic Review of the Literature Corresponding Author : *JMIR mHealth and uHealth* 3, 1: e20. https://doi.org/10.2196/mhealth.3335

77. Juanan Pereira. 2019. Using health chatbots for behavior change: a mapping study. *Journal of Medical Systems* 43, 5: 135.

78. Brandon Perry and Risto Uuk. 2019. AI Governance and the Policymaking Process : Key Considerations for Reducing AI Risk. 2017. https://doi.org/10.3390/bdcc3020026

79. Kathleen H Pine. 2018. Data Work in Healthcare: Challenges for Patients, Clinicians and Administrators. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 433–439.

80. John Powell and John Powell. 2019. Trust me, I'm a chatbot: how artificial intelligence in health care fails the turing test. *Journal of medical Internet research* 21, 10: e16222. https://doi.org/10.2196/16222

81. Lucy Powley, Graham Mcilroy, Gwenda Simons, and Karim Raza. 2016. Are online symptoms checkers useful for patients with inflammatory arthritis ? *BMC Musculoskeletal Disorders* 17, 1: 362. https://doi.org/10.1186/s12891-016-1189-2

82. Alvin Rajkomar, Jeffrey Dean, and Kohane Isaac. 2019. Machine Learning in Medicine. *New England Journal of Medicine* 380, 14: 1347–1358. https://doi.org/10.1056/NEJMra1814259

83. Salman Razzaki, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, and Daniel Mullarkey. 2018. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv preprint arXiv:1806.10698*.

84. Claudia Rijcken. 2019. *Pharmbot canopies*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-

817638-2.00012-2

85. Jordan Rivera, Amy Mcpherson, Jill Hamilton, Michael Coons, Sindoora Iyer, and Arnav Agarwal. 2016. Mobile apps for weight management: a scoping review. *JMIR mHealth and uHealth* 4, 3: e87. https://doi.org/10.2196/mhealth.5115

86. Claudia Daudén Roquet. 2018. Evaluating mindfulness meditation apps. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6.

87. Tarek Sakakini, Renato F L Azevedo, Victor Sadauskas, Kuangxiao Gu, Yang Zhang, Ann Willemsen-dunlap, Donald J Halpin, and James Graumlich. 2017. Dr. Babel Fish: a machine translator to simplify providers' language. In *AMIA*.

88. Marcel Salathé, Thomas Wiegand, and Markus Wenzel. 2018. Focus Group on Artificial Intelligence for Health. *arXiv preprint arXiv:1809.04797*.

89. Patrick Sanger, Andrea Hartzler, William B Lober, Heather L Evans, and Wanda Pratt. 2014. Design Considerations for Post-Acute Care mHealth: Patient Perspectives. In *AMIA Annual Symposium Proceedings*, 1920.

90. Anna-maria Seeger, Armin Heinzl, Anna-maria Seeger, and Armin Heinzl. 2017. When Do We Need a Human? Anthropomorphic Design and Trustworthiness of Conversational Agents. In *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS, AISeL, Seoul, Korea*.

91. Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. 2015. Evaluation of symptom checkers for self diagnosis and triage : audit study. *bmj* 351: h3480. https://doi.org/10.1136/bmj.h3480

92. B Y R A J Sheth. 2003. Avatar Technology: Giving a Face to the e-Learning Interface. *The eLearning Developers' Journal*: 1–10.

93. Keng Siau and Weiyu Wang. 2018. Building Trust in Artificial Intelligence , Machine Learning , and Robotics. *Cutter Business Technology Journal* 31, 2: 47–53.

94. Mohammed Slim, Ben Mimoun, Ingrid Poncin, and Marion Garnier. 2012. Case study—Embodied virtual agents: An analysis on reasons for failure. *Journal of Retailing and Consumer Services* 19,

6: 605–612. https://doi.org/10.1016/j.jretconser.2012.07.006

95. Sneha Baxi Srivastava. 2013. The patient interview. *Fundamental Skills for Patient Care in Pharmacy Practice* 1: 1–36.

96. Priyadarshi Tiwari, Jim Warren, and Karen Day. 2011. Empowering older patients to engage in self care: designing an interactive robotic device. In *AMIA Annual Symposium Proceedings*, 1402.

97. Rita Toader, M Mara, Cezar Toader, and Diana-cezara Toader. 2019. The Effect of Social Presence and Chatbot Errors on Trust. *Sustainability* 12, 1: 1–24.

98. Lucia Vaira, Mario A Bochicchio, Matteo Conte, Francesco Margiotta Casaluci, and Antonio Melpignano. 2018. MamaBot : a System based on ML and NLP for supporting Women and Families during Pregnancy. In *Proceedings of the 22nd International Database Engineering \& Applications Symposium*, 273–277.

