

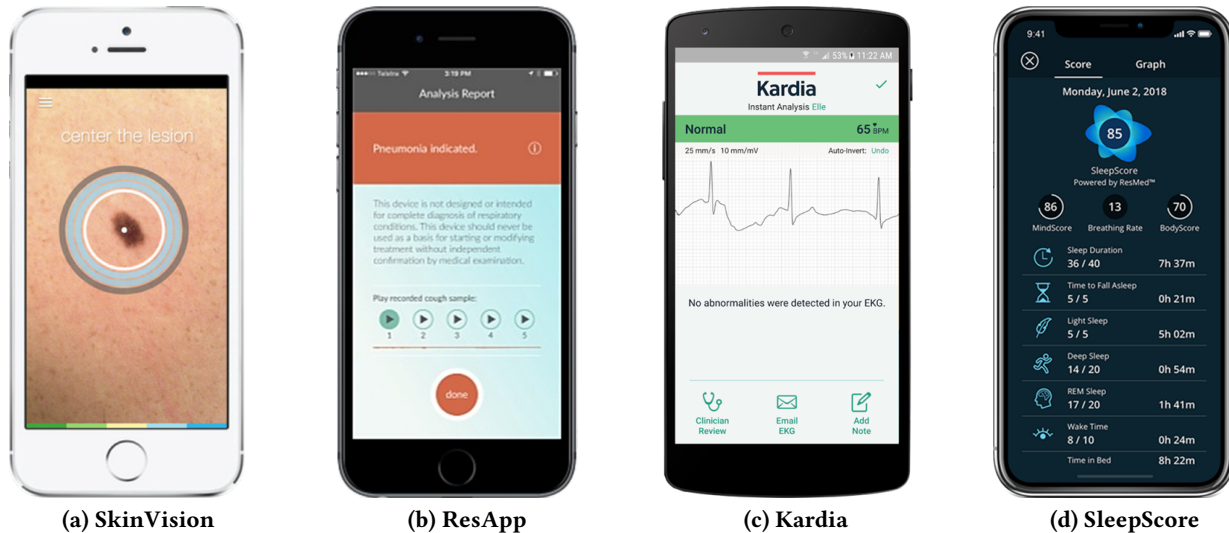


# Trust Me, I'm a Doctor – User Perceptions of AI-Driven Apps for Mobile Health Diagnosis

Matthias Baldauf  
matthias.baldauf@ost.ch  
Eastern Switzerland University of  
Applied Sciences  
St.Gallen, Switzerland

Peter Fröhlich  
peter.froehlich@ait.ac.at  
AIT Austrian Institute of Technology  
Center for Technology Experience  
Vienna, Austria

Rainer Endl  
rainer.endl@ost.ch  
Eastern Switzerland University of  
Applied Sciences  
St.Gallen, Switzerland



**Figure 1: Recent examples of AI-driven mhealth apps: (a) *SkinVision* detects malicious skin alterations from smartphone photos, (b) *ResApp* analyzes cough noises captured by the smartphone microphone, (c) *Kardia* detects ECG abnormalities (on data from an additional sensor), (d) *SleepScore* measures breath and body movements during the sleep and provides advice.**

## ABSTRACT

First consumer-facing apps for medical self-diagnosis through Artificial Intelligence have hit the market only recently. These promise to detect malicious skin changes from photos or respiratory diseases from cough noises captured by the smartphone microphone, for example. While there is a large body of research on HCI-related aspects of mobile health applications, knowledge about the user perceptions of such novel AI-driven self-diagnosis apps and factors affecting their acceptance and adoption is scarce. In an online survey, we investigated the participants' overall willingness-to-use (considering four types of captured and processed data) and identified trust factors and desirable features. We found that more than half of the participants would use AI-driven self-diagnosis apps, yet mainly integrated into prevailing general practitioner care. Based

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MUM 2020, November 22–25, 2020, Essen, Germany

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8870-2/20/11...\$15.00

<https://doi.org/10.1145/3428361.3428362>

on the results, we draw conclusions which can guide the design, development, and launch of AI-driven self-diagnosis apps.

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI); Empirical studies in ubiquitous and mobile computing.**

## KEYWORDS

mhealth, health assessment, self-diagnosis, artificial intelligence

### ACM Reference Format:

Matthias Baldauf, Peter Fröhlich, and Rainer Endl. 2020. Trust Me, I'm a Doctor – User Perceptions of AI-Driven Apps for Mobile Health Diagnosis. In *19th International Conference on Mobile and Ubiquitous Multimedia (MUM 2020)*, November 22–25, 2020, Essen, Germany. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3428361.3428362>

## 1 INTRODUCTION

Artificial Intelligence (AI) is expected to have tremendous impact on the healthcare sector [14, 25, 41] - with an estimated market volume of USD 23.85 billion by 2025 [23]. Driven by the steadily increasing amount of healthcare data and advanced methods for

natural language understanding and big data analytics, AI enables various beneficial health-related applications. Examples include the extraction of information from narrative texts such as physical examinations and clinical laboratory reports and the assistance of physicians with disease and treatment suggestions. Particularly this latter field is receiving a lot of attention from both academia and industry with a current focus on diagnostic imaging [14]. A recent systematic review and meta-analysis on the diagnostic accuracy of health-care professionals versus AI algorithms using medical imaging [22] found respective algorithms to have equivalent sensitivity and specificity to health-care professionals.

Only recently, first consumer-facing AI-driven mobile health apps hit the market. In the past few years, the functionality of so-called *mhealth* apps [20, 27] has continuously grown by utilizing data collected from the smartphone camera and microphone as well as from connected wearable devices such as smart watches and chest straps. In contrast to related mobile telemedicine services (cf. [40]) (enabling diagnosis support through videoconferencing with a remote physician, for example), latest *mhealth* apps plainly rely on AI techniques to evaluate submitted sensor data and derive a corresponding health assessment – without any human expert in the loop. Examples include *SkinVision*<sup>1</sup> (Figure 1a) and *Tibot*<sup>2</sup> which both analyze photos of the user’s skin to detect skin diseases. It is noteworthy that such apps do not promise a concrete diagnosis but provide a “health assessment”, for example in the form of a risk calculation. Another example is *ResApp*<sup>3</sup> (Figure 1b). The *ResApp* platform is able to detect respiratory diseases from the sound of the user’s cough or breathing captured through the smartphone microphone.

Due to the novelty of such AI-driven health assessments for consumers, knowledge about the perception of respective apps by users and factors affecting their acceptance and adoption is scarce. While a vast amount of research studies the quality of AI-driven health assessments from a purely technical perspective ([22, 28] e.g.), we deliberately investigate the users’ view on respective apps. To gain insights into users’ willingness to use mobile AI-driven health assessment apps and into their reliance in respective results, we conducted an online survey.

Inspired by prior work in the field of *mhealth* [17, 30], we particularly explored the impact of different data types (such as photo of the skin and cough noises) and compared users’ willingness-to-use to related telemedicine services with a human expert being responsible for the assessment. Furthermore, we studied relevant trust factors, desirable additional features of such apps as well as potential integration practices with medical professionals.

The contribution of our work is twofold: (1) We present the results of our survey which provide novel insights into the user perceptions of AI-driven self-diagnosis apps. (2) Based on these insights, we draw conclusions which can guide the design, development, and launch of AI-driven self-diagnosis apps. For enhanced readability, we refer to this kind of apps as “self-diagnosis apps” hereafter (being aware that formally they provide a health assessment and implying that their results are solely AI-based).

<sup>1</sup><https://www.skinvision.com/>

<sup>2</sup><https://tibot.ai>

<sup>3</sup><https://www.resapphealth.com.au/>

## 2 RELATED WORK

Our research builds on three strands of prior work: 1) general adoption and acceptance of *mhealth* applications, 2) human-centered *mhealth* privacy and safety as well as 3) trust in AI-based applications. In the following, we give an overview of relevant previous research in these three fields.

