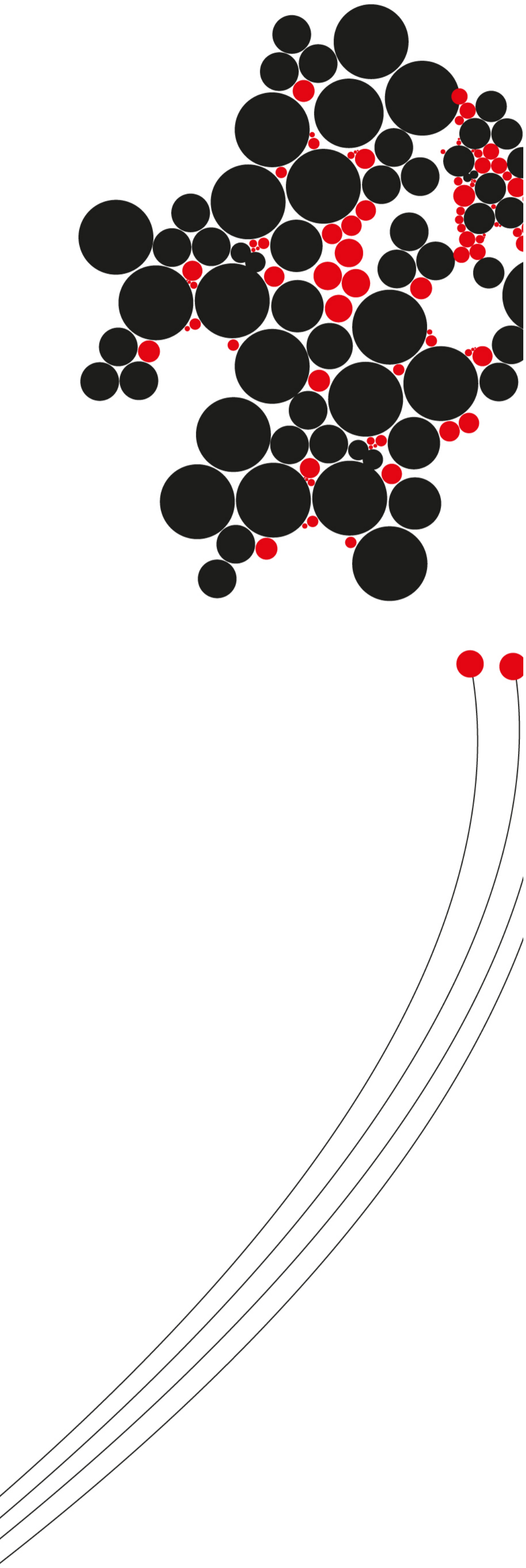


Master Thesis Interaction Technology

A CLOSED DOMAIN QUESTION ANSWERING SYSTEM FOR A GENETIC COUNSELOR WITHIN THE ECG FAMILY CLINIC

Maaike Keurhorst

Supervised by:
Randy Klaassen
Lorenzo Gatti
Marcus Gerhold

Faculty of Electrical Engineering,
Mathematics and Computer Science (EEMCS)

UNIVERSITY OF TWENTE.

6TH JULY 2023

Acknowledgements

This thesis has been a long process and one which was not always smooth sailing. Many people were involved in getting to this stage and I would sincerely like to thank all of them.

Firstly, I would like to express my appreciation to my supervisors Randy Klaassen, Lorenzo Gatti, and Marcus Gerhold for their guidance and support. Their feedback and insights throughout my research helped me shape this thesis and its contents. I am really grateful for all the time and effort you spent on our meetings and giving extensive feedback.

Secondly, I want to thank Jan Kolkmeier for helping with the integration of my algorithm in the eCG Family Clinic. Additionally, with the help of Thijs Dortmann I was able to set up a server to host the application. Moreover, he was always there to listen, motivate me and help me brainstorm. Furthermore, many thanks to Sybille Franken for her feedback and text editing.

Also, I thank all my study buddies for their support, input and the interesting discussions we have had. With a special thanks to Michael Bui who was always willing to help out with tips and tricks or to just be there to lend an ear.

Lastly, I would like to thank my friends, family and all people who contributed in any other way or participated in my user testing.

Abstract

Background: The eCG Family Clinic is a platform developed for online genetic counseling to help family members make informed decisions regarding the testing of hereditary cardiovascular diseases.

Objective: This research explores how an open question answering feature can be developed for a virtual assistant for genetic counseling. We aim to answer this by: 1) exploring which techniques for question answering systems could be used and 2) examining the users' attitude towards this open question answering feature.

Method: Using literature, a retrieval-based approach was selected. This approach aimed to detect similarities between user input and potential answers in the knowledge base. Named entities, shared bigrams, semantic similarity, and shared part-of-speech tags are used as features to detect the similarity. The knowledge base consisted of excerpts extracted from websites approved by specialists. Additionally, the prototype was evaluated with users by an qualitative study guided by the Technology Acceptance Model (TAM). User evaluation consisted of a set of tasks using the eCG Family Clinic and a semi-structured interview. The user study was analysed using a thematic analysis. In total 8 participants, all master's students, were recruited.

Results: We found that the algorithm had a top-1, top-2, and top-3 accuracy of 19.2%, 25.3%, and 26.3%. This accuracy is lower than related open question features found in the literature. Thematic analysis using the themes of 'Perceived Usefulness', 'Dialogue Management', and 'Ease of Use' showed us that participants see the value of the addition of the open question feature. However, more accuracy in this feature is needed to increase its usefulness.

Conclusion: While the accuracy of the methods needs to be improved, this study forms a foundation for future research. User testing demonstrated that users recognized the value and potential of the open question feature but indicated that an increase in accuracy and personalization is needed. For future research, it is suggested to focus on increasing the accuracy and addressing user suggestions.

Contents

1 Introduction 9

2 Background 11

2.1 ECG Family Clinic 11

2.2 Hypertrophic Cardiomyopathy 16

2.3 Conversational Agent 16

2.3.1 Categorization of QA Systems 17

2.3.2 Related QA Systems 19

2.4 Conclusion 24

3 Design Requirements 25

4 Design 27

4.1 Algorithm of the Open Question Answering Feature 27

4.1.1 Knowledge Source 28

4.1.2 Content Word Similarity 29

4.1.3 Structure Similarity 31

4.1.4 Context Similarity 31

4.1.5 Answer Selection 32

4.2 Implementation in the eCG Family Clinic 32

5 User Evaluation 37

5.1 Theoretical Framework 37

5.2 Methodology 38

5.3 Data Collection 40

5.4 Data Analysis 41

6 Results 43

6.1 Thematic Analysis 43

6.1.1 Perceived Usefulness 45

6.2 Dialogue Management . . . 50
6.2.1 Engagement . . . 50
6.2.2 Dialogue Flow . . . 51
6.2.3 Ease of Use . . . 52
6.3 Algorithm Accuracy . . . 54
6.4 Discussion of the Results . . . 56
6.4.1 Thematic Analysis . . . 56
6.4.2 Algorithm Accuracy . . . 58
6.5 Implementation of Requirements . . . 58

7 Discussion . . . 61
7.1 Techniques for QA Systems . . . 61
7.1.1 Limitations . . . 62
7.1.2 Future Work . . . 63
7.2 Users’ Attitude Towards the QA Feature . . . 64
7.2.1 Limitations . . . 64
7.2.2 Future Work . . . 66
7.3 Relevance . . . 66

8 Conclusion . . . 69

Bibliography . . . 71

A Research Protocol . . . 75

B Informed Consent Form . . . 83

C Information Form . . . 85

1. Introduction

Cardiovascular disease is the cause of many deaths throughout the world. In the United States, 39 percent of all deaths can be led back to cardiovascular diseases [28]. In the Netherlands, this was 24.6 percent of all deaths in 2019 [35]. The proper medication and treatment could have prevented many of these deaths. Cardiovascular diseases are heritable at a high rate [25]. Hypertrophic cardiomyopathy (HCM), for instance, is hereditary with a 50% chance of passing the disease from parent to child [12, 14, 15].

At risk family members diagnosed with HCM in the Netherlands are asked to inform family members by giving them flyers with information. These family members will then get an appointment with the genetic counselor to discuss risks and possible treatments. After testing for cardiovascular diseases, patients with positive test results can start medication and treatments if needed. However, less than half of the family members of patients diagnosed with heritable cardiovascular diseases start this genetic counseling and testing process.

The eCG Family Clinic is a project currently in development that aims to increase the informed decision-making of family members in testing for cardiovascular diseases. The project is a cooperation between the University Medical Center Utrecht (UMCU), KAESS, and Human Media Interaction (HMI) at the University of Twente. To support the informed decision-making of family members, information is offered through a virtual environment, a home DNA test, and appointments when mutations in the DNA are detected. This virtual environment includes a virtual assistant, which will allow family members to ask questions and receive personalized counseling. The agent will use preset conversation options to communicate with the user. While this method is reliable for a conversational agent, it is also limited since it takes much manual programming [13]. Additionally, using preset conversation options limits the information a user can receive. An additional function to expand the use case of the virtual assistant might be beneficial within the eCG Family Clinic virtual assistant. With this function, the agent could answer open questions from the user.

This additional function still needs to be able to reliably answer a user's question. For example, machine learning-based methods do not always return correct answers. Firstly, these methods train with large amounts of data. These datasets are often generic since, otherwise, the dataset will need to be even larger. Additionally, the information we want to give a user should always be correct. The answers the agent would need to return should adhere to preapproved information sources since we know these will contain the correct information. Additionally, this research will examine the extent to which a user would accept the open question answering module. These goals give us the following main research question:

> RQ: How to design an open question answering feature for a virtual assistant in a genetic counseling platform?

To answer this question, we have formulated multiple sub-questions. This research will consist of two parts. First, the open question answering function needs to be developed. Next, this function will be implemented to examine how it affects user satisfaction. This brings us to the following two subquestions:

> RQ1: What techniques for question answering systems can be used to develop a retrieval-based agent for open question answering in the context of a genetic counseling virtual assistant?
>
> RQ2: What is the users' attitude towards the open question answering feature in a genetic counseling virtual assistant?