99. Wendell Wallach and Gary E Marchant. 2018. *An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics*. The Hastings Center.

100. Weiyu Wang. 2018. Artificial Intelligence: A Study on Governance, Policies, and Regulations.

101. Weiyu Wang and Keng Siau. Living with Artificial Intelligence – Developing a Theory on Trust in Health Chatbots.

102. Pierre Wargnier, Adrien Malaisé, Julien Jacquemot, Samuel Benveniste, Maribel Pino, Anne-sophie Rigaud, and Pierre Jouvelot. 2015. Towards attention monitoring of older adults with cognitive impairment during interaction with an embodied conversational agent. In *2015 3rd IEEE VR International Workshop on Virtual and Augmented Assistive Technology (VAAT)*, 23–28.

103. David S Watson, Jenny Krutzinna, Ian N Bruce, Iain B Mcinnes, and Michael R Barnes. 2019. Clinical applications of machine learning algorithms : beyond the black box. *BMJ* 364: l886. https://doi.org/10.1136/bmj.l886

104. Arthur Willem, Gerard Buijink, Benjamin Jelle Visser, Louise Marshall, Arthur Willem, and Gerard Buijink. 2013. Medical apps for smartphones: lack of evidence undermines quality and safety. *BMJ Evidence-Based Medicine* 18, 3: 90–92.

105. Maria Klara Wolters, Fiona Kelly, and Jonathan Kilgour. 2016. Designing a spoken dialogue

interface to an intelligent cognitive assistant for people with dementia. *Health informatics journal* 22, 4: 854–866. https://doi.org/10.1177/1460458215593329

106. David Wong. 2018. Safety of patient-facing. *The Lancet* 392, 10161: 2263–2264. https://doi.org/10.1016/S0140-6736(18)32819-8

107. Deshraj Yadav. 2019. EvalAI : Towards Better Evaluation Systems for AI Agents. *arXiv preprint arXiv:1902.03570*.

108. Makoto Yamashita, Katsuki Fujisawa, and Masakazu Kojima. 2003. Implementation and evaluation of SDPA 6.0 (semidefinite programming algorithm 6.0). *Optimization Methods and Software* 18, 4: 491–505. https://doi.org/10.1080/1055678031000118482

109. JINGCONG ZHAO. 2019. What is ISO 27001 and the Benefits of Getting Certified. *hyperproof*. Retrieved from https://hyperproof.io/iso27001-certification/

110. Bin Zhu, Anders Hedman, and Haibo Li. 2017. Designing digital mindfulness: presence-in and presence-with versus presence-through. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2685–2695.

111. Timeline of artificial intelligence. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Timeline_of_artificial_intelligence#2010s

112. Ask NHS. Retrieved from https://www.sensely.com/asknhs/

113. Basics of the seal. *Bundesverband der Internetmedizin*. Retrieved from https://bundesverbandinternetmedizin.de/home/siegel/

114. About FDA. *U.S. Food & Drug Administration*. Retrieved from https://www.fda.gov/about-fda

115. Medicines and Healthcare products Regulatory Agency. *GOV.UK*. Retrieved from https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency/about

116. ISO/IEC 27001:2013. *International Organization for Standardization*. Retrieved from https://www.iso.org/standard/54534.html

117. Diseases and conditions dictionary. *MedicineNet, Inc*. Retrieved from https://www.medicinenet.com/diseases_and_conditions/article.htm

118. 2016. CE marking. *European Union*. Retrieved from https://ec.europa.eu/growth/single-market/ce-marking/

119. 2016. ISO 13485:2016. *International Organization for Standardization*. Retrieved from https://www.iso.org/standard/59752.html

120. 2018. Mobile Medical Applications. *Food and Drug Administration*. Retrieved from https://www.fda.gov/medicaldevices/%0Ddigital-health/mobile-medical-applications

121. 2019. Health Insurance Portability & Accountability Act. *California Department of Health Care Services (DHCS)*. Retrieved from https://www.dhcs.ca.gov/formsandpubs/laws/hipaa/Pages/1.00WhatisHIPAA.aspx

122. 2019. What is sickle cell disease? *Centers for Disease Control and Prevention*. Retrieved from https://www.cdc.gov/ncbddd/sicklecell/facts.html

123. 2020. Babylon Health UK. Retrieved from https://www.babylonhealth.com/us/what-we-offer

124. 2020. Care Quality Commission. Retrieved from https://www.cqc.org.uk/about-us/our-purpose-role/who-we-are