### 2.1 Adoption and Acceptance of *mHealth*

A rich body of literature studied adoption and acceptance of *mhealth* applications by different population groups in various regions (often developing countries) and tried to derive corresponding impact factors.

For example, Hoque and Sorwar [13] and Quaasar et al. [31] studied elderly users’ intention to adopt and use *mhealth* services. They developed a corresponding model based on the Unified Theory of Acceptance and Use of Technology (UTAUT [39]). The studies determined that performance expectancy, effort expectancy, social influence, and technology anxiety had a significant impact on the users’ behavioral intention to adopt *mHealth* services. Aligned with the UTAUT, Nunes et al. [26] examined the moderating roles of age, gender, and smartphone experience for the acceptance of *mhealth* applications. According to their results, the intention to use such apps is determined by performance expectancy moderated by age and smartphone experience. Cocosila and Archer [6] conducted a one-month experiment where participants were exposed to a health promotion application delivered through their mobiles. They found that the perceived overall risk was a significant deterrent to intention to use the technology while intrinsic motivation had a positive influence. Regarding gender differences, Zhang et al. [42] presented an empirical study with Chinese participants and found that males enjoy a higher level of *mhealth* adoption intention compared with females. An extensive systematic review on *mhealth* adoption (in developing countries) was conducted by Kruse et al. [19]. For an overview and comparison of different models for *mhealth* acceptance, we refer to the work of Sun et al. [37].

The knowledge adoption of *mhealth* apps among general practitioners was studied by Byambasuren et al. [5]. Through a technology survey they found that more than half of the participants recommended apps for patients either daily, weekly, or monthly. Mindfulness and mental health apps were recommended most often, followed by diet and nutrition, and exercise and fitness apps.

Other research approaches addressed UI design improvements to foster adoption of *mhealth* apps. For example, Katz et al. [16] investigated users’ interactions with diabetes self-help apps. They identified two principal areas of failure: excessive cognitive demands on users to extract value as well the need for emotional sensitivity given the affective potential of these interactions. Features and actions of *mhealth* apps which impact their trustworthiness were investigated by van Haasteren et al. [38]. Based on a literature review and focus group sessions they identified desirable features in five major categories: informational content, organizational attributes, societal influence, technology-related features, and user control factors. Identified mistrust features included, among others, incessant tracking and lengthy privacy policies. Related to the focus of our work, Zhao et al. [43] presented initial insights into design issues of AI-driven self-diagnosis app. From interviews they

conclude, that explicitly presenting a percentage risk as well as the rationale behind an assessment increase the users' perceived usefulness of such an app.

## 2.2 Studying mHealth Privacy and Safety

Several prior studies addressed privacy concerns for mhealth applications, studied practices of exchanging health data with various stakeholders, and explored ways to create trustworthy, privacy-protecting mhealth services [4, 15, 17, 18, 24, 30, 34].

In early work, Klasnja et al. [17] studied personal sensing by an mhealth application and corresponding privacy concerns, in particular with regard to different sensor types. The participants of the three-month field study tended to consider GPS data sensitive; continuous audio recordings were rated nearly unanimously negative. Prasad et al. [30] focused on the sharing of personal health information with family, friends, third parties, and the public. In a field study, the participants wore *Fitbit* devices and could change their sharing preferences for the collected data in a custom web portal. The researchers found that people share certain health information less with friends and family than with strangers and that information, people were less willing to share, could be information that is indirectly collected by the mobile devices.

Serrano et al. [34] conducted an extensive survey on patient-clinician communication with mobile devices and investigated the patients' willingness to exchange different types health-related information via mobile devices. They found that participants were less willing to exchange information that may be considered sensitive or complex such as symptoms or digital images/video, in comparison to less critical information such as appointment reminders or general health-related tips. In particular, adults aged 50 or above and participants with a bachelor's degree or higher were found more reluctant to exchange personal medical information through mobile devices.

More recently, researchers have started to understand user attitudes towards sharing personal health data across the health ecosystem. For example, through a questionnaire survey Karampela et al. [15] found that the majority of users are willing to share their health data, most often preferring to share it for scientific research. Age, education level, and occupation of the participants, in addition to the level of digitalization in their country were found to be associated with data sharing attitudes. Akbar et al. [2] presented an extensive review summarizing the kinds of clinical safety concerns with consumer-facing mhealth apps and their consequences. They found that the majority of concerns are related to quality of the content, i.e. incorrect and incomplete information as well as incorrect output of apps that provide calculations and diagnostics.

Complementing aforementioned surveys on user attitudes and practices, researchers have been working on providing principles and guidelines for building trustworthy mhealth infrastructures and applications. Kotz et al. [18] and Avancha et al. [24] presented an extensive conceptual privacy framework for supporting privacy-sensitive mhealth systems. Improving the users' control over sensitive data in mhealth services by providing meaningful information about how their personal data are processed and shared with other parties, was studied by Murmann [24].

## 2.3 Trust in Automation and AI

Investigating user interaction and experience with machines or services with (semi-)automated behavior has a long history (cf. [10]). In particular trust in automation was investigated by Lee and See [21], for example, who considered trust from the organizational, sociological, interpersonal, psychological, and neurological perspectives. Their implications for creating trustable automation contain showing the past performance of the automation as well as the process and algorithms of the automation by revealing intermediate results in a comprehensible way. A corresponding integrated framework for assessing automation adoption, the Automation Acceptance Model (AAM), was introduced by Ghazizadeh et al. [9].

Driven by recent advances in AI, particularly in Machine and Deep Learning, and their steadily increasing application in mass market products and services, we observe academia's recurring interest in exploring automation from a more human-centric than tech-oriented perspective (cf. [7]). However, in contrast to earlier work, today's users of AI-driven automated systems and services are no longer (only) trained professionals (such as operators or pilots), yet increasingly are non-experts (cf. [8]), such as in the fields of autonomous driving and medical self-diagnosis.

To support developers in designing and building understandable and trustworthy AI-driven applications, Amershi et al. [3] presented 18 generally applicable design guidelines for human-AI interaction. This collection contains guidelines such as *Make clear how well the system can do what it can do* and *Make clear why the system did what it did*. Hengstler et al. [11] investigated AI case studies of medical products (e.g., *IBM Watson*) and explored how firms systematically foster trust in applied AI. Among the crucial factors to initiate and foster trust in these products they found operational safety (e.g., through certification), data security, trialability, usability of the product, transparency of the development process, and the gradual introduction of a novel disruptive technology.

While early automation features might have been comprehensible, advanced AI methods are usually applied in a black box manner. Since understanding of an intelligent system's reasoning is crucial for the user's trust in its results, *Explainable Artificial Intelligence* (XAI), i.e. supporting methods for visualizing, explaining, and interpreting AI models, is gaining increasing interest (cf. [1, 33]). XAI approaches for the medical domain have been introduced by Holzinger et al. [12], for example, who emphasize both the complexity and relevance of explainable AI systems, partly due to the European General Data Protection Regulation (GDPR). Similarly, Reimer et al. [32] discussed XAI to build (physicians') trust in medical AI systems and highlighted the importance of official certifications. However, today's XAI approaches are not suitable for non-expert users, since they do still require expert knowledge.

## 3 RESEARCH QUESTIONS

Based on the review and analysis of related work, we identified several research gaps regarding modern self-diagnosis apps at the crossroads of user-centered mhealth and AI research. In this paper, we aim at answering the following four research questions:

**RQ1: Would users be willing to use AI-driven apps to conduct a medical self-diagnosis?**

We are interested in the users' overall willingness to use AI-driven apps for a medical self-diagnosis. To identify the impact of the solely AI-based assessment on the users' perception, we contrast such novel apps to related telemedicine services where the submitted health parameters are analyzed by medical professionals. Complementing the willingness-to-use, we want to explore the central advantages and disadvantages users attribute to self-diagnosis apps.