This thesis is structured as follows. Chapter 2 will explain what the eCG Family Clinic, HCM, and question answering conversational agents are. Additionally, often-used techniques are discussed together with examples from literature. In Chapter 3 the design requirements are shown. Next, Chapter 4 will explain how the open question answering feature and its implementation are realised. The design of the user evalaution is discussed in Chapter 5 and followed by the results in Chapter 6. In this same chapter a discussion of those results can be found, while Chapter 7 discusses the overal limitations and future work of the thesis. The last chapter is the conclusion in Chapter 8.

2. Background

To determine how an open question answering feature can be built we first need some background on the project and the existing question answering systems. This chapter will first decribe what the eCG family clinic project entails. Secondly, we will look at Hypertrophic Cardiomyopathy (HCM, a cardiovascular disease the project will mainly target). An explanation of question-and-answer conversational agents will follow. Lastly, we will discuss often-used techniques observed within conversational question-and-answering agents together with examples from literature.

2.1 ECG Family Clinic

In this thesis, a function is added to the eCG virtual assistant in order to support open question answering. To understand the goal of the agent and the constraints the agent has to deal with, it is essential to know what the eCG family clinic entails.

The eCG family clinic (e-Cardiovascular Genetics Family Clinic) contains multiple elements: (1) a virtual environment that offers accessible and personalized information on cardiovascular diseases and genetic testing, (2) a home DNA test for HCM, and (3) cardiological checks if the mutation is found in a patient.

Currently, a patient receives a letter that can be handed over to relatives. This letter includes information such as the symptoms, the testing, the consequences of the testing procedure, the treatment, and more. After reading this letter, the at risk family member can participate in a genetic counseling session to support their decision-making on predictive DNA testing and regular check-ups. Next, a family member can let their DNA be tested through a blood sample if the decision has been made. The result of this test will be communicated by phone and/or letter.

In the proposed workflow by the project, at risk family members will receive a login code for the eCG family clinic platform where they can fill in the information of their relatives. This will ensure that also the relatives will receive a login to access the virtual environment. This environment is currently in development, and current plans suggest that this entire environment will consist of a virtual assistant who will walk the user through all the information they would need for making a well-advised decision whether they want to have a DNA test or regular check-ups. If a family member chooses to take this predictive DNA test, they can take it at home with a swab kit. The results will be sent via an automatically generated letter through the virtual environment. If the family member does have the genetic mutation, they will be invited for regular cardiological checks. Then the flow will restart with relatives from that family member. This new workflow aims to increase the informed decision-making regarding DNA testing by 40 % and cardiogenetic care by 70%. The aim is that this proposed workflow will replace the current workflow by about 70%. Important to note is that within the proposed workflow, there is always a possibility to talk to a human or fall back on the current workflow. Family members and patients with low digital skills, low health literacy, or illiteracy might still use the current workflow.

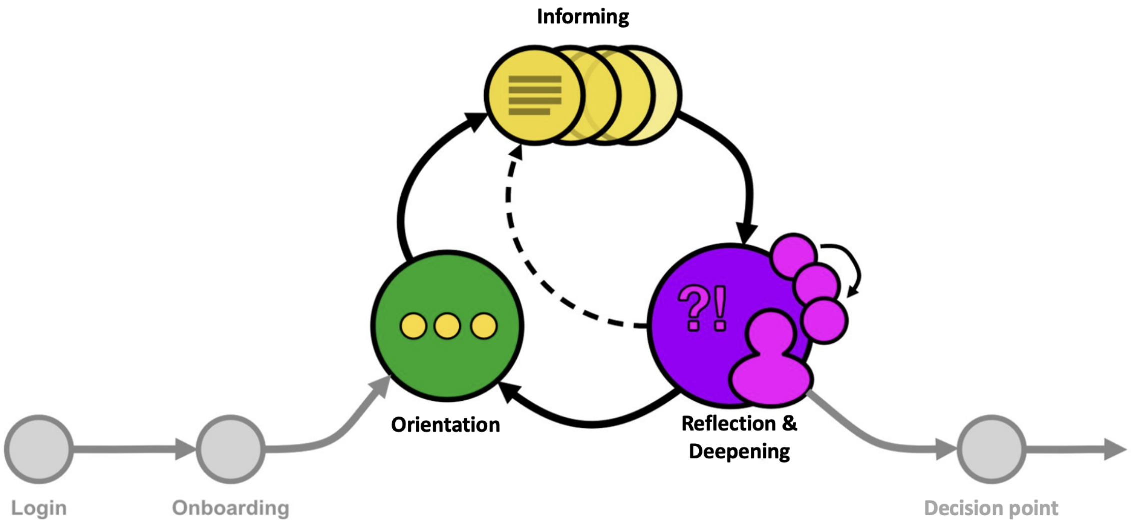
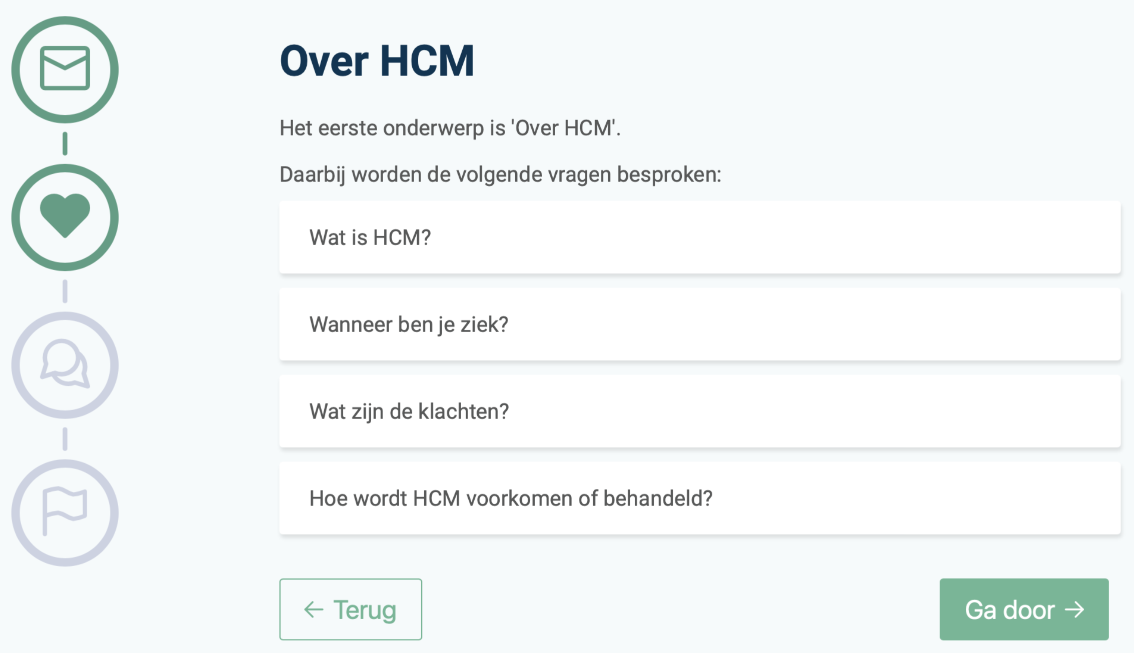


Figure 2.1: Interaction Flow in the eCG virtual assistant showing the steps of orientation, informing, and reflecting & deepening. Adapted from Mid-Term update slides HMI (University of Twente).

With experts, content has been developed that should be implemented in the virtual assistant, called Anna. This content is split into multiple categories in the virtual environment. Each category is discussed by the virtual assistant before moving to the next. In Figure 2.1, the flow of such an interaction can be seen. Currently, a limited prototype of this interaction flow is programmed and implemented in the eCG Family Clinic, which includes the topics 'Symptoms' and 'HCM & Sport.' These are also the topics which will be considered in this thesis.

The interaction flow starts with a login and onboarding. Here users get an introduction to the application, explaining what the platform offers. Once this general introduction is done, the categories are handled one by one.



Over HCM

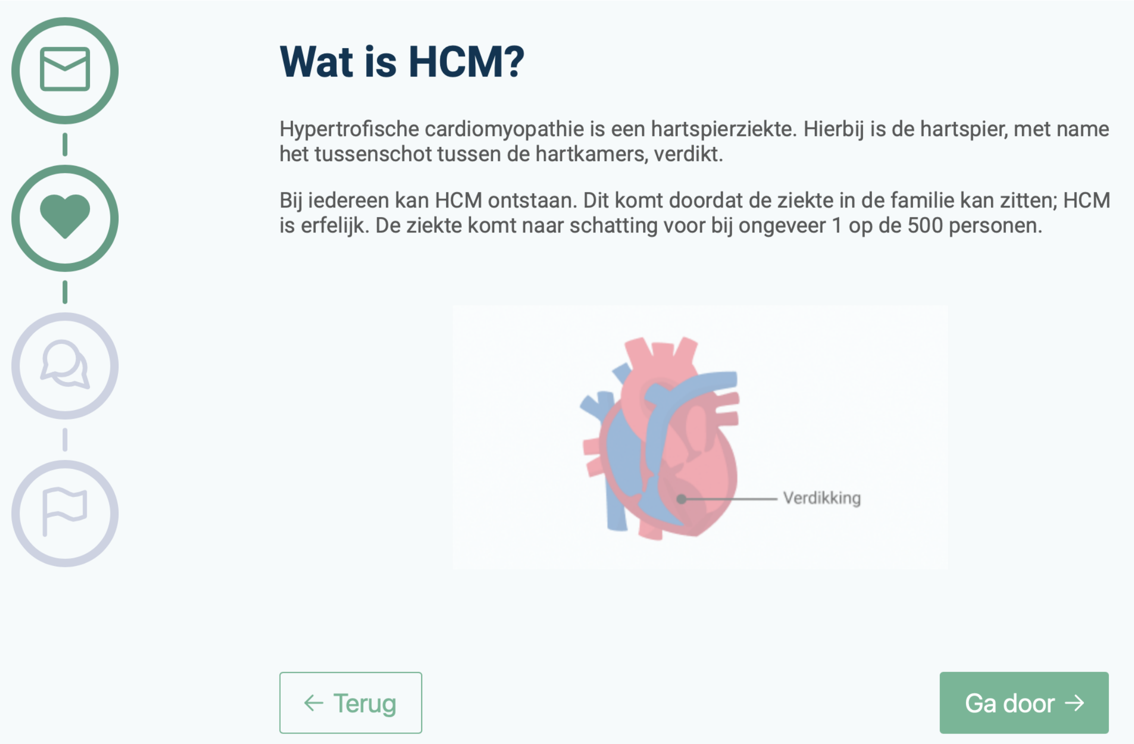
Het eerste onderwerp is 'Over HCM'.

