Human Computer Interaction Thesis (2021)

Development and Evaluation of a Diagnosis and Triage Healthcare Chatbot

A.J.B. Kockx (6967310)

June 25, 2021

Utrecht University

Abstract

This thesis consists of three parts: a systematic review of evaluations of non-mental healthcare chatbots, the development of a triage and diagnosis chatbot, and the evaluation of the chatbot. The systematic review provides an overview of chatbots in various healthcare contexts and highlights commonly used outcome variables used to evaluate the chatbots. The chatbot was built by leveraging Azure Health Bot service and related Azure services. The evaluation was carried out by 44 participants, whose demographics were skewed towards lower age and higher education. The chatbot was evaluated with a questionnaire, consisting of questions related to outcome variables commonly used to evaluate chatbots, namely user experience, chatbot usage, health behavior, user characteristics, and system quality. The chatbot performed well on user experience, but improvements related to personality and error handling are necessary. The health behavior of the participants after using the chatbot was found to be inconclusive. The analysis of meta-data of the chatbot usage showed that the conversation was relatively quick, but with a high drop-off rate.

Acknowledgements

Throughout the writing of this thesis, I have received a great deal of support and assistance.

I would first like to thank my first supervisor, prof. dr. ir. J.F.M. Masthoff who took the time every single Thursday morning to help me write this thesis. Her experience in the whole field of academia was invaluable, and she pushed me to make this thesis a bigger success. Even when there was not that much to discuss, we could fill up the time by just having a chat, which was more than welcome considering this thesis was written entirely while working from home in the midst of the pandemic.

I want to thank my second supervisor Dr. A.A. Akdag, who was willing to jump in and replace my initial second supervisor. We connected in a later phase of the thesis, but your expertise and enthusiasm were very valuable to me.

I had the privilege to work with my colleagues from the Customer Success Unit (CSU) of Microsoft Netherlands, which is a team of intelligent, ambitious, and kind people with a passion for technology who do not hesitate to apply their expertise to help customers achieve more with their businesses. I would like to particularly single out my manager Dennis Mulder and mentor Susanne Morsy. They helped me find my way and gave me the freedom to explore all possibilities within this company.

In addition, I would like to thank my parents and brother for never hesitating to help me, give me advice, and dragging me away from writing my thesis when it was time to relax. You are always there for me. Finally, I could not have completed this thesis without the support of my friends, who would happily listen to me going on and on about chatbots, but most importantly, allow for memorable and happy distractions outside of working on my research.

By reading this thesis, I hope you will feel as excited and inspired about the use of chatbots in healthcare as I have become over the last months. I cannot wait to see what the future holds for this technology!

Enjoy!

Anne Kockx

Contents

List of Figures

List of Tables

Introduction and Background

Introduction 1

This chapter starts with an introduction and background information on chatbots in healthcare. Next, the research questions and sub-questions of this thesis are explained. The thesis is written as part of an internship at Microsoft.

Chatbots in Healthcare

Interest in the usage of chatbots has increased due to recent developments in **Artificial Intelligence** (AI) [1]. As with other AI techniques, chatbots allow for automation of repetitive tasks and gaining insight into complex data. Chatbots can imitate human conversation by leveraging an AI sub-field known as natural language processing.

Definition 1.0.1 *A conversational interface or **chatbot** is a computer program that communicates smartly with a user on a text or spoken ground. [2]*

[2]: Dahiya (2017), 'A Tool of Conversation: Chatbot'

Because mobile phones have become commonplace, SMS text messaging-based applications are used everywhere, which has now become a natural way of interacting with a device or other person [3]. **ELIZA**, presented by Weizenbaum [4] in 1966, was the first chatbot that enabled a natural language conversation between user and computer. Since ELIZA, the distribution of highly interactive chatbots has been greatly facilitated in various environments, such as e-learning [5] and customer support [6].

Literature distinguishes three types of chatbots: rule-based, AI-based and hybrid chatbots. **Rule-based chatbots** abide by a rule-based orientation, where the chatbot's response is processed through a prepared set of rules [7]. Rule-based chatbots often offer a set of answer options to the user. This is different from **AI-based chatbots** where the conversation between chatbot and user is based on textual or spoken input [8]. **Hybrid chatbots** are a combination of the two where there are rule-based tasks, but they also allow for free-text input.

Chatbots allow for two **mediums**: voice-based or text-based. Although voice-based chatbots are implemented in some healthcare contexts, a systematic review by Bérubé et al. [9] showed that research on voice-based chatbots is still in its infancy. Therefore, this thesis focused specifically on Text-Based Healthcare Chatbots (TBHC).

Definition 1.0.2 ***Text-Based Healthcare Chatbots (TBHC)** are chatbots that support healthcare delivery in an automated manner with simple text-based messages and, in some cases, media objects such as videos and podcasts [10] .*

[10]: Hauser-Ulrich et al. (2020), 'A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): Pilot randomized controlled trial'

Aside from more established chatbot contexts such as customer support, chatbots are also slowly adopted in **healthcare** and have promising potential. By using these technologies, patients are supported by enabling self-management and providing advice [11]. As with many electronic health (eHealth) interventions, chatbots offer scalability and 24-hour availability to provide unmet health needs. In addition, Lucas et al. [12] showed that people might feel more comfortable disclosing personal information to a chatbot than to a person because chatbots do not think on their own or form judgements.

The benefits of developing and implementing chatbots in healthcare are evident, and they may be ready to replace components of healthcare as we know it. However, one important factor to consider in applying chatbots in healthcare is the **perceptions of physicians**. In a study by Palanica et al. [13] the opinions of 100 practicing physicians on chatbot technology in healthcare were investigated. They found that physicians believed in the benefits associated with chatbots, however, this perception differs per implementation. In addition, the physicians stated that chatbots have a beneficial role to play in healthcare to support, motivate and coach patients, and support organizational processes. In essence, chatbots could function as non-medical caregivers. However, there are two major perceived disadvantages of chatbots: firstly, a chatbot cannot comprehend a user's emotional state, and secondly, chatbots lack expert medical knowledge and intelligence. Although, even with these disadvantages, based on this paper, physicians are open to accepting chatbots in some healthcare roles.

[13]: Palanica et al. (2019), 'Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey'

Because chatbots are proven to be successful in supporting, motivating and coaching patients, most chatbots have been implemented in **mental healthcare**. There already are chatbots for different mental disorders and purposes as assessed in multiple literature reviews [14–18]. Mental healthcare chatbots have also been developed for specific mental illnesses such as depression [19]. Also, the perceptions and opinions of mental health patients on chatbots have been reviewed [20]. According to these reviews, the research field of healthcare chatbots for mental illnesses is already quite mature and ahead of their application in other medical fields.

This means that there are research opportunities for other health domains where fewer reviews have been conducted. The available literature reviews often focus on technical aspects of healthcare chatbots [21–23] and not on the impact or usability of the implementation. After performing a scoping review, Car et al. [24] identified a need for **robust evaluations of diverse healthcare chatbots**. Therefore, this thesis proposes research to fill that gap with a systematic review on healthcare chatbots and the implementation and evaluation of a diagnosis and triage healthcare chatbot.

Diagnosis and Triage Chatbots

The number of people searching for health information online increases dramatically [25]. This lead to diagnosis and triage chatbots being developed at increasing rates. Diagnosis and triage chatbots are often referred to as chatbot-based symptom checker (CSC) apps in literature.

Definition 1.0.3 *A **chatbot-based symptom checker** (CSC) assesses medical symptoms based on the users' input in the chat and provides the patient with the most likely diagnosis [26] .*

[26]: Montenegro et al. (2019), 'Survey of conversational agents in health'

CSCs can increase healthcare quality by, for example, minimizing over and under-triage[1] and preventing unnecessary visits to the doctor. Even though there are CSC apps that have been downloaded more than one million times from the Google Play app [27], these applications have received little attention in the research domain. You and Gui [27] conducted interviews and performed an analysis on the features and user reviews of eleven CSC apps. Using these methods, they found **deficiencies** in five areas, namely:

1: Under-triage occurs when the symptoms or injuries from a patient are underestimated, and the patient needed more resources. Over-triage is the opposite, where the symptoms are overestimated and too many resources were used to treat the patient.

1. Insufficient consideration of health history;
2. Strict input requirements;
3. Vague probing questions;
4. Incomplete set of health conditions;
5. Lack of functions for follow-up treatments.

Given these problems, it is clear that there are still opportunities to improve the quality of diagnosis and triage chatbots.

The benefits of chatbots in healthcare are evident, such as the ability to provide instant access to information and **increased efficiency** [28]. The efficiency of healthcare is an important topic, as pressure on healthcare systems has been increasing due to aging populations, rising healthcare costs, and recently the COVID-19 pandemic [29]. Therefore, the rise in the development of these types of healthcare chatbots during the COVID pandemic did not come unexpected [30–33]. One thing these chatbots have in common is the lack of evaluation due to the necessary speed of the implementations. However, an evaluation is crucial to determine the impact and usability of chatbot technologies in healthcare before they can be implemented in even more healthcare facilities.

Azure Health Bot

For the development of the chatbot in this thesis, the Azure Health Bot service [34] was used. As the name suggests, this is a Software as a Service (SaaS) solution and part of **Microsoft Azure** [35]. Microsoft Azure is a collection of Microsoft's cloud computing services to help people and businesses build, deploy and manage applications through a global network of data centers. Microsoft Azure offers eight service categories: compute services, cloud storage, networking, application hosting, artificial intelligence, Internet of Things (IoT), integration, and security. Azure Health Bot is part of the AI category, together with the other bot services.

Figure 1.1: Azure Health Bot logo

Azure bot services [36], launched in 2017, provide developers with a method to accelerate bot creation with an integrated environment and templates for common scenarios [37]. In addition, as these services are part of the Azure landscape, Microsoft also includes analytic tools for obtaining metrics such as the number of active users and user retention. Whereas the bot services are used in a range of sectors, Microsoft introduced Azure Health Bot in February 2019 [38] which is designed specifically for the development of chatbots applied in a healthcare context. In addition to the general functionalities of the bot services, Azure Health Bot is extended with templates for healthcare use cases, easy access to medical databases, and compliance with a range of security certifications [39]. This allows for quick and easy implementation of a healthcare chatbot, and this has been done increasingly often during the COVID-19 pandemic.

Infermedica [34] is one of the included knowledge bases that can be used for Triage and Symptom checking. Infermedica provides an advanced triage engine that is based on a clinically evaluated probabilistic model. It contains simplistic decision tree models based on the mapping of thousands of symptoms and conditions to accurately determine the likelihood of a condition given a set of symptoms. By using feedback loops, the triage engine is continually improving based on real clinical outcomes. When Infermedica is applied in a Health Bot implementation, the user is asked a range of questions based on their symptoms and, ultimately, it determines the most likely conditions and severity of these conditions by providing a triage level [34]. The triage levels advise the user to help them take the next step, such as requesting emergency care or to apply self-care only. The combined features of Azure Health Bot and the Infermedica knowledge base have a huge opportunity to provide triage and diagnosis processes efficiently, for example, in primary care [40].

Research Questions

Concluding, this thesis presents research in the field of healthcare chatbots. The thesis is made up out of three parts. The first part comprises a **systematic review** on non-mental healthcare chatbots. A systematic review in this area has not yet been conducted, as most review focus on either technical aspects of the chatbots (e.g. frameworks used) instead of focusing on a user's evaluation of the chatbot. Therefore, the review focused on evaluations of the chatbot to analyze the effect and usability of the chatbots. The focus on non-mental healthcare chatbots was chosen explicitly because multiple (systematic) reviews have already been carried out in the mental health care domain. The systematic review covered the first research question of this thesis (**RQ1**). Three sub-questions covered the first research question. First, the purpose of **RQ1a** was to determine in which healthcare contexts the chatbots were implemented. Then, the purpose of **RQ1b** was to determine which outcome variables were commonly used to evaluate the healthcare chatbots. Lastly, the purpose of **RQ1c** was to determine common drawbacks and benefits of healthcare chatbots.

RQ1: Systematic Review

What is the current state of the art of chatbots in non-mental health-care?

- RQ1a: In what healthcare contexts are chatbots most often implemented?
- RQ1b: What outcome variables are often used to evaluate health-care chatbots?
- RQ1c: What are the drawbacks and benefits of these chatbots?

CSCs are already popular in the general population but have received little attention in the research domain. The analysis by You and Gui [27] already showed areas for improvements to highlight the benefits even more. The second part of this thesis, therefore, comprises the **development of a triage and diagnosis chatbot** (**RQ2**). Three sub-tasks supported this aim. Firstly, **RQ2a** focused on what functionalities need to be implemented in the chatbot. This comprised both the needs of the users and the deficiencies found by You and Gui. Another important factor of a chatbot's implementation was how the chatbot is hosted, which was addressed by **RQ2b**.

RQ2: Chatbot Development

How can a chatbot be developed to perform diagnosis and triage?

- RQ2a: What functionalities need to be implemented for a chatbot to perform diagnosis and triage?
- RQ2b: How can the chatbot be hosted to be easily accessible for the users?

Car et al. [24] presented the need to perform robust evaluations of diverse healthcare chatbots. Therefore, the third and last part of this thesis comprises the **evaluation of the chatbot** (**RQ3**). The subtasks of this part were based on the commonly used outcome variables from the systematic review. Therefore, the chatbot was evaluated on chatbot usage (**RQ3a**), user experience (**RQ3b**), and health behavior (**RQ3c**).

RQ3: Chatbot Evaluation

How do potential end-users evaluate a diagnosis and triage chatbot on commonly used outcome variables?

- RQ3a: What are the quantitative characteristics of Chatbot Usage?
- RQ3b: How do participants evaluate the User Experience with the chatbot?
- RQ3c: To what extent does the chatbot influence the Health Behavior of the participant?

Methodology

2 Methodology

In this chapter, the methods used in this thesis are explained. It is divided into three parts: first, the methods for the systematic review of evaluated non-mental healthcare chatbots are covered, followed by the methods for developing a diagnosis and triage healthcare chatbot, and lastly, the methods used to evaluate the chatbot.

2.1 Systematic Review

The first part of this thesis was a systematic review with the primary goal to **provide an overview of evaluated non-mental healthcare chatbots** (RQ1). The sub-goals of the review were to identify in which healthcare contexts (RQ1a) the chatbots were most often implemented, which outcome variables were commonly used to evaluate the chatbots (RQ1b), and the drawbacks and benefits of healthcare chatbots (RQ1c). The review focused on non-mental healthcare chatbots because a pilot literature search showed a range of published systematic reviews on mental healthcare or mental interventions. Thus, it focused on other healthcare contexts, such as chronic care, healthy lifestyle, and oncology. The review followed the **PRISMA protocol** [41] and thus contained an identification, screening, eligibility, and included phase (Figure 2.1).

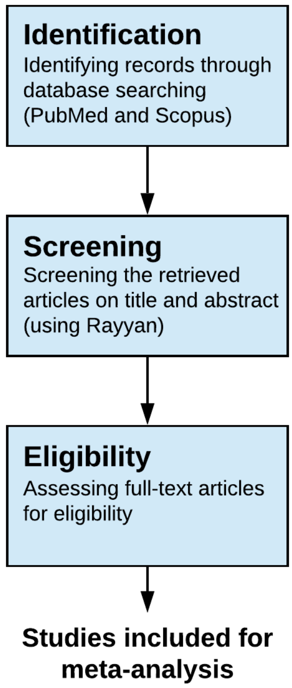


Figure 2.1: Steps in systematic review protocol

Identification

In the identification phase, papers were obtained using **Scopus** and **PubMed** databases in December 2020. The searches were restricted to queries that contained terms related to conversational interfaces and healthcare. With a pilot search through the literature, the list of terms to be included was composed. The list of terms is shown in Table 2.1. The exact search strings for both databases are shown in Appendix A. Papers that contained the keywords in the title or abstract were included.

Papers with a **publication date** before 2015 (I1) were excluded from the search, as chatbots are a relatively new and rapidly developing technique. Also, papers were filtered on **language** (I2), and only papers written in either English or Dutch were included, as the reviewer only speaks those languages on a professional level. Scopus allowed for additional filters, such as scanning the **document type** (I3), which was used to filter on journal papers, conference papers, and conference reviews. Another

Table 2.1: Terms used for database searches in the Identification phase

Term: Chatbot		Term: Healthcare
conversational interface OR **chatbot** OR **conversational chatbot** OR **bot**	AND	**medical** OR **health** OR **healthcare** OR **health care** OR **delivery of health care**

Inclusion Criteria
I1: The paper is published after 2015
I2: The paper is written in English or Dutch
I3: The paper is either a journal paper, conference paper, or conference review.
I4: The paper is published

Table 2.2: Inclusion criteria

essential inclusion criterion, also only on Scopus, was that the papers were already **published** (I4), so they were peer-reviewed and more reliable. The papers that adhered to the described inclusion criteria were loaded into **Rayyan** [42] , a systematic review tool that can be used to easily and orderly include and exclude papers. An overview of the inclusion criteria is displayed in Table 2.2.

[42]: Ouzzani et al. (2016), 'Rayyan—a web and mobile app for systematic reviews'

Screening

The first part of the screening phase entailed detecting duplicates as the papers were identified using multiple sources. Rayyan [42] allowed for automatic duplicate detection by calculating an equality score, which then was used to remove or include detected papers manually.

After removal of duplicate papers, all remaining papers were scanned on title and abstract, and papers were included and excluded for the eligibility phase using the list of predefined **exclusion criteria** (Table 2.3). When a paper matched one of the exclusion criteria, the paper was excluded and labeled with the exclusion criterion. A total of eight exclusion criteria ensured the correct inclusion of papers in this systematic review. As the systematic review revolved around healthcare chatbots, the first two exclusion criteria excluded papers where a **chatbot is not central** (E1) in the paper and papers where the chatbot was **not implemented in healthcare** (E2). The third exclusion criterion excluded all papers on **mental-healthcare** (E3) chatbots, because the focus of this review was on non-mental healthcare. The fourth exclusion criterion excluded papers on **voice-based chatbots** (E4), as these do not fit the predefined definition for TBHC[1] . The fifth exclusion criterion excluded chatbots implementations that were **not evaluated** (E5) with (potential) end-users. The sixth criterion excluded papers with **wrong publication types** (E6), as only journal papers and conference papers were included. The seventh criterion excluded papers without the availability of a **full-text** (E7) version, as the next phase (eligibility) comprised of assessing the full-text papers for eligibility. The last criterion excluded papers **shorter than five pages** (E8) to filter out low-quality papers.

1: The definition of TBHC that was used in this thesis is explained in Definition 1.0.2.

Eligibility

The next round comprised assessing the **full-text papers** for eligibility. The eligibility of a study was determined by reading the full-text paper and checking that the paper passed based on the exclusion criteria as presented in Table 2.3. The papers that were not rejected in this phase were included in the review.

Table 2.3: Exclusion criteria

Exclusion Criteria	Explanation
E1: A chatbot is not central in the paper	The paper must revolve around a conversational interface.
E2: The chatbot is not applied in healthcare	The chatbot was applied in healthcare or supportive of a healthcare process.
E3: The chatbot is applied in mental healthcare or supports a mental healthcare intervention.	As multiple systematic reviews on mental healthcare chatbots have been conducted, chatbots implemented in mental healthcare were excluded.
E4: The chatbot is completely voice-based	Voice-based chatbots were excluded because this review focused on text-based healthcare chatbots (TBHC).
E5: The study is not about the evaluation of a chatbot by its (potential) end-users	The chatbot in the paper should be evaluated on either the usability, effect, acceptability, or perception by (potential) end-users.
E6: The paper is not a journal paper or a conference paper	The inclusion criteria on paper type (I3) was only applied on papers retrieved from Scopus, therefore all remaining papers with other publication types (e.g. conference papers) were excluded with this exclusion criterion.
E7: There is no full-text available	In the eligibility phase, full-text papers were assessed for eligibility. Therefore, it was required that a full-text version of the paper was available to download.
E8: The paper is less than five pages long	Papers shorter than five pages (references included) were filtered out to eliminate low-quality papers.

The references from the included papers were used to perform **backwards snowballing** [43] to find important older papers to include in the review. Especially as the chosen time frame in the inclusion criteria was relatively narrow, it was important to identify more papers to include. The process for snowballing corresponded with the phases as described above; thus, first scanning titles and abstracts, followed by retrieving full-texts and assessing those for eligibility using the exclusion criteria. The papers that passed the exclusion criteria were included in the review, in addition to the papers included in the eligibility phase.

Data Extraction

For every included paper, data was extracted to create two tables with information from the paper. The first provides **background information** of the included papers. For this table, data was extracted to list the author and year, the aim of the chatbot, the country where the research took place, the number of participants, some characteristics of the participants (gender, age, cultural backgrounds and other characteristics such as participants with a particular disease), study design, chatbot type, and publication type (see Table 2.4). When the chatbot type was not mentioned in the paper, an assumption was made. Assumptions were marked with an asterisk in the table.

The second table provides a **structured summary** of the included papers. Data for this table were extracted to determine background information, outcome variables, methods, outcomes, and limitations to the study reported in the paper (see Table 2.5). These items were chosen to follow the PRISMA protocol [41], which prescribes that a structured summary should include (if applicable): background, objectives (the aim of chatbot), data sources, participants, interventions (methods), results, limitations and conclusions (outcome).

Some data items were **categorized** to allow for easy comparison. For the overview table, the aim of the chatbot was categorized by assigning a **healthcare context** the chatbot was applied in. The healthcare context was categorized as antenatal healthcare, chronic healthcare, health information, hereditary health, oncology, or healthy lifestyle. These types of healthcare contexts are further explained in the results section (see: 3.1). The **participants' characteristics** were separated into three categories, namely gender (either male or female, no non-binary people were identified as participants in any of the studies), age (adults, children, or elderly), and cultural background (Asian, African American, Caucasian, or Hispanic). In addition, the **study design** of the evaluation was categorized as experimental, observational analytic, qualitative research, or survey. The final two classifications were applied to **chatbot type** (CT) (AI-based, Rule-based or Hybrid) and **publication type** (PT) (conference paper or journal paper). In the structured summary table, **outcome variables** were categorized as behavior change, system quality, usability and user experience, user characteristics or other. In addition, **limitations** were categorized as no baseline or control group; limitations related to used methods; limitations related to participants; low response rate or low activation rate or other.

Table 2.4: Data in background information

Data	Explanation
Aim	Type of healthcare and goal of chatbot
Country	Notation based on [44]
Participants	Number, gender, age, cultural backgrounds, and other characteristics
SD	Experimental, observational, qualitative or survey
CT	AI-based, rule-based or hybrid
PT	Conference paper or journal paper

SD: study design, CT: chatbot type, PT: publication type

Table 2.5: Data in structured summary

Data	Explanation
Background	Background information of the healthcare context the chatbot is applied in
Outcome variables	Measured outcome variables to evaluate the chatbot
Method	Used methods
Outcomes	Key findings of the evaluation
Limitations	Limitations listed by the authors

2.2 Chatbot Development

After conducting the systematic review and analyzing its outcomes, the chatbot was developed and implemented. As explained in the introduction, the goal of the chatbot was to **assist a (potential) patient during triage and diagnosis** before visiting a general practitioner or healthcare provider (RQ2). This section will describe what functionalities were implemented in the chatbot (RQ2a) and how the chatbot was hosted (RQ2b).

Functionalities

The chatbot can be defined as a CSC[2] . As mentioned in the introduction, You and Gui [27] found **five deficiencies** of CSCs, namely: (I) insufficient consideration of health history, (II) strict input requirements, (III) vague probing questions, (IV) incomplete set of health conditions, and a (V) lack of functions for follow-up treatments. Therefore, the deficiencies that fall in the scope of this research will also be addressed in the implementation.

2: The definition of CSC used in this thesis was explained in Definition 1.0.3.

With the goal and deficiencies found by You and Gui in their analysis of CSC apps in mind, five user stories were developed. The deficiencies were used as a guideline, because there were no evaluations of CSCs found in the systematic review. Therefore, the chatbot in this thesis could not build upon suggested future research by those papers. These user stories are shown in Table 2.6. Only potential patients or their family members were in the scope of this conversational interface and were referred to as the user in these user stories. The first user story (U1) revolved around discovering the possible causes for the symptoms the patient is experiencing, which required a functionality that the chatbot showed a list of **possible conditions** after the triage procedure. The second user story (U2) described that the user should have insight into which conditions are most likely, and the requirement attached to this user story was that a level of **likelihood** was provided for every suggested condition. The third user story (U3) described the need for a **triage level**. Every user was provided with a triage level to either self-care at home, visit the doctor for a routine check-up, visit the doctor today, visit the emergency department or call an ambulance. The fourth user story addressed that the user needed to be informed about the severity of their symptoms (U4), which can also be achieved by providing triage levels. Lastly, the fifth user story (U5) explained that the patient should be asked to fill out some information on their medical profile (gender, age and medical risk factors) to address one of the common problems with CSCs (I).

In addition to the user stories, the chatbot allowed for free text entry (II), used clear language (III) and contained a comprehensive set of conditions (IV) by using a third-party database (Infermedica) with thousands of symptoms and conditions [45]. Unfortunately, the last deficiency fell out of the scope of this research, as the thesis was only focused on triage and diagnosis before visiting a healthcare provider.