### RQ2: Do different types of health data have any impact on the users' willingness to use a self-diagnosis app?

Prior research in the field of mhealth investigated the impact of different types of personal or sensed information and sensor types on sharing preferences and privacy concerns [17, 30]. Building on these studies, we are interested in the user perception of entirely AI-driven self-diagnosis apps with regard to different health data types such as a photo of the skin or captured cough noises.

### RQ3: What are crucial factors for users to trust the diagnosis of an AI-driven mhealth app?

As guidance for designers, developers, and publishers of medical self-diagnosis apps, we aim at unveiling the central reasons for users to rely on such "intelligent" apps and their results. We are interested whether the type or location of the app publisher (such as a major tech company, a hospital, or a national company) matter.

### RQ4: How would users combine self-diagnoses through an AI-driven app and visiting a medical professional?

AI-driven mhealth apps for consumers enable anytime and anywhere medical self-diagnoses in high quality and thus may change current healthcare routines. Considering the users' trust in such apps, we are interested in how users would integrate AI-driven self-diagnosis apps into their healthcare routines with medical professionals.

## 4 METHOD

In this section, we outline our study design. We describe the chosen method and the questionnaire in detail and present the participant characteristics.

### 4.1 Overall Setup

The questionnaire was implemented using the popular survey tool *SurveyMonkey*. It consisted of 35 questions and took approximately 15 minutes to complete. Its structure and content are described in detail in the following section.

The study considered four mhealth application categories capturing and analyzing different data types. These categories were selected due to illustrative publicly available apps with partly extensive recent media coverage. We considered the following data types:

- *Photos of the skin*. As mentioned in the introduction, several apps offer an AI-based visual analysis of skin photos to detect skin changes and diseases. Examples include *SkinVision* (Figure 1a) and *Tibot*.
- *Cough noise*. Other apps perform auditory analyses of cough noises to detect diseases such as pneumonia. An example is the aforementioned *ResApp* (Figure 1b).

- *ECG data*. Taking an electrocardiogram (ECG) by an additional external device allows for detecting heart abnormalities by an app. Recent examples include the ECG functionality of *Apple Watch Series 4* or the app *Kardia* (Figure 1c) connecting to a custom ECG accessory.
- *Motion and noise data*. Several apps promise to analyze the user's sleep for reasons of insomnia, etc. by capturing noise data and body movement (through accelerometers of either the smartphone or a connected smartwatch). A recent example applying AI algorithms for assessing the user's sleep is *SleepScore* (Figure 1d).

As we were interested in an overall analysis on future users' acceptance of mobile health diagnosis, we aimed at addressing the unrestricted breadth and diversity of the future adult population of Switzerland. We decided for this national restriction due to several references to available medical services within the survey which made equal prerequisites for all participants necessary to draw sound conclusions. For the dissemination of the survey we broadly distributed the participant invitation both through health-related and non-health-related social media and e-mail lists.

As remuneration, participants could optionally take part in a raffle for three vouchers for an online shop by submitting their e-mail address. Data were collected during three weeks in November 2019. Data collection was in line with the national data protection regulation and were used with the participants' consent. At the beginning of the questionnaire, participants were informed about the target group, furthermore about the anonymity of data capture, the involved data types, as well as the organisation processing the data. They were also made aware of their right to abort the completion of the questionnaire.

### 4.2 Questionnaire

In our questionnaire we collected quantitative as well as qualitative data. After an introduction where we asked for the participants' demographic data, the questions were grouped by topic as follows:

**4.2.1 Definition and Own Experiences.** First, we clearly stated, that this survey does not target typical tracking apps without any diagnosis functionality (e.g., sports and activity trackers, calorie counters, blood pressure checkers) as well as remote diagnosis apps involving human physicians.

We asked whether the participant currently uses a respective diagnosis/self-assessment app (and in case he or she did, which one). Furthermore, we were interested in whether the participant had used such an app at all in the past. If yes, we wanted to know which one and the reason the participant uses this app no longer (bad usability, missing features, incorrect diagnosis, no time, data protection, other).

**4.2.2 Willingness to Use.** After the participants had reported on prior experiences with diagnosis apps, we asked them regarding their overall willingness to use such apps. We introduced four categories of health assessment apps (cf. Section 4.1), differing with regard to the analyzed data: taking a photo of a skin part, coughing into the device microphone, capturing ECG data by a smartwatch, as well as capturing noise and motion data for sleep analysis.

Each type of app was introduced by a short example. Then, the participants were asked whether they are willing to use an app, which provides a solely AI-driven diagnosis on the submitted data (on a five-point Likert scales from 1 – no to 5 – yes) and state the reasons for their assessment. Addressing typical users without any particular AI knowledge, we deliberately did not focus on a specific AI approach (such as machine learning) or elaborate on any AI implementation details, yet emphasized that the results are computed without a human expert in the loop. In addition, we asked whether they would use a related app, which sends the collected data to a hospital for analysis and diagnosis by a human physician. For each app category, we asked participants to assume a comparable performance of the AI algorithm and a human expert (cf. [22, 36]).

**4.2.3 Trust Factors.** In the following section, we focused on potential factors establishing trust in a health assessment app. We asked the participants, to which extent they would trust such an app with regard to the type of publisher (well-known technology concern such as Apple, Google, or Microsoft; a renowned hospital; a start-up). Furthermore, we asked whether they would trust an app if it is certified as a medical product, if it transfers and analyzes data in anonymized form, and if the user knows other apps of the publisher. All factors could be assessed on a five-point Likert scales (from 1 – no to 5 – yes). Furthermore, we asked to optionally reason their assessment in a free text box.

**4.2.4 Interplay with Physician.** Using an AI-driven self-diagnosis app could be combined with a traditional doctor’s visit in several ways. In this section, we let the participants rate the following statements on five-point Likert scales (from 1 – I do not agree at all to 5 – I fully agree). Each decision could be explained by an optional comment in a free text box.

- *For a medical diagnosis, I prefer visiting a doctor. However, for consequent control examinations I prefer a self-diagnosis app.*
- *I prefer an app for a first diagnosis, without involving a doctor. However, for consequent control examinations I prefer visiting a doctor.*
- *I prefer using an app for both diagnosis and consequent controls. However, I would get its results confirmed by a doctor.*
- *I prefer using an app for both diagnosis and consequent controls. In case its results are as reliable as those of a human expert, I would not involve a doctor.*

**4.2.5 Expected Features.** Besides the actual AI-driven health assessment, such apps might provide additional helpful features. We asked the participants about the subjective relevance of the following potential features on five-point Likert scales (from 1 – not important to 5 – important): listing doctors (with relevant information) in the vicinity, recommending a medication after the diagnosis, and creating a treatment plan after the diagnosis. Furthermore, the participants could optionally add further useful app features.

**4.2.6 Subjective Advantages and Disadvantages.** Finally, we asked the participants to state the most relevant advantages and disadvantages of self-diagnosis apps. They could choose from two lists of advantages (location-independent, always available, faster diagnosis, no waiting times, etc.) and disadvantages (no trust in AI-based

diagnosis, privacy concerns, missing personal contact to doctor, etc.), all derived from prior literature on mobile or AI-driven health solutions). Since all options represented actual advantages and disadvantages, participants were asked to select the (up to three) most relevant ones for them (including an “other” option with free text).

### 4.3 Participants

Overall, 120 participants took part in the online survey. As we excluded the data of 14 participants who had not completed the survey, we had a remaining set of 106 participants. This sample comprised 62 (58.5%) male and 44 (41.5%) female participants, with a mean age of 34 years (Median = 35 years), ranging from 19 to 60 years). As intended, participants came from very diverse educational and professional backgrounds. Two participants had experience with AI-driven apps for sleep analysis, one participant stated to use the health assessment app *Ada*.