Daarbij worden de volgende vragen besproken:

- Wat is HCM?
- Wanneer ben je ziek?
- Wat zijn de klachten?
- Hoe wordt HCM voorkomen of behandeld?

Figure 2.2: An orientation screen in the eCG virtual assistant.

The first screen a user sees is an orientation screen (cf. Figure 2.2), which shows the subjects and topics to discuss. The first subject, "About HCM" is introduced with the topics "What is HCM?", "When are you sick?", "What are the symptoms?" and "How is HCM treated?". If a user presses continue, the application will continue with the informative screens.



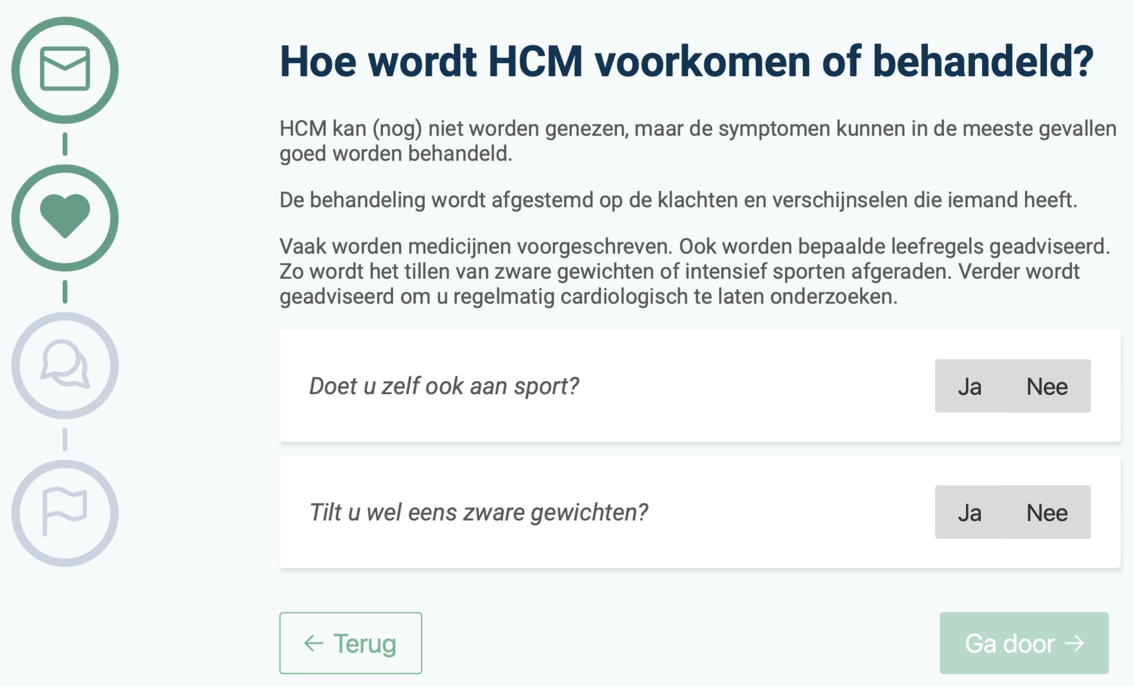
Wat is HCM?

Hypertrofische cardiomyopathie is een hartspierziekte. Hierbij is de hartspier, met name het tussenschot tussen de hartkamers, verdikt.

Bij iedereen kan HCM ontstaan. Dit komt doordat de ziekte in de familie kan zitten; HCM is erfelijk. De ziekte komt naar schatting voor bij ongeveer 1 op de 500 personen.

Figure 2.3: An information screen in the eCG virtual assistant.

In the informative screens, each topic is explained. Figure 2.3 shows an example of such an informative screen. In this screen, HCM and its symptoms are described. There can be multiple informative screens to discuss all the topics in a category.



Hoe wordt HCM voorkomen of behandeld?

HCM kan (nog) niet worden genezen, maar de symptomen kunnen in de meeste gevallen goed worden behandeld.

De behandeling wordt afgestemd op de klachten en verschijnselen die iemand heeft.

Vaak worden medicijnen voorgeschreven. Ook worden bepaalde leefregels geadviseerd. Zo wordt het tillen van zware gewichten of intensief sporten afgeraden. Verder wordt geadviseerd om u regelmatig cardiologisch te laten onderzoeken.

Doet u zelf ook aan sport? Ja Nee

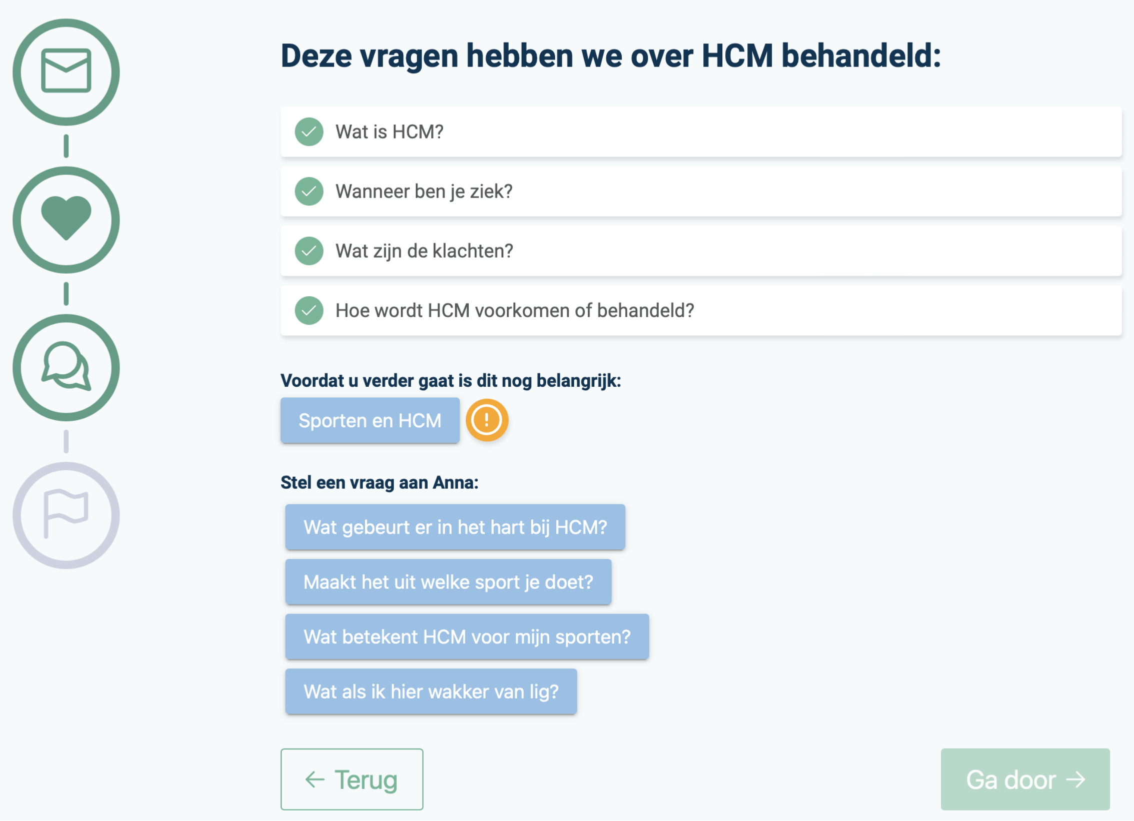
Tilt u wel eens zware gewichten? Ja Nee

Figure 2.4: An information screen with questions to the user in the eCG virtual assistant.

At the last informative screen (cf. Figure 2.4), regarding the treatment of HCM, several questions are asked by the virtual environment to the user. The first question is if the user practices sports, while the second question targets potential heavy weight lifting by the user. Users can answer these questions with a yes or no. The answers to these questions determine the follow-up questions within a mandatory conversation in the reflection & deepening screen.

There are currently two different designs for the last type of screen (cf. Figure 2.5). The goal of this screen is to answer questions a user might still have and to ask follow-up questions if the questions in the previous screen triggered a 'flag'. For example, intensive sports can increase the risks of dangerous symptoms, so if a user indicates to practice sports, the eCG system would ask more in-depth questions to examine the extent of the risk and pass this to a doctor.

In the reflection & deepening screen, the virtual assistant appears for the first time. In the first design (cf. Figure 2.5a), there are buttons with potential questions and a mandatory button the user must press when a flag is triggered. All buttons lead the user to a screen where the virtual assistant is present and where they will interact through a closed-button option with the assistant regarding the topic they just selected. The second design (cf. Figure 2.5b) leads the user immediately to the conversation with Anna, where within that conversation, they can press different buttons to steer the conversation and ask their questions. For both designs, a user has to have a conversation about the potential flags they have triggered before continuing with the next category.



Deze vragen hebben we over HCM behandeld:

✓ Wat is HCM?

✓ Wanneer ben je ziek?

✓ Wat zijn de klachten?