Table 2.6: User stories for development

User stories
U1: As a user, I want to know the **possible causes** for my symptoms so that I can have a better insight in my health.
U2: As a user, I want to see **which conditions are most likely** so that I can consider which condition fits my symptoms best
U3: As a user, I want to be provided a **triage level**, so that I know the next steps in my health journey.
U4: As a user, I want to be **informed** about the severity of my symptoms, so that I can take measures accordingly.
U5: As a user, I want to be sure that my **medical profile** is taken into account during triage, so that I have more reliable possible conditions presented to me.

Azure Resources

For the development of the chatbot, the Azure Health Bot service was used. As this service is part of Microsoft's cloud solutions, it is easy to integrate the chatbot with many other tools such as AI-based services, analytics, and web applications. For the development phase of the thesis, three resources were used: **Azure Health Bot** [34], **Application Insights** [46], and **App Service** [47]. To be able to create these resources, some preparation was needed. Firstly, an **Azure subscription** [48] was needed. This is a logical entity that gives the user entitlement to deploy and consume Azure resources. There are different subscriptions, namely free subscriptions, pay-as-you-go subscriptions or prepaid credit carrying subscriptions. In this case, Microsoft provided a prepaid credit subscription as this research was carried out as part of an internship at the company. When a subscription was established, a **resource group** [49] was created, which is a container that allows for the management of multiple applications within the same solution or group. For example, roles and people can be given access to an entire resource group instead of giving every person individually access to every application. Within a resource group, **resources** were deployed. Three services were used in this implementation. Firstly, the chatbot was built with Azure Health Bot, deployed to a web environment with App Service, and the metadata of chatbot usage was quantitatively analyzed with Application Insights.

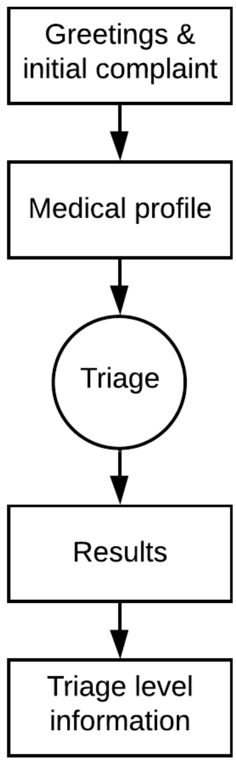


Figure 2.2: Conversation flow (simplified)

Conversation Flow and Azure Health Bot

To adhere to the required functionalities (described above), the conversation flow as shown in Figure 2.2 was used when developing the chatbot. The conversation started with an **onboarding** where the functionalities of the chatbot were explained, a scope was set, and tips on how to interact were given. This follows the suggestions by [50], who describe that proper onboarding can increase the user experience with a chatbot. The patient was also prompted to fill out an initial complaint. Then, the participant was asked to fill out information regarding their **medical profile**, such as gender, age and medical risk factors. The next step was the bulk of the conversation, namely the **triage**. First, the chatbot asked which other symptoms the patient is experiencing and suggested symptoms often in tandem with the inputted symptom(s) and symptoms that can indicate emergencies. After that, the chatbot asked follow-up questions based on the inputted symptoms, e.g. "how long has your cough lasted?" or "how high is your fever?". Finally, the combination of the information derived from the triage was used to determine the **results** in terms of possible conditions and triage levels. The possible triage levels were self-care at home, visit the doctor for a routine check-up, visit the doctor today, visit the emergency department or call an ambulance. The chatbot then provided **additional information** about the specific triage level to conclude the conversation.

The healthcare chatbot was built with a **hybrid** approach, as it allowed for free-text input when entering symptoms, based the follow-up questions on the inputted symptoms, however also followed a scenario script to give structure to the conversation. The handling of free language is an artificial intelligence-powered task, whereas the script and buttons with answer suggestions are derived from a rule-based approach [22]. Therefore, there was freedom in the use of the chatbot, but it was restricted to a certain amount, as spontaneous utterances often have lower accuracy. This was also the case for a study by Davis et al. [51] where scripted questions were answered correctly 97% of the time, compared to 21% when spontaneous exchanges were made. Therefore, this still presents a need for a scripted conversation, as a healthcare chatbot's accuracy is essential.

The conversation was scripted by developing **conversation flows** in Azure Health Bot, which was the protocol that the chatbot will follow. Although the service allows for multiple scenarios, only one scenario was implemented (triage). This scenario already covered the scope of diagnosis and triage. The initial conversation flow was set up using the triage and handoff template in the Azure Health Bot service. This template has pre-build processes and extends the default symptom checker for more reliable diagnosing. For example, Infermedica is already set as the chatbots medical knowledge in this template. **Infermedica** [34] is, as explained in the introduction, one of the included knowledge bases in Azure Health Bot that can be used for triage and symptom checking. It contains simplistic decision tree models based on the mapping of thousands of symptoms and conditions to accurately determine the likelihood of a condition given a set of symptoms. When Infermedica is applied in a HealthBot implementation, the user is asked a range of questions based on their symptoms and ultimately determines the most likely conditions and severity of these conditions by providing a triage level. The template was subsequently adjusted to adhere to the requirements. For example, the template contains a handoff feature that connects the user with a live nurse if the symptoms are severe, but that feature is out of the scope of this research. Components were added to the template such as the additional information on triage level and the onboarding at the start of the conversation. The final result of the chatbot conversation flow is shown in the result section (see Figure 3.6). The configurations of the chatbot were adjusted as well, such as specific language usage, assessment type, the channel for hosting, and the analytics settings. With every adjustment to the chatbot, a test was carried out by running a (part of the) conversation with the chatbot to ensure everything still worked properly. Conversation logs were captured to allow for qualitative assessment in the evaluation (see Section 2.3).

Hosting and Analytics

After the development of the chatbot, it needed to be **hosted** to allow users to access and use the chatbot. Azure Health Bot offers several hosting options, called channels, such as Microsoft Teams, Twilio, Facebook, Telegram, Alexa and Whatsapp. However, these all require a sign-in from the user and sometimes that the app is already downloaded. Therefore, the chatbot was hosted on a website using Azure App Services to lower the chatbot's usage barrier. A GitHub repository [52] was used as an example to create a container for the Health Bot that allowed users to communicate with the Azure Health Bot through a WebChat. That way, it was not necessary to log in or download an application and therefore it also guaranteed the anonymity of the participants. Because the chatbot was hosted on a website, it was also accessible for all devices with a browser.

The final Azure service used in this implementation was Application Insights to allow for **analytics**, where the meta-data of the conversation of the users with the chatbot were captured using the Custom Telemetry functionality of the chatbot. This functionality leverages the native connection between Health Bot Service and Application Insights, which allows Application Insights to capture Custom Events via an App Insights Instrumentation key which can be requested in the Health Bot Service. The captured Custom Events can subsequently be queried in the service, using **Kusto Query Language** (KQL), which is used amongst all Azure services. KQL is a read-only query language designed to ensure that the syntax is easy to read and understand, provide high-performance through scaling, and smooth transitions from simple to complex queries [53].

2.3 Chatbot Evaluation

The goal of this part was to **evaluate the Diagnosis and Triage chatbot with (potential) end-users with commonly used outcome variables** to identify the current quality of the chatbot, and pinpoint possible improvements (RQ3). Given the commonly used outcome variables identified in the systematic review, the chatbot was evaluated on chatbot usage (RQ3a), user experience (RQ3b), and influence on health behavior (RQ3c). To achieve this goal, a questionnaire with a link to the chatbot was sent out where participants filled out questions related to the outcome variables to test the chatbot. The questionnaire was generated using Qualtrics software, version May 2021.

Participants

As the goal was to evaluate the chatbot with potential end-users, the target group was broad because everyone can receive primary care. Therefore, there were only two inclusion criteria, namely the participants needed to be at least **18 years** old and be able to speak **English** at an intermediate level as both the chatbot and questionnaire were in English. The participants were recruited through **convenience sampling** [54]. This method has multiple advantages; firstly, it ensures that the included participants adhere to the participation criteria. In addition, it is a cheap and fast method to gather participants. Due to the limited time frame of this thesis, this was an important characteristic of the sampling method. Participants were mostly recruited through social media such as LinkedIn and Whatsapp. Five pilot tests were performed to find deficiencies in the questionnaire, and all were included in the final results as there were no (big) shortcomings.

Questionnaire

The questionnaire was based on the outcome variables that were found in the systematic review (Table 3.3). These were user characteristics, user experience, system quality, behavior change, chatbot usage, and other. The first five outcome variables were included in the questionnaire to evaluate the chatbot, other was excluded as those outcome variables were specific for that implementation or study (e.g. genetic test results or correlation between different evaluation measures). The complete questionnaire is shown in Appendix B. Table 2.7 shows an overview of how each outcome variable is implemented in the questionnaire.

Table 2.7: Outcome variables in the Questionnaire

Outcome Variable	Measured by
UC	Demographics
U	Metadata analysis, conversation logs
UX	CUQ, free-text question
SQ	CUQ (Response factor) and free-text question
B	Advice from chatbot match and extent of influence

UC = User Characteristics, U = Usage, UX = User Experience, SQ = System Quality, B = Behaviour change

Consent Form

The questionnaire started with a consent form. Here, the participant was first informed of the project, the procedure of the evaluation, and that their participation is entirely voluntary. Also, the researcher's and supervisor's contact details were provided for any issues or questions. The consent form ended with a summary of the information, after which consent was asked. The consent form is included in the Questionnaire in Appendix B.

Demographics

Next, participants were asked to fill out their demographic data. Here, age, gender and highest degree of education were collected. With this information, an overview of the sample of participants used in the evaluation and information on the **user characteristics** were gathered. In the results from the systematic review (see Table 3.4), multiple studies reported demographics of their participants [55–57]. This is also the case for this thesis.

Chatbot Usage

In the next section of the questionnaire, the participants were guided to use the chatbot. To refrain participants from filling out their own symptoms, medical information, or medical history[3] , they were asked to use their imagination when talking to the chatbot. The participants could also use one of the three **scenarios** that were included in the questionnaire (see Table 2.8). This is a scenario-based technique, which is an appropriate method for evaluating systems because they represent a concrete example of system use [58]. However, to also ensure that the chatbot is not only tested on those scenarios, it is also possible to fill out any other set of symptoms. Before starting the conversation with the chatbot, the participant was provided with a participant number (a random number between 0 and 10,000). The chatbot subsequently asked for the participant number in one of the first messages. This way, it was possible to connect questionnaire responses to the conversation logs.

3: For ethical, safety, and anonymity reasons.

Table 2.8: Scenarios in the questionnaire

Scenarios
Scenario 1 (Migraine): *Judy is a 42-year-old woman and sometimes suddenly has very severe headaches, usually on one side of the head. When these headaches occur, it sometimes is accompanied by other symptoms such as nausea and vomiting. The headaches can last for hours or sometimes even days. Judy is very worried about her symptoms and wants to visit a doctor but decides to consult the chatbot first.*
Scenario 2 (Tuberculosis): *Jerry (male, 76 years old) has been coughing for several weeks, and first expected that it was just another cold, but it keeps getting worse. In addition to his cough, he also experiences pain when coughing or breathing. This week he even coughed up blood and that made him very worried. Jerry is not a huge fan of hospitals, so he decides to consult the chatbot before visiting the doctor.*
Scenario 3 (Depression): *Jennifer is a 26-year-old woman who has felt sad, empty, and hopeless for a longer period now. In addition, she keeps falling out with her boyfriend as she seems to be irritated by everything he does. Her friends are also very worried, as she keeps rejecting normal activities such as exercising or picking up her hobbies. In addition, she cannot remember the last time she had a good night of sleep. She decides to consult the chatbot with these symptoms.*

All messages of the conversation between the chatbot and the participant can be retrieved by downloading the **conversation logs**, which were automatically captured when activating Azure Health Bot. To adhere to GDPR and reduce unnecessary data storage, there is an option to delete the data after a certain period automatically. In the case of this research, the logs were deleted after the publication of the results.

In addition, metadata of the messages, such as timestamps and identification numbers, were captured with the Custom Telemetry functionality as explained in the previous section. This was used to determine outcomes related to **chatbot usage**, which was also often done in the papers in the systematic review. Following the examples in those papers, the metadata was used to determine the elapsed **time** (median duration) [55–57, 59–62], **drop-off rate** [59–61, 63], and **number of messages** [55, 62–64].

User Experience

After using the chatbot, the participants were prompted to fill out questions regarding their **user experience** with the chatbot. Three papers in the systematic review used quantitative measures to determine the user experience with the chatbots [56, 57, 62]. In this research, the **Chatbot Usability Questionnaire (CUQ)**, created and validated by [56], was used to evaluate the usability of the diagnosis and triage chatbot. The CUQ is a relatively new questionnaire designed specifically to measure the usability of chatbots by scoring the chatbot on factors that are retrieved from the commonly used ALMA Chatbot Test Tool [50]. Factors included in the CUQ are **personality, onboarding, purpose, navigation, understanding, responses, error handling, and ease of use**. Two statements were presented to the user with different polarity for every factor: one positive statement and one negative statement. The level of agreement with the sixteen statements was ranked on a five-point Likert scale [65], from "Strongly Disagree" to "Strongly Agree". The CUQ statements are listed in Table 2.9.

In addition to the CUQ, two free-text input questions were asked to receive comments on the experience and the chatbot's responses. The input for comments on the chatbot's responses was added to cover the **system quality** outcome variable, in addition to statements 11 and 12 from the CUQ (statements from the Responses factor). In the studies included in the systematic review, quality of the response was often used as a measure for System Quality [55, 60, 64, 66–68].

Advice from the Chatbot

The last part of the questionnaire was included to determine whether participants would follow the advice from the chatbot and to what extent the advice from the chatbot influenced them. Therefore, the first two questions asked what the chatbot advised and what the participant's following action would be considering the inputted symptoms and the advice from the chatbot. The answer options were the five triage levels from the chatbot (self-care at home, doctor routine check-up, doctor today, emergency department or ambulance). With the third and last question from the survey, participants were asked to what extent the advice from

Table 2.9: Statements in the Chatbot Usability Questionnaire (CUQ) [56]

Factor		Statement
Personality	1	The chatbot's personality was realistic and engaging
	2	The chatbot seemed too robotic
Onboarding	3	The chatbot was welcoming during initial setup
	4	The chatbot seemed very unfriendly
Purpose	5	The chatbot explained its scope and purpose well
	6	The chatbot gave no indication as to its purpose
Navigation	7	The chatbot was easy to navigate
	8	It would be easy to get confused when using the chatbot
Understanding	9	The chatbot understood me well
	10	The chatbot failed to recognise a lot of my inputs
Responses	11	Chatbot responses were useful, appropriate and informative
	12	Chatbot responses were not relevant
Error Handling	13	The chatbot coped well with any errors or mistakes
	14	The chatbot seemed unable to handle any errors
Ease of use	15	The chatbot was very easy to use
	16	The chatbot was very complex

the chatbot influenced them. Here, the five answer options ranged from "a great deal" to "none at all" (a great deal, a lot, a moderate amount, a little or none at all). These questions were related to the **behavior change** outcome variable, where the aim was to measure adherence to the chatbot advice. Adherence to provided information from a chatbot, such as a treatment, diet, or physical plan was often measured in the papers from the review [51, 62, 63, 69].

Analysis of the Results

Statistical Analysis

SPSS (Version 26) was used to analyze the demographics, behavior change, and part of the CUQ sections. For the **demographics**, SPSS was used to calculate the descriptive statistics of gender, age and education levels. Also, the chart builder was used to create visualizations of the data. This provided a better overview of the participants that were included in the study. A binary variable was created to determine whether the participants followed the advice from the chatbot in the **behavior change** section. The variable was valued 1 when there was a match between the chatbot's advice and the next action from the participant, and 0 otherwise. Then, the descriptive statistics of this variable were calculated and visualized with the chart builder. The descriptive statistics were also calculated and visualized for the question that measured the extent to which the chatbot's advice influenced the participant.

CUQ Scores

SPSS was also used to score the chatbot on the eight **factors of the CUQ**. For this, points were allocated for the different answer possibilities (strongly disagree: 0 points, somewhat disagree: 1 point, neither agree nor disagree: 2 points, somewhat agree: 3 points, and strongly agree: 4 points). Then, a score per factor was calculated by subtracting the score of the negative question from the score of the positive question, then scaling it to a score from 0 to 10 (see Formula 2.1). The scores per factor were subsequently visualized with the chart builder.

$$CUQFactor = (positive\ question - negative\ question\ + 4) * 1.25 \tag{2.1}$$

Formula for scoring the CUQ factors on a scale from 0 to 10, for example: PersonalityScore = (score statement 1 - score statement 2 + 4) * 1.25

The CUQ score was calculated by first assigning scores to the answer possibilities, as described above. Then, a CUQ score per participant was calculated using Formula 2.2 as presented by [56]. The CUQ score calculator [70] was used to allow for quicker calculation of the results. The scores were then ranked from high to low and matched to the participant number. Different colours were added to distinguish excellent scores (CUQ > 85), very high scores (CUQ > 75), high scores (CUQ > 65), OK scores (CUQ > 50), poor scores (CUQ > 40), and awful scores (CUQ <= 40). The median and average scores were calculated to determine the usability score of the chatbot.

$$CUQ = ((\sum_{n=1}^{m} 2n - 1) - 5) + (25 - (\sum_{n=1}^{m} 2n)) * 1.6 \tag{2.2}$$

Formula CUQ Score: where m = number of questions (16) and n = individual question score per participant.

The **conversation logs** of the participants who gave CUQ Scores lower than 50 were downloaded for qualitative assessment of the conversation. This way, possible bugs and mistakes can be identified. In addition, the free-text answers from those participants (regarding user experience and response) were also inspected to determine possible deficiencies.

Free-text Analysis

The free-text input from the participants was first separated into two categories: positive and negative comments regarding the chatbot. Then, the negative comments were coded to determine possible improvements to the chatbot. For this, the process of Open Coding [71] was used, where the concepts emerge from raw data and were later grouped into conceptual categories. These categories were the areas where the chatbot has room for improvement.

Analysis of Metadata

The metadata, captured with custom telemetry, was saved to Application Insights that saved the data as custom events and allowed querying of the events using the Kusto Query Language (KQL). With the data, the median dialog duration, drop-off rate, and the number of messages were calculated using the KQL queries as displayed in Appendix C.

Median dialog duration: The median time of the finished conversations was calculated to determine how long a conversation usually lasted. As there were outliers on both sides, the median time of the finished conversations was most suitable. The query started with filtering the data on the period in which the evaluation was performed, which was from 6 May to 24 May 2021. Then, the UserID was retrieved to separate different conversations. The messages were joined together on UserID to retrieve the timestamps of the start and end of the conversation. The start and end of the conversations were automatically labeled as "ScenarioStart" and "ScenarioEnded". Then, the duration was calculated by subtracting timestamp (start) from timestamp1 (end). The data were then summarized to determine the median duration.

Drop-off rate: The drop-off rate was determined by first filtering on the evaluation period, followed by separating the conversations with the UserID. Then, the number of started conversations were counted when a message labeled "ScenarioStart" was encountered. Finally, the number of completed conversations were determined by counting only the messages that were labeled "ScenarioEnded". These numbers can be used to calculate the drop-off with the equation below.

$$DropOffRate = (1 - \frac{\#\,ended\ conversations}{\#\,started\ conversations}) * 100\% \tag{2.3}$$

Formula to calculate the drop-off rate

Number of messages: Lastly, the average and the median number of messages per finished conversation were calculated with the KQL query below. This query also started with filtering on the period. Then, custom events with the label "ScenarioStart", "Message", or "ScenarioEnded" were filtered to only count messages and not other events such as database calls. Then, it was determined whether a message was ended by creating a binary variable that was 1 when the scenario was ended and 0 when there was no end to the conversation. Only messages of completed conversations, thus when the binary value was 1, were counted by using the countif function. Finally, the summarize function retrieved the final result.

Results

Results | 3

3.1 Systematic Review

Included papers

Using the query in Table 2.1, 1005 papers on Scopus and 275 papers on PubMed were identified, a total of 1280. These papers were first pre-screened using the filter functionalities of the databases following the inclusion criteria. 319 papers were not included because they were **published before 2015** (I1), leaving 961 papers. 28 papers were not included because they were not written in **English or Dutch** (I2). Another 46 papers from Scopus were not included because of a **wrong publication type** (I3). Lastly, papers that were **not yet published** (I4) were not included. Ultimately, 863 papers were included in the screening phase.

As a first step in the **screening** phase, 167 papers were excluded after scanning for duplicates. The remaining 696 papers were screened on title and abstract using the exclusion criteria. Using these predetermined requirements, 385 papers were excluded because the paper was not about a chatbot or the **chatbot is not central** in the research (E1). For example, this was the case when the search term "BOT" was used for an unrelated topic, or when a chatbot was suggested as one of the possible AI implementations. 77 papers were excluded because the chatbot in the paper was **not applied in a healthcare context** (E2). 34 papers were excluded as they were applied in **mental healthcare** (E3) or supported a mental healthcare intervention. 26 papers on **voice-based chatbots** (E4) were excluded as this thesis focuses on text-based chatbots. 48 papers were excluded because the **chatbot was not evaluated** (E5). Lastly, 84 papers were excluded because of a **wrong publication type** (E6). In total, 654 papers were excluded in the screening phase of the PRISMA protocol, leaving 42 papers to be assessed in the eligibility phase.

In the eligibility phase, the 42 full-text papers were retrieved and assessed for **eligibility**. Five were excluded because the chatbot was **applied in mental healthcare** (E3) or supported a mental health intervention as a treatment of a non-mental health problem. One paper was excluded because the chatbot was **voice-based** (E4). Sixteen were excluded because the chatbot was not **evaluated** (E5). Another was excluded because of a **wrong publication type** (E6). Lastly, three were excluded because they were **shorter than 5 pages** (E8). Overall, 26 papers were excluded in the eligibility phase and 16 papers were included in the systematic review.

The references of the 16 included papers were scanned to perform **backwards snowballing**. A total of 420 references were scanned on title to identify additional papers on chatbots applied in healthcare. From these references, 50 abstracts were retrieved as the title indicated a research related to chatbots. The abstracts of these papers were screened, which excluded 42 papers. Reasons for exclusion were:

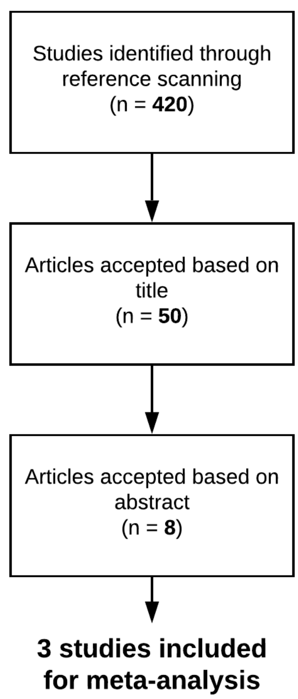


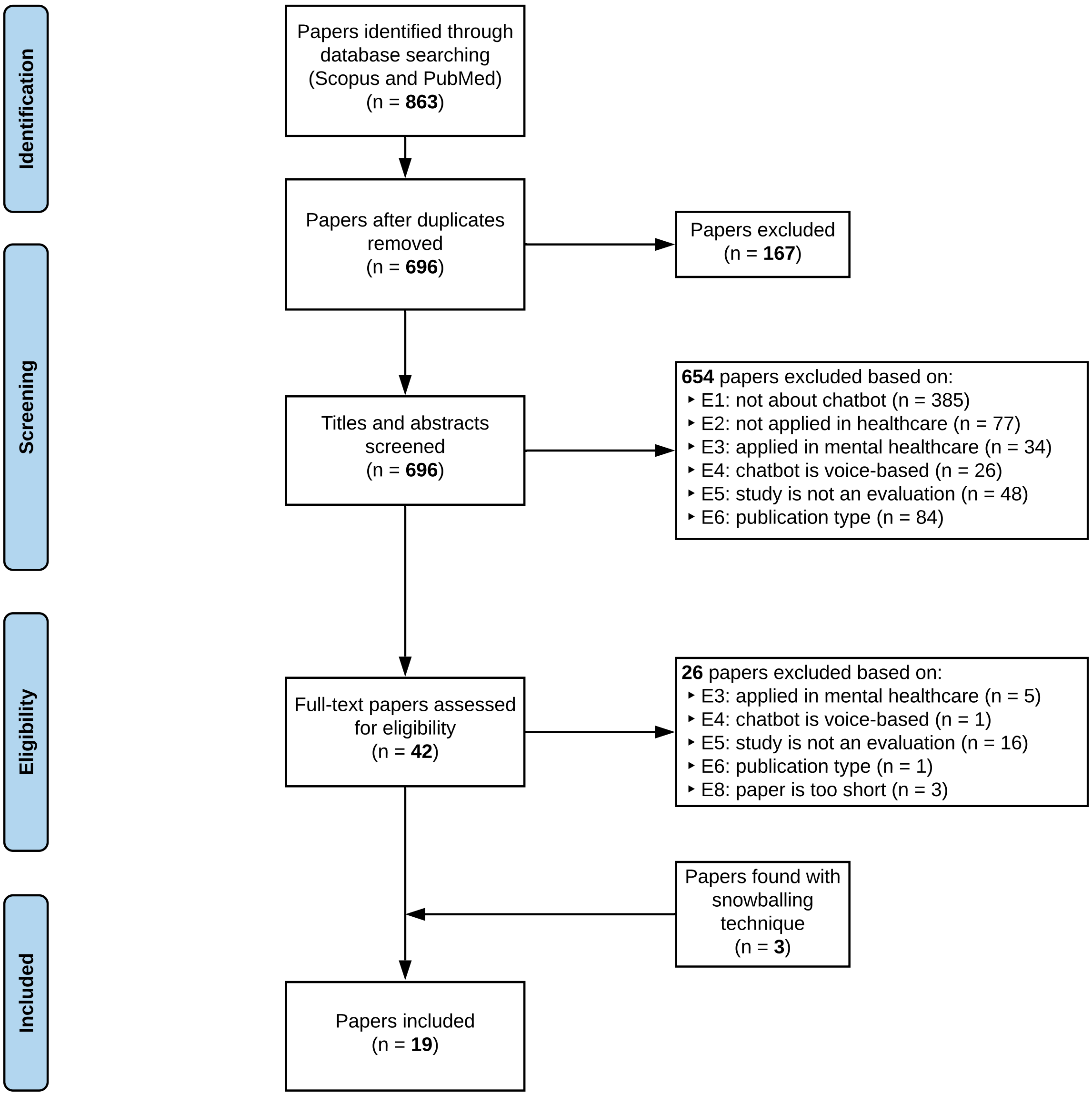
Figure 3.1: Flowchart for backwards snowballing

- The paper was a duplicate of papers in the initial search (n = 25)
- E1: The paper is not about chatbots (n = 9)
- E2: The chatbot was not applied in healthcare (n = 4)
- E3: The chatbot is applied in mental healthcare (n = 2)
- E4: The chatbot is voice-based (n = 1)
- E5: The chatbot is not evaluated (n = 7)
- E6: The publication type is not journal papers or conference paper (n = 1)
- E7: The full-text is not available (n = 1)

The full-text of the 8 remaining papers was assessed, from which 5 papers were excluded because one chatbot was not applied in healthcare (E2), two were based in mental healthcare (E3), one was not about the evaluation of a chatbot (E5), and one chatbot was voice-based (E4). The remaining 3 papers were included for meta-analysis. Figure 3.1 provides an overview of the snowballing process. Data was extracted of a total of 19 papers.