## 5 RESULTS

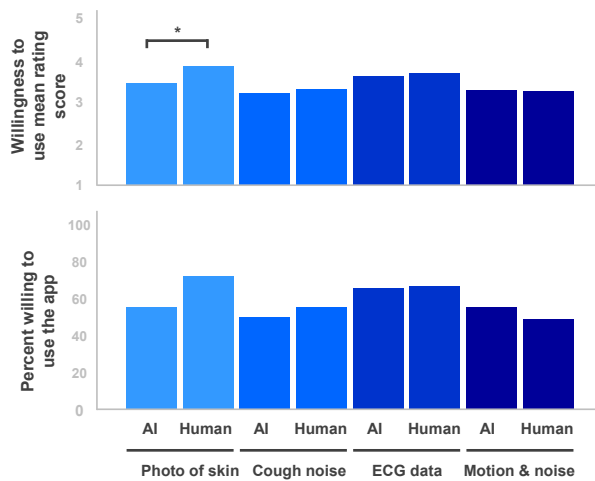
The results presentation is structured according to the thematic questionnaire blocks, namely willingness to use, trust factors, interplay of self-diagnosis app and physician, expected features, and overall advantages and disadvantages. For each of these, statistical analysis of the quantitative results as well as a qualitative analysis of the free text responses are described. As regards the statistical analysis, we conducted non-parametric tests (Wilcoxon for paired samples comparisons, Friedman for comparisons with more than two fields), and Spearman rank correlations (each with two-tailed significance levels). When running multiple comparisons, we applied Bonferroni corrections, and correspondingly we report the resulting corrected error probabilities.

### 5.1 Willingness to Use

Overall we found a remarkable consistency throughout individual subjects’ answers across the different willingness-to-use questions. Correlations between the realization by medical experts and AI algorithms as well as between the four health data types were significant ( $p < .01$ ) and at a level between  $r = 0.45$  and  $r = 0.8$ .

As can be seen in Figure 2, the mean willingness to use results were rather similar, with mean rating scores ranging between 3 and 4 (see upper bar chart within the figure). However, as is illustrated in the lower bar chart of Figure 2, for all variants more than half of the respondents would like to use the respective application, having rated it with either a 4 or a 5. As indicated by the brackets in the upper part of Figure 2, the only significant pairwise difference between realization by medical experts versus AI algorithms was for the photo of the skin, where the human expert variant was rated more positively than the realization by an AI algorithm ( $Z = -3.24$ ,  $p = 0.007$ ). The rating difference between the composite variables of medical experts versus AI algorithms across all health data types was significant ( $Z = 2.71$ ,  $p = 0.007$ ).

An analysis of age differences, by splitting the sample into 2 or 3 equally distributed groups, did not result in significant differences. However, we found that males generally provided higher willingness-to-use rating scores than females. Significant differences were obtained for AI health applications based on skin photos ( $M = 3.73$  for males vs.  $M = 2.95$  for females;  $Z = -2.81$ ;  $p = 0.04$ ) and



**Figure 2: Top: Mean scores for willingness to use, displayed separately for the four different types of health data and for their realization by an AI algorithm (“AI”) versus a medical professional (“Human”). A significant difference between the realization by a medical professional versus an AI algorithm is indicated by a bracket above the bars ( $p < 0.01$ ). Bottom: Percentage of study participants who would use or would surely use the respective application (i.e. who rated with a 4 or a 5 on the willingness-to-use rating scale.**

cough noise ( $M=4.19$  vs.  $M=3.36$ ;  $Z=-2.98$ ,  $p=0.02$ ). No differences between males and females were found for the other two types of health data.

In the following subsections, we summarize the participants’ qualitative responses regarding the four investigated data types.

**5.1.1 Photo of Skin.** Some participants appreciated the more holistic perspective of a human dermatologist compared to the limited focus of an app analyzing images. For example, P24 stated: “A doctor does not only see, but is also able to feel and touch”. P85 emphasized the role of human experience: “Why should I trust the app? I rather confide in a doctor with a lot of experience.”

Several participants positively mentioned the social impact of visiting and talking to a human doctor – which is obviously missing when using a diagnosis app. For example, P34 assumed that “even the social interaction with a human doctor may support the healing”. Similarly, P69 reported that “the human interaction is important for me when visiting a doctor”. Furthermore, the participant mentioned that the personal contact enables to pose follow-up questions.

Several participants mentioned they would first test the performance of such an app themselves. For example, P55 said that he “[...] would compare the diagnosis of the app with the one of a dermatologist. If they match, my trust in the app would be higher.

Participants’ arguments for preferring a respective app over a human dermatologist included the avoidance of waits. For example, P115 reported that “getting an appointment with my dermatologist takes very much time, about 3 months. With such an app I would get

a first diagnosis result promptly and I could, if needed, get a doctor’s appointment with priority.

Only a few participants pointed out the power of a large underlying set of health data. P113 asked: “Why should I trust the experience of only one person if there is a huge database and a smart algorithm available?” P25 reported on bad prior experiences with doctors: “I have lost my trust in doctors and would like to see such apps promoted.”

**5.1.2 Cough Noise.** A lot of participants raised concerns regarding the overall technical feasibility of auditory diagnosis approaches via a smartphone and concluded that they would not be confident in a respective diagnosis.

P28 compared such apps to recent voice assistants: “Siri and others or navigation systems with voice are not sufficiently accurate - even with words. How should a microphone detect different diseases only by coughing with so many different noises?”. A person concerned confirms: “I suffer from an allergic bronchitis and the coughing noise varies very much” (P108). In a similar vein, P54 stated that “a diagnosis only from coughing is not trustworthy. Lung sounds must be considered as well. P110 and P118 have their doubts about the quality of the smartphone microphones and potential background noise. Furthermore, P110 pointed out that many further attributes need to be considered: “What if the user is a heavy smoker?”

Positive remarks mainly addressed the cost and time savings by assessing cough noises by an app.

**5.1.3 ECG Data.** Among the positive statements regarding the usage of a smartwatch-based ECG check, several addressed the convenience of this approach. For example, P46 mentioned: “The smartwatch collects this data anyway, so that’s promising.” Similarly, P62 noted: “I already use a smartwatch, it’s a good thing to record heart frequency.” P8 noted that the smartwatch enables a continuous assessment: “In my opinion, the analyzed data is reliable since it is collected over a longer period.”

Some participants pointed out that potential cardiac diseases are critical and life-threatening. Thus, they would prefer to be treated and informed by a human doctor. P4 commented: “If you fear to have problems with your heart, you need to contact a doctor immediately and get examined professionally.” Similarly, P78 considered “heart diseases [...] dicey... I would visit a doctor if I suspect one.”

Several participants emphasized that they would not like an app to present a message on such a critical assessment. For example, P102 commented: “I don’t think that I suffer from a heart disease - and if I do, I don’t want to know that from an app!” P110 expected “sensitive feedback and explanation of the further progression of the disease.”

Several remarks contained doubts about the feasibility of detecting heart diseases using smartwatch ECG. An exemplary statement from P35: “Certainly not! 12 electrodes need to be directly attached to the body for a measurement. A watch that measures only at the wrist can’t do that.” Similarly, P52 is convinced that “a diagnosis needs more indications.” P71 raised concerns about false positives: “If the application would really work - yes. But if it only had a low success rate I wouldn’t because of the fear-mongering.”

Privacy concerns regarding the ECG data were mentioned by several participants. For example, P121 found these “[...] highly

sensitive data. I have doubts about the privacy.” Similarly, P71 mentioned: “I don’t want these data to fall into the hands of companies such as Apple or Samsung which might sell them”. P120 did not like the continuous monitoring: “I don’t want to have my heart observed all the time.”