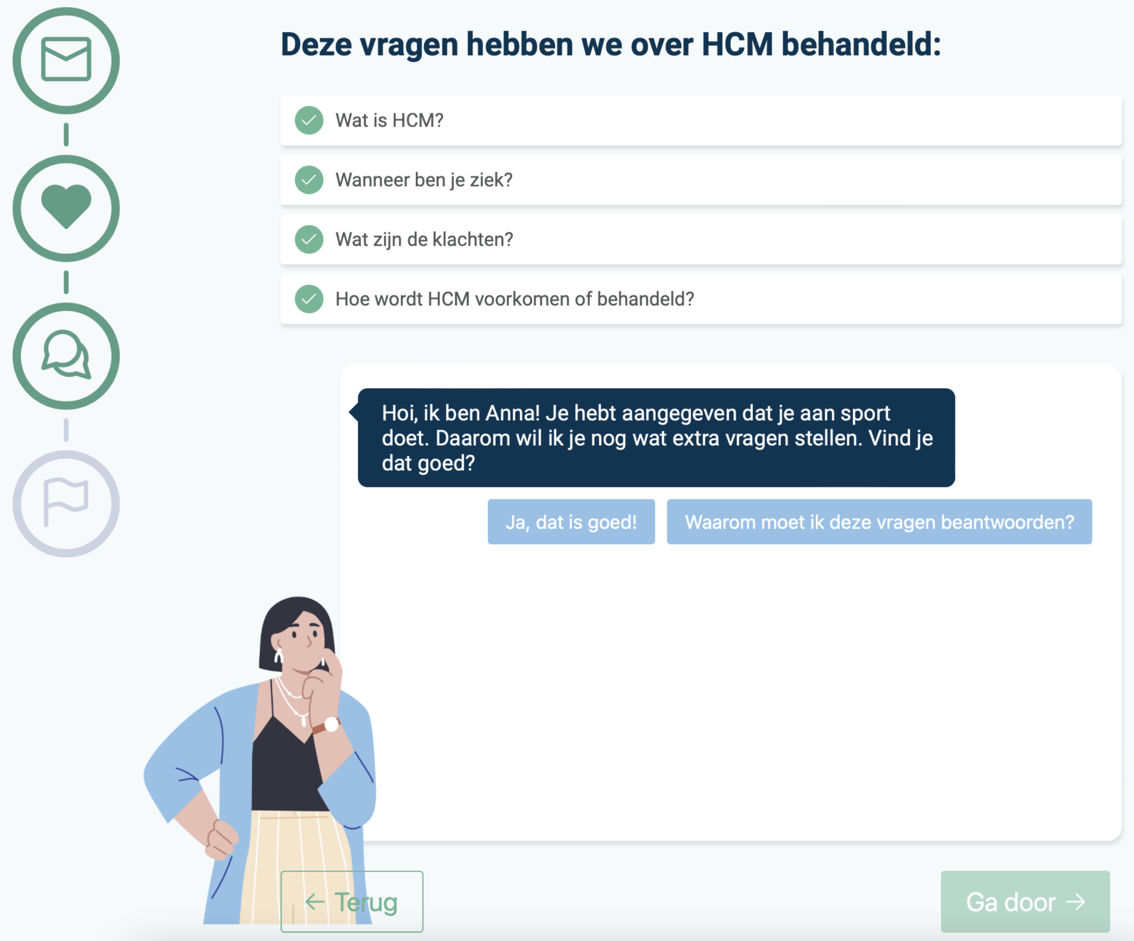
✓ Hoe wordt HCM voorkomen of behandeld?

Voordat u verder gaat is dit nog belangrijk:

Stel een vraag aan Anna:

Sporten en HCM
!
Wat gebeurt er in het hart bij HCM?
Maakt het uit welke sport je doet?
Wat betekent HCM voor mijn sporten?
Wat als ik hier wakker van lig?
← Terug
Ga door →

(a) A reflection & deepening screen with a button type design in the eCG virtual assistant.



Deze vragen hebben we over HCM behandeld:

✓ Wat is HCM?

✓ Wanneer ben je ziek?

✓ Wat zijn de klachten?

✓ Hoe wordt HCM voorkomen of behandeld?

Hoi, ik ben Anna! Je hebt aangegeven dat je aan sport doet. Daarom wil ik je nog wat extra vragen stellen. Vind je dat goed?

(b) A reflection & deepening screen with a conversation design in the eCG virtual assistant.

Figure 2.5: Two designs of the reflection & deepening screens.

2.2 Hypertrophic Cardiomyopathy

While the eCG Family Clinic will initially target the cardiovascular disease hypertrophic cardiomyopathy (HCM), the possibility to extend the application to other cardiovascular diseases should be kept open. For this reason, this thesis will mainly focus on HCM while keeping in mind that the proposed techniques should be adaptable to more types of cardiovascular diseases.

Genetic hypertrophic cardiomyopathy is a cardiovascular disease where changes in genes cause a thickening of the heart muscle, which can make it more difficult for the heart to pump blood around [12, 14, 15]. Genetic HCM is hereditary with a 50% chance of passing the disease from parent to child [12, 14, 15].

Symptoms of HCM can include chest pain, vertigo, fainting, heart murmur, the sensation of fast, fluttering or pounding heartbeats, reduced fitness and stamina, and shortness of breath [12, 14, 15]. All these symptoms are more likely to occur during exercise or, in case of fainting, also just after exertion [15]. The complications which a patient can encounter are fast or irregular heartbeats, an increased risk of developing blood clots, a blocked blood flow leaving the heart, mitral valve disease, dilated cardiomyopathy, heart failure, fainting, and sudden cardiac death [12, 14, 15].

While it is not possible to cure or prevent HCM, there are treatments available [12, 14, 15]. A patient can, for instance, get prescribed medicine and will get the advice to live healthy [12, 14]. In the case of dangerous symptoms or severe cases, an implantable cardioverter defibrillator, a supportive heart, or a heart transplant might be required [12, 14]. Not every patient with the HCM gene mutation will experience symptoms [14]. Some patients will never experience any symptoms throughout their lives. Through the examination of the heart with an ultrasound or MRI, it can be determined what the risk is that a patient will experience dangerous symptoms [12, 14]. This cardiologic investigation can be repeated regularly to determine if treatment is needed [12, 14].

Since HCM is hereditary, family members of a patient are at risk of getting the disease. To determine if a family member has the gene mutation, genetic testing can be done [12, 14, 15]. In the Netherlands, this testing is covered by insurance, except for some budget health insurances [14]. However, the testing does use the family members' deductible excess [14]. Unfortunately, not all gene mutations causing HCM can be found thus far [12, 14, 15]. While with testing, a family member will know for sure whether they are a carrier and can start discussing a treatment plan with their health care professional, there are also downsides to being tested. A patient can, for instance, feel less healthy if HCM is detected. Additionally, insurance or mortgage applications might be affected. These applications can require the disclosure of HCM [11, 37]. Therefore, it is essential that within the eCG family clinic, users are adequately informed on the pros and cons of getting tested and undergoing regular check-ups to support decision-making.

2.3 Conversational Agent

Conversational agents are often used within e-health and education. Some examples are: an agent for automatic cancer genetic counseling [45], health education and counseling regarding breastfeeding [26], affectively aware depression counseling [32], or atrial fibrillation education and counseling [22]. Many of these agents are limited in their functionality.

Often the agent is in charge of the conversation, and the user can only interact through restricted options [9, 19, 26, 31, 32, 45], which is also the case in the current eCG Family Clinic. This could be because conversational agents within the medical field have to deal with certain restrictions since the agents need to be specific and mindful of the information provided to the user. However, in this research, we will explore a mixed dialogue initiative where the user can take the initiative.

There are multiple types of conversational agents and multiple ways to categorize them. Agents can, for instance, be categorized on their tasks or the used approach. For the latter, Singh and Beniwal [34] present three approaches. The first approach is rule-based methods, where a set of predetermined rules and patterns leads to responses. The second method is generative-based. In this method, NLP algorithms are used to understand the inquiry, gather the intent, and react with an appropriate response. This method often makes use of deep neural networks, with an example being seq-to-seq models. Retrieval-based approaches are the last category, where heuristics from the ongoing conversation are used to pick a response from the corpus. This approach is considered more simplistic than the generative-based approach, even though it too can use deep neural networks.

Zaib et al. [44] focus on task-based categorization and present again three different type of systems: task-oriented dialogue systems, chat-oriented dialogue systems, and Question Answering (QA) dialogue systems. Task-oriented dialogue systems are mostly agents which can complete specific tasks for a user, such as ordering food and making a reservation in a restaurant. The chat-oriented dialogue systems focus mainly on having natural and interactive conversations with a user. The last dialogue system, the QA dialogue system, should be able to answer users' questions with information from different data sources. Examples of data sources can be text documents or knowledge bases. While the first two types are often researched, the last dialogue system is relatively new and still being extensively researched [44]. Conversational Question Answering (CQA) techniques use QA dialogue systems and can interact with a user through a multi-turn conversation where the agent can ask for more detailed information regarding the user's question. The traditional QA system receives unrelated questions from a user, while a CQA system would be able to handle drilling down, topic shifts, topic returns, and clarification questions. Since this thesis aims to research the effects of having users in control and their ability to direct the conversation through questions, this thesis will mainly focus on the QA techniques. While a CQA system could create a better conversational flow, it still needs to be determined if this complex system is more desired by users than the more straightforward QA system. Therefore, the next section will explore both QA and CQA systems.

2.3.1 Categorization of QA Systems

Within CQA, we can use a categorization regarding the data domains, types of questions, types of data sources, and the type of systems. Each type of system within this categorization might have its advantages or disadvantages. Additionally, a specific type of system might better fit a particular domain or purpose. In order to clarify which types of systems we can distinguish, this section will take a closer look into the categorization mentioned above based upon the research by Zaib et al. [44] This categorization consists of (1) data domain, (2) questions, (3) data sources, and (4) types of CQA systems. Let us now discuss these in further detail.

1) Data Domain

Within the data domains, we can make a distinction between open and closed domains. An agent within the open domain will be able to answer a broad range of questions, and users are free to ask whatever they want. The closed domain agents are more restrictive since questions the agent will be able to answer are limited to a specific application domain. In the case of this thesis, the agent is a closed domain agent, with the domain being specific information regarding HCM.

2) Questions

User questions can be classified into seven categories: factoid, confirmation, simple, complex, casual, listing, and unanswerable questions.