The entire process of identification, screening, eligibility assessment and snowballing is displayed in a **PRISMA flowchart** in Figure 3.2.

Figure 3.2: PRISMA Flowchart for paper selection



Background Information

Table 3.2 provides an overview of the background information of the included papers in this systematic review. The papers were published between **2011 and 2020**, however most papers (n = 14) were published in 2019 or later.

Table 3.1: Frequency of chatbots in a healthcare context

Context	#	Citation(s)
Healthy Lifestyle	5	[51, 56, 62, 69, 72]
Health Information	4	[55, 59, 64, 68]
Oncology	3	[61, 63, 73]
Hereditary	3	[57, 60, 74]
Antenatal	2	[66, 67]
Medication	1	[75]
Chronic	1	[76]

Healthcare Contexts

The chatbots were applied in various healthcare contexts (see Table 3.1), but most chatbots had a special focus on a promoting a **healthy lifestyle** (5 chatbots). In this healthcare context, the aim was to coach users towards a healthy lifestyle by promoting meal quality [62], physical activity [72], or both [51, 56, 69].

Four chatbots aimed to provide **health information** to the user. One assisted people with low health and computer literacy in finding relevant clinical trials on the internet [59]. Another answered questions on sex, drugs, and alcohol [55]. Another aimed to improve information provision regarding the treatment of kidney stones [64]. Lastly, one aimed to provide the right ICD-10 code for a disease, health issue, or public health service [68].

Three chatbots were applied in **oncology**. Two chatbots were developed to aid breast cancer patients [63, 73]. One chatbot was developed for follow-up with cancer patients receiving chemotherapy [61].

Three chatbots were applied in **hereditary health** for various conditions. One had a specific focus on hereditary cancer [60]. The other two focused on more general hereditary conditions [57, 74].

Then, two chatbots were applied in **antenatal healthcare**. The chatbots were focused on information provision around family planning [66] and fertility knowledge [67].

One chatbot assisted with **medication** by improving drug information access and increasing awareness of drug risk minimization measures among physicians [75].

Lastly, one chatbot was implemented in **chronic healthcare** with a focus on food recommendations for diabetes patients [76].

When inspecting these chatbot aims, there are **three general reasons** for using a chatbots in healthcare, namely **information provision** [55, 60, 64, 66, 67, 73, 75, 76], **coaching/promoting** towards health behavior [51, 56, 62, 63, 69, 72] or **supporting** a user in a task [57, 59, 61, 68, 74].

General Information of the Papers

The evaluations of the chatbots were conducted in various **countries** and continents, as shown in the country column of Table 3.2. Nine studies were conducted in the United States [57, 59–62, 64, 66, 69, 74]. Five in European countries (3 in France [63, 73, 75], one in the United Kingdom [56], and one in the Netherlands [55]). Four in Asian countries (2 in Thailand [68, 76], one in Japan [67] and one in South-Korea [72]). Lastly, one study was carried out in Australia [51].

The **number of participants** ranged from 9 to 958 (M = 212, SD = 340). Most papers included both male and female participants, except for the papers focusing on female fertility [67] or breast cancer [73] where only women participated. Most papers included adults, however [55] focused especially on adolescents and [61] only included elderly people aged 65 or older. Most papers included participants with a specific user profile [55, 59–64, 66, 67, 72–75], e.g. Bibault et al. [73] only included patients with breast cancer (in treatment or remission) as their aim was to provide answers to patients with breast cancer. Papers on chatbots with a more general purpose, such as promote physical activity included participants without any specific user profile characteristics [51, 56, 57, 69]. Two papers included participants with both a general and specific user profile [68, 76].

The majority of the papers used either an experimental (6 papers: [57, 59, 67–69, 73]) or survey (6 papers: [51, 55, 60, 61, 72, 76]) **study design**. Five [56, 64, 66, 74, 75] used a qualitative research study design. Two used an observational study design [62, 63].

The **chatbot technique** column described whether the chatbot was rule-based, AI-based, or hybrid. However, in 6 papers [55–57, 63, 73, 74], the chatbot technique was not mentioned in the paper or easily deductible by looking at the way the chatbot functions. In those cases, an assumption was made and marked with an asterisk (*) in the column. Eleven chatbots used a rule-based approach [55, 57, 59, 60, 62, 64, 66, 67, 72, 74, 76] (3 assumptions). One paper used an AI-based approach [68]. Seven papers used a hybrid approach [51, 56, 61, 63, 69, 73, 75] (3 assumptions).

The last column of the table describes the **publication type** of the paper, which was either a journal paper or conference paper[1] . Fourteen papers were journal papers (n = 14) and five were conference papers. The average number of participants in conference papers is significantly lower than the average number of participants in journal papers (31 compared to 208).

1: As a result of exclusion criterion E6 (see 2.3)

Table 3.2: Background information of the included papers

Source	Aim of chatbot	Country	#	Gender	Age (years)	Cultural background	Other	SD	CT	PT
Bibault et al. (2019) [73]	[O] Provide answers to patients with breast cancer with a level of satisfaction similar to that given by a group of physicians.	FR	142	142 F	[A] M = 42 (SD = 19)		Women with breast cancer in treatment or remission	E	H*	J
Bickmore et al. (2013) [69]	[L] Promote both physical activity and vegetable consumption.	US	122	61% F	[A] 21-69 (M = 33, SD = 12.6)	[C,A] 52% C, 33% A	BMI ranged from 18.8 to 46.4 (M=27.8)	E	H	J
Bickmore et al. (2016) [59]	[I] Allow individuals with low health and computer literacy to identify and learn about clinical trials on the internet.	US	89	48 F	[A] M = 59.2 (SD = 9.8)	[C,AA] 54% C, 46% AA	98% of participants had a current cancer diagnosis	E	RB	J
Chaix et al. (2019) [63]	[O] Empower patients with breast cancer and their relatives and reinforce medication adherence.	FR	958	88.90% F	[A] M = 48		Women with breast cancer or in remission	O	H*	J
Crutzen et al. (2011) [55]	[I] Answer questions about sex, drugs and alcohol.	NL	929	64% F	[C] M = 15		Adolescents	S	RB*	J
Davis et al. (2020) [51]	[L] Assist users to undertake lifestyle changes such as physical activity or diet.	AU	28	68% F	[A] 45-75, M = 56.2 (SD = 8)			S	H	J
Goldenthal et al. (2019) [64]	[I] Deliver information to patients post-ureteroscopy, which is a common procedure to treat kidney stones.	US	20		[A] 31-69		Endured a ureteroscopy	Q	RB	J
Heald et al. (2020) [60]	[H] Identify who is at risk of acquiring hereditary cancer and educate patients who are scheduled for colonoscopy.	US	487	57.9% F	[A] M = 56.6 (SD = 12.5)	[C,AA,H,A] 86% C, 9.3% AA, 1% H, 1% A	Scheduled for colonoscopy	S	RB	J
Holmes et al. (2019) [56]	[L] Function as a self-help motivational tool for weight loss maintenance by encouraging self-reporting, personalised feedback, and motivational dialogues.	UK	30		[A]		Healthy adults	Q	H*	C

Table 3.2 Background information of the included articles (continued)

Source	Aim of chatbot	Country	#	Gender	Age (years)	Cultural background	Other	SD	CT	PT
Hussain et al. (2019) [66]	[A] Provide information about family planning and contraceptives.	US	49		[A] M = 31 (SD = 8.7)	[C,AA] 69% C, 8% AA	Married, living together or engaged	Q	RB	C
Koman et al. (2020) [75]	[M] Improve drug information access and awareness of drug risk minimization measures among physicians.	FR	10	7 M	[A] M = 51.5		8 general practitioners, 2 specialists	Q	H	J
Maeda et al. (2020) [67]	[A] Improve fertility knowledge and intention to optimise preconception health without increasing anxiety.	JP	927	927 F	[A] 20-34, M = 28.8 (SD = 3.6)		Hoping to have children (now or in the future)	E	RB	J
Piao et al. (2020) [72]	[L] Coach office workers towards a healthy lifestyle.	KR	20	16 F	[A] 20-59		Office workers	S	RB	J
Piau et al. (2019) [61]	[O] Follow-up older patients with cancer using a chatbot to free up nurses' time.	US	9	5 M	[E] M = 83.4 (SD = 2.1)		Patients with cancer, 65 years and older	S	H	J
Ponathil et al. (2018) [57]	[H] Collect family health history to identify risks for common chronic diseases.	US	25		[A] M = 26.15 (SD = 2.64)			E	RB*	C
Schmidlen et al. (2019) [74]	[H] Facilitate communication with participants receiving clinically actionable genetic variants from the MyCode® Community Health Initiative.	US	62	68% F	[A] Mostly >58 (84%)	[C] 94%C	Current active enrollment in MyCode	Q	RB*	J
Siangchin & Samanchuen (2019) [68]	[I] Provide the right ICD-10 code for a disease, health issue or public health service.	TH	26				Two groups: experienced and inexperienced in ICD-10 coding	E	AI	C
Stein & Brooks (2017) [62]	[L] Coach users towards weight loss and changes in meal quality.	US	70	74.5% F	[A] M = 47		Overweight and obese (BMI >= 25)	O	RB	J
Thongyoo et al. (2020) [76]	[C] Provide food recommendations for diabetes patients.	TH	24	13 F	[A] 14-72		Both users with diabetes and users not having diabetes	S	RB	C

Empty cells in the Table indicate that the information was not provided in the article

Goal of chatbot: A = antenatal healthcare, C = chronic healthcare, I = health information, H = hereditary health, M = medication, O = oncology, L = promote a healthy lifestyle

Country: based on ISO Country Codes for Selected Countries [44]

Gender: F = Female, M = Male (No non-binary people were included in any of the studies)

Age: A = adults (18 - 65), C = children (age < 18), E = elderly (age > 65)

Cultural background: A = Asian, AA = African American, C = Caucasian, H = Hispanic

SD (Study Design): E = experimental, O = observational analytic, Q = qualitative research, S = survey

CT (Chatbot Technique): AI = AI-based, RB = rule-based, H = hybrid

PT (Publication Type): C = conference paper, J = journal article

Structured Summary

Table 3.4 presents a structured summary of the papers included in this review. Background information, outcome variables, methods, primary outcomes and limitations are listed for every reference. The results in this section will be summarized per column.

Table 3.3: Frequency of outcome variables in evaluation

Outcome variable	#	Citations
User experience	15	[51, 55–57, 59, 61–63, 66, 67, 72–76]
Chatbot usage	9	[55–57, 59–64]
System quality	8	[51, 55, 60, 64, 66–68, 76]
Health behavior	6	[51, 62, 63, 66, 67, 69]
User characteristics	5	[55–57, 66, 67]
Other	2	[56, 60]

Outcome Variables

The outcome variables were categorized as either behavior change, system quality, user experience (including usability), usage, user characteristics or other. Below, the reported outcome variables are described in order of frequency.

Fifteen papers reported on **outcomes variables** related to **user experience** [51, 55–57, 59, 61–63, 66, 67, 72–76]. Six measured the perception of or attitude towards the chatbot [63, 66, 67, 69, 74, 75]. Four measured the user satisfaction [51, 63, 69, 76]. Four measured the perceived quality or performance of the chatbot [55, 66, 72, 73]. Lastly, three used quantitative measures, such as the system usability scale [56, 57, 62].

Nine papers reported on outcome variables related to **chatbot usage** [55–57, 59–64]. These used a quantitative approaches to measure characteristics related to the usage of chatbots. Seven measured the elapsed time [55–57, 59–62]. Four measured retention and/or drop-out rates [59–61, 63]. Lastly, six measured variables related to the chatbot conversation (e.g. number of queries) [55, 57, 61–64].

Eight papers reported on outcome variables related to **system quality** [51, 55, 60, 64, 66–68, 76]. Six measured outcome variables related related to quality of the response [55, 60, 64, 66–68]. Four measured technical performance, such as speed [51, 55, 68, 76].

Six papers measured outcome variables related to change in **health behavior** [51, 62, 63, 66, 67, 69]. All papers with this outcome variable were chatbots with the general aim to coach the user towards certain health behavior, as explained in the previous section. Four measured behavior change as adherence to treatment, diet or physical activity plan [51, 62, 63, 69]. The other behavior related outcome variables revolved around health intention [66, 67].

Six papers measured outcome variables related to **user characteristics** [55–57, 66, 67]. In these papers, participants were asked to fill out either their demographics or related medical information.

Lastly, two papers measured outcome variables that fell outside the categories and are therefore labeled as '**other**' [56, 60]. In these papers, genetic test results and variables related to the effectiveness of evaluation methods were retrieved.

Methods

The **methods** used varied. Six papers used an **experimental** approach [57, 59, 67–69, 73]. In most of these papers, participants were split into (at least) an intervention and control group. One study used a within-subject design [57]. Four papers used a **longitudinal** approach [51, 61, 62, 72]. The chatbot was then used for 3 to 12 weeks with either continuous follow-up [61, 62, 72] or with predetermined phases and times of measuring [51]. Four papers used **qualitative** methods. Two papers used semi-structured interviews [64, 75] to explore users' impressions of the chatbot. The other two papers used focus groups [63, 74]. Three papers used **questionnaires** to either measure medication adherence [63], compare the chatbot to conventional methods [55], measure the acceptance of the chatbot [66], or measure the usability [56, 76]. Lastly, [60] measured the effectiveness of the chatbot by defining risk of hereditary cancer and evaluating the system quality by comparing the calculated risk to the genetic test results.

Outcomes

For description of the outcomes of the evaluations, the chatbots will be grouped by healthcare context[2] .

2: An overview of the chatbots applied in the various healthcare contexts is provided in Table 3.1

Papers on chatbots implemented to promote a **Healthy Lifestyle** [51, 56, 62, 69, 72] were used to coach towards healthy behavior and therefore promoted healthy diets and physical activity. Four of these chatbot evaluations measured the effectiveness of the chatbot in terms of users' physical activity, diet adherence and/or weight loss. Papers [77], [51] and [62] all reported an increase in diet adherence. For physical activity, only [72] finds that the chatbots increased repetitive physical activity, whereas [77] only found an insignificant increase. In addition, [51], [56], and [72] reported on a positive user experience with the associated chatbots. Another chatbot, though implemented in **chronic healthcare** [76], also noted a positive user experience when coaching toward good nutrition for diabetes patients.

Papers on chatbots implemented to support **health information** provision also reported positive user experiences. Crutzen et al. [55] reported that the information from the chatbot was faster, more anonymous, conciser and of higher quality than traditional methods of information provision. The use of the chatbot in [59] resulted in more satisfaction, increased trust and pleasure, and decreased frustration, in addition to increased task completion. However, this paper also noted that the task completion time was greater for the chatbot. Siangchin and Samanchuen [68] especially measured the system quality and found strong evidence of good performance and accessibility of the chatbot. There were also positive remarks on the user experience in [64], where the chatbot was used to address concerns around ureteroscopy. The chatbot received positive remarks from the interviews.

From the 3 chatbots applied in **oncology**, two papers positively reported on the effect of the chatbot. Firstly, [73] found that the perceived information quality of the chatbot was non-inferior to the information provided by a group of physicians. [63] saw that the usage of a chatbot increased the compliance with treatment. In addition to promising effects of the chatbot, [63] and [61] reported on high user satisfaction.

In **hereditary health**, two papers reported on positive user experiences with the chatbots [57, 74], however [57] noted the increased time and clicks to complete a task. This was however compensated by a decreased workload and mental demand, in addition to a more usable experience than with the baseline. [60] found the chatbot to be a feasible method to identify people with hereditary cancer.

Two chatbots were applied in **antenatal care**, but measured different outcomes. Firstly the paper by Hussain et al. [66] mostly focused on finding the main reasons for a participant to reuse the chatbot and saw a high intention to reuse and positive attitude towards the chatbot in the measurements. In the paper by Maeda et al. [67], a chatbot was compared to two control groups, namely one where participant were provided with a booklet with fertility information (CG1), and one where the participants were provided with a booklet with irrelevant information (CG2). It was found that the chatbot significantly improved fertility knowledge compared to CG2, however this effect was smaller than the effect of CG1. However, the chatbot saw a reduced anxiety compared to both control groups.

Lastly, the **medication** chatbot [75] was perceived to be a useful and innovative tool that can help users seek information on different drugs and risk minimization measures.

When comparing evaluations outcomes of chatbots across various healthcare contexts, there were some frequent **effects** of chatbots usage. The most frequent outcome was a positive user experience, which was measured in 15 of the 19 papers and positively evaluated in all 15 [51, 55–57, 59, 61–63, 66, 67, 72–76]. However, two usability issues were reported multiple times. Firstly, problems with spontaneous utterances for AI-based and hybrid chatbots [51, 75] were reported in 2 of the 8 chatbots that allowed for free-text communication. Also, the use of a chatbot resulted in longer task completion time compared to traditional methods [57, 59].

Limitations

The limitations listed in the included papers were categorized as either no baseline or control group; limitations related to methods or used measures; limitations related to participants; low response rate or low activation rate and/or other. Of the 19 papers included in this review, 13 listed limitations in the discussion. Six papers did not mention limitations. Of these 6, 4 were conference papers [56, 66, 68, 76] and two journal papers [63, 72]. From the 5 conference papers included in the review[3] only one paper listed limitations [57].

3: The publication type for every paper included in this review is listed in Table 3.2.

Twelve papers listed limitations related to the **participant** sample [51, 55, 57, 59–61, 64, 67, 69, 73–75]. Six papers explained that the limitations were caused by a small number of participants [51, 55, 59, 61, 69, 75]. Another six papers explained that the limitations were caused by the sample not being representative of the population [51, 57, 60, 67, 74, 75], more specifically two papers reported that patients were previously involved in trials or studies [59, 75].

Ten papers listed limitations related to **methods or used measures** [51, 57, 59, 60, 62, 64, 67, 69, 73, 74]. The limitations in this category were very diverse, however always as a result of the chosen methods. Examples are scarce collection of demographics [62, 73], outcome variables that cannot be quantified (e.g. opinions) [51, 74] and that the task in the evaluation did not completely correspond with real-life usage [57, 74].

Three papers listed limitations related to **low response rate or low activation rate** [55, 60, 64]. Two listed the limitation that a **control group** would have made the research more reliable [62, 69]. Lastly, three listed other limitations [60, 62, 75].

Table 3.4: Structured summary of the papers included in the review. For every paper, the background, outcome variable(s), methods, outcomes and limitations are listed.

Source	Background	Outcome variables	Methods	Outcomes	Limitations
Bibault et al. (2019) [73]	The data regarding to the use of conversational agents in oncology are scarce.	• [UX] Perceived quality of answers on frequently asked questions measured with QLQ-INFO25 [78] questionnaire	Two conditions: • 71 patients received information from a medical committee • 71 patients received information from VIK chatbot	• **Quality** of information scores from chatbot were found to be non-inferior to the scores of the group of physicians	• [M] No demographic features of the participants • [P] Due to recruitment method, results of participants may not be generalized to average population
Bickmore et al. (2013) [69]	Behavioural health modifications are important to manage and prevent chronic diseases such as diabetes, cancers, and obesity.	• [B] Physical activity (PA) • [B] Servings of fruits and vegetables (intake)	4-arm randomized trial of a two-month daily contact intervention: • ACT: Exercise group (n=31) • DIET: fruit and vegetable group (n=30) • ACT + DIET: both (n=30) • CONTROL: control group (n=31)	• ACT on **PA**: an insignificant increase • DIET on **intake**: significant increase • ACT+DIET on **PA**: no support • ACT+DIET on **intake**: an insignificant increase	• [P] Small convenience sample and relatively short duration reduces the generalizability of the results • [C] CONTROL group does not represent a true non-intervention control, as providing pedometers alone has been shown to increase physical activity • [M] Pre-intervention baseline behaviour was not measured
Bickmore et al. (2016) [59]	The majority of US adults look for health information online, however keyboard-based search engines can present a significant barrier for many disadvantaged adults.	• [UX] User satisfaction • [UX] Trust • [UX] Pleasure • [UX] Frustration • [U] Task completion • [U] Elapsed time	Two conditions: • 43 participants used the conversational search engine interface • 46 participants used the conventional keyboard- and facet-based interface Two tasks per participant: • Find a clinical trial for yourself • Find a trial that meets specified criteria	Use of agent resulted in: • More **satisfaction** • Increased **trust** • Increased **pleasure** • Decreased **Frustration** **Task completion**: None of the low-literacy participants were able to find a correct clinical trial using the conventional interface, compared to 36% using the conversational interface Conversational interface takes more **time**.	• [P] Small number of participants • [P] 21% of users were previously involved in clinical trials and therefore are not representative of general population • [M] REALM [79] was used to measure health literacy, which could have been more refined and should have included computer literacy

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Chaix et al. (2019) [63]	Improved treatment adherence could have a big impact in terms of global health. In combination, the number of cancer patients are increasing exponentially.	• [B] Medication Adherence Rate • [UX] User satisfaction • [U] Retention Rate • [U] Number of conversations	• A prospective study was conducted by analyzing the users' data, their usage time, their interest in the various themes proposed, and their level of interactivity. • One question on a weekly basis • Focus groups	• VIK (the chatbot) increased patients' **compliance** with their treatment • High **user satisfaction** (93.95%) • High **retention** • Some emotional **attachment**	None mentioned
Crutzen et al. (2011) [55]	Internet-delivered health promotion initiatives may be particular suitable to reach the present generation of adolescents.	• [UC] Retrieval of user characteristics • [U] Frequency of conversations • [U] Duration • [U] Number of queries • [UX] Perceived anonymity • [SQ] System speed • [SQ] Conciseness • [UX] Perceived quality and quantity of information	• Questionnaire to evaluate the chatbot and compare to information lines and search engines • Interviews as a pilot study	• The chatbot reached **adolescent users** who were high school attendees, varying in level of urbanization, education, and experience with sex, drugs and alcohol • **Usage** peaked at the start, but was stable thereafter (M = 45 minutes) • Chatbot was perceived as **faster** and more **anonymous** than information lines and search engines. • The information provided by the chatbot was perceived to be more **concise** and of **higher quality**.	• [RR] The non-optimal response rates may have resulted in a selective sample of adolescents who filled out the questionnaire • [P] Some participants may never have used information lines, which was used to compare to the chatbot.