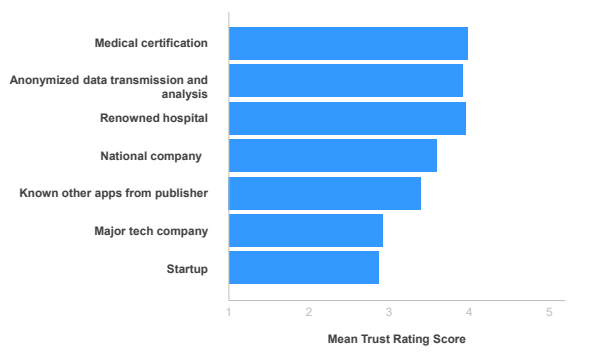
**5.1.4 Motion and Noise Data.** Some participants had gained first experiences with apps for sleep analysis and were not satisfied with the quality of their results. For example, P4 used related apps, yet considered “these apps [...] not mature enough”. Similarly, P33 expressed his experiences: “The app was not precise enough.” Others raised questions regarding the overall feasibility of this app-based assessment approach in case the user does not sleep alone. P55 questioned “How should the app know who the assessment is about?”

The most often mentioned argument for not using an app for sleep assessment was that the participants did not want to sleep close to their smartphone. Exemplary statements include “I don’t like it to have my smartphone next to my bed.” (P52), “I don’t want to spend the entire night next to my smartphone.” (P55), or “I don’t want to have my mobile next to me during the whole night and collecting because of the radiation.” (P64). Others leave their smartphones in another room during the night: “My smartphone is not in my bedroom”, P28 pointed out, for example.

A few participants noticed differences of this continuous measurements to singular assessment through taking of photo of the skin or coughing in the smartphone microphone. P33 asked “How is the practical usage? Do I have to use this every day?”

**5.2 Trust Factors**

Medical certification, anonymized data transmission and analysis received the highest scores related to their contribution to trust in a self-diagnosis app (Figure 3). These factors were deemed significantly more important than the publishing through a major tech company or a startup, or the knowledge about other known apps from the publisher ( $p < 0.5$ ). When comparing male and female responses, we only found a significantly higher rating for national company difference by males ( $M = 3.79$  vs.  $3.32$ ,  $Z = 9.10$ ,  $p = 0.02$ ). No other differences were found with regard to gender and age.



**Figure 3: Mean rating scores on trust factors. The potential trust factors listed on the y-axis were rated by the participant as to whether they would trust a self-diagnosis app in the respective case.**

In addition to the options above, participants provided several further reasons to rely on the results of a self-diagnosis app and potential measures to foster the users’ trust, respectively. From their answers, we derived the following additional factors:

**Data Protection.** Several participants explicitly emphasized the importance of protecting the users’ health data. They mentioned the need for strong encryption and authentication with state-of-the-art technologies. One participant recommended processing the health data and determining the assessment locally on the device in order to avoid the transmission of the data at all.

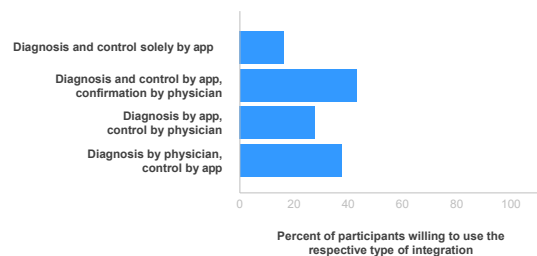
**Recommendation by General Practitioner.** A few participants mentioned to trust an AI-driven self-diagnosis app if it was recommended and explained by their general practitioner and/or their local hospital.

**Recommendations by Renowned Expert.** Similarly to recommendations by a doctor they know, some participants considered the recommendation by a renowned experienced medical professional a reason to trust a respective app.

**User Reviews and Ratings.** Some participants related to evaluation mechanisms of existing software stores such as reviews and ratings: User-provided positive experience reports and good ratings were mentioned as another reason to trust a respective app.

**5.3 Interplay of App and Physician**

Figure 4 shows that none of the proposed forms of interplay between medical experts and a self-diagnosis app would be used by more than 50% of the participants. The applied Friedman test revealed a significant difference between these four options ( $\chi^2 = 26.4$ ;  $p < .001$ ). Pairwise comparisons revealed that assessment and control solely by a self-diagnosis app received a significantly lower score than the assessment and control by a self-diagnosis app with confirmation by a physician ( $Z = 3.31$ ,  $p = 0.006$ ), and it was also significantly lower than the diagnosis by a physician and the subsequent control by a self-diagnosis app ( $Z = 3.02$ ,  $p = 0.015$ ). No further pairwise differences were found.

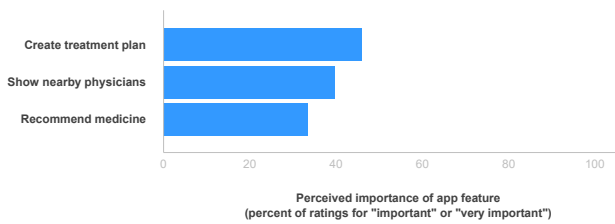


**Figure 4: Percent of participants who would (strongly) agree to the statement that they would use the listed types of interplay between human experts and a self-diagnosis app.**

**5.4 Additional App Features**

As can be seen from Figure 5, none of the proposed app features was perceived important or very important by more than 50% of the respondents. The creation of a treatment plan was rated as more

important than the recommendation of medicine ( $Z=-2.45$ ,  $p=0.014$ ). No other significant difference could be obtained.



**Figure 5: Percent of participants who consider the respective feature important or very important.**

When asked for additional features for self-diagnosis apps, the study participants came up with several ideas. Again, we grouped the participants' answers topic-wise:

**Explanation of the Analysis.** A lot of the participants mentioned the need for an explanation on how the app analyzed the data and derived the current assessment. Participants explained that they would like to comprehend the app's reasoning. They considered this understanding not only crucial for the trust by non-expert users but also for a solid interpretation by a physician (the user might visit in case of a worrying diagnosis). The reliability of the diagnosis result was a frequently mentioned essential information.

**Disease Information.** Many participants wished for further information on the diagnosed disease. This includes its detailed explanation, the typical course of the disease, incubation time, as well as guidance information. Information on recommended medicine was expected to include adverse effects and intolerances.

**Treatment Costs.** Related to the former category of disease information is the demand for information on the treatment costs. Participants expected an estimation of the costs of the suitable treatment, the involved drugs, etc. Some suggested to also indicate the availability of affordable generic drugs.

**Recommending Visiting a Doctor.** Participants demanded clear advice after the AI-based assessment whether visiting a doctor is recommended or required, respectively. Visiting a doctor might not only be recommended when the app supposes a severe disease, yet also when no clear result could be determined. For indeterminate cases, one participant suggested completing the AI-based assessment by information gained from additional relevant questions to the user. Irritating and worrying users through inconclusive assessments must be avoided, as stated by several participants.

**Alternative Therapies.** Several participants emphasized that such apps should provide treatment information beyond drugs of the conventional medicine. This could include alternative therapies, household remedies, and physiotherapeutic exercises.

**Prevention Information.** A few participants suggested to complement treatment information with information on the prevention of the detected disease in order to avoid another outbreak in future.

**Treatment Companion.** In case a health assessment app provides a treatment plan, the app could also accompany the user during this treatment and support tracking the progress. Suggested features include reminders for follow-up examinations by a medical

expert or instructions and hints for supporting the user's recovery in daily life.

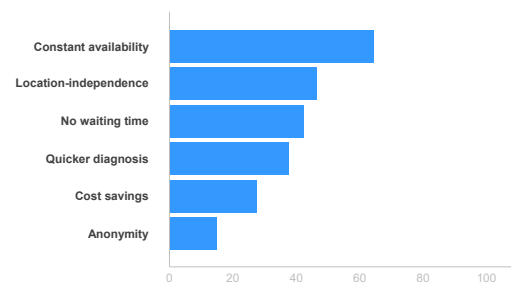
**Mental Support.** Several participants emphasized the emotional impact of a severe health assessment. Since users would probably be alone when they receive the critical message, the participants recommended that such an app must provide mental support. Yet, they did not suggest concrete features how to achieve this mental user support.