Factoid questions are questions where the system has to retrieve a simple and factual answer. The questions often begin with 'wh' words, such as "what", "where", and "why". Examples of factoid questions are: "What symptoms could I experience with HCM?" and "Why do I want to know if I have the HCM mutation?". Techniques to answer these questions include structure matching and reasoning.

For confirmation questions a binary answer needs to be returned. These types of questions are often still challenging since a strong inference mechanism is needed, and the answers are complex text extractions. Examples of confirmative questions are: "Is it possible my brother also has the HCM gene?" and "Can I plan an appointment through this application?".

Answering simple questions requires a small piece of text where the answer can be found. With simple questions, it should be easy to deduce what type of answer is required. An example of a simple question is: "How often should I get a check-up?" Here, it only requires one type of information to answer the question.

Complex questions require multiple steps to determine the answer. An example question is: "How do the HCM symptoms compare to the symptoms of other cardiovascular diseases?" To answer this question, much information is needed. First, the symptoms of other cardiovascular diseases and HCM must be determined. Next, they need to be compared. Some complex techniques are iterative query generation, multi-hop reasoning, decomposition into sub-questions, and combining cues from multiple documents.

Causal questions are questions where the answer is longer and calls for more details. These questions usually start with "why" or "how". An example of such a question is: "Why does the mutation take place?" To answer this question, a longer answer is needed.

As the name suggests, a listing question requires a list of entities or facts as the answer. For listing questions, approximately the same methods can be used as for factoid questions. An example of such a question is: "Can you list the symptoms?".

Lastly, some questions are unanswerable. The answers to these questions can simply not be found, and the CQA should, in this case, be able to indicate that the question is not answerable. An example is: "What is the meaning of life?".

3) Data Sources

The data for a CQA system can be organized in multiple manners, namely structured, semi-structured, and unstructured. Structured data is stored through entities in a separate table. Each entity in a table can have multiple attributes. For accessing the data, a query can be used. A semi-structured data source has no clearly defined boundaries between stored data and its schema. This way of structuring is often labor-intensive since much manual input is needed. Lastly, an unstructured data source has no rules for how data is stored. Often advanced natural language processing and information retrieval are used to retrieve relevant information. This way of structuring data negatively affects the reliability of finding the correct information. However, this structure does allow for more flexibility in the way it is used.

4) Types of CQA Systems

There are two types of CQA systems, sequential knowledge-base (KB) and conversational machine reading comprehension machines (CMRC). In knowledge-based systems, the information is retrieved through querying those knowledge bases, whereas CMRC methods extract information from sets of documents written in natural language. A sequential KB-QA system can remember previous dialogue states and ask questions to help users ask their next question, whereas a single-turn KB-QA system only uses a semantic parser and knowledge base reasoning. The advantage of using a sequential KB-QA system over CMRC methods is the flexibility and ease of use [44].

2.3.2 Related QA Systems

CQA and QA allow the user to steer the conversation and to get answers to questions they find meaningful. In this section, we will take a look at existing literature and relate this to this thesis. Since this thesis focuses on a closed-domain, the literature will only focus on closed-domain solutions.

Often closed-domain question-answering systems have similar stages within their proposed method. This can be split into the data part and the answer retrieving part. Depending on the type of data a method uses, there is a requirement for data pre-processing before the data can be used for answer retrieval. Additionally, the pre-processing of data can increase the performance of a system [18]. The answer retrieval part starts with a question a user asks. This question is then analyzed and processed in such a way that we end up with a query. With this query, we can retrieve documents, passages, sentences, and/or words that contain the answer. From this set, an answer or multiple answers are selected. Lastly, the answer needs to be translated into the desired format and displayed to the user. This section will compare a set of closed-domain QA systems by taking a closer look at methods applied during each node in the general system architecture depicted in Figure 2.6. In total, seven related QA systems will be considered.

Dataset and Data Pre-processing

Multiple types of data sources can be used within QA systems and within the related QA research, both structured and unstructured data sources are present. In total, two systems used unstructured data [5, 8], four structured data [4, 18, 20, 21], and one both [38]. As can be seen, fewer proposed methods make use of unstructured data. This could be because structured data is more reliable and easy to use.

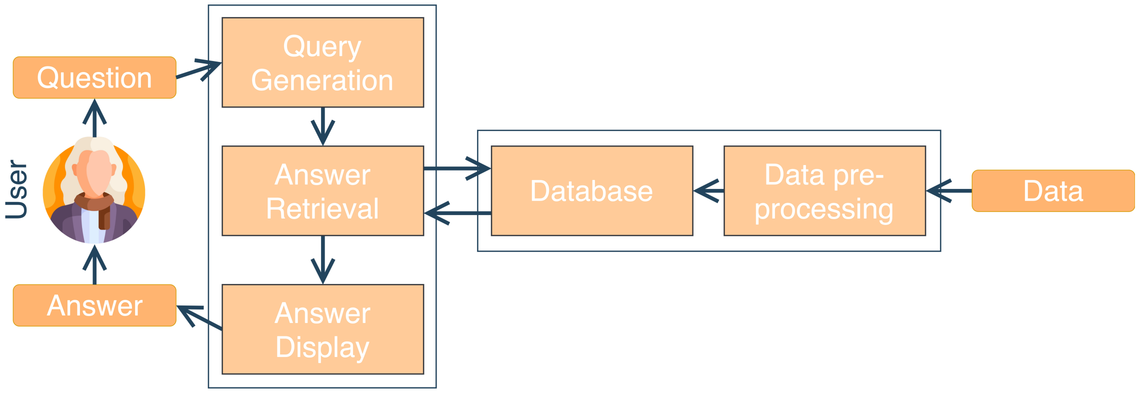


Figure 2.6: Overall architecture QA system based upon architectures found in related research.

Three of the QA systems used sets of documents written in natural language as their database. Derici et al. [8] made a QA framework for answering Turkish and English questions of students on biology and geography. Based upon information from geography and biology books and questions from students, they selected 31 websites that contained either biology or geography information, from which the plain text was extracted. The website evaluation checklist developed by the University of California Berkeley Library was used to determine which websites would best function as a source for their system. With a manual check, possible spelling errors were detected and corrected. The other system using unstructured data sources was the system on cultural heritage by Damiano et al. [5]. Their documents were annotated with metadata, which includes synonyms and ad-hoc rules for determining dates and places. While the work of Vargas-Vera and Lytras [38] does include a method of searching the web for natural language answers, it first tries to retrieve the answer in an ontology-based dataset. If this fails, however, it will continue the search for an answer through the sources found in a web search.

The other sources used a knowledge base to retrieve the answers in their system. Two of the systems made use of an ontology-based data source. The work of Latifi, Rodriguez Hontoria, and Sànchez-Marrè [20] uses an already established enterprise ontology. Additionally, the work of Vargas-Vera and Lytras [38] uses an ontology source that contains people, organizations, research areas, projects, publications, technologies, and events. Lende and Raghuwanshi [21] used another structured method to store and use their data. They retrieved information on education acts such as schools, funding authorities, school areas, teachers' duties related to students, and the duties of parents to secure children's education from one specific website. A database table was created with a list of documents that contain a specific term. This term was extracted from the information through stop word removal and stemming. An even more manual method can be seen in the work of Cuteri, Reale, and Ricca [4] where templates are used. These templates are pre-defined by the application designer. The most intensive method of data pre-processing can be seen in the system by Kia et al. [18], which is adaptable to different closed domains. In their research, they tested their system on four different domains, namely the Tesla, California, EU-law, and COVID-QA domains. The natural language files were pre-processed using Stanford CoreNLP[1] through sentence splitting,

---

[1] https://stanfordnlp.github.io/CoreNLP/

tokenization, full parsing, and POS tagging. Additionally, they saved the sentences based on the different types of questions they might be able to answer. The Candidate Answer Identifier classifies the sentences, using the Giveme5W1H tool[2], into the question type categories when, where, who, what, why, and how. This tool is able to identify patterns in text and label them as being able to answer a certain type of question.

Query Formulation

In the query formulation step, the goal is to end up with a query that can be used for retrieving the correct data required for answer formulation. That means the question must be analyzed and rephrased into this query. Every research found implements a different overall approach, although some shared methods are found. We can for example see POS tagging [4, 5, 20, 21, 38], stop words removal [5, 21], stemming [5, 21], named entity recognition [5, 20], lemmatization [5, 20, 21], answer type detection [5, 18], and dependency parsing [4, 20, 21].

Firstly, the work of Lende and Raghuwanshi [21] is an example that uses some of the aforementioned methods. They used POS tagging with the Stanford POS tagger, stop words removal, and stemming to get to their search terms.

Secondly, another research that uses many of the methods mentioned above is that by Latifi, Rodriguez Hontoria, and Sànchez-Marrè [20]. In their research, they use the Semantic-based Interchange Format. In this format, they used tokenization, and morphological analysis, including lemmatization, POS tagging, named entity recognition, and dependency parsing (with Stanford CoreNLP).

Thirdly, Damiano et al. [5] used multiple steps for their query formulation, such as answer type detection, named entity extraction, keyword extraction, NLP metadata extraction, and query formulation. The answer type detection aims to find the required Lexical Answer Type (LAT) by looking at which type of question is asked. This task is realized by implementing a deep learning approach that uses the Natural Language Classifier (NLC) by IBM Bluemix. Next, persons, places, and dates are extracted through NER using a dictionary tailored to the domain. In their metadata extraction, an additional layer of information is built through POS tagging (Apache OpenNLP). They mention that stemming and lemmatization are not done since the information retrieval part of their system applies it to their data documents. One last step is keyword extraction, where they extract dates, nouns, and verbs. Trivial nouns, common words, and auxiliary verbs are excluded in this step. The LAT, named entities, keywords, verbs, and dates, are then combined into a query with the help of a set of transcribers.

Fourthly, another approach is seen in the system by Derici et al. [8], where they first start with a dependency parser. Next, important information is extracted, such as subject, proper nouns, focus, modifiers, and the class of the question. The focus is the type of entity expected in the answer. This focus is identified by combining two methods, a trained statistical hidden Markov model and a rule-based model. Additionally, for the question of class recognition, rule-based and statistical methods were used.