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Davis et al. (2020) [51]	There is a global rise in disease burden from cardiovascular disease and diabetes—chronic conditions which are predominantly modifiable by healthy diet and lifestyle choices.	• [B] Mediterranean Diet Adherence • [B] Physical Activity Adherence • [SQ] Technical Performance • [UX] User Engagement • [UX] User feedback	Single-arm 12 week study with three phases where each session, a self-reported study outcome measures and feedback questionnaire was filled out.	• Mean **dietary adherence** was 91% • The **step goal** was achieved 59% of the time • **Correctly answered** scripted questions 97% of the time, however spontaneous exchanges only 21% of the time • Participants **enjoyed** the program and found Paola **useful**, but not for spontaneous exchanges	• [P] Small sample size • [M] No accurate quantified user engagement • [P] More females, who were more likely to engage with the chatbot
Goldenthal et al. (2019) [64]	Patients experience common symptoms and/or complications after undergoing ureteroscopy, which is a common procedure used to treat kidney stones.	• [U] Chatbot Usage • [SQ] Ability to deliver information • [O] Reasons not to activate chatbot	Semi-structured interviews were the following was assessed: • Overall impressions of their recovery • Activation of chatbot • If not activated, a reason was asked. • If activated, they asked if the chatbot met their needs	• Seven of the twenty participants **activated** the chatbot • Patients who did activate the chatbot found it a **convenient** method to find information about their symptoms • Patients experienced some **usability** issues	• [P] Few participants • [RR] Low activation rate • [M] Single-institutional study, cannot be generalized to other practices.
Heald et al. (2020) [60]	Hereditary colorectal cancer (HCRC) syndromes account for 10% of colorectal cancers but remain under-diagnosed.	• [U] Genetic Counsellor (GC) Time • [U] Genetic counselling assistant (GCA) Time • [U] Progression through the chat • [SQ] Identify HCRC risk factors • [O] Genetic test results	• Patients used the Colon Cancer Risk Assessment Tool (CCRAT) [80] to screen for HCRC syndromes • Those with one or more positive responses to a CCRAT question received chatbot-deployed genetic education and the option to receive genetic testing • For those consenting, blood was drawn on the day of the colonoscopy to determine genetic risk	• **GC Time**: average is 14.3 (SD 7.3) minutes • **GCA time**: average is 19.2 (SD 9.8) minutes • 96.2% **completed** the chat with the chatbot • **Test results**: 12 participants were found to have a germ-line pathogenic variant	• [RR] Although a high completion rate, the initiation rate was quite low • [P] Half of the participants reported to be typically eager to try new technology • [M] Reported history was not checked with medical records for all subjects • [O] The study did not identify any patients with Lynch syndrome • [O] Chatbot did not replace the care of a healthcare provider

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Holmes et al. (2019) [56]	Self-reporting, personalized feedback and motivation have been shown to be beneficial for weight loss maintenance in the short term.	Evaluation of chatbot: • [UX] SUS scores [81] • [UX] UEQ metrics [82] • [UX] CUQ score (developed and validated in this paper) Effectiveness of evaluation methods: • [O] Correlation between different evaluation measures • [O] Optimum number of users to identify usability issues • [O] Number of task repetitions to reach optimum task performance • [UX] Found usability issues using this method • [U] Task completion time • [UC] Participant demographics	• A pre- and post-task Single Ease question was asked to the participant ("how easy do you think it will be to complete this task?") • The completion of the task is recorded with audio and video • After all tasks, each participant completed Post-Test usability surveys including System Usability Scale (SUS) survey, User Experience Questionnaire (UEQ) and a Chatbot Usability	• WeightMentor is highly usable (**SUS**: M = 84.83 $\pm$ 12.03, **UEQ**: each scale above +0.8, **CUQ**: M = 76.20 $\pm$ 11.46) Effectiveness of evaluation methods: • **Correlation** between the three main questionnaires was generally strong and was highest between the CUQ and UEQ • The **optimum number** of users is 21 to identify usability issues • In general, **task completion** times did improve with each repetition of a task	None mentioned
Hussain et al. (2019) [66]	Women with unintended pregnancies often receive inadequate prenatal care along with poor health outcomes for their children.	UTAUT variables: • [UX] Performance Expectancy (PE) • [UC] Attitude towards Tech (AT) • [UX] Social Influence (SI) • [SQ] Facilitating Conditions (FC) • [UX] Self-Efficacy (SE) • [UX] Anxiety (AX) • [B] Behavioural Intention (BI) • [UX] Effort Expectancy (EE)	UTAUT Survey [83] was filled out by the participant using Qualtrics. Survey questions were based on the outcome variables.	• EE increased **PE** • PE increased **AT** • AT increased **BI** • Positive attitude towards using the chatbot was determined by the **effort** required, which in turn determined the value of information gained.	None mentioned

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Koman et al. (2020) [75]	Seeking medical information can be an issue for physicians.	[UX] Physicians perceptions such as: • Relevance of chatbot • Appreciation • Needs for such a chatbot • Expectations	Individual, in-depth, semi-structured interviews to explore physicians' opinions and perceptions of a chatbot.	**Perceptions** (+): • Relevant concept • Ergonomic tool • Diagnosis helping • Time-saving **Perceptions** (-): • Natural language comprehension issues • Medical data issue • Threatens health care professionals	• [P] Small sample size • [P] Majority of the participant were male • [P] Recruited participants had already participated in other digital epidemiological studies carried out by the same research team • [O] Due to confidentiality, the drug associated with the chatbot could not be cited
Maeda et al. (2020) [67]	Fertility awareness is of growing interest and importance in the world, for example because many people postpone parenthood because of career, education, relationship and financial issues.	• [SQ] Improvement of fertility knowledge (FK) • [UC] Preconception health behaviour (smoking, supplements, vaccins etc.) • [UC] Preconception health status (weight, height and information on period) • [B] Change in health-related intentions • [UX] Anxiety	Three-armed RCT • Intervention: conversation with chatbot • Control group 1 (CG1): booklet with fertility information • Control group 2 (CG2): booklet with irrelevant information	• Intervention group saw a significant improvement of **FK**, although smaller than in CG1 • Effects on **behaviour modification** were equivalent between the groups • The level of **fertility knowledge** improved considerably immediately after exposure to fertility information in the intervention group and CG1 • The online education improved **intention to change health behaviour** • Intervention group showed significantly lower **Anxiety**	• [M] The use of social research panels could have caused selection bias towards participants with higher education • [P] Prevalence of participants who reported taking folic acid and oral contraceptives was higher than reported in national data • [M] Participants could in theory access the educational content during the post-test survey • [M] The study measured mostly self-reported measurements • [M] The developed chatbot was not for men, thus not for all people in the study • [M] Study was conducted in Japan and therefore cannot be generalized to other countries and culture

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Piao et al. (2020) [72]	It is crucial to find ways to fit regular exercise into the daily lives of office workers. Non-exercise activity thermo-genesis has been introduced as an effective form of daily exercise.	• [UX] Perceived usefulness • [UX] Perceived ease-of-use • [UX] Usage intention	Participants tested Healthy Lifestyle Coaching Chatbot (HLCC) for three weeks where usability issues were identified, the technology acceptance model (TAM) [84] was used to test the usability.	• The HLCC improved the **effectiveness** of habitually performing simple, repetitive exercises as part of daily life • All users found the program **useful** (3.98 ± 0.77) and **easy to use** (3.79 ± 0.84). The **usage intent** was rated high (3.86 ± 0.89) as well. All outcome variables were rated on a five-point Likert scale.	None mentioned (however many more females participants and no evidence for the claimed improved effectiveness due to HLCC)
Piau et al. (2019) [61]	Almost two thirds of patients diagnosed with cancer are age 65 years or older. In order to follow up on older patients with cancer receiving chemotherapy at home, remote phone monitoring is currently conducted by skilled oncology nurses.	• [U] Number of completed questionnaires per patient per week of follow-up • [U] Refusal and drop-out rates • [U] Filling rate for a set of questions • [U] Average time to answer a set of questions (completion time) • [U] Person filling out the information • [U] Use of free-text communication • [U] Questionnaire compliance rate • [UX] Attitude	During the 7 weeks of continuous follow up, participants were asked to fill out the questionnaires, provided through a chatbot, about their symptoms and the outcome variables.	• Participants answered on average 6 **questionnaires** in the 7 week follow-up period (52 in total) • 100% **completion** rate (no drop-out) • Completion **time** was 3min 27s on average • Respondents were either the **patient** (44%) or the family **caregiver** (56%) • **Free-text communication** was used in 58% of the conversations • 86% **compliance** rate • **Satisfactory** rates ranged from 8-10 (on a 10-point scale)	• [P] Small number of participants

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Ponathil et al. (2018) [57]	Early diagnosis is vitally important in the treatment of diseases, notably those with genetic links like cancer.	• [UC] Demographics • [U] GOMS ideal completion time • [UX] Perceived usefulness (TAM) • [UX] Perceived ease-of-use (TAM) • [UX] Workload (NASA-TLX) • [UX] Mental demand (NASA-TLX) • [O] Preferred application • [U] Clicks • [U] Completion time	Within-subject experimental design where a conversational interface (CI) is compared to traditional interface (TI) with: • Pre-test demographic questionnaire • NASA-TLX workload instrument [85] • Technology Acceptance Model (TAM) [84] • Retrospective think aloud session • Post-test questionnaire ranking the applications • GOMS ideal completion time [86]	• Completion **time** (calculated): CI = 640s, TI = 473s • CI had a higher perceived **usefulness** score • CI had a higher perceived **ease of use** score • TI resulted in a higher **workload** • TI resulted in a higher **mental demand** • 16 of the 20 participants (80%) **preferred** CI over TI • CI required more **clicks** • CI completion **time** was higher	• [P] Participants were all well-educated • [M] Participants were asked to fill out fictional data, this may have lead to increased task completion times and errors
Schmidlen et al. (2019) [74]	Interest in, access to, and demand for genomic testing continues to increase and the demand for genetic counselors to help interpret and integrate genomic information increases in parallel.	• [UX] Acceptability • [UX] Usability • [UX] Understanding of chatbot	Focus groups on three chatbots: consent-chatbot, follow-up chatbot and cascade chatbot in three regions. Six focus groups (two in each region) to gather data on usability, acceptability, functionality, and understanding of chatbots.	• Participants overwhelmingly **supported the use of chatbots** for coordination and sharing genetic risk information with their relatives. • Participants found the consent chatbot to be **more informative** than their previous in-person, paper-based consent experience • Several participants expressed willingness to **reuse** the tool should they ever receive such a result from the study	• [P] Although the included 62 participants, only three focus groups reviewed each chatbot, thus complete saturation may not have been reached. • [P] Most participants were Caucasian, non-Hispanic, college-educated, older than 50, and generally interested in genomics. • [M] Participants had not received genetic testing results, although potential patients would have when they will use the chatbots. • [M] Opinions can't be ranked or quantified. • [M] Because a compensation was offered, there might be a incentive-based bias.

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Siangchin & Samanchuen (2019) [68]	Completely and correctly ICD-10 coding requires basic medical knowledge, pathology, anatomy, and so on to understand the nature of the diseases, procedures, and treatments from the diagnosis report and therefore takes a lot of time.	• [SQ] Performance • [SQ] Humanity • [SQ] Affect • [SQ] Accessibility	The chatbot was compared with the conventional ICD-10 application by using Analytic Hierarchy Process (AHP): • Performance • Humanity • Affect • Accessibility Participants with experience with and responsibility for ICD-10 coding were distinguished.	The chatbot has clearly displayed its strengths to provide ICD-10 codes especially for **Performance** and **Accessibility**. The difference for **Humanity** and **Affect** is small.	None mentioned
Stein & Brooks (2017) [62]	Type 2 diabetes is the most expensive chronic disease in the United States. Two-thirds of US adults have prediabetes or are overweight and at risk for type 2 diabetes.	• [B] Weight loss • [B] Changes in meal quality • [U] Average duration of app use • [U] Number of sessions • [UX] Satisfaction score (SS) • [UX] Net promotor score (NPS) • [UX] Disappointment score (DS) • [UX] Health outcome score (HOS)	Longitudinal observational study among overweight and obese participants who used the Lark Weight Loss Health Coach AI (HCAI), measured: • Weight loss, meal quality, physical activity, and sleep data (through user input and automatic detection by the user's mobile phone • User engagement: assessed through duration and amount of app use • Usability and Acceptability: measured through a four question trust survey	• **Weight loss** was 2.38% of baseline weight • Percentage of **healthy meals** increased by 31%. • The average **duration** of app use was 15 (SD 1.0) weeks • Users averaged 103 **sessions** each • The average **SS**, **NPS**, **DS**, and **HOS** scores were 87, 47, 68, and 60, respectively	• [C] No control group • [M] Scarcity of collected demographic information • [M] Observational study, thus unable to determine causality • [M] Physical activity was tracked with mobile sensors, therefore inaccuracies could have occurred • [O] Incomplete or incorrect classification of foods and therefore meals.

Table 3.4 Overview of the articles included in the review (continued)

Source	Background	Outcome variables	Method	Outcomes	Limitations
Thongyoo et al. (2020) [76]	When patients with diabetes neglect good nutrition this can cause many health problems.	[UX] Content satisfaction: • Relevant information for decision making • Accuracy and trustworthiness • Sufficient information needed [UX] Design satisfaction: • Aesthetics of chatbot • Pattern easy to use • Performance speed • Clear content [UX] Implementation satisfaction • Usefulness • Information resources meet users needs • Further usefulness [SQ] Chatbot performance compared to old system	System satisfaction assessments were conducted using online questionnaires for users both with and without diabetes. The assessment topics were divided into three main areas: • Content • Design • Usage Each scale is from 1 (least satisfied) to 5 (most satisfied). Maximum score is 120 (24 participants times the maximum score).	• **Content**: average of satisfaction scores in all 3 topics is 91.66 points • **Design**: average of the satisfaction results is 99.25 points • **Implementation**: average of user satisfaction scores is 109 points • **Performance**: the chatbot is faster and more stable than the old system	None mentioned

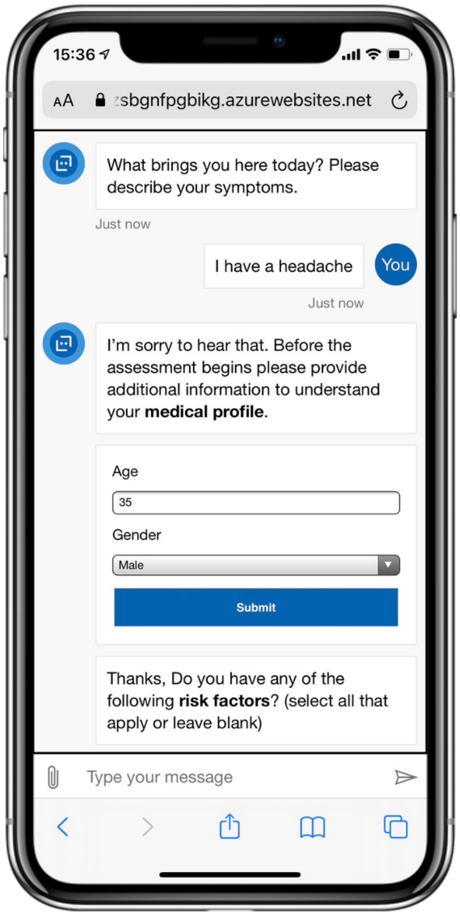
Outcome variables: B = behaviour change, SQ = system quality, UX = usability and user experience, U = usage, UC = user characteristics, O = other

Limitations: C = no baseline or control group, M = related to methods or used measures, P = related to participants, RR = low response rate or low activation rate, O = other

3.2 Chatbot implementation

As explained in the Methodology, the chatbot was developed with predefined user stories in mind. These user stories were implemented into one conversation flow. The goal of this conversation flow was to enable a chatbot to perform diagnosis and triage. Five user stories were developed to achieve this goal (see Table 2.6) and implemented in the conversation flow.

Figure 3.3: The webchat with the chatbot on a phone



Implementation of the User Stories

The first user story prescribed the need to show the **possible causes** for the symptoms of the user (U1). U1 was implemented by showing the user a list of possible conditions. Usually, one to five conditions were presented to the user. However, the maximum number of conditions was capped at ten. The second user story explained that the user should be informed on the **likelihood** for certain conditions (U2). This user story was implemented by displaying the cause frequency next to each possible cause in the triage results. The frequency was rated as common, less common, rare, or very rare. The frequencies were retrieved from Infermedica, which has a frequency level for every condition. The third and fourth user stories showed the need for providing a **triage level** (U3, U4), which was assigned to the user at the end of the conversation. By inspecting the possible conditions, one of the following triage levels was assigned: self-care at home, visit the doctor for a routine check-up, visit the doctor today, visit the emergency department or call an ambulance. In addition to providing a triage level, information was provided to help the user take the next steps in their health journey. Then, the last user story suggested the need for taking the **medical profile** into account during triage. Therefore, the chatbot asked the user to fill out their age and gender and medical risk factors. The risk factors the chatbot queried were pregnancy, after menopause, recent physical injury, diabetes, hypertension, high cholesterol, heart disease, and smoking. These were already included in the template used to build the chatbot (see Chapter 2).

Aside from implementing the user stories, another goal was to address the **deficiencies** presented by You and Gui [27]. The first common deficiency with chatbot-based symptom checkers (CSC) is that **health history** is rarely considered when deciding upon possible conditions. U5 was designed to address this deficiency and was successfully implemented in the chatbot. The second deficiency described that CSCs often have **strict input requirements**. The chatbot had some input requirements, for example, when the answer options were either yes or no. However, there was a wide range of input freedom when, for example, stating the initial complaint or providing follow-up symptoms. The follow-up questions were designed to be as straightforward as possible to combat the third deficiency. Also, often **clear answer options** were provided to guide the user in the right direction. Lastly, the fourth deficiency often found in CSC's is an **incomplete set of health conditions**. The chatbot in this thesis had a wide range of health conditions, as it leveraged the Infermedica database with thousands of symptoms and conditions [45].

Conversation Flow

The conversation started with **onboarding**, where the chatbot explained to the user how it can help them find the possible causes for their symptoms and that it can advise them on the next steps. Then, it asked for a participant number to match the conversation log to the survey in the evaluation. After the user's response, the chatbot asked for their initial complaint to kick off the triage flow.

The next phase in the conversation revolved around gathering **background information** of the participant. First, it asked for age and gender. Then, the chatbot presented a list of possible medical risk factors and asked the participant to select those that apply. The list was dependent on the inputted age and gender because, for example, pregnancy and after menopause were not shown as options to male participants (see Appendix D). The age, gender and selected risk factors were saved as variables and passed to the triage engine.

During the **triage** phase, the chatbot suggested symptoms related to the initial complaint and red flag symptoms. These are provided by leveraging the Infermedica database. The red flag symptoms were symptoms related to the initial complaint that may indicate emergencies. The chatbot constantly echoed the symptoms to verify if the inputted symptoms were registered correctly. By suggesting these related symptoms, the chatbot tried to collect at least three symptoms if possible. After the symptoms had been gathered, the chatbot would ask follow-up questions on the symptoms to rule out more conditions. The chatbot would only skip the follow-up questions and immediately advice the user to seek emergency care if the user selected multiple red flag symptoms. Then, the most likely conditions were determined and presented with the frequency that the condition occurs in the general population. Based on the conditions and symptoms, a triage level was presented. Also, a summary of the inputted symptoms was displayed. This is shown in Figure 3.4.

Then, information related specifically to the **triage level** was presented to the user (see Appendix E). The chatbot conversation ended with thanking the user for their participation and reminding them to fill out the questionnaire for the evaluation. A screenshot of the triage flow is presented in Figure 3.6 and annotated with the different phases. The triage flow contains different types of statements, which are further explained in Figure 3.5.

Figure 3.4: The results from the triage as presented to the user

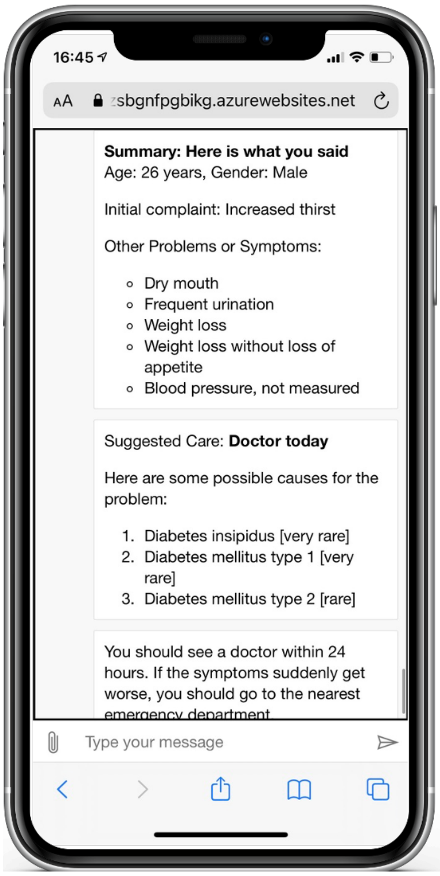
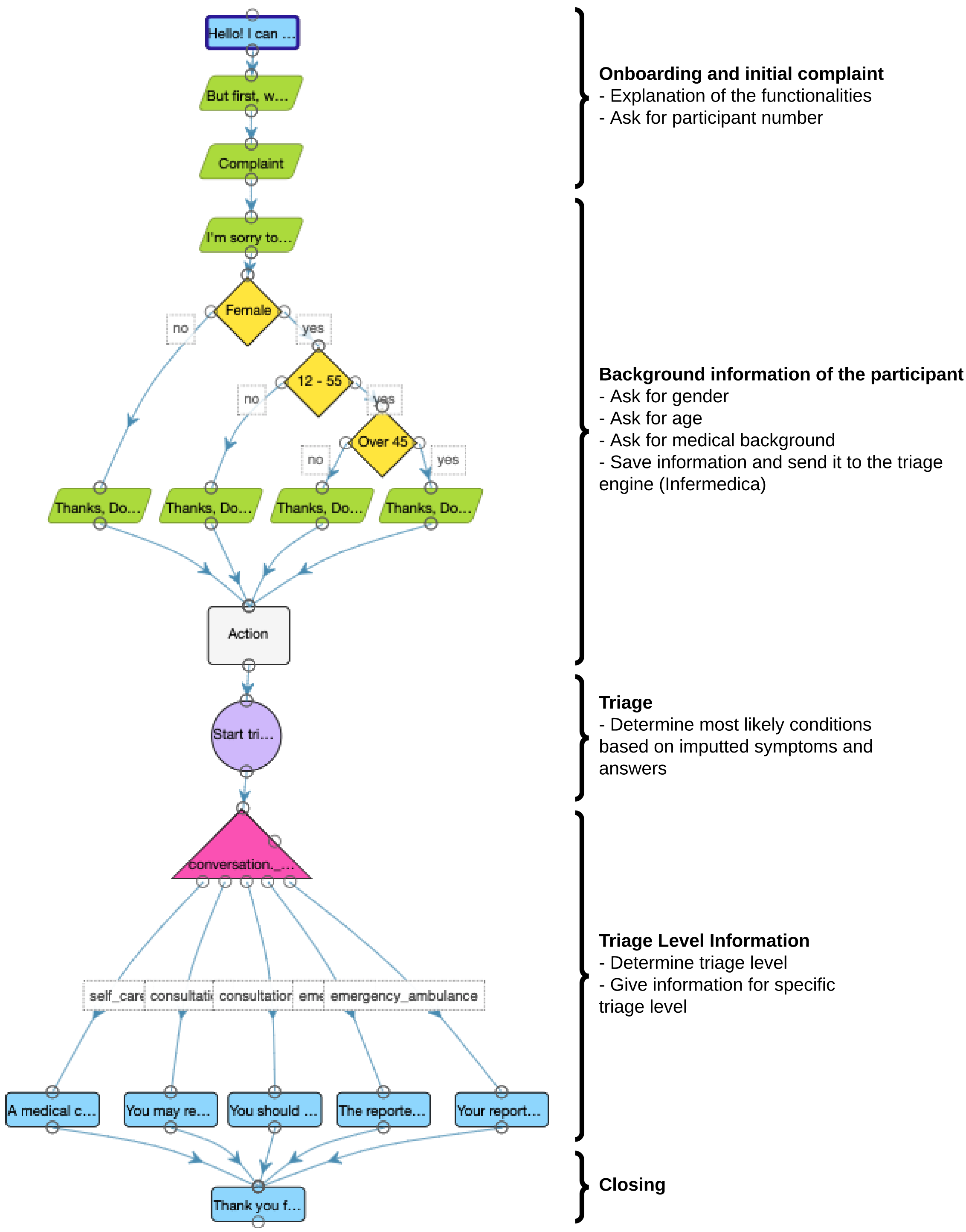


Figure 3.5: The different types of statements in the conversation flow

Statement	Statements are used to **display** and **send** messages to the user without requiring any feedback or interaction from the user.
Prompt	Prompts are statements where the chatbot **expects an answer** from the user. The response can be saved in a variable.
Branch	A branch is a **Boolean prompt** combined with a **decision** step. Depending on the response from the chatbot, a path is chosen.
Action	An action is a snippet of **JavaScript code** running in the context of the conversation session as part of a conversation dialog step.
triage	This dialog step **triggers a sub-scenario**, in this case triage. It can receive input parameters and return output to the calling scenario.
Switch	The Switch element defines a **multi-way split** in the flow of the scenario. Instead of using a Branch multiple times, this can be used just once.

Figure 3.6: Conversation flow of the chatbot with annotations



3.3 Chatbot Evaluation

The following section describes the evaluation results from Thursday 6 May (1 PM) until Tuesday 24 May (9 AM) 2021. In this period, **44 participants** completed the survey. The chatbot was tested on user characteristics, user experience, system quality, chatbot usage, and health behavior.