**Objectivity and Independence.** No concrete functional feature, yet a request for the overall app was its independence of a medial or pharmaceutical company. A few participants assumed that recommendations for drugs, for example, would not be objective, in case such a self-diagnosis app would be published or sponsored by a respective company.

## 5.5 Overall Advantages and Disadvantages

In the following, we report on the participants' opinions on advantages and disadvantages of self-diagnosis apps.

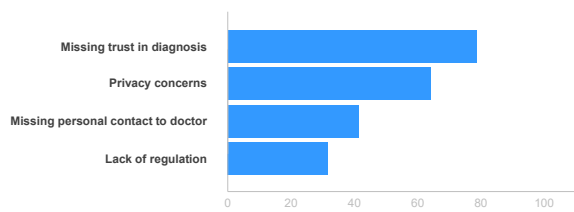
**Subjective Advantages.** Figure 6 depicts the results of subjective advantages. Mentioned by 64% of the participants, the constant availability was the most important advantage of mobile diagnosis apps for the study participants. Second-most mentioned advantage was the location independence (45%). A quicker diagnosis and no waiting times were chosen by 38% and 42%, respectively. Furthermore, cost savings were chosen by 28% of the participants among their top three advantages and anonymity by 15%.



**Figure 6: Percentage of study participants who considered the respective property a significant advantage.**

A few additional advantages and remarks were submitted by participants. For example, P55 stressed the relevance of the anonymity and considered self-diagnosis apps "particularly useful for embarrassing and awkward incidents where no personal contact might be more convenient". P94 emphasized that, in contrast to visiting a doctor for a diagnosis, an app could be "easily used multiple times in short intervals to identify trends or influencing factors". 7% of the participants could not see any advantage of such apps at all.

**Subjective Disadvantages.** Figure 7 shows the participants' ratings of the disadvantages. 78% of the participants considered a lack of trust a major disadvantage of self-diagnosis apps. 65% had concerns about the privacy of their personal health data. Furthermore, the lack of personal face-to-face contact with a human expert was mentioned by 41% of the participants. Finally, 31% of them considered the lack of regulations for such apps a major disadvantage.



**Figure 7: Percentage of study participants who considered the respective property a significant disadvantage.**

In a remark, P110 explained his rating regarding the “missing trust”: In contrast to a medical examination by a human expert, *“the treatment [by an app] is not holistic which may lead to misdiagnosis”*.

Not disadvantages of self-diagnosis apps at all were seen by 2% of the study participants.

## 6 DISCUSSION

In this section, we refer back our original research questions, discuss the results, and derive conclusions and recommendations.

### 6.1 RQ1: Willingness to Use

Overall, the results for the four investigated data categories indicate that more than half of the participants would use respective AI-driven apps. Surprisingly, we did not find a significant difference in the users’ willingness to use an app with AI-driven assessment vs. one with a human diagnosis in three of four categories (cough noise, ECG data, motion and noise data). One reason for the significant difference for the skin analysis could be the severity of a bad diagnosis in combination with doubts about the reliability of the AI-based visual detection. Regarding gender, our results are in line with related research on mobile (non-AI) health apps (cf. [42]): The male participants were generally more willing to use AI-driven self-diagnosis apps than the female ones.

Central arguments in favor (collected through free text fields for giving reasons and the rating of advantages) were the constant availability as well as the location-independence. Furthermore, time savings through eliminated waiting times and quicker diagnoses were relevant advantages for our participants. The anonymity (in terms of avoiding a face-to-face visit to a physician) turned out to be an insignificant advantage of self-diagnosis apps.

Based on recent study results (e.g., [22, 28]) we professed a comparable diagnosis performance of the AI algorithm and a medical professional in the survey. While this could explain the results of the quantitative analysis for three data types (no significant differences with regard to an analysis by a human expert), the qualitative feedback of the participants showed a more diverse picture. Several participants had serious doubts about the technical feasibility of such self-diagnosis apps (see also RQ2 below). Furthermore, some participants referred to the “experience of a human expert” as an advantage over an AI-based solution. This illustrates the lack of knowledge of non-expert users on how today’s machine learning algorithms work, since their outcomes are based on large data sets of historical data (i.e. formalized experience of lots of experts and

confirmed diagnoses). We conclude that self-diagnosis apps must clearly illustrate and communicate their effectiveness and recognition rates for laypersons. While some publishers of existing apps provide scientific papers about the underlying technical realization and the medical validations on their websites (e.g., *ResApp*), improved evidence presentations which are easily understandable and reasonable for non-experts within the apps are recommended.

### 6.2 RQ2: Impact of Data Types

Our second research question was whether different types of health data have any impact on the users’ willingness to use a health assessment app. Our quantitative analysis did not show any significant differences between the data types. Instead, we found a strong consistency throughout individual participants’ answers. This means, participants, in general, were either positive or negative about the self-diagnosis apps overall. Their opinion did not depend on the type of captured and processed data.

While we did not find significant differences, the highest willingness-to-use received the AI-driven assessments of skin photos and ECG data. The lower value for the cough analysis could be ascribed to doubts on technical feasibility mentioned by several participants. The sleep analysis (through motion detection and noise analysis) also raised questions regarding the technical feasibility, yet also was not accepted by many participants since they refused to place their smartphone next to their bed.

Another insight was a certain fear to receive a serious health assessment by a technical gadget, not a physician (*“I don’t want to know that from an app!”*). In such cases, it seems that respective participants appreciate personal contact and feel to be in better hands at the doctor’s office. There, either the physician or staff can recommend measures, give consolation and provide mental support, immediately. Designers of self-diagnosis apps which are able to detect (life-)threatening diseases (in our study this would be malicious skin changes and critical heart abnormalities), need to consider this psychological aspect. Respective notifications should be designed with particular attention and/or alternative responses such as recommending a visit to the doctor be considered.

### 6.3 RQ3: Trust Factors

From the results of our survey three relevant factors for the users’ trust in AI-driven self-diagnosis apps emerged: an official medical certification, the guarantee of anonymized transmission and analysis of the users’ personal health data as well as a trustworthy app publisher in form of a renowned hospital. In many countries, respective certificates for medical products are a legal requirement for health apps providing diagnoses or treatments. In the US, for example, such apps are regulated by the Office for Civil Rights (OCR) of the Department of Health and Human Services, the Food and Drug Administration (FDA), and the Federal Trade Commission (FTC) [35]. Such regulations seem to build trust, however, are currently only national initiatives.

Protecting the users’ sensitive health data and corresponding assessments was not only rated highly in our survey but also explicitly emphasized by several participants. Therefore, respective information on how the self-diagnosis app anonymizes, encrypts,

and protects critical data should not only be hidden in comprehensive terms of use. Designers of AI-driven mhealth apps should clearly communicate respective protection and security features in an understandable and appealing way. For example, web browsers visualize a secure HTTPS connection by a lock icon. Similar approaches to indicate anonymization, encryption, etc. for mhealth applications could be designed and agreed upon by developers.

A renowned hospital as app publisher seems to well-represent the professional and scientific background of the product. Startups as publishers of a respective app received a negative trust score on average. We ascribe this to doubts regarding the overall diagnosis quality, the business model, and long-term availability of the product. For AI-medtech startups to proof evidence of their app's performance, it seems beneficial to seek for and promote collaborations with hospitals, if not even jointly publish the app. It is noteworthy, that these results do not permit any conclusions about the interplay of the factors. Several of the trust factors could be combined (e.g., a startup could acquire a medical certification).

#### 6.4 RQ4: Interplay of AI App and Physician

The fourth research question addressed potential combinations of AI-driven health assessments through an app with traditional healthcare practices involving a physician. Only a minority of our participants stated that they would rely solely on an AI-driven app for assessing their health. This is in line with prior expectations that medical AI systems will not replace human experts, but will act as a complement [29, 32, 36]. As reasons, several participants mentioned prior bad experiences with a physician and mistrust in one single professional's opinion.