Next, in their question processing, Vargas-Vera and Lytras [38] use an NLP parser to classify specific sentence segments into subject, verb, prepositional phrases, adjectives,

---

[2] https://github.com/fhamborg/Giveme5W1H

and objectives. Additionally, when the answer cannot be found in the ontology dataset, the question is reformulated, and what/which (a noun that describes the category of the entity involved), who (person), when (date), why (reason), and where (location) are classified.

Also, Cuteri, Reale, and Ricca [4] use some often used methods to formulate their query. They start with named entity recognition by implementing a supervised model based on conditional random fields (CRF++). Next, tokenization and parts-of-speech tagging are used to determine the adjectives, nouns, pronouns, verbs, and/or articles. For this part, Apache OpenNLP with pre-trained models was used. Through dependency parsing with MaltParser, lexical dependencies of words were determined. Next to these often occurring methods, they also implemented templates that the designer defined. Questions were matched to one of these pre-defined templates to classify the question from a syntactic point of view. Lastly, the intent was determined through hyponyms or synonyms with the help of BabelNet.

The last paper by Kia et al. [18] stated that they used WordNet to expand on the search term retrieved through their search terms. The best search terms were determined by ranking and filtering these expansion terms.

Answer Retrieval and Display

While answer retrieval and answer display are two different steps, this section will discuss both. The format of a retrieved answer and the format in which the answer should be displayed differ per system, but both steps heavily rely on each other. This means that the background of the retrieved information is needed to understand how it is translated into an answer. In the seven different systems researched, there are many differences in the approach each system takes. In this section, those approaches will be discussed.

Lende and Raghuwanshi [21], for example, searched their database table for documents that match every search term of their query. Next, the documents are ranked based on a calculated score. This score is calculated as $(A \cap B)/(A \cup B)$, where $A =$ set of extracted keywords and $B =$ the set of keywords from the documents. To extract the answer from the selected documents, the authors extracted the paragraph in the same manner.

Another method within answer retrieval is that from Damiano et al. [5] where document and passage retrieval are combined. Through the execution of the query, a list of documents is returned. Each of these documents is paired with its annotated metadata. Next, the documents are matched to the question by stemming and lemmatization. As a result of these matches, a set of top documents are determined. Passage retrieval is applied within these top documents, giving a list of sentences as output. For this process, the IBM Content Miner is used. These sentences are then scored and filtered. This is done through answer type filtering, named entity matching, overlapping keywords, local proximity prioritizer, and score normalization. Here the answer type filtering is done by removing sentences that do not match the expected LAT. The named entity matching looks at similarities between the query and the potential answer. The overlapping keywords matching does the same, but then regarding the keywords. However, it first filters the stopwords. The local proximity prioritizer scores the potential answers higher if the answer shows a denser distribution of query matches within the text. From this process, all potential answers receive a score that is then normalized. The highest score is, of course, picked as the best answer.

Yet another method is depicted in the research of Derici et al. [8] where the answer generation is done through text summarization. They determined which sentences per document scored highest from their data concerning the query. This was done as a sum of several scores: term frequency, similarity in words between query and sentence, and lexical chains. Since the authors now had a set of important sentences per document, the next step was to combine all these documents into one answer. This was done by creating a multi-document summary, where sentences were removed based on relative similarities of sentence pairs.

The research by Vargas-Vera and Lytras [38] used a different method. They first extracted passages from their documents and then selected an answer by clustering and scoring answers.

The system by Latifi, Rodriguez Hontoria, and Sànchez-Marrè [20] shows a different method from the aforementioned systems. They generated an ontology QGraph which is then used as search space. This is done by enriching the extracted keywords with morphological variants and WordNet synonyms. Additionally, they matched the keywords with ontology dictionaries, where a match receives a certain score. These scores then function as weights in the QGraph notes. The QGraph is then expanded with as many paths as possible.

Cuteri, Reale, and Ricca [4] work with templates generated by the program designer. The intents and templates used as queries are matched to pre-determined answer templates. Next, the template is translated into an NLP answer by filling in a natural language pattern with the question intent and extracted terms from the dataset.

The last work by Kia et al. [18] uses a CNN-attention-based answer selector, where a pre-trained language BERT model was used and then finetuned for answer retrieval.

Comparison of Methods

It is not only important to look at how the related research was executed but also at the results of those approaches presented by the authors. Unfortunately, not all the authors presented results [4, 21, 38] from which we can derive the effectiveness of the designed system. The results of the other presented approaches are available, but vary considerably in scores and the type of measurements. The system presented in the research by Latifi, Rodriguez Hontoria, and Sànchez-Marrè [20] has an accuracy of 33%. Damiano et al. [5] state that they have found a top-5 accuracy of 81-86% and a top-1 accuracy of 60-66%. Derici et al. [8] found that 59% of the summarizations were related to the content of the questions. Lastly, Kia et al. [18] presented an exact match score between 55,6-92,6% depending on which dataset it was used.

Even though we have now discussed multiple methods and discussed some of their results, it can not be said which approach outperforms the others. The results strongly depend on which dataset it is tested on and how the accuracy or match is measured. For instance, the research of Kia et al. [18] where the exact match percentage can drop from around 92,6% to 55,6% with the same method but another data source. The differences in methods of measuring results can be due to the lack of a standardized way of measuring the results within the closed-domain QA field. Moreover, the discussed approaches differ in the type of answer they produce, their flexibility, and the type of data they can handle. For this thesis, we will consider which methods can be used for which types of systems. Additionally, the approach of Vargas-Vera and Lytras [38], where a user could indicate agreement or disagreement with the presented answer, could be interesting to investigate further within this thesis.

2.4 Conclusion

Through the literature in this chapter, we can determine which techniques for QA systems would be most applicable in the context of this thesis. In this section, we will discuss the approach taken for the development of the open question answering feature.

Firstly, we must determine which type of QA dialogue system would be most applicable. The requirement for our feature is that we want to select a paragraph or excerpt from a website that specialists have approved. This is to ensure that all information retrieved is correct and that no false information can be generated. Additionally, the feature must be a minimal effort for specialists and flexible enough to be applied to other hereditary diseases with minimal changes. This means that the generative and rule-based method would not be appropriate within the categorization of Singh and Beniwal [34]. Thus we can determine that we will focus on retrieval-based approaches.

Moreover, through the categorization by Zaib et al. [44], we decide that we will develop a Question Answering system. The other two options in this categorization do not fit the goal of the desired system since a task-oriented dialogue system is focused on agents that can complete specific tasks, and a chat-oriented dialogue system is focused on agents that can have natural and interactive conversations with a user. Thus the QA system would be the most applicable since it is focused on answering user questions with information from different data sources.

When comparing related research of QA systems within a closed domain, we found multiple approaches to using datasets, pre-processing, query formulation, answer retrieval, and the answer display. If we take into account the requirement of reducing the effort for the experts within the development of the knowledge base, we found that the work of Derici et al. [8] and Damiano et al. [5] used methods that are in line with this requirement. Both studies used websites or documents with natural language. Additionally, we found that POS tagging [4, 5, 20, 21, 38], stop words removal [5, 21], stemming [5, 21], named entity recognition [5, 20], lemmatization [5, 20, 21], answer type detection [5, 18], and dependency parsing [4, 20, 21] are often used method within question answering systems. The answer retrieval and selection differed. However, the overall method was to attach scores to potential answers [5, 8, 20, 21, 38]. We base the exploration of our method on the related research we found; we can determine that we will develop an open question feature using (1) websites with natural language, (2) techniques such as POS tagging, stop words removal, stemming, named entity recognition, answer type detection, and dependency parsing, and (3) we return an answer based upon scoring and filtering of all potential answers.

Additionally, this thesis will not focus on CQA, since these are multi-turn and

3. Design Requirements

For this research several requirements and limitations have already been defined in the Background Section. The conversational agent in the eCG environment should include a feature which uses structured data and QA systems in order to allow users to ask open questions in regards to genetic cardiovascular diseases. This function should be implemented in a logical spot in the eCG application and should be tested with users. In order to give a structured overview with the importance of each system requirements the MoSCoW method is used [43].

Label	Requirement	Priority	Source	Implementation
1	The user can type their open question.	Must have	Research question - necessary to function	A textbox where the user can type their questions.
2	The algorithm is able to retrieve an appropriate answer from a list of website passages related to the subject just discussed by the virtual assistant.	Must have	Research question - necessary to function	Implement an algorithm that is able to return answers based upon the question.
3	The algorithm makes use of keyword matching.	Must have	Background [5, 20, 21]	Research which way this could be implemented and implement in the answer selection algorithm.
4	The algorithm makes use of sentence structure.	Must have	Background [5, 21]	Research which way this could be implemented and implement in the answer selection algorithm.
5	The agent should be able to display the chosen answer.	Must have	Research question - necessary to function	Adapt way it will be displayed to the current prototype of eCG Family Clinic.
6	The open question answer function should be hosted in such a way that it will be able to communicate with the eCG virtual assistant.	Must have	Research question - necessary to function	Research how this could be hosted and implement this.
7	The eCG virtual assistant should have an option to switch between the predetermined question options and open questions.	Must have	Discussions with the programmer of the eCG Family Clinic prototype	Have a separate option which a user can click on with 'More questions' to get to a text box where an user can write their questions.
8	A communication protocol should be specified between the assistant and hosted function.	Must have	Discussions with the programmer of the eCG Family Clinic prototype	Discuss with the programmer of the eCG Family Clinic what the protocol will be and implement this.
9	The algorithm makes use of expected answer type.	Should have	Background [5, 18]	Try to implement the giveme5W1H algorithm used by Kia et al. [18]
10	The algorithm should be able to indicate if an appropriate answer is not found.	Should have	Research question - gives overview of what is useful for a user	Discuss with the programmer of the eCG Family Clinic what the protocol will be and implement this.
11	The algorithm should give another response if appropriate answer is not found.	Should have	Research by Damiano et al. [5] shows that looking at the top several answers has a higher accuracy.	Discuss with the programmer of the eCG Family Clinic what the protocol will be and implement this.
12	The algorithm should adapt responses bases upon the users situation.	Won't have	Personalising information is one of the goals of the eCG Family Clinic	This is out of the scope of this project.
13	The algorithm is able to translate a web-page into a list of passages.	Won't have	Would give the possibility to use unstructured data, would decrease manual labor, would decrease reliability [44]. Used in research by Derici et al., Vargas-Vera and Lytras [8, 38]	This is out of the scope of this project and could negatively influence the reliability.