User Characteristics

The questionnaire started with questions on the participants' demographics to get an indication of user's demographics. For this purpose, gender, age, and highest education were measured.

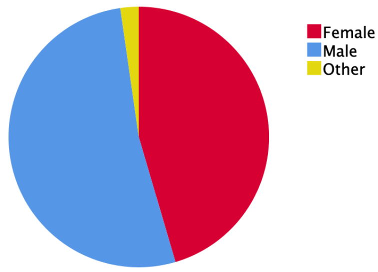


Figure 3.7: Distribution of Gender

Gender. Of the 44 participants, 20 were female (45.5%), 23 were male (52.3%), one participant classified their gender as "other". The gender of the participants was equally distributed. This is clearly shown in Figure 3.7.

The **Age** of the participants was on average 30.4 years (SD = 11.4). Most participants (72.7%) were younger than 30 years old. The youngest participant was 21 years old, and the oldest participant was 59 years old. The distribution of age is shown in Figure 3.8.

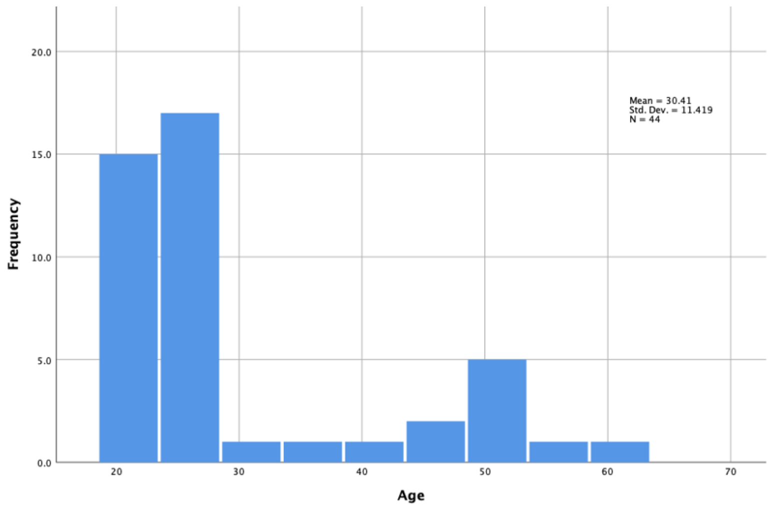


Figure 3.8: Distribution of Age

Participants were also asked to provide their **Education Level** by stating their highest completed education level. Most participants completed a Bachelor's degree (28 participants, 63.6%) or a Master's degree (12 participants, 27.3%). One person completed a degree at a trade school, two completed high school, and one participant had some high school as their highest education. The education levels of the participants are visualized in Figure 3.9.

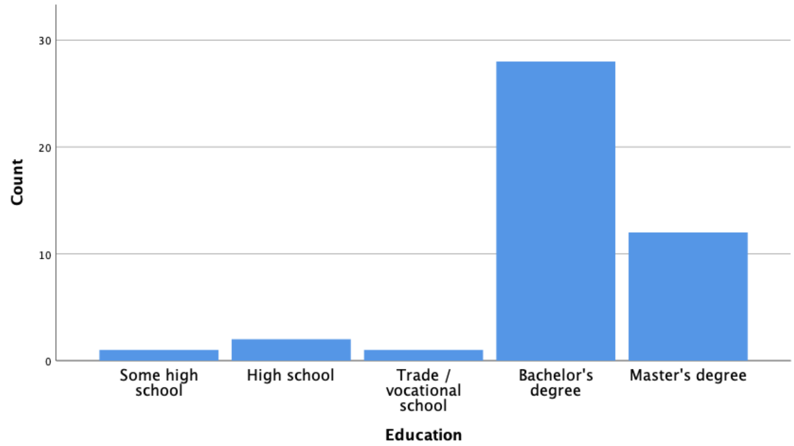


Figure 3.9: Distribution of Education Levels

User Experience

CUQ Scores

The Chatbot Usability Questionnaire (CUQ) scores ranged from 17.2 to 95.3 on a scale from 0 to 100. The **average CUQ score was 68.9** (Median = 70.3, SD = 16.6). Four participants (9%) rated the chatbot with a score of 85 or higher, which is an excellent score. Twelve participants (27%) rated the chatbot with very high scores, which are scores greater than 75. 15 participants (34%) rated the chatbot with high scores, which are scores higher than 65 and lower than 75. Then, 7 participants (16%) rated the chatbot with OK scores, which are scores between 50 and 65. 4 participants (9%) gave the chatbot a poor score, which are scores lower than 50. Lastly, two participants (4,5%) rated the chatbot with awful scores, which are scores lower than 40. An overview of the scores per participant is shown in Table 3.5. The distribution of the scores is displayed in Figure 3.10.

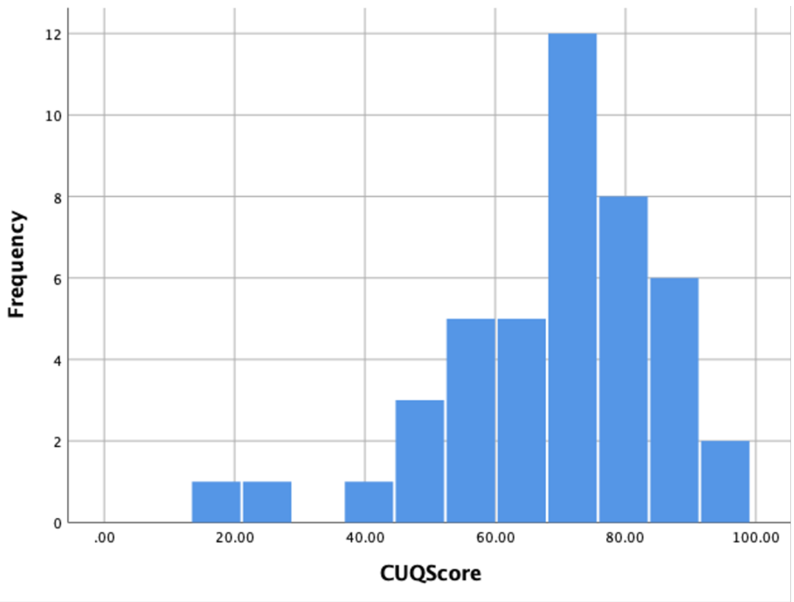


Figure 3.10: Distribution of CUQ scores

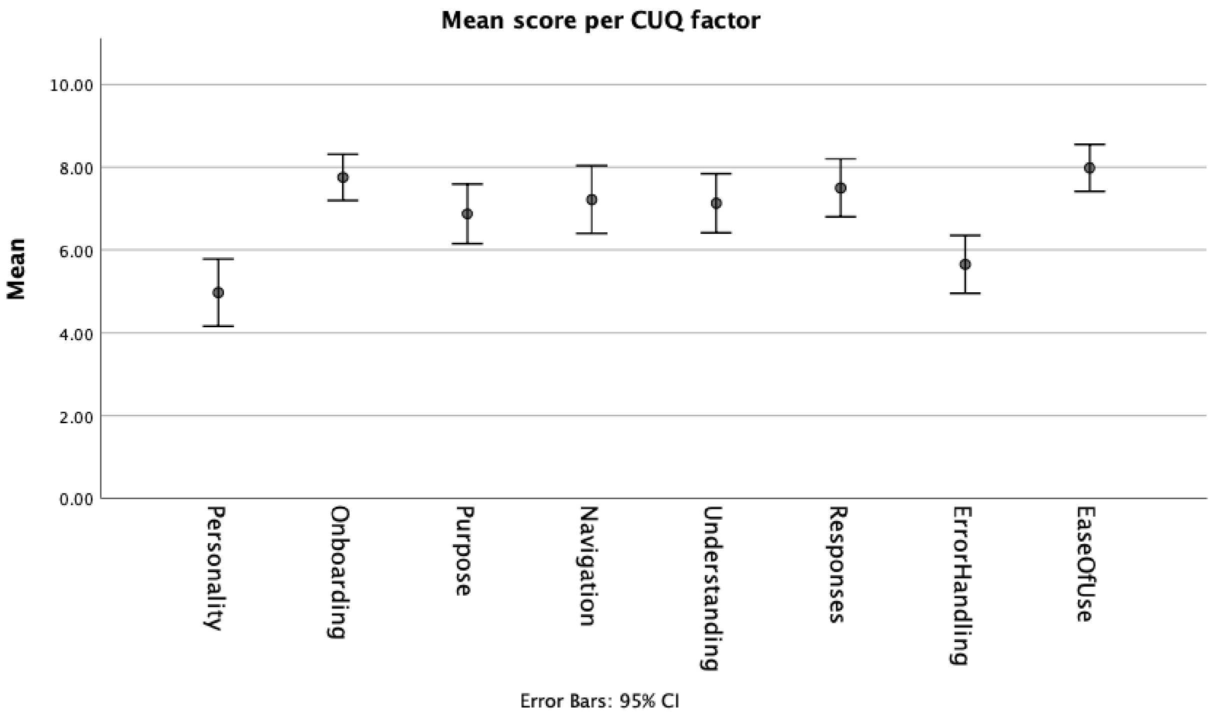
Table 3.5: CUQ Score per participant (see Table 2.9 for the CUQ statements)

								Chatbot being tested:		*Diagnosis and Triage chatbot*							
Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**CUQ Score**
Participant #	*Question scores*																
2844	4	2	5	1	5	1	5	1	5	1	5	1	4	1	5	1	**95.3**
3110	5	1	5	1	5	1	5	1	5	1	5	1	3	3	5	1	**93.8**
7945	5	2	4	1	5	1	5	2	5	1	5	1	5	4	5	1	**90.6**
1061	5	1	4	1	5	2	5	1	5	1	5	2	3	3	5	1	**89.1**
2260	4	3	5	1	5	1	5	1	4	1	4	2	3	3	5	1	**84.4**
6988	4	2	5	1	4	2	5	2	5	1	4	2	3	2	5	1	**84.4**
1100	4	3	5	1	5	1	5	1	4	2	5	1	3	3	5	2	**84.4**
4268	4	4	4	3	5	1	5	1	4	1	4	2	5	1	5	1	**84.4**
1086	4	3	4	2	4	1	5	2	5	2	5	1	4	2	5	2	**82.8**
1089	2	1	4	2	5	2	5	2	5	1	5	2	4	2	4	1	**82.8**
1898	4	3	5	1	4	1	5	2	4	1	4	1	3	3	5	1	**82.8**
8922	4	3	4	3	2	3	5	1	5	1	4	1	5	1	5	1	**81.3**
210	2	4	5	1	4	1	5	1	3	1	4	1	3	1	4	1	**79.7**
3330	5	2	4	2	3	2	5	2	4	2	4	1	3	2	4	1	**78.1**
9050	5	1	5	1	2	2	5	1	4	2	4	2	3	3	4	2	**78.1**
2033	2	4	2	2	4	1	5	1	4	2	4	1	4	1	5	1	**76.6**
5322	3	2	4	2	4	2	4	2	4	2	3	2	4	2	4	1	**73.4**
4957	3	3	4	1	4	3	4	2	4	1	5	1	3	3	4	2	**73.4**
1235	3	3	4	1	4	2	5	3	4	1	4	2	3	3	4	2	**71.9**
8224	4	3	5	2	4	2	4	2	4	2	4	2	3	3	4	2	**71.9**
79	4	4	5	1	4	1	5	4	3	4	5	2	3	3	5	2	**70.3**
4128	4	4	5	2	4	3	4	2	4	3	4	2	4	2	4	2	**70.3**
4228	2	3	4	3	5	2	4	3	3	2	5	2	4	2	4	1	**70.3**
4950	4	3	4	1	4	3	5	2	3	4	4	2	2	1	4	1	**70.3**
9607	3	3	3	3	3	1	4	2	4	2	5	1	3	2	4	2	**70.3**
2334	4	2	4	1	4	4	4	2	4	2	4	2	3	3	4	2	**70.3**
1147	3	4	5	2	4	3	4	2	4	2	4	2	3	2	4	2	**68.8**
6096	4	4	5	2	3	3	4	2	4	1	4	2	3	3	4	2	**68.8**
1728	2	4	4	2	3	5	5	2	4	1	4	1	3	1	4	2	**67.2**
811	4	2	4	3	4	4	4	2	4	2	4	2	3	4	5	2	**67.2**
9578	4	4	5	1	5	2	4	1	2	3	4	2	1	5	4	1	**65.6**
7122	4	2	4	3	4	2	4	2	4	2	2	4	3	3	4	2	**64.1**
517	2	4	4	1	4	1	2	4	2	1	4	1	4	2	2	2	**62.5**
4407	2	3	3	2	2	3	4	3	4	2	3	3	4	2	3	3	**56.3**
4728	2	4	5	4	5	1	4	5	2	4	4	2	1	4	5	1	**54.7**
6894	1	5	4	2	3	3	3	4	2	2	4	2	3	2	4	2	**53.1**
2609	2	4	2	1	2	3	4	4	3	2	3	2	3	3	5	3	**53.1**
3000	2	5	4	2	3	2	2	4	4	2	4	2	3	3	2	2	**53.1**
4113	2	4	3	2	3	4	2	5	4	1	4	2	4	2	2	4	**50.0**
2113	1	5	3	4	2	3	4	1	3	3	4	2	3	4	4	2	**50.0**
4120	2	5	4	2	2	3	4	4	2	3	2	2	3	3	4	2	**48.4**
618	3	4	2	3	3	2	2	4	2	4	4	5	2	3	4	2	**42.2**
9975	1	3	4	2	1	4	2	5	1	1	1	5	1	5	2	3	**26.6**
1507	1	5	2	3	2	2	1	4	1	5	1	5	1	5	2	3	**17.2**

The mean score per CUQ factor was calculated and is displayed in Figure 3.11[4] . The chatbot was rated on all factors on a scale from 0 to 10. Here, 0 meant that the chatbot did not perform well at the specific factor at all, and 10 that the chatbot performed perfectly for that factor. The chatbot performed best on the **Ease Of Use** factor with an average score of 7.98 (SD = 1.85). The second highest-rated CUQ factor was **Onboarding** with an average score of 7.76 (SD = 1.84). The third highest-rated factor was **Responses** with an average score of 7.50 (SD = 2.30). The fourth highest-rated factor was **Navigation**, with an average score of 7.21 (SD = 2.69). This factor had the highest standard deviation for all factors. The fifth highest-rated factor was **Understanding** with an average score of 7.13 (SD = 2.36). The sixth highest-rated factor was **Purpose** with an average score of 6.88 (SD = 2.36). **Error Handling** did not receive a very high score with an average score of 5.65 (SD = 2.30). The lowest rated factor was **Personality**, which received an average score of 4.98 (SD = 2.68).

4: The questions from the CUQ per factor are displayed in Table 2.9

Figure 3.11: Score per CUQ factor



Qualitative Assessment

The conversation logs of conversations with six participants who rated the chatbot with a CUQ score lower or equal to 50 were downloaded and qualitatively assessed. These conversation logs are shown in Appendix F. The free-text comments were also retrieved to determine what went wrong in those conversations. The conversation logs for participants who gave the chatbot a score lower than 65 were scanned through; however, no bugs or abnormalities were found. During the qualitative assessment of the logs, three issues were detected by assessing six conversation logs.

One conversation showed problems with the **understanding of the inputted complaint**. One time, the initial complaint was not understood

("I have to laugh during the [whole] day"). However, an error was given later in the conversation, and therefore the participants did not know where it went wrong. The participants were also not provided with an option to re-enter the complaint. The two other times, the input during the assessment of symptoms was not recognized ("I am happy" and "Bad hair"). In these cases, the chatbot did provide the participant with an option to rephrase the input. Another input issue was that the chatbot did not understand "sudden pain shots"; however, it did register "sudden pain in both knees" as knee pain.

One participant ran into an issue where there was **no response on the "start over"** prompt. Start over is suggested to the user when the chatbot gets stuck. However, the user was not notified when the chatbot received the prompt and restarted the conversation.

Lastly, two participants ran into some issues where the input to a question did not result in a **proper follow-up**. One participant received an error after answering "no" when the chatbot asked whether the recognized symptoms were correct. This was probably a programming bug, and was only mentioned by this participant. The other participant answered "yes" to "what else would you like to report?" which was not recognized as a correct input. Therefore the chatbot said, "Sorry, I did not understand. Please try to describe the symptoms using an alternative phrasing." and the participant answered, "Nothing elze". Eventually, the user skipped the question, and the conversation was correctly completed.

Free-text Comments

The free-text comments on the user experience and the chatbot's responses indicated possible improvements to the chatbot. All comments are displayed in Appendix G (user experience) and Appendix H (responses). The comments were separated into positive and negative remarks, and from those ten themes were identified.

4 participants commented on **problems with language recognition**. 2 participants noted that the chatbot missed out on keywords such as "sudden" and "unexpected" and that the chatbot did not pick up on the time factor of the symptoms. Another complaint in this category was that the chatbot was not able to handle spelling mistakes; the participant commented: "*A thing that can be improved is the ability to handle small orthographic mistakes. For example, I wrote panic(k) a(t)tack and the chatbot didn't recognise the symptoms.*". Lastly, one participant commented that the chatbot did not understand everything in the conversation.

3 participants noted the need for **more elaboration on the triage levels**. One participant indicated that some tips would be helpful before visiting the doctor. About this, the participant said: "*It would be nice if the chatbot would give some tips or things to do in the meantime until you go to the doctor, as now it only states the (obvious) cause and that you may need a doctors appointment.*" Another participant stated that the possible causes and triage levels are vague and that it could include a question that asks if you have already been at the doctor. Lastly, one participant suggests adding information about a suicide line in the case of depression-like symptoms (e.g. when the user engages in self-harm).

4 participants commented on **time-related subjects**. However, the opinions of the participants differed on this subject. 2 participants suggested that the triage could be quicker by noting that the process was "*a bit too long*" and that the "*response seemed slow*". Though, two other participants noted the quickness of the chatbot by saying "*It went too fast*" and "*It was a very nice experience. I liked that [it]*[5] *was quite quick and to the point whilst still feeling friendly.*".

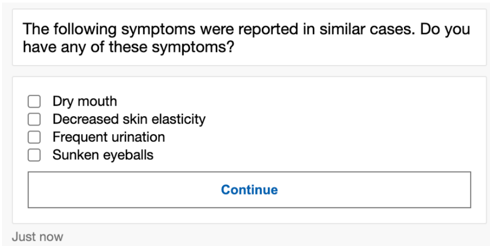
5: Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, as is the case here.

4 participants noted on the **humanness of the chatbot**, where one person stated that they liked that the chatbot was not too human: "*I like using a chatbot that is not too human, as you intuitively adapt your own lines into a more simplistic way so the bot can understand.*". However, 2 other participants reflected negatively on the chatbots humanness, for example, when touching upon mental health issues. A participant stated: "*A chatbot feels [too] [impersonal] when you're experiencing mental issues. The questions about self-harm and potential suicide are quite harsh and should be asked with more care.*" Lastly, one participant suggested giving the chatbot a name to improve the experience.

Then, 2 participants commented that the chatbot used **difficult language**. A participant noted that some questions were too complex, and another participant said, "*Triage is a difficult word, maybe replace it with something simpler? Same for some other medical terms, I'd try to keep it simple*".

Two participants noted that questions with checkboxes (an example is shown in Figure 3.12) had **no "none of these" option**. Their absence confused the participants; for example, one participant stated, "*When none of the options the chatbox provided was relevant, I searched for a checkbox option with the text 'none of these'. When this option wasn't there, I was a bit confused by whether I should just click 'continue', without checking one (or multiple) of the boxes.*".

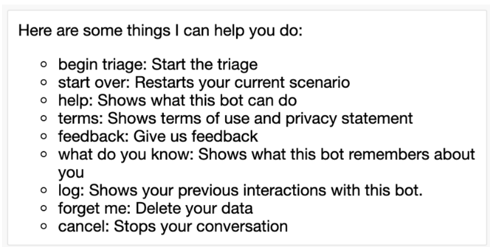
Figure 3.12: A question with checkboxes as answer options



As identified in the previous section, the chatbot gave **no response to "start over"**. Two participants commented on this problem. A participant commented: "*Bot asks for risk factors, then asks to confirm symptoms. I answered 'No', because the risk factors are not the symptoms I (imaginatively) have or call for and then the [chatbot] got stuck. I [type] 'start over' and nothing happens*".

One participant noted that the commands which are shown after the **"help" prompt should be clickable** (see Figure 3.13). The participant said: "*It was difficult to memorize the available commands that show up with the help command. Then I also had to type them in myself instead of clicking on one that I want. This took time and I wanted to make sure to have no spelling mistake (because I thought the bot would not recognize my command then)*".

Figure 3.13: The list of options after the "help" prompt



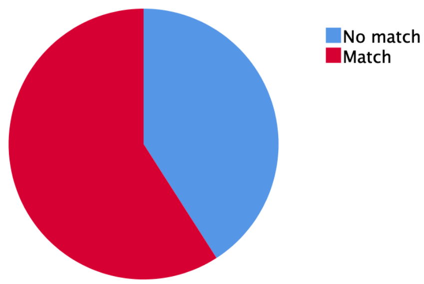
Then, five participants experienced some **bugs** during the conversation with the chatbot. Two participants stated that the chatbot showed multiple messages at once. For example, one participant explained: "*The option answers came too quickly. I got multiple 'select all that apply/select an option' at the same time so it was confusing to answer them; I could not keep track of the dialogue.*". Then, one participant noted that sometimes the chatbot showed that it was typing when that was not the case. Also, in one of the conversations, the chatbot added random symptoms that the participant did not select. Lastly, one participant stated that some information from the chatbot was not displayed in its entirety.

There were also some comments related to **the chatbot in general**. For example, one participant said that they found that using the chatbot for this purpose is scary "*It still feels a bit scary that it says what's wrong with my health based on some simple textual input that I provide.*". Another person stated that a chatbot for this purpose was not necessary: "*It was basically just a checklist of symptoms, [it] doesn't really need a chatbot*". Lastly, one person wanted more reassurance related to privacy.

Health Behavior

The majority of the participants (26 participants, 59.1%) stated that their **next action would be the same** as what the chatbot suggested the user to do. For example, when the chatbot suggested visiting the doctor for a routine check-up, the user would follow the advice and select that triage level in the questionnaire. Eighteen participants (40.9%) chose **another triage level than recommended by the chatbot**. For example, when the chatbot would recommend a doctor routine check-up as the triage level, the user chose one of the other four triage levels as the next action (self-care, doctor today, emergency department or ambulance). An overview of match and no match is shown in Figure 3.14.

Figure 3.14: Distribution of match and no match of advice and next action



The participants were asked to what degree the advice from the chatbot influenced them. The answers differed a lot. Ten participants (22.7%) stated that they were **not influenced at all** by the chatbot's advice. Another ten participants stated that they were only influenced **a little** by the advice. 12 participants (27.3%) stated that they were influenced **a moderate amount** by the advice. 11 participants (25.0%) stated that the advice influenced the participant's choice **a lot**. Lastly, only one participant (2.3%) stated that the advice from the chatbot influenced the participant **a great deal**.

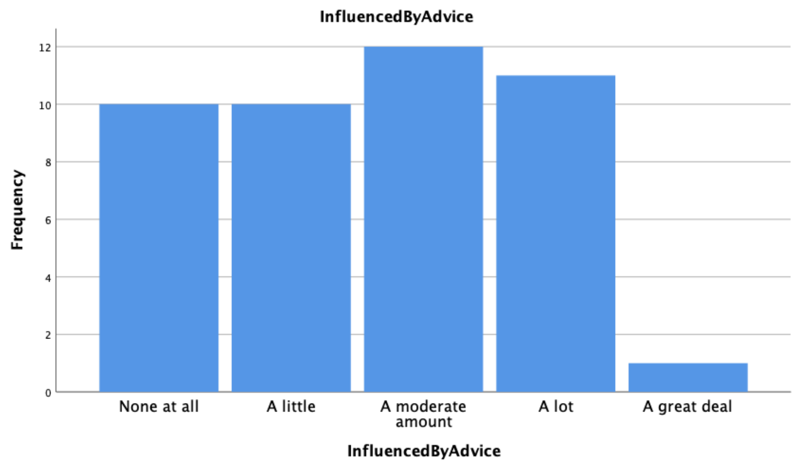


Figure 3.15: Degree of influence of advice on the next action of the participant

Chatbot Usage

In the evaluation period, a total of 3,579 custom events were registered by the chatbot. The median dialog duration, drop-off rate, and the median number of messages were calculated using the custom events. For these metrics, the median was chosen to mitigate the effect of outliers in the data. The **median duration of a dialog** was 3 minutes and 14 seconds. From the 65 participants that started a conversation, 47 reached the end of the triage dialog. Therefore, the **drop-off rate** was 27.7 percent. Lastly, the **median number of messages** of finished conversation was 47, with an average of 47.9 messages (SD = 11.8). The number of messages ranged from 29 to 83.