Significantly more positive ratings received two variants combining self-diagnosing through an app and visiting a physician: 80% (strongly) agreed to prefer either receiving the diagnosis by a physician and conducting follow-up control examinations with an AI app or diagnosing and controlling their health by an AI app, yet get results confirmed by a physician.

We conclude that consumer-facing AI-driven health apps should be designed with this interplay with medical professionals in mind. *Kardia* (Figure 1c), for example, offers first related features: it tracks data over time and enables sharing of medical-grade recordings by e-mail. To avoid sending privacy-sensitive data over insecure channels, self-diagnosis apps could provide dedicated views for experts in addition to their primary screens for medical laypersons. For example, these could show raw sensor data, relevant key figures, their development over time in a much more fine-grained and detailed view than for the non-expert user. The expert might check these screens during the user's visit with his or her permission. Another example are personalized features: certain functions of a self-diagnosis app (such as a threshold for a parameter) might be customized by the physician for his or her patient.

#### 6.5 Limitations

We focused on four health data types motivated by four recent mhealth app categories increasingly applying AI-driven analysis. Obviously, this study is non-exhaustive and follow-up investigations should explore upcoming application types capturing and processing additional health data. Furthermore, the perception of

conversational mhealth apps (such as *Ada*<sup>4</sup>) which capture a user's medical history (by asking questions and collecting the user's answers) and provide a symptom assessment could be considered.

Our participants covered a broad age range from 19 to 60 years. Still, with a mean age of 34 years, the age distribution is slightly skewed towards younger persons. In contrast to previous work on mhealth acceptance and adoption (cf. [15, 34]), our results did not show significant age-related differences regarding the willingness to use AI-driven apps. Related follow-up studies should focus on an older participant group to provide additional findings regarding the impact of the user's age on self-diagnosis apps.

Due to the recent emergence of consumer-facing self-diagnosis apps and their currently low penetration, we decided for an online survey to gain first insights into consumers' perceptions and guide the further design and development of such apps. As soon as feedback from long-term users is available, the results of this survey should be validated and extended.

## 7 CONCLUSION AND OUTLOOK

The application of Artificial Intelligence in medicine is growing rapidly, currently with a focus on decision-support systems for physicians. Only recently, related features found their way into consumer-facing apps for self-diagnosis. While, in general, there is a large body of research on HCI-related aspects of mhealth and telemedicine applications, a deeper understanding of the user perception of such novel solely AI-driven self-diagnosis apps is required to guide their design and development.

In this paper, we introduced a survey exploring user perceptions and attitudes towards such self-diagnosis apps. We investigated the participants' overall willingness-to-use (considering different types of captured and processed sensor data) and identified relevant trust factors such as an official medical certification, the guarantee of anonymized transmission and analysis of the users' personal health data as well as a trustworthy app publisher. Desirable features beyond the diagnosis function turned out to be an explanation of the analysis (comprehensible for non-experts), information about the detected disease as well as a treatment plan (considering alternative therapies). Furthermore, we studied the participants' preferred ways to integrate these apps into prevailing general practitioner care.

This research represents first steps towards understanding users' reliance, expectations, concerns as well as potential usage behavior of self-diagnosis apps. Based on our results, we identified several promising directions for future research. Corresponding questions address appropriate application behavior for sensitively informing about serious health assessments and measures during self-diagnosis as well as UX strategies to communicate data security and technical effectiveness of self-diagnosis apps. Furthermore, we argue for investigating physician-oriented features in consumer-facing self-diagnosis apps to be utilized for medical insights at the doctor's office or for personalized and continuous monitoring between visits. Participatory design approaches involving non-expert users and physicians seem promising to identify and consider requirements of both user groups and derive generalizable guidelines supporting the interplay of consumer-facing self-diagnosis apps and medical professionals.