Table 3.1: Overview design requirements

4. Design

The design requirements in the previous chapter have to be translated into a system that can answer open questions from the user. This open question answering feature consists of multiple components which will be combined to retrieve an answer. This feature is then implemented in the already existing eCG virtual assistant. Both steps will be discussed in Section 4.1 and 4.2, respectively.

4.1 Algorithm of the Open Question Answering Feature

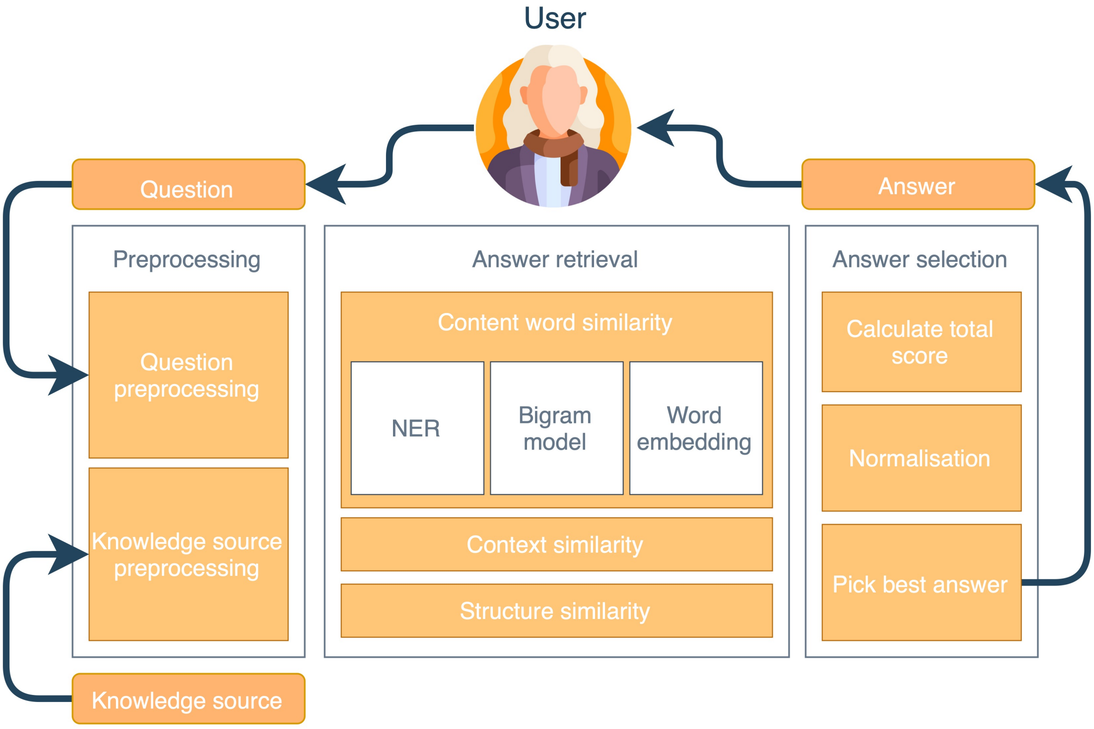


Figure 4.1: System design of the open question answering feature.

This section discusses the open question answering feature in depth. Figure 4.1 shows the system design and its components. It can be seen that the algorithm follows five steps to translate the question into a set of three ranked answers. These steps involve: (1) the algorithm receives a question from the user, (2) the algorithm preprocesses this question, (3) the algorithm gives an individual score to each potential answer for structure similarity, context similarity, and keyword similarity, (4) based upon filtering and the individual scores the answers are ranked, and (5) top three answers are returned. Each step, including a description of the dataset and its preprocessing, is explained in more detail below.

The algorithm was implemented using Python 3.10 and can be found on GitHub[1].

4.1.1 Knowledge Source

In Section 2.3.2, we described multiple ways to set up a dataset in such a way that we can use it in a QA system. One possibility could be to do a web search for natural language answers related to the user's question, as seen in the research of Vargas-Vera and Lytras [38]. However, this method would not suffice for this QA system. Having a web search does not allow you to control the sources information is extracted from. Especially for the closed domain of genetic HCM counseling and information provision, we want to ensure that users are provided with the correct information. Another method found in related research is using an ontology-based data source [20, 38]. While this method gives a structured overview of the information available and makes searching this information easy, the manual effort needed to translate natural language text into this ontology structure is high. For our system, we want to keep the human effort needed within the system to a minimum. For the same reason, also the use of predefined templates used in the research by Cuteri, Reale, and Ricca [4] will not be used within this project. In order to minimize the effort needed to add new information to the system, it has been decided to use a CSV file where paragraphs from reliable sources are separated and are in plain text format.

All paragraphs within the knowledge source were extracted from Dutch medical websites[2,3,4,5] recommended by Dutch cardiologists. To determine which information needs to be present in the knowledge source, we looked at the topics discussed in the prototype of the eCG Family Clinic. These topics include 'Symptoms' and 'HCM & Sport.' All information provided to the user regarding these topics was analyzed, and potential questions a user might ask were determined. This was done by looking at any unclarities in the text or clarifications needed. Answers were found on these recommended websites and added to the knowledge source. Additionally, during pilot testing, described in Chapter 5, the pilot participant was asked to list any questions they might have after using the eCG Family Clinic. Where applicable, answers to those questions were added to the knowledge source. In order to avoid redundancy within the knowledge source, answers to questions were combined if needed. For every knowledge source entry, there was a Dutch version and an English version since the eCG Family Clinic is in Dutch, and the algorithm can only handle English texts. The overall knowledge base consists of 27 potential answers related to the topics 'Symptoms' and 'HCM & Sport,' where each answer is available in both Dutch and English. This knowledge base is then used for all the steps described below.

---

[1] https://github.com/maaikexk/open-question-answering

[2] https://www.erfelijkehartziekten.nl

[3] https://harteraad.nl

[4] https://www.umcutrecht.nl/nl/ziekte/erfelijke-hartziekten

[5] https://www.patienteninformatietool.nl/hypertrofische-cardiomyopathie

4.1.2 Content Word Similarity

This system component starts with the extraction of content words from the question and compares them to the knowledge source content words. Content words in this component are all the lemmatized, lowercase non-stopwords which can be found within a user question and a potential answer from the knowledge source. The comparison between these questions and potential answer content words is shown in Figure 4.2 and done in four steps: (1) the named entities are extracted and combined into one token, (2) the bigram model is applied, (3) similarities between question content words and potential answer content words are found through word embedding, and (4) a similarity score is calculated. All these steps are individually explained in this section.

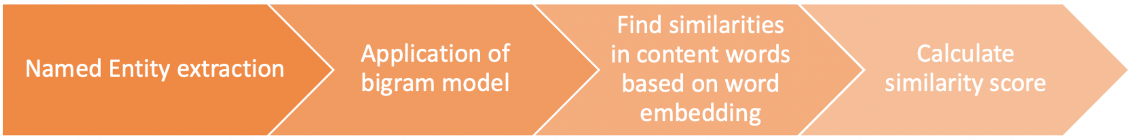

Named Entity extraction
Application of bigram model
Find similarities in content words based on word embedding
Calculate similarity score

Figure 4.2: Steps in content word similarity

Both the bigram model and the word embedding models are trained with a dataset. Since the dataset from which we extract possible answers is somewhat limited, it has been decided to use a different dataset. For this dataset, we could not use just a random dataset. For example, a dataset containing the text of all McDonalds' menus would be less likely to recognize terms used in questions regarding HCM. Additionally, a dataset on HCM, which is used within the medical field, might also not suffice. It might differentiate too much on medical jargon and may thus be too specific for our target audience. The challenge is to find a balanced dataset that is both general and contains information regarding HCM. The dataset was selected through a manual search of Kaggle[6]. This manual search found a dataset of 7.8 million sentences from the August 2018 English Wikipedia dump. The creator has already cleaned up this dataset to a certain extent. For example, sentences shorter than three letters and longer than 255 characters, duplicate sentences, and sentences requiring citations have been deleted. Upon examining whether the selected dataset contained information on HCM, 130 matches for the word cardiomyopathy, 778 matches for the word cardiovascular, and 29 matches for the term hypertrophic cardiomyopathy were found. Based on these findings, the dataset was deemed satisfactory for this research.

Of course, also this data needs to be preprocessed for the training steps of the bigram and word embedding models. This dataset is preprocessed by identifying the named entities as explained below. Next, the stopwords are removed, and the words are lemmatized. For use within training for the word embedding, it is also first passed through the bigram model.

1) Named Entity Recognition

Named Entity Recognition (NER) is an NLP task that can extract specific types of information from a text called a Named Entity (NE) [24]. Examples of types of NEs are: person, organization, place, time/date, and percentage. In the specific context of HCM, some examples of NER can be seen in Table 4.1.

---

[6] www.kaggle.com

English	Dutch	Type of NE
High blood pressure	Hoge bloeddruk	Medical object/illness/medication
Hiking	Wandelen	Sport
Patient association Harteraad	Patiëntenvereniging Harteraad	Organization
Cardiologist	Cardioloog	Profession of a working person

Table 4.1: Overview Named Entities in eCG Family Clinic context.

The goal of including named entity recognition is to group together the named entity and treat them as one content word. Spacy[7] is used for this step in the system and applied to the user question, the training data for the bigram model and word embedding, and the potential answers in the knowledge source. This way, all the data we use within the system will be comparable since they all have the same multi-word expressions. Within an expression, all words are combined in one token by replacing spaces with an underscore. An example of such a translation is 'high blood pressure' which will be translated into 'high_blood_pressure.' Treating NEs as one content word will ensure all content words will receive equal weight in further steps of the algorithm. Content words which are not part of a multi-word expression, such as 'cardiologist', are saved without any alteration.