Discussion and Conclusion

Discussion 4

Principal Findings

Systematic Review

The systematic review resulted in a structured summary of 19 papers where non-mental healthcare chatbots were evaluated. The studies were mostly carried out in the United States and European countries. The chatbots were implemented in various healthcare contexts, namely promoting a healthy lifestyle, health information, oncology, hereditary health, antenatal health, medication, and chronic healthcare. The chatbots were evaluated using either experimental, observational, qualitative, or quantitative approaches (surveys). Different outcome variables were measured in these papers, namely user experience, chatbot usage, system quality, health behavior, user characteristics, and other. The researchers reported limitations related to no baseline or control group; limitations related to methods or used measures; limitations related to participants; low response rate or low activation rate, and other.

Each of the chatbots either had the aim to **provide information** to the end-user, to **coach the user towards healthier behavior**, or **support the user in a task** related to the medical field. This is in line with the study by Palancia et al. [13], where the perceptions of physicians on healthcare chatbots were investigated. These physicians also noted that chatbots could be beneficial in healthcare to support, motivate, and coach patients as well as to support organizational processes. Most chatbots were implemented as a **preventative** measure, such as the hereditary chatbots, or to function in **specialised healthcare**, such as the chatbots with a focus on oncology. Only a few chatbots had a broad and general view of the user's health, such as the chatbots implemented to promote a healthy lifestyle. No Triage and Diagnosis chatbots were implemented and evaluated in any of the papers included in the review.

Most of the chatbots aimed to promote a **healthy lifestyle** by coaching the user towards a healthy diet and increasing physical activity. Chatbots that aimed to promote a healthy lifestyle seem to be effective in terms of **diet adherence**, as all chatbots that used this as an outcome variable were effective. Although this effect has not been researched for other chatbots before, a systematic review and meta-analysis by Robert et al. [87] of other eHealth interventions indicated a similar effect. They found that eHealth nutritional interventions improved anthropometric (such as height, weight and BMI) and clinical outcomes. For **physical activity**, the effect was less convincing, as only one of two chatbots reports on significant physical activity increase. Muellmann et al. [88] conducted a systematic review on the effect of eHealth interventions to promote physical activity for adults aged 55 years and above. Here, the researchers found that the eHealth interventions did lead to increased levels of physical activity. However, there was no significant increase in the results compared to non-eHealth interventions (such as printed folders). The effect of eHealth interventions on physical activity, therefore, is still under-researched and inconclusive.

The fact that chatbots, and thus eHealth interventions, sometimes perform worse than non-eHealth interventions was also apparent in the study by Maeda et al. [67]. In this study, a chatbot was compared to two control groups: one where participants were provided with a non-eHealth intervention aimed to improve fertility knowledge (CG1), and one where the participants were provided with a booklet with irrelevant information (CG2). It was found that the chatbot significantly improved fertility knowledge compared to CG2; however, this effect was smaller than the effect of CG1.

From the results of the systematic review, there is powerful evidence that healthcare chatbots provide users with a **positive user experience** because all 15 papers that reported on user experience or usability measures had a positive outcome. This overwhelming positive result is in line with a systematic review by [89] on the effect of healthcare chatbots (mostly applied in mental healthcare), where 27 of the 30 papers that reported on the usability of the chatbot showed positive outcomes.

Alongside the positive effects of using a chatbot, the papers included in the review also reported some downsides. Firstly, in two papers from the review [51, 75] **problems with spontaneous utterances** were identified. This is in line with other research on healthcare chatbots, as presented in the systematic review by [90]. The review focused specifically on healthcare chatbots with unconstrained natural language input capabilities. They found that the most frequent user experience issues were related to spoken language understanding or dialogue management problems. Another problem with the use of healthcare chatbots is that the use of a chatbot results in **longer task completion time** compared to traditional methods [57, 59]. An increase in dialog duration was also noted by Hill et al. [91] in a study where the researchers compared human-human conversations to human-chatbot conversations.

Chatbot Evaluation

In the chatbot's evaluation, variables related to user characteristics, user experience, health behavior and chatbot usage were measured in a questionnaire. 44 participants contributed to the evaluation. Compared to the median number of participants in surveys in the systematic review (26 participants) [51, 55, 60, 61, 72, 76], this is a relatively large sample.

When looking at the **user characteristics**, it is clear that the participants in the sample do not represent the general Dutch population. Although gender is almost equally distributed (45.5% was female), the data from education levels and age were heavily skewed. The average age in the participant sample was 30 years, and 72.7% of the participants were younger than 30 years. These numbers differ from the age of the Dutch society, collected by the CBS [92], where the average age is 41.8 years, and only 16.6% of the population is older than 18 and younger than 30 years old[1] . In the sample, 63.6% completed a Bachelor's degree and 27.3% completed a Master's degree. These percentages are also much

1: Older than 18, because only participants older than 18 years old were included in the study.

higher compared to the Dutch population [93], where 19% completed a Bachelor's degree, and 11% completed a Master's degree.

The **user experience** was measured using the **CUQ**. The chatbot scored on average a 68.9 on a scale from 0 to 100, which indicates that the chatbot's user experience is good. However, as the CUQ has no benchmark (see the limitations section), it is currently impossible to define an objective rating for this. The score per factor of the chatbot was calculated. The chatbot scored high on ease of use, onboarding, responses, navigation, understanding, purpose, and responses. However, the chatbot scored lower on error handling and personality. This indicates that the user experience can be improved by altering the chatbot on those factors. A study by [94] found that personality has a significant positive effect on the user experience of the chatbot interface. Aside from the low score on personality, four participants commented on the humanness of the chatbot, proving the urgency to improve the chatbot on this factor. The chatbot also scored lower on the error handling factor, however this may be caused by the fact that not all participants faced errors to be handled. Therefore, a more mid-point of scale can be normal. However, it is important to improve the error handling abilities of the chatbot were necessary, as recovery from errors is included as one of Nielsen's heuristics [95]. These heuristics are principals towards usable interface designs. The low score on error handling can be improved by giving the user better instructions when the chatbot runs into an error.

The **free-text comments** on the user experience and the chatbot's responses and the **qualitative assessment** of the conversation logs showed areas for improvements. Participants mostly commented on problems with **language recognition**. A hybrid approach for the chatbot's development was chosen not solely to rely on natural language understanding but to still allow for more input freedom. However, in the parts of the conversation where spontaneous utterances were allowed, such as inputting the initial complaint, problems arose when the chatbot did not recognize the symptoms or when the user made small spelling mistakes. Problems with language recognition are common with chatbots, as was also apparent in the systematic review (as explained above). In addition, Davis et al. [51] found that scripted questions were answered correctly 97% of the time, compared to 21% when spontaneous exchanges were made. Although the number of correctly answered questions was much higher for the chatbot in this thesis, this finding describes one of the most prominent drawbacks of using chatbots with free-text input. This drawback may be mitigated by using better prompts and giving the user an example (e.g. show pain in the knee as a possible input symptom) before entering a symptom. Another problem related to this was that an error was only shown later in the conversation when the initial complaint was not recognised. This problem might have influenced the low score on the error handling factor.

Another problem mentioned by the participants, which also could have influenced the error handling score, was that the chatbot gave **no response on the start over** prompt. This is in contradiction with the first usability heuristic by Nielsen [95], which describes that the system status of the interface should be visible at all times. As the user is not notified that the system restarted the part of the conversation, users assumed that the chatbot stopped working. Therefore, the user could not recover from mistakes, as "start over" is suggested by the system when an error occurs.

Four participants commented on **time-related topics**. Although the opinions differed on this topic, chatbots are known to result in longer task completion time compared to traditional methods [57, 59, 91]. This result was also found in the systematic review. However, when comparing the chatbot to the chatbots in the systematic review, the conversation with the chatbots was quite short (3 minutes and 14 seconds). This is further elaborated upon below.

Although the participants were generally highly educated, two participants commented on the **difficult language** that was sometimes used by the chatbot. Complex language usage is a common problem in eHealth services, especially for multinational patients[2] . Mizera-Pietraszko and Swiatek [96] researched the communication gap between patients and nurses or healthcare professionals in eHealth technologies. They found that the complexity and readability were far over the standard threshold (a measure to indicate complexity, the standard threshold was established so that a high school junior could comprehend the text). Therefore, across many eHealth services, as well as for the chatbot in this thesis, there is still a need for language simplification.

2: e.g. immigrants, foreign students, and temporary workers.

Two participants noted that there was no "**none of these**" option for checkboxes. However, the Nielsen Norman Group describes that checkboxes are used when there are lists of options and the user may select any number of choices, including zero, one, or several [97]. Clicking on continue (without selecting any symptoms) or typing "no" would have worked, but the chatbot could have explained this to remove this problem altogether.

Other problems noted by the participants are the need for more extensive elaboration on the **triage levels**. The **help prompt** should be clickable, and some **bugs** need to be addressed in a future version of the chatbot.

The evaluation also measured outcomes related to **health behavior**. Most participants followed the advice from the chatbot and indicated that they would seek care as the chatbot suggested. However, when looking at to what extent the chatbot influenced the participant, the results are inconclusive. Health behavior was measured six times in the systematic review [51, 62, 63, 66, 67, 69], and showed mostly positive results. For example, [77], [51] and [62] all found an improvement in diet adherence, but there are inconclusive results for improvement of physical activity [72, 77]. In a systematic review by [98], the authors also found that adherence to eHealth technology is inconclusive.

The metadata of **chatbot usage** were retrieved and queried. The **drop-off rate** for the chatbot was 27.7%, this is much higher than in the papers by Heald et al. [60] (3.8%) and Piau et al. [61] (0%). However, the chatbot by Piau et al. was only used by 9 participants. The high drop-off rate can also be caused by the fact that the chatbot usage was part of a survey. Surveys, in general, have a high drop-off rate, and this increases with the length of the survey [99]. This may have influenced the drop-off rate.

The **median duration** of a conversation with the chatbot was 3 minutes and 14 seconds. Compared to the papers from the systematic review [55–57, 59, 61] this is is a short task completion time. The median task completion time of these papers was 6 minutes and 8 seconds, almost twice as long as the chatbot in this thesis. Only the chatbot in [56] was quicker. The median number of messages from the chatbot and user was 47.

Limitations

Although the **systematic review** is the first of its kind to present an overview of evaluated non-mental healthcare chatbots, some limitations still need to be discussed. Firstly, only **one reviewer** was involved for the whole review process. Secondly, **indexing bias** may have occurred. Indexing bias [100] occurs when the indexing or retrieving of the papers is compromised. In this research only two databases were searched (PubMed and Scopus). Therefore, the review might have missed out on some papers. Also, the included papers were **not assessed on quality**, although only published conference and journal papers of over five pages were used to make the inclusion of higher quality papers more likely. Before using, assessing the methodological quality of the study is important, as results from a poor quality study can be skewed by bias and thus influence the interpretation of the results [101].

The **development and evaluation** of the novel diagnosis and triage **chatbot** presented in this thesis also has its limitations. First of all, **only products from Microsoft** were considered when building, hosting and analysing the chatbot. Also, the chatbot is currently only available in English. Thus, people who do not speak English were excluded from participating in the study.

Also, as explained above, the participants in the sample were **younger and higher educated** than the general Dutch population. This may have been a result of the sampling method, as the sample retrieved from convenience sampling is often not representative of the population [54]. As the participants are young and highly educated, they often have more **internet skills** than the general population [102]. This could have resulted in the participants producing better results with using the chatbot. Unfortunately, this was not measured in the evaluation.

Another limitation is the generally **high drop-off rate**, which could be an indication that the chatbot's duration or complexity was not optimal. Although the drop-off rate was commonly listed as a limitation by papers in the systematic review [55, 60, 64], this may still have resulted in a selective sample of participants. For example, more interested participants could have completed the conversation, whereas less interested participants could have abandoned the conversation. However, the evaluation did not report on such measures.

In the evaluation, the **CUQ** was used to measure user experience. However, the quality of the CUQ score cannot be rated objectively, as the questionnaire has not been benchmarked yet. In addition, the CUQ only measured two items per factor, whereas Worthington and Whittaker suggest to include at least three items per factor [103]. The CUQ was also not validated with a conformity factor analysis [104] to determine the independence of the factor. Despite the limitations to the questionnaire, the CUQ was chosen as it is developed especially for chatbots. Other questionnaires, such as the system usability scale [81] could have been used, but those questionnaires are not specifically created to evaluate chatbots.

Lastly, although participants were encouraged to fill out any set of symptoms, the evaluation was not applied in a **real-life context** with patients who wanted to check their symptoms. Therefore, the results cannot be generalized, and future research is necessary to validate the effect of the chatbot in practice. The health behavior was only measured by asking the user what they would do instead of measuring compliance with the chatbots advice.

Future Research

Future research is necessary to address the limitations mentioned in the previous section. First of all, the chatbot should also be tested by a **more representative** participant sample, thus also include lower-educated and older participants. The participants should be sampled in such a way that the user profile is representative of the general population, as the chatbot aimed to assist users before receiving primary care.

As an extension, future research could identify whether lower-educated users benefit more from the chatbot than conventional methods when searching for health information. This was the case in the paper by Bickmore et al. [59] who asked participants to find clinical trials that adhered to certain criteria. The participants performed the task using a chatbot and via the conventional method (keyboard- and facet-based interface). The use of the chatbot resulted in higher task completion compared to the conventional method. Therefore, there is an opportunity to research whether this is also the case when searching for possible conditions that can explain the symptoms of a user. Health information is increasingly often obtained via the internet [105], however, it may be easier for low-educated individuals to obtain this information via a chatbot, as was found in the study by Bickmore et al. The example by the paper could be followed to discover whether this effect is also apparent with the diagnosis and triage chatbot.

Although this thesis showed that the chatbot was able to identify a wide range of conditions by a set of symptoms, the evaluation showed multiple **possible improvements**, as described in previous sections. In future research, the improvements could be implemented, and another round of evaluations can be carried out. The chatbot can also be **extended**, for example, by implementing a hand-off feature to enable a healthcare provider to take over the conversation. Another addition can be to extend the chatbot to a voice-based chatbot. These additions also need to be validated before implementing the functionalities in practice.

The chatbot scored worst on **personality**, therefore, future research is needed to address this deficiency. As stated before, personality has a significant positive effect on the user experience of the chatbot interface [94]. Therefore, it is important to build a personality for the chatbot. This can be based on a framework by [106] that describes that a chatbot's personality should contain four components: the brand mission, an understanding of the users and their needs, the role of the chatbot, and an appropriate personality model. In future research, different types of personality descriptions can be tested to define which personality is most suitable for the chatbot. For this, the example by [94] can be followed, who studied the effect of the chatbot with an agreeable personality compared to a conscientious personality. After choosing the appropriate personality, the effect of the personality compared to no personality (the current version of the chatbot) can be studied. In addition to building a chatbot personality, the chatbot can be provided with an appearance. Currently, the chatbot did not have an icon or other visual representation. However, a chatbot's appearance ensures that users can more quickly form impressions of its personality and therefore experience a lower threshold to communicate with it [107]. Studying the effect of giving the chatbot an appearance can also result in interesting insights [108].

Conclusion

This thesis describes the development and evaluation of a diagnosis and triage chatbot built upon a solid foundation with a systematic review of non-mental healthcare chatbots applied in various healthcare contexts. The chatbot was built by leveraging Azure Health Bot service and related Azure services. The evaluation, carried out by 44 participants (skewed towards lower age and higher education), consisted of testing outcome variables commonly used to evaluate chatbots as in the systematic review. The chatbot performed well on user experience, but improvements related to personality and error handling are necessary. The health behavior of the participants after using the chatbot was found to be inconclusive. The analysis of the chatbot usage meta-data showed a high drop-off rate, but also that the conversation was relatively short compared to other chatbots. Future research is needed to address the limitations of this thesis and the implementation of the suggested improvements.

References

Here are the references in citation order.

[1] P. B. Brandtzaeg and A. Følstad. 'Why people use chatbots'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10673 LNCS. Springer Verlag, Nov. 2017, pp. 377–392. DOI: 10.1007/978-3-319-70284-1{_}30 (cited on page 2).

[2] M. Dahiya. 'A Tool of Conversation: Chatbot'. In: *Article in International Journal of Computer Sciences and Engineering* (2017) (cited on page 2).

[3] R. Whittaker et al. 'Mobile phone-based interventions for smoking cessation'. In: *Cochrane Database of Systematic Reviews* 2016.4 (Apr. 2016). DOI: 10.1002/14651858.CD006611.pub4 (cited on page 2).

[4] J. Weizenbaum. 'ELIZA-A computer program for the study of natural language communication between man and machine'. In: *Communications of the ACM* 9.1 (1966), pp. 39–45. DOI: 10.1145/365153.365168 (cited on page 2).

[5] F. Colace et al. 'Chatbot for e-learning: A case of study'. In: *International Journal of Mechanical Engineering and Robotics Research* 7.5 (2018), pp. 528–533. DOI: 10.18178/ijmerr.7.5.528-533 (cited on page 2).

[6] L. Cui et al. 'Superagent: A customer service chatbot for E-commerce websites'. In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations* (2017), pp. 97–102. DOI: 10.18653/v1/P17-4017 (cited on page 2).

[7] J. Singh, M.H. Joesph, and K.B.A. Jabbar. 'Rule-based chabot for student enquiries'. In: *Journal of Physics: Conference Series*. Vol. 1228. 1. Institute of Physics Publishing, June 2019, p. 012060. DOI: 10.1088/1742-6596/1228/1/012060 (cited on page 2).

[8] M. Adam, M. Wessel, and A. Benlian. 'AI-based chatbots in customer service and their effects on user compliance'. In: *Electronic Markets* (2020), pp. 1–19. DOI: 10.1007/s12525-020-00414-7 (cited on page 2).

[9] C. Bérubé et al. 'Voice-based conversational agents for the prevention and management of chronic and mental health conditions: Systematic literature review'. In: *Journal of Medical Internet Research* 23.3 (Mar. 2021), e25933. DOI: 10.2196/25933 (cited on page 2).

[10] S. Hauser-Ulrich et al. 'A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): Pilot randomized controlled trial'. In: *JMIR mHealth and uHealth* 8.4 (Apr. 2020), e15806. DOI: 10.2196/15806 (cited on page 2).

[11] A.A. Abd-Alrazaq et al. 'Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review'. In: *Journal of Medical Internet Research* 22.6 (June 2020), e18301. DOI: 10.2196/18301 (cited on page 3).

[12] G.M. Lucas et al. 'It's only a computer: Virtual humans increase willingness to disclose'. In: *Computers in Human Behavior* 37 (Aug. 2014), pp. 94–100. DOI: 10.1016/j.chb.2014.04.043 (cited on page 3).

[13] A. Palanica et al. 'Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey'. In: *Journal of Medical Internet Research* 21.4 (Apr. 2019), e12887. DOI: 10.2196/12887 (cited on pages 3, 56).

[14] A.A. Abd-Alrazaq et al. 'An overview of the features of chatbots in mental health: A scoping review.' In: *International journal of medical informatics* 132 (2019), p. 103978 (cited on page 3).

[15] A.N. Vaidyam, D. Linggonegoro, and J.B. Torous. 'Changes to the Psychiatric Chatbot Landscape: A Systematic Review of Conversational Agents in Serious Mental Illness: Changements du paysage psychiatrique des chatbots: une revue systématique des agents conversationnels dans la maladie mentale sérieuse.' In: *Canadian journal of psychiatry. Revue canadienne de psychiatrie* (2020), p. 706743720966429 (cited on page 3).

[16] H. Gaffney, W. Mansell, and S. Tai. 'Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review.' In: *JMIR mental health* 6.10 (2019), e14166 (cited on page 3).

[17] S. Hoermann et al. 'Application of synchronous text-based dialogue systems in mental health interventions: Systematic review'. In: *Journal of Medical Internet Research* 19.8 (2017) (cited on page 3).

[18] A.N. Vaidyam et al. 'Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape'. In: *Canadian Journal of Psychiatry* 64.7 (2019), pp. 456–464 (cited on page 3).

[19] A. Teles et al. 'Mobile mental health: A review of applications for depression assistance'. In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* 2019 (2019), pp. 708–713 (cited on page 3).

[20] A.A. Abd-Alrazaq et al. 'Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review.' In: *Journal of medical Internet research* 23.1 (2021), e17828 (cited on page 3).

[21] Z. Safi et al. 'Technical Aspects of Developing Chatbots for Medical Applications: Scoping Review.' In: *Journal of medical Internet research* 22.12 (2020), e19127 (cited on page 3).

[22] S.B. Ahire and J.R. Prasad. 'A review on chatbot for healthcare system using recurrent neural network'. In: *International Journal of Advanced Science and Technology* 29.5 (Apr. 2020), pp. 336–340 (cited on pages 3, 14).

[23] A.R.W. Tjiptomongsoguno et al. 'Medical Chatbot Techniques: A Review'. In: *Proceedings of the Computational Methods in Systems and Software* 1294 (2020), pp. 346–356 (cited on page 3).

[24] L. Tudor Car et al. 'Conversational Agents in Health Care: Scoping Review and Conceptual Analysis.' In: *Journal of medical Internet research* 22.8 (2020), e17158 (cited on pages 3, 6).

[25] N. Xiao et al. 'Factors influencing online health information search: An empirical analysis of a national cancer-related survey'. In: *Decision Support Systems* 57.1 (Jan. 2014), pp. 417–427. DOI: 10.1016/j.dss.2012.10.047 (cited on page 3).

[26] J.L.Z. Montenegro, C.A. da Costa, and R. da Rosa Righi. 'Survey of conversational agents in health'. In: *Expert Systems with Applications* 129 (Sept. 2019), pp. 56–67. DOI: 10.1016/j.eswa.2019.03.054 (cited on page 4).

[27] Y. You and X. Gui. 'Self-Diagnosis through AI-enabled Chatbot-based Symptom Checkers: User Experiences and Design Considerations'. In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2020 (Jan. 2020), pp. 1354–1363 (cited on pages 4, 6, 12, 44).

[28] G. Cameron et al. 'Towards a chatbot for digital counselling'. In: *HCI 2017: Digital Make Believe - Proceedings of the 31st International BCS Human Computer Interaction Conference, HCI 2017*. Vol. 2017-July. BCS Learning and Development Ltd., July 2017. DOI: 10.14236/ewic/HCI2017.24 (cited on page 4).

[29] W.E. Hautz, A. Exadaktylos, and T.C. Sauter. 'Online forward triage during the COVID-19 outbreak'. In: *Emergency Medicine Journal* (2020). DOI: 10.1136/emermed-2020-209792 (cited on page 4).

[30] J. Espinoza, K. Crown, and O. Kulkarni. 'A guide to chatbots for COVID-19 screening at pediatric health care facilities'. In: *JMIR Public Health and Surveillance* 6.2 (Apr. 2020). DOI: 10.2196/18808 (cited on page 4).

[31] K.E. Morse et al. 'Use characteristics and triage acuity of a digital symptom checker in a large integrated health system: Population-based descriptive study'. In: *Journal of Medical Internet Research* 22.11 (Nov. 2020). DOI: 10.2196/20549 (cited on page 4).

[32] A. Martin et al. 'An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot'. In: *Scientific Reports* 10.1 (Dec. 2020). DOI: 10.1038/s41598-020-75912-x (cited on page 4).

[33] M. Almalki. 'Perceived utilities of COVID-19 related chatbots in Saudi Arabia: A cross-sectional study'. In: *Acta Informatica Medica* 28.3 (2021), pp. 219–224. DOI: 10.5455/AIM.2020.28.219-224 (cited on page 4).

[34] *Built-in Medical Intelligence - Health Bot | Microsoft Docs*. URL: https://docs.microsoft.com/en-us/healthbot/bot_docs/triage_symptom_checking (cited on pages 4, 5, 13, 14).

[35] *Cloud Computing Services | Microsoft Azure*. URL: https://azure.microsoft.com/nl-nl/ (cited on page 4).

[36] Azure. *Azure Bot Service | Microsoft Azure*. 2020. URL: https://azure.microsoft.com/en-us/services/bot-service/ (cited on page 4).

[37] J. Edqvist Rasmus Lennartsson. 'Chat Bots & Voice Control Applications and limitations of combining Microsoft's Azure Bot Service and Cognitive Services' Speech API'. In: *Karlstad University* (2019) (cited on page 4).

[38] *Microsoft Healthcare Bot brings conversational AI to healthcare | Azure-blog en -updates | Microsoft Azure*. URL: https://azure.microsoft.com/nl-nl/blog/microsoft-healthcare-bot-brings-conversational-ai-to-healthcare/ (cited on page 4).

[39] *Overview - Health Bot | Microsoft Docs*. URL: https://docs.microsoft.com/en-us/healthbot/ (cited on page 5).

[40] L. Ni et al. 'MANDY: Towards a smart primary care chatbot application'. In: *Communications in Computer and Information Science*. Vol. 780. Springer Verlag, Nov. 2017, pp. 38–52. DOI: 10.1007/978-981-10-6989-5{_}4 (cited on page 5).

[41] D. Moher, L. Shamseer, and M. Clarke. 'Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement'. In: *Revista Espanola de Nutricion Humana y Dietetica* 20.2 (Dec. 2016), pp. 148–160. DOI: 10.1186/2046-4053-4-1 (cited on pages 8, 11).