<sup>4</sup><https://ada.com/>

## REFERENCES

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. <https://doi.org/10.1109/access.2018.2870052>
- [2] Saba Akbar, Enrico Coiera, and Farah Magrabi. 2019. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. *Journal of the American Medical Informatics Association* (Oct. 2019). <https://doi.org/10.1093/jamia/ocz175>
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, and et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI'19). Association for Computing Machinery, New York, NY, USA, Article 3, 13 pages. <https://doi.org/10.1145/3290605.3300233>
- [4] Sasikanth Avancha, Amit Baxi, and David Kotz. 2012. Privacy in Mobile Technology for Personal Healthcare. *ACM Comput. Surv.* 45, 1, Article 3 (Dec. 2012), 54 pages. <https://doi.org/10.1145/2379776.2379779>
- [5] Oyungerel Byambasuren, Elaine Beller, and Paul Glasziou. 2019. Current Knowledge and Adoption of Mobile Health Apps Among Australian General Practitioners: Survey Study. *JMIR mHealth and uHealth* 7, 6 (June 2019), e13199. <https://doi.org/10.2196/13199>
- [6] Mihail Cocosila and Norm Archer. 2010. Adoption of mobile ICT for health promotion: an empirical investigation. *Electronic Markets* 20, 3–4 (Nov. 2010), 241–250. <https://doi.org/10.1007/s12525-010-0042-y>
- [7] Peter Fröhlich, Matthias Baldauf, Thomas Meneweger, Ingrid Erickson, Manfred Tscheligi, Thomas Gable, Boris de Ruyter, and Fabio Paternò. 2019. Everyday Automation Experience: Non-Expert Users Encountering Ubiquitous Automated Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI EA '19). Association for Computing Machinery, New York, NY, USA, Article W25, 8 pages. <https://doi.org/10.1145/3290607.3299013>
- [8] Peter Fröhlich, Matthias Baldauf, Thomas Meneweger, Manfred Tscheligi, Boris de Ruyter, and Fabio Paternò. 2020. Everyday automation experience: a research agenda. *Personal and Ubiquitous Computing* (Oct. 2020). <https://doi.org/10.1007/s00779-020-01450-y>
- [9] Mahtab Ghazizadeh, John D. Lee, and Linda Ng Boyle. 2011. Extending the Technology Acceptance Model to assess automation. *Cognition, Technology & Work* 14, 1 (Oct. 2011), 39–49. <https://doi.org/10.1007/s10111-011-0194-3>
- [10] Peter A. Hancock, Richard J. Jagacinski, Raja Parasuraman, Christopher D. Wickens, Glenn F. Wilson, and David B. Kaber. 2013. Human-Automation Interaction Research. *Ergonomics in Design: The Quarterly of Human Factors Applications* 21, 2 (April 2013), 9–14. <https://doi.org/10.1177/1064804613477099>
- [11] Monika Hengstler, Ellen Enkel, and Selina Duelli. 2016. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change* 105 (April 2016), 105–120. <https://doi.org/10.1016/j.techfore.2015.12.014>
- [12] 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? *CoRR abs/1712.09923* (2017). <http://arxiv.org/abs/1712.09923>
- [13] Rakibul Hoque and Golam Sorwar. 2017. Understanding factors influencing the adoption of mHealth by the elderly: An extension of the UTAUT model. *International Journal of Medical Informatics* 101 (May 2017), 75–84. <https://doi.org/10.1016/j.ijmedinf.2017.02.002>
- [14] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. Artificial intelligence in health-care: past, present and future. *Stroke and Vascular Neurology* 2, 4 (June 2017), 230–243. <https://doi.org/10.1136/svn-2017-000101>
- [15] Maria Karampela, Sofia Ouhbi, and Minna Isomursu. 2019. Connected Health User Willingness to Share Personal Health Data: Questionnaire Study. *Journal of Medical Internet Research* 21, 11 (Nov. 2019), e14537. <https://doi.org/10.2196/14537>
- [16] Dmitri S. Katz, Blaine A. Price, Simon Holland, and Nicholas Sheep Dalton. 2018. Data, Data Everywhere, and Still Too Hard to Link: Insights from User Interactions with Diabetes Apps. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, Article 503, 12 pages. <https://doi.org/10.1145/3173574.3174077>
- [17] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring Privacy Concerns about Personal Sensing. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 176–183. [https://doi.org/10.1007/978-3-642-01516-8\\_13](https://doi.org/10.1007/978-3-642-01516-8_13)
- [18] David Kotz, Sasikanth Avancha, and Amit Baxi. 2009. A Privacy Framework for Mobile Health and Home-Care Systems. In *Proceedings of the First ACM Workshop on Security and Privacy in Medical and Home-Care Systems* (Chicago, Illinois, USA) (SPIMACS '09). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/1655084.1655086>
- [19] Clemens Kruse, Jose Betancourt, Stephanie Ortiz, Susana Melissa Valdes Luna, Inderdeep Kaur Bamrah, and Narce Segovia. 2019. Barriers to the Use of Mobile Health in Improving Health Outcomes in Developing Countries: Systematic Review. *Journal of Medical Internet Research* 21, 10 (Oct. 2019), e13263. <https://doi.org/10.2196/13263>
- [20] Santosh Kumar, Wendy Nilsen, Misha Pavel, and Mani Srivastava. 2013. Mobile Health: Revolutionizing Healthcare Through Transdisciplinary Research. *Computer* 46, 1 (Jan. 2013), 28–35. <https://doi.org/10.1109/mc.2012.392>
- [21] John D. Lee and Katrina A. Sec. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan. 2004), 50–80. [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392)
- [22] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shandas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 1, 6 (2019), e271 – e297. [https://doi.org/10.1016/S2589-7500\(19\)30123-2](https://doi.org/10.1016/S2589-7500(19)30123-2)
- [23] Mordor Intelligence. 2020. Artificial Intelligence in Medicine Market - Growth, Trends, and Forecast (2020 - 2025). <https://www.researchandmarkets.com/r/e4hy1e>. Accessed: 2020-06-25.
- [24] Patrick Mummert. 2018. Usable Transparency for Enhancing Privacy in Mobile Health Apps. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) (MobileHCI'18). Association for Computing Machinery, New York, NY, USA, 440–442. <https://doi.org/10.1145/3236112.3236184>
- [25] Nariman Noorbakhsh-Sabet, Ramin Zand, Yanfei Zhang, and Vida Abedi. 2019. Artificial Intelligence Transforms the Future of Health Care. *The American Journal of Medicine* 132, 7 (July 2019), 795–801. <https://doi.org/10.1016/j.amjmed.2019.01.017>
- [26] Andreia Nunes, Teresa Limpo, and São Luís Castro. 2019. Acceptance of Mobile Health Applications: Examining Key Determinants and Moderators. *Frontiers in Psychology* 10 (Dec. 2019). <https://doi.org/10.3389/fpsyg.2019.02791>
- [27] Phillip Olla and Caley Shimskey. 2015. mHealth taxonomy: a literature survey of mobile health applications. *Health and Technology* 4, 4 (Jan. 2015), 299–308. <https://doi.org/10.1007/s12553-014-0093-8>
- [28] Paul Porter, Udantha Abeyratne, Vinayak Swarnkar, Jamie Tan, Ti wan Ng, Joanna M. Brisbane, Deirdre Speldewinde, Jennifer Choveaux, Roneel Sharan, Keegan Kosasih, and Phillip Della. 2019. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respiratory Research* 20, 1 (June 2019). <https://doi.org/10.1186/s12931-019-1046-6>
- [29] 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 (Oct. 2019), e16222. <https://doi.org/10.2196/16222>
- [30] Aarathi Prasad, Jacob Sorber, Timothy Stablein, Denise Anthony, and David Kotz. 2012. Understanding Sharing Preferences and Behavior for MHealth Devices. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society* (Raleigh, North Carolina, USA) (WPES '12). Association for Computing Machinery, New York, NY, USA, 117–128. <https://doi.org/10.1145/2381966.2381983>
- [31] G.M. Azmal Ali Qaosar, Md. Rakibul Hoque, and Yukun Bao. 2018. Investigating Factors Affecting Elderly's Intention to Use m-Health Services: An Empirical Study. *Telemedicine and e-Health* 24, 4 (April 2018), 309–314. <https://doi.org/10.1089/tmj.2017.0111>
- [32] Ulrich Reimer, Edith Maier, and Beat Tödtli. 2020. Going Beyond Explainability in Medical AI Systems. In *Proceedings of the Workshop "Modelle in der KI" at Modellierung 2020*.
- [33] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *CoRR abs/1708.08296* (2017). [arXiv:1708.08296](http://arxiv.org/abs/1708.08296)
- [34] Katrina J. Serrano, Mandi Yu, William T. Riley, Vaishali Patel, Penelope Hughes, Kathryn Marchesini, and Audie A. Atienza. 2016. Willingness to Exchange Health Information via Mobile Devices: Findings From a Population-Based Survey. *The Annals of Family Medicine* 14, 1 (Jan. 2016), 34–40. <https://doi.org/10.1370/afm.1888>
- [35] Karandeep Singh, Kaitlin Drouin, Lisa P. Newmark, JaeHo Lee, Arild Faxvaag, Ronen Rozenblum, Erika A. Pabo, Adam Landman, Elissa Klinger, and David W. Bates. 2016. Many Mobile Health Apps Target High-Need, High-Cost Populations, But Gaps Remain. *Health Affairs* 35, 12 (Dec. 2016), 2310–2318. <https://doi.org/10.1377/hlthaff.2016.0578>
- [36] Kayt Sukel. 2017. With a little help from AI friends. *New Scientist* 235, 3134 (July 2017), 36–39. [https://doi.org/10.1016/s0262-4079\(17\)31376-3](https://doi.org/10.1016/s0262-4079(17)31376-3)
- [37] Yongqiang Sun, Nan Wang, Xitong Guo, and Zeyu Peng. 2013. Understanding the acceptance of mobile health services: A comparison and integration of alternative models. *J. Electr. Commerce Res.* 14, 2 (2013), 183–200.
- [38] Afua van Haasteren, Felix Gille, Marta Fadda, and Effy Vayena. 2019. Development of the mHealth App Trustworthiness checklist. *DIGITAL HEALTH* 5 (Jan. 2019), 205520761988646. <https://doi.org/10.1177/2055207619886463>

- [39] Venkatesh, Morris, Davis, and Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425. <https://doi.org/10.2307/30036540>
- [40] Ronald S. Weinstein, Ana Maria Lopez, Bellal A. Joseph, Kristine A. Erps, Michael Holcomb, Gail P. Barker, and Elizabeth A. Krupinski. 2014. Telemedicine, Telehealth, and Mobile Health Applications That Work: Opportunities and Barriers. *The American Journal of Medicine* 127, 3 (March 2014), 183–187. <https://doi.org/10.1016/j.amjmed.2013.09.032>
- [41] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2, 10 (Oct. 2018), 719–731. <https://doi.org/10.1038/s41551-018-0305-z>
- [42] Xiaofei Zhang, Xitong Guo, Kee hung Lai, Feng Guo, and Chenlei Li. 2014. Understanding Gender Differences in m-Health Adoption: A Modified Theory of Reasoned Action Model. *Telemedicine and e-Health* 20, 1 (Jan. 2014), 39–46. <https://doi.org/10.1089/tmj.2013.0092>
- [43] Jason Chen Zhao, Ngai-Man Cheung, Ricardo Sosa, and Dawn Chin-Ing Koh. 2015. Design Self-Diagnosis Applications for Non-Patients. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) (*CHI EA '15*). Association for Computing Machinery, New York, NY, USA, 1433–1438. <https://doi.org/10.1145/2702613.2732865>