[42] M. Ouzzani et al. 'Rayyan—a web and mobile app for systematic reviews'. In: *Systematic Reviews* 5.1 (2016), p. 210. DOI: 10.1186/s13643-016-0384-4 (cited on page 9).

[43] C. Wohlin. 'Guidelines for snowballing in systematic literature studies and a replication in software engineering'. In: *ACM International Conference Proceeding Series* (2014). DOI: 10.1145/2601248.2601268 (cited on page 10).

[44] Marshall Islands. 'Appendix D . ISO Country Codes for Selected Countries'. In: Md (Oct. 2007), pp. 1–4 (cited on pages 11, 30).

[45] *Available conditions v3 – Infermedica for Developers*. URL: https://developer.infermedica.com/docs/v3/available-conditions (cited on pages 12, 44).

[46] *Wat is Azure Application Insights? - Azure Monitor | Microsoft Docs*. URL: https://docs.microsoft.com/nl-nl/azure/azure-monitor/app/app-insights-overview (cited on page 13).

[47] *App Service | Microsoft Azure*. URL: https://azure.microsoft.com/nl-nl/services/app-service/#features (cited on page 13).

[48] Microsoft. *Create your Azure free account today | Microsoft Azure*. 2019. URL: https://azure.microsoft.com/en-us/free/search/ (cited on page 13).

[49] *Manage resource groups - Azure portal - Azure Resource Manager | Microsoft Docs*. URL: https://docs.microsoft.com/en-us/azure/azure-resource-manager/management/manage-resource-groups-portal (cited on page 13).

[50] J. Martín, C. Muñoz-Romero, and N. Ábalos. *GitHub - chatbottest-com/chatbottest: Landing for the project chatbottest.com*. URL: https://github.com/chatbottest-com/chatbottest (cited on pages 13, 18).

[51] C.R. Davis et al. 'A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant'. In: *International Journal of Environmental Research and Public Health* 17.23 (2020), pp. 1–14 (cited on pages 14, 19, 26–28, 31–34, 37, 57–59).

[52] *GitHub - microsoft/HealthBotContainerSample: Healthcare Bot Container Sample*. URL: https://github.com/Microsoft/HealthBotContainerSample (cited on page 15).

[53] Microsoft. *Overview - Azure Data Explorer | Microsoft Docs*. URL: https://docs.microsoft.com/en-us/azure/data-explorer/kusto/query/ (cited on page 15).

[54] P. Sedgwick. 'Convenience sampling'. In: *BMJ* 347.oct25 2 (Oct. 2013), f6304–f6304. DOI: 10.1136/bmj.f6304 (cited on pages 16, 60).

[55] R. Crutzen et al. 'An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study.' In: *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 48.5 (2011), pp. 514–519 (cited on pages 17, 18, 26–28, 31–34, 36, 57, 59, 60).

[56] S. Holmes et al. 'Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?' In: *Proceedings of the 31st European Conference on Cognitive Ergonomics.* (2019), pp. 207–214 (cited on pages 17–20, 26–28, 31–33, 38, 59, 60).

[57] A. Ponathil et al. 'New approaches to collecting family health history - A preliminary study investigating the efficacy of conversational systems to collect family health history'. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62.1 (2018), pp. 277–281 (cited on pages 17, 18, 26, 27, 29, 31–34, 41, 57, 59).

[58] S.R. Haynes, S. Purao, and A.L. Skattebo. 'Situating evaluation in scenarios of use'. In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2004), pp. 92–101. DOI: 10.1145/1031607.1031624 (cited on page 17).

[59] T.W. Bickmore et al. 'Improving Access to Online Health Information With Conversational Agents: A Randomized Controlled Experiment'. In: *Journal of medical Internet research* 18.1 (2016), e1 (cited on pages 18, 26–28, 31–33, 35, 57, 59, 61).

[60] B. Heald et al. 'Using chatbots to screen for heritable cancer syndromes in patients undergoing routine colonoscopy.' In: *Journal of medical genetics* (2020) (cited on pages 18, 26–28, 31–34, 37, 57, 59, 60).

[61] A. Piau et al. 'A smartphone Chatbot application to optimize monitoring of older patients with cancer'. In: *International Journal of Medical Informatics* 128 (Aug. 2019), pp. 18–23. DOI: 10.1016/j.ijmedinf.2019.05.013 (cited on pages 18, 26, 27, 29, 31–33, 40, 57, 59).

[62] N. Stein and K. Brooks. 'A Fully Automated Conversational Artificial Intelligence for Weight Loss: Longitudinal Observational Study Among Overweight and Obese Adults.' In: *JMIR diabetes* 2.2 (2017), e28 (cited on pages 18, 19, 26, 27, 29, 31–34, 42, 59).

[63] B. Chaix et al. 'When Chatbots Meet Patients: One-Year Prospective Study of Conversations Between Patients With Breast Cancer and a Chatbot.' In: *JMIR cancer* 5.1 (2019), e12856 (cited on pages 18, 19, 26–28, 31–33, 36, 59).

[64] S.B. Goldenthal et al. 'Assessing the feasibility of a chatbot after ureteroscopy.' In: *mHealth* 5 (2019), p. 8 (cited on pages 18, 26–28, 31–34, 37, 60).

[65] I.E. Allen. 'Likert Scales and Data Analyses'. In: *Quality Progress* 40.7 (2007), pp. 64–65 (cited on page 18).

[66] S.A. Hussain, F. Ogundimu, and S. Bhattarai. 'Mobile Phone-Based Chatbot for Family Planning and Contraceptive Information'. In: *International Conference on Human-Computer Interaction* 11582 (2019), pp. 342–352 (cited on pages 18, 26, 27, 29, 31–33, 38, 59).

[67] E. Maeda et al. 'Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial'. In: *Reproductive BioMedicine Online* 41.6 (2020), pp. 1133–1143 (cited on pages 18, 26, 27, 29, 31, 33, 39, 57, 59).

[68] N. Siangchin and T. Samanchuen. 'Chatbot Implementation for ICD-10 Recommendation System'. In: *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)* (2019), pp. 1–6 (cited on pages 18, 26, 27, 29, 31–33, 42).

[69] T.W. Bickmore, D. Schulman, and C. Sidner. 'Automated interventions for multiple health behaviors using conversational agents'. In: *Patient Education and Counseling* 92.2 (2013), pp. 142–148. DOI: https://doi.org/10.1016/j.pec.2013.05.011 (cited on pages 19, 26–28, 31–35, 59).

[70] S. Holmes and R. Bond. *The Chatbot Usability Questionnaire (CUQ) calculator - Ulster University*. URL: https://www.ulster.ac.uk/__data/assets/excel_doc/0010/478810/CUQ-Calculation-Tool.xlsx (cited on page 20).

[71] S.H. Khandkar. 'Open coding'. In: *Basics of qualitative research: ...* 1998, pp. 101–121 (cited on page 20).

[72] M. Piao et al. 'Development and usability evaluation of a healthy lifestyle coaching chatbot using a habit formation model'. In: *Healthcare Informatics Research* 26.4 (2020), pp. 255–264 (cited on pages 26, 27, 29, 31–33, 40, 57, 59).

[73] J.E. Bibault et al. 'A chatbot versus physicians to provide information for patients with breast cancer: Blind, randomized controlled noninferiority trial'. In: *Journal of Medical Internet Research* 21.11 (Nov. 2019). DOI: 10.2196/15787 (cited on pages 26–28, 31–33, 35).

[74] T. Schmidlen et al. 'Patient assessment of chatbots for the scalable delivery of genetic counseling'. In: *Journal of Genetic Counseling* 28.6 (2019), pp. 1166–1177 (cited on pages 26, 27, 29, 31–34, 41).

[75] J. Koman et al. 'Physicians' Perceptions of the Use of a Chatbot for Information Seeking: Qualitative Study.' In: *Journal of medical Internet research* 22.11 (2020), e15185 (cited on pages 26, 27, 29, 31–34, 39, 57).

[76] P. Thongyoo et al. 'A Personalized Food Recommendation Chatbot System for Diabetes Patients'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12341 LNCS. Springer Science and Business Media Deutschland GmbH, 2020, pp. 19–28. DOI: 10.1007/978-3-030-60816-3{_}3 (cited on pages 26, 27, 29, 31–33, 43, 57).

[77] A. Bick, A. Blandin, and K. Mertens. 'Work from Home After the COVID-19 Outbreak'. In: *Federal Reserve Bank of Dallas, Working Papers* 2020.2017 (June 2020). DOI: 10.24149/wp2017 (cited on pages 32, 59).

[78] K. Bjordal et al. 'A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H and N35) in head and neck patients'. In: *European Journal of Cancer* 36.14 (Sept. 2000), pp. 1796–1807. DOI: 10.1016/S0959-8049(00)00186-6 (cited on page 35).

[79] T. C. Davis et al. 'Rapid estimate of adult literacy in medicine: A shortened screening instrument'. In: *Family Medicine* 25.6 (June 1993), pp. 391–395 (cited on page 35).

[80] F. Kastrinos et al. 'Development and validation of a colon cancer risk assessment tool for patients undergoing colonoscopy'. In: *American Journal of Gastroenterology* 104.6 (June 2009), pp. 1508–1518. DOI: 10.1038/ajg.2009.135 (cited on page 37).

[81] J. Brooke. 'SUS: a "quick and dirty" usability scale'. In: *Usability Evaluation in Industry*. London: Taylor and Francis, 1996. Chap. 21, pp. 189–194 (cited on pages 38, 60).

[82] B. Laugwitz, T. Held, and M. Schrepp. 'Construction and evaluation of a user experience questionnaire'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 5298 LNCS. Springer Verlag, Nov. 2008, pp. 63–76. DOI: 10.1007/978-3-540-89350-9{_}6 (cited on page 38).

[83] V. Venkatesh et al. 'User acceptance of information technology: Toward a unified view'. In: *MIS Quarterly: Management Information Systems* 27.3 (2003), pp. 425–478. DOI: 10.2307/30036540 (cited on page 38).

[84] V. Venkatesh and F.D. Davis. 'Theoretical extension of the Technology Acceptance Model: Four longitudinal field studies'. In: *Management Science* 46.2 (2000), pp. 186–204. DOI: 10.1287/mnsc.46.2.186.11926 (cited on pages 40, 41).

[85] S.G. Hart and L.E. Staveland. 'Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research'. In: *Advances in Psychology* 52.C (Jan. 1988), pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9 (cited on page 41).

[86] S.K. Card, T.P. Moran, and A. Newell. *The Psychology of Human Computer Interaction*. Hillsday: New Jersey: Lawerence Erlbaum Asso- ciates. Inc., 1983 (cited on page 41).

[87] C. Robert et al. 'Effectiveness of eHealth Nutritional Interventions for Middle-Aged and Older Adults: Systematic Review and Meta-analysis'. In: *Journal of Medical Internet Research* 23.5 (May 2021), e15649. DOI: 10.2196/15649 (cited on page 56).

[88] S. Muellmann et al. 'Effectiveness of eHealth interventions for the promotion of physical activity in older adults: A systematic review'. In: *Preventive Medicine* 108 (Mar. 2018), pp. 93–110. DOI: 10.1016/j.ypmed.2017.12.026 (cited on page 56).

[89] M. Milne-Ives et al. 'The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review'. In: *Journal of Medical Internet Research* 22.10 (Oct. 2020), e20346. DOI: 10.2196/20346 (cited on page 57).

[90] L. Laranjo et al. 'Conversational agents in healthcare: A systematic review'. In: *Journal of the American Medical Informatics Association* 25.9 (Sept. 2018), pp. 1248–1258. DOI: 10.1093/jamia/ocy072 (cited on page 57).

[91] J. Hill, W. Randolph Ford, and I.G. Farreras. 'Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations'. In: *Computers in Human Behavior* 49 (Aug. 2015), pp. 245–250. DOI: 10.1016/j.chb.2015.02.026 (cited on pages 57, 59).

[92] CBS. *Population pyramid (age)*. URL: https://www.cbs.nl/en-gb/visualisations/dashboard-population/population-pyramid (cited on page 57).

[93] *Education - Figures - Society | Trends in the Netherlands 2018 - CBS*. URL: https://longreads.cbs.nl/trends18-eng/society/figures/education/ (cited on page 58).

[94] T.L. Smestad and F. Volden. 'Chatbot personalities matters: Improving the user experience of chatbot interfaces'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11551 LNCS. Springer Verlag, 2019, pp. 170–181. DOI: 10.1007/978-3-030-17705-8{_}15 (cited on pages 58, 61, 62).

[95] J. Nielsen. *Heuristic Evaluation Ten Usability Heuristics*. Tech. rep. 2005 (cited on page 58).

[96] J. Mizera-Pietraszko and P. Swiatek. 'Access to eHealth language-based services for multinational patients'. In: *2015 17th International Conference on E-Health Networking, Application and Services, HealthCom 2015*. Institute of Electrical and Electronics Engineers Inc., 2015, pp. 232–237. DOI: 10.1109/HealthCom.2015.7454504 (cited on page 59).

[97] Nielsen Norman Group. *Checkboxes vs. Radio Buttons*. URL: https://www.nngroup.com/articles/checkboxes-vs-radio-buttons/ (cited on page 59).

[98] F. Sieverink, S.M. Kelders, and V. Gemert-Pijnen. 'Clarifying the concept of adherence to ehealth technology: Systematic review on when usage becomes adherence'. In: *Journal of Medical Internet Research* 19.12 (Dec. 2017). DOI: 10.2196/jmir.8578 (cited on page 59).

[99] M. Hoerger. 'Participant dropout as a function of survey length in internet-mediated university studies: Implications for study design and voluntary participation in psychological research'. In: *Cyberpsychology, Behavior, and Social Networking* 13.6 (Dec. 2010), pp. 697–700. DOI: 10.1089/cyber.2009.0445 (cited on page 59).

[100] D.T. Felson. 'Bias in meta-analytic research'. In: *Journal of Clinical Epidemiology* 45.8 (Aug. 1992), pp. 885–892. DOI: 10.1016/0895-4356(92)90072-U (cited on page 60).

[101] X. Zeng et al. 'The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: A systematic review'. In: *Journal of Evidence-Based Medicine* 8.1 (Feb. 2015), pp. 2–10. DOI: 10.1111/jebm.12141 (cited on page 60).

[102] A.J.A.M. Van Deursen, J.A.G.M. Van Dijk, and O. Peters. 'Rethinking Internet skills: The contribution of gender, age, education, Internet experience, and hours online to medium- and content-related Internet skills'. In: *Poetics* 39.2 (Apr. 2011), pp. 125–144. DOI: 10.1016/j.poetic.2011.02.001 (cited on page 60).

[103] R.L. Worthington and T.A. Whittaker. 'Scale Development Research: A Content Analysis and Recommendations for Best Practices'. In: *The Counseling Psychologist* 34.6 (June 2006), pp. 806–838. DOI: 10.1177/0011000006288127 (cited on page 60).

[104] K.G. Jöreskog. 'A general approach to confirmatory maximum likelihood factor analysis'. In: *Psychometrika* 34.2 (June 1969), pp. 183–202. DOI: 10.1007/BF02289343 (cited on page 60).

[105] S. Christmann. *Health literacy and Internet. Recommendations to promote Health Literacy by the means of the Internet.* Tech. rep. 2005 (cited on page 61).

[106] T.L. Smestad. 'Personality matters! Improving the user experience of chatbot interfaces'. In: June (2018), pp. 1–89 (cited on page 61).

[107] M. Bar, M. Neta, and H. Linz. 'Very first impressions'. In: *Emotion* 6.2 (May 2006), pp. 269–278. DOI: 10.1037/1528-3542.6.2.269 (cited on page 62).

[108] H. Nguyen and J. Masthoff. 'Is it me or is it what I say? Source image and persuasion'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 4744 LNCS. Springer Verlag, 2007, pp. 231–242. DOI: 10.1007/978-3-540-77006-0{_}29 (cited on page 62).

[109] *Triage v3 – Infermedica for Developers*. URL: https://developer.infermedica.com/docs/v3/triage (cited on page 81).

APPENDICES

Appendix A: Search Strings of the Systematic Review

Date last searched: 11 January 2020

Search string on PubMed: (("chatbot"[tiab] OR "conversational interface"[tiab] OR "conversational chatbot"[tiab] OR "bot"[tiab]) AND ("delivery of health care"[MeSH Terms] OR "Medical"[tiab] OR "Health"[tiab] OR "Health Care"[tiab] OR "Healthcare"[tiab]))

Search string on Scopus: TITLE-ABS-KEY ((("chatbot" OR "conversational interface" OR "conversational chatbot" OR "bot") AND ("delivery of health care" OR "Medical" OR "Health" OR "Health Care" OR "Healthcare")))

Appendix B: Questionnaire Evaluation

Utrecht University

Informed consent

With this informed consent form the researcher requests your consent for participation in the chatbot evaluation. This evaluation is part of an ongoing study on chatbots applied in healthcare.

You are being asked to use the chatbot by filling out symptoms derived from user stories and walk through the series of questions asked by the chatbot. After using the chatbot, some questions will be asked to evaluate the chatbot. The outcomes of this evaluation will be used to pinpoint opportunities for further improvement to the chatbot.

Your participation in this evaluation session is voluntary. You have the right to withdraw at any point during the session. Any materials produced by the evaluation may be used for publication but will be fully anonymized.

If you have any queries, please email a.j.b.kockx@students.uu.nl.
In case of any issues, please contact my supervisor J. Masthoff via: j.f.m.masthoff@uu.nl.

By clicking the 'I consent to participate' button below, you acknowledge:

- Your participation in the study is voluntary.

- You are at least 18 years of age.
- You are aware that you may choose to terminate your participation at any time for any reason.
- You consent to allow the fully anonymized data to be used for future publications and scholarly means of disseminating the findings of the study.

◯ I consent to participate in this study
◯ I do not consent to participate in this study

Demographics

What gender do you identify as?

◯ Male
◯ Female
◯ Other
◯ Prefer not to answer

What is your age?

What is the highest degree or level of education you have completed?

◯ Some high school

- ◯ High school
- ◯ Trade / vocational school
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ PHD or higher
- ◯ Prefer not to say

Use the chatbot

Time to use the chatbot!

After clicking on the link (see bottom of this page), the chatbot will ask you to fill out your symptoms. Then, it will ask follow-up questions to determine the possible conditions and give you appropriate advice. You are completely free to use your imagination and fill out any set of symptoms fitting a condition. However, to help you, here are some **user stories** you could use with the chatbot:

User story 1: *Judy is a 42-year-old woman and sometimes suddenly has very severe headaches, usually on one side of the head. When these headaches occur, it sometimes is accompanied by other symptoms such as nausea and vomiting. The headaches can last for hours or sometimes even days. Judy is very worried about her symptoms and wants to visit a doctor but decides to consult the chatbot first.*

User story 2: *Jerry (male, 76 years old) has been coughing for several weeks, and first expected that it was just another cold, but it keeps getting worse. In addition to his cough, he also experiences pain when coughing or breathing. This week he even coughed up blood and that made him very worried. Jerry is not a huge fan of hospitals, so he decides to consult the chatbot before visiting the doctor.*

User story 3: *Jennifer is a 26-year-old woman who has felt sad, empty, and hopeless for a longer period now. In addition, she keeps falling out with her boyfriend as she seems to be irritated by everything he does. Her friends are also very worried, as she keeps rejecting normal activities such as exercising or picking up her hobbies. In addition, she cannot remember the last time she had a good night of sleep. She decides to consult the chatbot with these symptoms.*

The chatbot will ask you for a **participant number**, yours is: **${e://Field/RandomID}**

You can access the chatbot via **this link** (opens in a new tab), after using the chatbot you can move to the next page.

(or copy-paste the following link into your browser: https://healthcare-bot-qzsbgnfpgbikg.azurewebsites.net/).

Have you used the chatbot?

- ◯ Yes
- ◯ No

Experience with the chatbot

Please rate your experience with the following questions:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The chatbot's personality was realistic and engaging	◯	◯	◯	◯	◯
The chatbot seemed too robotic	◯	◯	◯	◯	◯
The chatbot was welcoming during initial setup	◯	◯	◯	◯	◯
The chatbot seemed very unfriendly	◯	◯	◯	◯	◯
The chatbot explained its scope and purpose well	◯	◯	◯	◯	◯

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The chatbot gave no indication as to its purpose	○	○	○	○	○
The chatbot was easy to navigate	○	○	○	○	○
It would be easy to get confused when using the chatbot	○	○	○	○	○
The chatbot understood me well	○	○	○	○	○
The chatbot failed to recognise a lot of my inputs	○	○	○	○	○
Chatbot responses were useful, appropriate and informative	○	○	○	○	○
Chatbots responses were not relevant	○	○	○	○	○
The chatbot coped well with any errors or mistakes	○	○	○	○	○
The chatbot seemed unable to handle any errors	○	○	○	○	○
The chatbot was very easy to use	○	○	○	○	○
The chatbot was very complex	○	○	○	○	○

Any comments on the experience with the chatbot?

Any comments on the responses from the chatbot?

Advice from the chatbot

What did the chatbot advise?

- ◯ Self-care at Home
- ◯ Doctor Routine Check-up
- ◯ Doctor Today
- ◯ Emergency Department
- ◯ Ambulance

Considering your inputted symptoms and the advice from the chatbot, what would be your next action?

- ◯ Self-care at Home
- ◯ Doctor Routine Check-up
- ◯ Doctor Today
- ◯ Emergency Department
- ◯ Ambulance

To what extent were you influenced by the advice from the chatbot?

- A great deal
- A lot
- A moderate amount
- A little
- None at all

Appendix C: KQL Queries for Meta-data Analysis

Median Dialog Duration

```
// Median dialog duration
let queryStartDate = datetime(2021-05-06T13:00:00Z);
let queryEndDate = datetime(2021-05-24T09:00:00Z);
customEvents
| where timestamp > queryStartDate
| where timestamp < queryEndDate
| extend userID = tostring(customDimensions.user_id)
| where name == "ScenarioStart"
| join kind=leftouter (customEvents | where name=="ScenarioEnded" | extend userID = tostring(customDimensions.user_id)) on userID
| extend duration = timestamp1 - timestamp
| summarize median = percentile(duration, 50)
```

Drop-off Rate

```
// Completed Dialog: shows completes relative to starts
let queryStartDate = datetime(2021-05-06T13:00:00Z);
let queryEndDate = datetime(2021-05-24T09:00:00Z);
customEvents
| where timestamp > queryStartDate
| where timestamp < queryEndDate
| where name=="ScenarioStart"
| extend convID = tostring(customDimensions.conv_id)
| join kind=leftouter (
    customEvents
    | where name=="ScenarioEnded"
    | extend convID = tostring(customDimensions.conv_id)
  ) on convID
| summarize started=countif(name=='ScenarioStart'), completed=countif(name1=='ScenarioEnded')
```

Number of Messages

```
//Count only finished
let queryStartDate = datetime(2021-05-06T13:00:00Z);
let queryEndDate = datetime(2021-05-24T09:00:00Z);
customEvents
| where timestamp > queryStartDate
| where timestamp < queryEndDate
| extend convID = tostring(customDimensions.conv_id)
| order by convID, timestamp asc
| where name == "ScenarioStart" or  name == "Message" or  name == "ScenarioEnded"
| join kind=leftouter (
    customEvents
    | where name=="ScenarioEnded"
    | extend convID = tostring(customDimensions.conv_id)
  ) on convID
| extend finished = case(name1 == "ScenarioEnded", 1, 0)
| summarize cnt=countif(finished == 1) by convID
| where cnt != 0
| summarize avgcnt = avg(cnt), median = percentile(cnt, 50), mincnt = min(cnt), maxcnt = max(cnt), std=stdev(cnt)
```

Appendix D: Medical Risk Factors

Gender is Male

- Recent physical injury
- Diabetes
- Hypertension
- High cholesterol
- Heart disease
- Smoking

Gender is Female, Age is lower than 12 or higher than 55

- Recent physical injury
- Diabetes
- Hypertension
- High cholesterol
- Heart disease
- Smoking

Gender is Female, age is lower than 45 and higher than 12

- Pregnancy
- Recent physical injury
- Diabetes
- Hypertension
- High cholesterol
- Heart disease
- Smoking

Gender is Female, age between 45 and 55

- Pregnancy
- After menopause
- Recent physical injury
- Diabetes
- Hypertension
- High cholesterol
- Heart disease
- Smoking

Appendix E: Information per Triage Level

Self-care at home: A medical consultation is not strictly required; you should observe your symptoms and consult a doctor only if symptoms worsen within 24 hours.

Visit the doctor for a routine check-up: You may require medical evaluation and may need to schedule an appointment with a doctor. If symptoms get worse, you should see a doctor immediately.

Visit the doctor today: You should see a doctor within 24 hours. If the symptoms suddenly get worse, you should go to the nearest emergency department.

Visit the emergency department: The reported evidence appears serious and you should go to an emergency department. If you can't get to the nearest emergency department, you should call an ambulance.

Call an ambulance: Your reported symptoms are very serious and you may require emergency care. You should call an ambulance right now [109].

Appendix F: Qualitative assessment

In this Appendix, the conversation logs of participants who rated the chatbot with a CUQ score lower than 50 are assessed.

Participant 1507, CUQ = 17.2

Discovered problems:

- No response after "start over" prompt
- Unknown issue after answering "No" to "is that correct?"

begin triage
[C]: Hello! I can help you with identifying **possible causes** for the symptoms you are experiencing. In addition, I can advise you on the **following steps** such as whether it is necessary to visit a doctor.
[C]: But first, what is your **participant number**? This number is provided to you in the survey.
[U]: 1507
[C]: What brings you here today? Please describe your symptoms.
[U]: I coughed up blood
[C]: I'm sorry to hear that. Before the assessment begins please provide additional information to understand your **medical profile**.
[U]: {"id":"1234567890","age":76,"gender":"M"}
[C]: Thanks, Do you have any of the following **risk factors**? (select all that apply or leave blank)
[U]: {"id":"1234567890","riskfactors":"Recent physical injury,High cholesterol,Heart disease"}
[C]: It sounds like you would like to report the following symptoms: - Recent physical injury - High cholesterol - Coronary disease
[C]: Is that correct?
[U]: **No**
[C]: OK. If you would like to report medical symptoms, please try to use an alternative phrasing.
[C]: **Oops. Something went wrong and we need to start over. Please enter 'start over'.**
[U]: **start over**

[C]: Message by the chatbot

[U]: Message by the user

Sometimes multiple questions are asked at once, such as age and gender. Then, the responses by the user are shown in the form of {"questionID":"input_by_user",...}. If it is just {}, then none of the answer options were chosen.

Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, e.g. [word].

Participant 9975, CUQ = 26.6

Discovered problems

- Did not understand initial complaint, an error is given later in the conversation.
- Initial complaint and additional complaints were not symptoms
- Asks for an ambulance at the end, but that falls out of the scope of the chatbot
- No response on ’start over’ prompt

[U]: Help
[C]: Here are some things I can help you do: * begin triage: Start the triage * start over: Restarts your current scenario * help: Shows what this bot can do * terms: Shows terms of use and privacy statement * feedback: Give us feedback * what do you know: Shows what this bot remembers about you * log: Shows your previous interactions with this bot. * forget me: Delete your data * cancel: Stops your conversation
[U]: Begin triage
[C]: Hello! I can help you with identifying **possible causes** for the symptoms you are experiencing. In addition, I can advise you on the **following steps** such as whether it is necessary to visit a doctor.
[C]: But first, what is your **participant number**? This number is provided to you in the survey.
[U]: 9975
[C]: What brings you here today? Please describe your symptoms.
[U]: **I have to laugh during the [whole] day**
[C]: I’m sorry to hear that. Before the assessment begins please provide additional information to understand your **medical profile**.
[U]: {"id":"1234567890","age":52,"gender":"F"}
Thanks, Do you have any of the following **risk factors**? (select all that apply or leave blank)
[U]: {"id":"1234567890"}
[C]: **Sorry, I do not understand. Please retry or enter ’start over’ to restart the conversation.**
[C]: Oops. Something went wrong and we need to start over. Please enter ’start over’.
[U]: Start over
[U]: **I am happy**

[C]: Sorry, I do not understand. Please retry or enter 'start over' to restart the conversation.
[U]: Start over
[U]: 'Start over'
[C]: Sorry, I do not understand. Please retry or enter 'start over' to restart the conversation.
[U]: Begin triage
[C]: Hello! I can help you with identifying **possible causes** for the symptoms you are experiencing. In addition, I can advise you on the **following steps** such as whether it is necessary to visit a doctor.
[C]: But first, what is your **participant number**? This number is provided to you in the survey.
[U]: 7788
[C]: What brings you here today? Please describe your symptoms.
[U]: Thirsty
[C]: I'm sorry to hear that. Before the assessment begins please provide additional information to understand your **medical profile**.
[U]: {"id":"1234567890","age":52,"gender":"F"}
[C]: Thanks, Do you have any of the following **risk factors**? (select all that apply or leave blank)
[U]: {"id":"1234567890","riskfactors":"Diabetes"}
[C]: The following symptoms have been registered: - Diagnosed diabetes - Increased thirst
[U]: To assess the reported symptoms we need to collect some additional information.
[C]: The following symptoms were reported in similar cases. Do you have any of these symptoms?
[U]: {}
[C]: The following symptoms may be related to potential emergency situations. Do you have any of these symptoms?
[U]: {"choices":"08"}
[C]: Additional symptoms have been identified. So far, the following symptoms have been reported: - Diagnosed diabetes - Increased thirst - Agitation
[C]: Would you like to report anything else?

[U]: Yes
[C]: What else would you like to report?
[U]: **Bad hair**
[C]: Sorry, I did not understand. Please try to describe the symptoms using an alternative phrasing.
[U]: Hair
[C]: Sorry, I did not understand. Please try to describe the symptoms using an alternative phrasing. You can type "skip" to continue to assessment without providing more symptoms.
[U]: Skip
[C]: OK, let's assess the reported symptoms.
[C]: Are your muscles weaker than usual?
[U]: Don't know
[C]: Do you often feel hot or have an intolerance to heat?
[U]: Don't know
[C]: Does your heart unexpectedly start to race or pound, or does it feel like it skips a beat?
[U]: Don't know
[C]: Have you recently been crying, or feeling more nervous than usual?
[U]: Don't know
[C]: Are you sweating more than usual?
[U]: Don't know
[C]: **Summary: Here is what you said** Age: 52 years, Gender: Female Initial complaints: - Increased thirst - Diagnosed diabetes Other Problems or Symptoms: - Agitation

[C]: Suggested Care: **Emergency Department** Here are some possible causes for the problem: 1. Hyperthyroidism [very rare] 2. Hypoglycemia [rare] 3. Graves' disease [very rare] 4. Diabetic ketoacidosis [very rare] 5. Food poisoning [common] 6. Stomach flu [common]
[C]: The reported evidence appears serious and you should go to an emergency department. If you can't get to the nearest emergency department, you should call an ambulance.
Thank you for using the chatbot! Please, do not forget to **fill out the survey**.
[U]: **Thank you. [Can you] call an ambulance for me too?**
[U]: Thank you. Can you call an ambulance for me too?
[C]: I understand. Unfortunately, it seems I can't answer this.
[U]: Thats a pity
[U]: That's a pity
[C]: I understand. Unfortunately, it seems I can't answer this.
[U]: You told me already
[C]: Sorry, I do not understand. Please retry or enter 'start over' to restart the conversation.

[C]: Message by the chatbot

[U]: Message by the user

Sometimes multiple questions are asked at once, such as age and gender. Then, the responses by the user are shown in the form of {"questionID":"input_by_user",...}. If it is just {}, then none of the answer options were chosen.

Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, e.g. [word].

Participant 628, CUQ = 42.2

- User states in comments on user experience field that the chatbot added multiple symptoms that the user did not submit.
- The chatbot received some unexpected input, e.g. "yes" to "what would you like to report".

[U]: begin triage
[C]: Hello! I can help you with identifying **possible causes** for the symptoms you are experiencing. In addition, I can advise you on the **following steps** such as whether it is necessary to visit a doctor.
[C]: But first, what is your **participant number**? This number is provided to you in the survey.
[U]: 618
[C]: What brings you here today? Please describe your symptoms.
[U]: Tiredness, swollen eyes
[C]: I'm sorry to hear that. Before the assessment begins please provide additional information to understand your **medical profile**.
[U]: {"id":"1234567890","age":22,"gender":"F"}
[C]: Thanks, Do you have any of the following **risk factors**? (select all that apply or leave blank)
[U]: {"id":"1234567890"}
[C]: The following symptoms have been registered: - Fatigue - Swelling around the eyes
[C]: To assess the reported symptoms we need to collect some additional information.
[C]: The following symptoms were reported in similar cases. Do you have any of these symptoms?
[U]: {"choices":"02,03"}
[C]: The following symptoms may be related to potential emergency situations. Do you have any of these symptoms?
[U]: {"choices":"02,03"}
[C]: Additional symptoms have been identified. So far, the following symptoms have been reported: - Fatigue - Swelling around the eyes - Headache - Shortness of breath - Confusion after an injury - Fast heartbeat Would you like to report anything else?

[U]: {"choices":"05"}
[C]: I didn’t understand. Please answer ’yes’ or ’no’.
[U]: Yes
[C]: What else would you like to report?
[U]: **Yes**
[C]: Sorry, I did not understand. Please try to describe the symptoms using an alternative phrasing.
[U]: **Nothing elze**
[C]: Sorry, I did not understand. Please try to describe the symptoms using an alternative phrasing. You can type "skip" to continue to assessment without providing more symptoms.
[U]: skip
[C]: OK, let’s assess the reported symptoms.
[C]: **Summary: Here is what you said** Age: 22 years, Gender: Female Initial complaints: - Fatigue - Swelling around the eyes Other Problems or Symptoms: - Headache - Shortness of breath - Confusion after an injury - Fast heartbeat
[C]: Suggested Care: **Emergency Department** Here are some possible causes for the problem: 1. Hypothyroidism [rare] 2. Pink eye [common] 3. Severe allergic reaction [very rare] 4. Cavernous sinus thrombosis [very rare] 5. Acute viral sinusitis [common] 6. COVID-19 [common] 7. Chalazion [less common] 8. Chronic kidney disease [very rare]
[C]: The reported evidence appears serious and you should go to an emergency department. If you can’t get to the nearest emergency department, you should call an ambulance.
[C]: Thank you for using the chatbot! Please, do not forget to **fill out the survey**.

[U]: **ok**
[C]: Sorry, I do not understand. Please retry or enter ’start over’ to restart the conversation.

[C]: Message by the chatbot

[U]: Message by the user

Sometimes multiple questions are asked at once, such as age and gender. Then, the responses by the user are shown in the form of {"questionID":"input_by_user",...}. If it is just {}, then none of the answer options were chosen.

Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, e.g. [word].

Participant 4120: CUQ = 48.8

Problem:

- User stated that the chatbot did not pick up on important keywords such as "sudden" and "unexpected"

[U]: begin triage
[C]: Hello! I can help you with identifying **possible causes** for the symptoms you are experiencing. In addition, I can advise you on the **following steps** such as whether it is necessary to visit a doctor.
[C]: But first, what is your **participant number**? This number is provided to you in the survey.
[U]: 4120
[C]: What brings you here today? Please describe your symptoms.
[U]: pain in the knees
[C]: I'm sorry to hear that. Before the assessment begins please provide additional information to understand your **medical profile**.
[U]: {"id":"1234567890","age":51,"gender":"M"}
[C]: Thanks, Do you have any of the following **risk factors**? (select all that apply or leave blank)
[U]: {"id":"1234567890"}
[C]: The following symptoms have been registered: - Knee pain
[C]: To assess the reported symptoms we need to collect some additional information.
[C]: The following symptoms were reported in similar cases. Do you have any of these symptoms?
[U]: {}
[C]: The following symptoms may be related to potential emergency situations. Do you have any of these symptoms?
[U]: {"choices":"08"}
[C]: Additional symptoms have been identified. So far, the following symptoms have been reported: - Knee pain - Knee pain when moving it
[C]: Reporting additional relevant symptoms is recommended. If there are other symptoms please report them, otherwise type "none" to begin the assessment.
[U]: **sudden** pain [shots]
[C]: Sorry, I did not understand. Please try to describe the symptoms using an alternative phrasing.

[U]: **sudden** pain in both knees
[C]: It sounds like you would like to report the following symptoms: - Knee pain
[C]: Is that correct?
[U]: Yes
[C]: No additional symptoms have been registered. So far, the following symptoms have been reported: - Knee pain- Knee pain when moving it
[C]: If there are other symptoms please report them, otherwise type "none" to begin the assessment.
[U]: none
[C]: OK, let's assess the reported symptoms.
[C]: Do you have a high BMI, above 30, i.e., you weigh too much in relation to your height?
[U]: Yes
[C]: Did your joint pain start suddenly?
[U]: Yes
[C]: Have you recently had any trauma or physical injury?
[U]: No
[C]: Does your joint pain worsen in cold or damp weather?
[U]: No
[C]: Does your joint hurt when you touch or press it?
[U]: No
[C]: Does the joint pain worsen with movement and weight-bearing, and is relieved by rest?
[U]: No
[C]: Do any of your toes hurt?
[U]: No
[C]: Do your joints crack or creak when you move?
[U]: No
[C]: Do you have pain in both of your lower limbs, e.g., both legs or feet?

[U]: No
[C]: Do either of your thumbs hurt?
[U]: No
[C]: Are you walking slower than usual?
[U]: No
[C]: **Summary: Here is what you said** Age: 51 years, Gender: Male Initial complaint: Knee pain Other Problems or Symptoms: - Knee pain when moving it - Sudden joint pain - Obesity
[C]: Suggested Care: **Self-Care at Home** Here are some possible causes for the problem: 1. Unspecific joint pain [common]
[C]: A medical consultation is not strictly required; you should observe your symptoms and consult a doctor only if symptoms worsen within 24 hours.
[C]: Thank you for using the chatbot! Please, do not forget to **fill out the survey**.

[C]: Message by the chatbot

[U]: Message by the user

Sometimes multiple questions are asked at once, such as age and gender. Then, the responses by the user are shown in the form of {"questionID":"input_by_user",...}. If it is just {}, then none of the answer options were chosen.

Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, e.g. [word].

Participant 2113, CUQ = 50,0

No problems found in the conversation, however the user says "I don't like you" at the end. There are no comments.

[U]: begin triage
[C]: Hello! I can help you with identifying **possible causes** for the symptoms you are experiencing. In addition, I can advise you on the **following steps** such as whether it is necessary to visit a doctor.
[C]: But first, what is your **participant number**? This number is provided to you in the survey.
[U]: 2113
[C]: What brings you here today? Please describe your symptoms.
[U]: headache
[C]: I'm sorry to hear that. Before the assessment begins please provide additional information to understand your **medical profile**.
[U]: {"id":"1234567890","age":35,"gender":"M"}
[C]: Thanks, Do you have any of the following **risk factors**? (select all that apply or leave blank)
[U]: {"id":"1234567890","riskfactors":"Diabetes"}
[C]: The following symptoms have been registered: - Diagnosed diabetes - Headache
[C]: To assess the reported symptoms we need to collect some additional information.
[C]: The following symptoms were reported in similar cases. Do you have any of these symptoms?
[U]: {"choices":"07"}
[C]: The following symptoms may be related to potential emergency situations. Do you have any of these symptoms?
[U]: {"choices":"05"}
[C]: Additional symptoms have been identified. So far, the following symptoms have been reported: - Diagnosed diabetes - Headache - Muscle pain - Depressed mood
[C]: Would you like to report anything else?
[U]: No

[C]: OK, let's assess the reported symptoms.
[C]: How strong is your headache?
[U]: Mild
[C]: Do you have a sore throat?
[U]: No
[C]: Is your neck twisted, and does it hurt when you try to straighten it?
[U]: Yes
[C]: Have you recently had any trauma or physical injury?
[U]: No
[C]: Where is your headache located?
[U]: Back of the head
[C]: Have you often had similar headaches within the last three months?
[U]: No
[C]: **Summary: Here is what you said** Age: 35 years, Gender: Male Initial complaints: - Headache - Diagnosed diabetes Other Problems or Symptoms: - Muscle pain - Depressed mood - Mild headache - Twisted neck - Pain in back of head - Headaches history shorter than 3 months
[U]: Suggested Care: **Doctor today** Here are some possible causes for the problem: 1. Neck strain [common]
[C]: You should see a doctor within 24 hours. If the symptoms suddenly get worse, you should go to the nearest emergency department.
[C]: Thank you for using the chatbot! Please, do not forget to **fill out the survey**.
[U]: **i [don't] like you**

[U]: i don’t like you
[C]: I understand. Unfortunately, it seems I can’t answer this.
[U]: **who are you**
[C]: Sorry, I do not understand. Please retry or enter ’start over’ to restart the conversation.

[C]: Message by the chatbot

[U]: Message by the user

Sometimes multiple questions are asked at once, such as age and gender. Then, the responses by the user are shown in the form of {"questionID":"input_by_user",...}. If it is just {}, then none of the answer options were chosen.

Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, e.g. [word].

Participant 4113, CUQ = 50.0

Conversation seemed to run fine. The user did not leave any comments on user experience or responses.

[U]: begin triage
[C]: Hello! I can help you with identifying **possible causes** for the symptoms you are experiencing. In addition, I can advise you on the **following steps** such as whether it is necessary to visit a doctor.
[C]: But first, what is your **participant number**? This number is provided to you in the survey.
[U]: 4113
[C]: What brings you here today? Please describe your symptoms.
[U]: stomach aches
[C]: I'm sorry to hear that. Before the assessment begins please provide additional information to understand your **medical profile**.
[U]: {"id":"1234567890","age":23,"gender":"M"}
[C]: Thanks, Do you have any of the following **risk factors**? (select all that apply or leave blank)
[U]: {"id":"1234567890","age":23,"gender":"M"}
[C]: The following symptoms have been registered: - Stomach pain
[C]: To assess the reported symptoms we need to collect some additional information.
[C]: The following symptoms were reported in similar cases. Do you have any of these symptoms?
[U]: {"id":"1234567890"}
[C]: I didn't understand. Please select all that apply.
[U]: {"choices":"02,03"}
[C]: The following symptoms may be related to potential emergency situations. Do you have any of these symptoms?
[U]: {}
[C]: Additional symptoms have been identified. So far, the following symptoms have been reported: - Stomach pain - Bloating - Diarrhea
[C]: Would you like to report anything else?

[U]: No
[C]: OK, let's assess the reported symptoms.
[C]: How long have you had diarrhea?
[U]: Less than 48 hours
[C]: Where is your stomach pain located?
[U]: All over the abdomen
[C]: How long has your stomach pain lasted?
[U]: Less than 2 days
[C]: Have you recently had any trauma or physical injury?
[U]: No
[C]: Do you experience any gastric symptoms shortly after you eat or drink dairy products, e.g., milk, ice cream, or cheese?
[U]: No
[C]: Do you have a fever?
[U]: No
[C]: How strong is your stomach pain?
[U]: Mild
[C]: How did your stomach pain start?
[U]: It came on gradually and reached its peak in several hours
[C]: Do you have muscle pain, aches, or soreness?
[U]: No
[C]: How would you describe your stomach pain?
[U]: Crampy
[C]: Have you eaten or drunk something that might not have been fresh?
[U]: Don't know

[C]: **Summary: Here is what you said** Age: 23 years, Gender: Male Initial complaint: Stomach pain Other Problems or Symptoms: - Bloating - Diarrhea - Diarrhea lasting less than 48 hours - Widespread stomach pain - Stomach pain, lasting less than two days - Mild stomach pain - Slow and gradual stomach pain - Crampy stomach pain
[C]: Suggested Care: **Self-Care at Home** Here are some possible causes for the problem: 1. Stomach flu [common] 2. Food poisoning [common] 3. Indigestion [common] 4. Giardiasis [very rare]
[C]: A medical consultation is not strictly required; you should observe your symptoms and consult a doctor only if symptoms worsen within 24 hours.
[C]: Thank you for using the chatbot! Please, do not forget to **fill out the survey**.

[C]: Message by the chatbot

[U]: Message by the user

Sometimes multiple questions are asked at once, such as age and gender. Then, the responses by the user are shown in the form of {"questionID":"input_by_user",...}. If it is just {}, then none of the answer options were chosen.

Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, e.g. [word].

Appendix G: Comments on User Experience

Tips	Praise
It would be nice if the chatbot would give some **tips** or things to do in the meantime until you go to the doctor, as now it only states the (obvious) cause and that you may need a doctors appointment.	Nice chatbot. Good way of asking for additional symptoms.
It was difficult to **memorize the available commands** that show up with the help command. Then I also had to type them in myself instead of clicking on one that I want. This took time and I wanted to make sure to have no spelling mistake (because I thought the bot would not recognize my command then)	It liked using the chatbot! Nice job Anne!
The bot missed **keywords** like "sudden" and "unexpected", which seem like indications that should have been investigated more with more questions.	It was a great experience overall.
It went **too fast**	The input was **correctly registered** by the chatbot which was very nice.
It did **not understand** everything I said and when I misclicked and wanted to tell the chatbot to continue it got confused.	Easy to use
Response seemed **slow** Order of questions strange	It was a very nice experience. I liked that it was quite **quick and to the point** whilst still feeling friendly.
A thing that can be improved is the ability to **handle small orthographic mistakes**. For example, I wrote panic(k) a(t)tack and the chatbot didn't recognize the symptom	Really nice idea and implementation, very easy to **navigate** and much easier than searching google for all of your symptoms. All in all a really nice chatbot, with some improvements on the possible causes and recommendations I think it can be of great help to users as well as to doctors :)
Triage is a **difficult word**, maybe replace it with something simpler? Same for some other medical terms, I'd try to keep it simple	I like it! I filled in some fictious things and liked its conclusion
It was basically just a checklist of symptoms, **doesn't really need a chatbot**	The overall experience was really nice, I was also impressed by the overall **medical knowledge** of the chatbot
Sometimes there were **4 messages at once** and I wasn't sure whether I should answer all checklists	Easy to use, structured and efficient

I have tested the chatbot and ran into two issues. 1. If no points are checked in the Medical History an **error** occurs. 2. The bot did not respond to the command **"start over"**.

Sometimes an **icon** popped up as if it was typing, but it wasn't.

The **name 'chatbot'** feels a bit robotic. I think another more personal name would help with a better experience.

When none of the options the [chatbot] provided was relevant, I searched for a checkbox option with the text **'none of these'**. When this option wasn't there, I was a bit confused by whether I should just click 'continue', without checking one (or multiple) of the boxes.

The first time I tried to use the chatbot, it added a lot of symptoms, without me clicking these. It "**glitched**" and my answers to its questions were not processed right. I had to begin again, but this time it did not behave weird.

Some questions are too difficult

A bit **too long**

What to do when **none of the symptoms** occurs

Whilst it picked up on symptoms individually very well, I did notice that it did not pick up on the **time factor** of symptoms and **possible triggers** of symptoms. It did get to the end diagnosis correctly (Lactose Intolerance) but only picked up the trigger of the symptom after it had ruled out other possibilities when it could have picked that up earlier when I mentioned it in the description.

The possible causes are still a bit vague. **More in-depth solutions** would be nice than just a 'doctor routine check-up'. Also, it can ask if you've already been at the doctor, and if that was of any help or not, so that it does not recommend another doctor routine check-up.

Added multiple symptoms that I did not submit myself

I liked the **checklist** to describe the symptoms. This prevented me from struggling to try to describe my symptoms in an understandable way.

It's quite **easy to use** and nice to have advice so quickly. Being able to check symptoms makes it easier to explain your problem

Nice to use it

I liked the **structured approach** of the chatbot. It was really seamless and easy to use.

Adjustments to spelling or grammar mistakes are noted by placing rectangle brackets, e.g. [word].

The comments are separated into praise or tips. Comments placed next to each other are not necessarily from the same participant.

Appendix H: Comments on Responses

Tips	Praise
Could sometimes be a **bit more ["human"]**, but overall really good.	Very clear and understanding responses. Especially the **summary of all the answers** and the symptoms until that time was a good experience.
The option answers came **too quickly**. I got multiple 'select all that apply/select an option' at the same time so it was confusing to answer them; I could not keep track of the dialogue.	Clear responses
Bot asks for risk factors, then asks to confirm symptoms. I answered 'No', because the risk factors are not the symptoms I (imaginatively) have or call for and then the chat bot got stuck. I asks to type **'start over' and nothing happens**	They were somewhat accurate based on the input
Some answers and options provided by the chatbot **were too long** and could therefore not be displayed correctly (sentences stopped too quickly..)	The chatbot seemed to understand my responses
It still feels a bit **scary** that it says what's wrong with my health based on some simple textual input that I provide. Maybe I can't find the right words.	I really liked that the chatbot **repeated the symptoms** that I put in because I can imagine that people might be skeptical whether the chatbot understood them correctly. Repeating the symptoms allows people to be reassured that they were indeed understood and allows for corrections if necessary. I think that this will increase trust in the chatbot, which seems very relevant given the topic of health.
I would advise **adding a suicide line** in case of depression-like symptoms (self-harm), but this may be outside of the scope of this research.	Very **analytical and factual**, but at least I know what I have and what to do very clearly
Maybe let the chatbot reassure you of the privacy involved	[A good structure of questions. A great foundation for further development.]
A chatbot feels to **unpersonal** when you're experiencing **mental issues**. The questions about self harm and potential suicide are quite harsh and should be asked with more care.	Nope. Good job!
A nice addition would be to have a **end message** or response to an end message.	Clear, well explained and very valuable that the chatbot summarizes responses and asks for **confirmation of input**. Clear instructions in the end that guide you on what to do next + helpful that the bot asks whether there are any additional comments

I like using a chatbot that is **not too human**, as you intuitively adapt your own lines into a more simplistic way so the bot can understand. I like that it gives you buttons and that it understands what I say.

Useful that at the diagnoses it was shown if the medical condition was **rare** or not. This could help in determining what the most likely problem is.

Clear responses, the answers were easy to understand

Really clear responses, works well!

Adjustments to spelling or grammar mistakes, or translations are noted by placing rectangle brackets, e.g. [word].

The comments are separated into praise or tips. Comments placed next to each other are not necessarily from the same participant.