2) Bigram Model

The Gensim[8] phrases model is used for the bigram model. This model uses the training data to calculate how often words are used together. This model can then be applied to new sentences. The model will calculate a score for a bigram, and if the score is higher than a predetermined threshold, they are assumed to be often used together. In that case, it will combine the content words into one. Example: hypertrophic cardiomyopathy would become hypertrophic_cardiomyopathy. We have trained this model in our system, as previously stated on the Wikipedia dump dataset. It is then applied to the Wikipedia dataset to prepare it for the word embedding step. This can be seen as the last step in preprocessing this data. Additionally, this model is applied in preprocessing the knowledge source and the user question.

3) Word Embedding

Users of this application can use different words and terms than the text from which the answers are extracted. We do not only want to check if specific words contained in the question appear in the potential answer from the knowledge source. We also want to check if other words might have the same meaning. For this check, word embedding is used. Word embedding is a technique that maps words to vectors. Vectors of words can be compared in order to determine the similarity between them. In order to get the correct vectors for words, we need to train a model. The only problem with this project is that a large amount of data is needed to train this model. This is where the Wikipedia dataset is used again. The implementation of the word embedding consists of two elements: (1) we have the training of the model on this Wikipedia dataset, and (2) we have the calculation of similarity between words of the knowledge base and the user question.

---

[7] https://spacy.io

[8] https://pypi.org/project/gensim/

For the training of the model, Gensim Word2Vec[9] is used. The minimum number of word occurrences used is five, with a vector size of 100. The similarity between two specific content words is calculated through cosine similarity. If two words have a higher similarity than 0.6, the words are considered similar and will be considered in the similarity score of potential answers in the knowledge source.

4) Similarity Score

The paragraphs are split into sentences to calculate the similarity of the potential answer paragraphs from the knowledge source. Next, each sentence receives a score between 0 and 1, which describes the similarity to the asked question. The sentence with the highest score will determine the overall score of the paragraph.

The sentence score is determined by comparing the question content words to the sentence content words. For each question content word, the highest semantic similarity score found is saved when compared to the sentence. Next, the semantic similarity scores of the question content scores are averaged. This will then give us a sentence score between 0 and 1.

4.1.3 Structure Similarity

The goal of the structure similarity component is to compare the question's syntactic dependencies to the answer's dependencies. Through this grammatical structure, we can find relationships in sentences between linguistic units. We can find a long list of dependencies, such as compound, nominal subject, determiner, root, object, and punctuation. If a user asks the question "What is HCM?", a potential answer could be: "HCM is a heart muscle disease. The heart muscle, especially the septum between the heart chambers, is thickened.". We can see that HCM in both the question and the answer has the same relation to the main verb, namely it is the nominal subject. While this does not hold for all good potential answers, an answer with the same dependency structure is more likely to be correct.

To determine the dependencies of words, the Spacy dependency parser[10] is used. Of course, there is an extra challenge since we need to compare words that are similar but not exactly the same. So we have to deal with similar words and find the most similar to compare these. Examples of these kinds of similar words are "HCM" and "Hypertrophic cardiomyopathy" or "symptoms" and "complaints." The best matching content words are labeled during the keyword matching, explained in Section 4.1.2. The dependencies are analyzed for all potential answers. The question and content words are compared, and for each token where the dependency matches, 1 point is given. These points are then normalized by dividing the number of points by the number of tokens in the user question. This gives us a score of sentence structure similarity between 0 and 1.

4.1.4 Context Similarity

To ensure that the algorithm does not prefer longer paragraphs of text over shorter ones, where the likeliness of content words being found is higher, context similarity was implemented. Without, a user could, for example, ask about the symptoms of HCM

---

[9] https://radimrehurek.com/gensim/models/word2vec.html

[10] https://spacy.io/api/dependencyparser

and get an answer regarding HCM and pregnancy since the answer dived into symptoms where a doctor should be called. While the keywords 'symptoms' and 'HCM' did match, the context did not. Thus the decision was made to consider the context of the entire potential answer.

The context similarity calculation follows the approximate steps as the content word similarity (cf. Figure 4.2). The two components differ in the calculation of the score. In the content word similarity, the score is calculated per sentence in a paragraph of a potential answer, and then the highest-scoring sentence determines the score for the entire paragraph. In the context similarity, the score is determined by calculating the semantic similarity to the most similar sentence content word for each content word in that paragraph. This is then averaged. This leaves us with a score between 0 and 1, indicating the similarity of the entire potential answer to the asked question.

4.1.5 Answer Selection

During the answer selection, all previously explained components are taken into account. However, this is done differently depending on the type of component. All the weights and thresholds have been set based upon manual and empirical testing using a set of test questions formulated by us. Figure 4.3 shows that the final answer selection consists of six steps: (1) each potential answer with a context score of below 0.5 is removed, (2) each potential answer with a content word similarity score of below 0.5 is removed, (3) the potential answers are ranked based upon the content word similarity, (4) check if there is an answer with a similar sentence structure score of 1 in the top five potential answers, if so move this to top position, and (5) return top three answers. Thus these six steps will determine the top three answers, which will be returned based on the user's question.

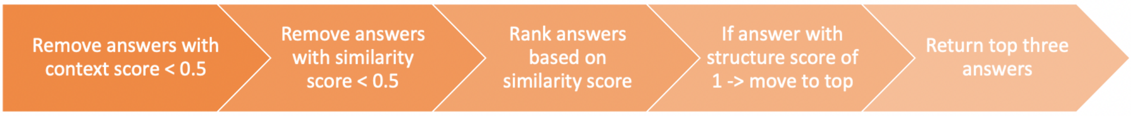

Remove answers with context score < 0.5
Remove answers with similarity score < 0.5
Rank answers based on similarity score
If answer with structure score of 1 -> move to top
Return top three answers

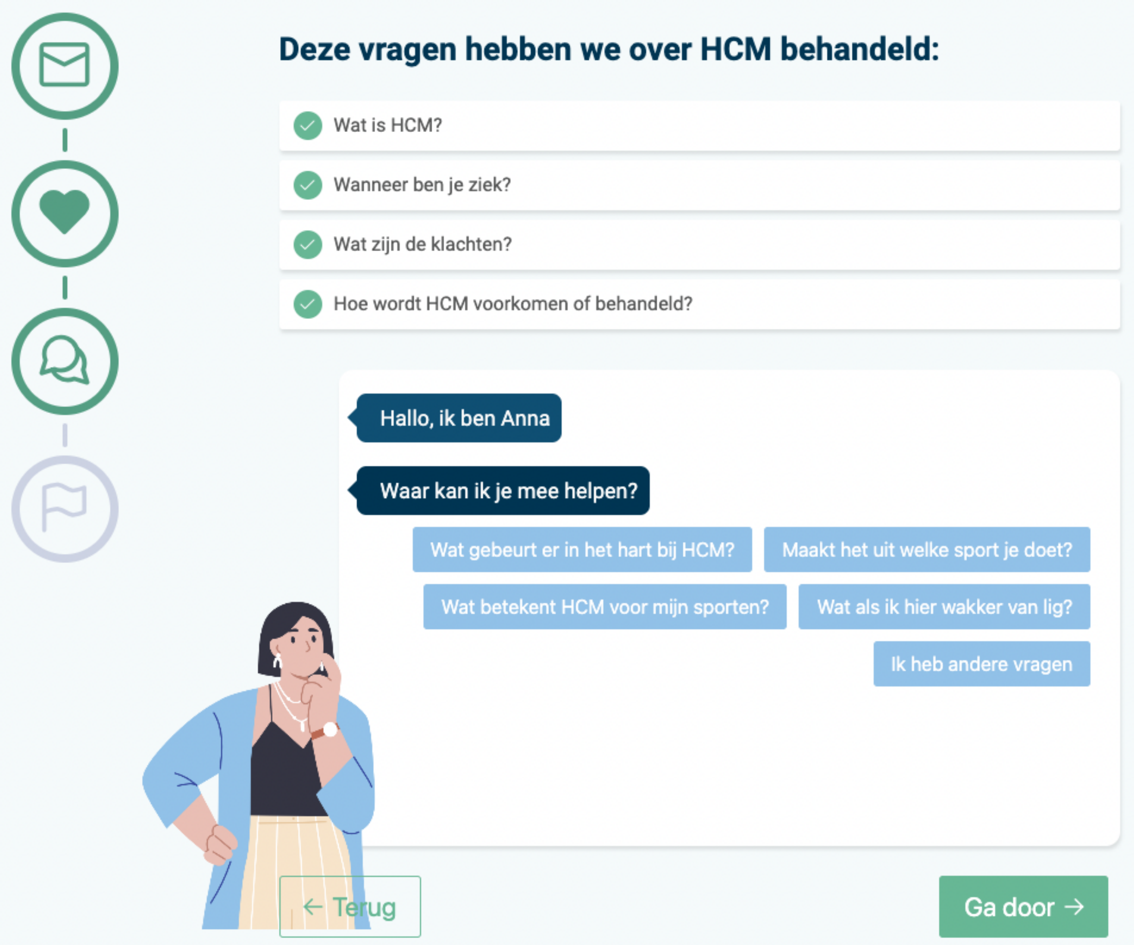
Figure 4.3: Steps in the top three answer selection.

4.2 Implementation in the eCG Family Clinic

To implement the open question answering feature, we need to decide in which of the two eCG Family Clinic versions, described in Section 2.1 and depicted in Figure 2.5, the feature should be implemented. Both the conversation design (cf. Figure 2.5a) and the button type design (cf. Figure 2.5b) would be suitable for the implementation of the open question answering feature. However, since one of the goals is to test the acceptability of adding the open question feature within a virtual assistant, the conversation design has been chosen..

One of the design considerations is at which moment of the interaction the open question answering feature should be implemented. Three options were considered: (1) have a button throughout the entire eCG flow that a user can press to ask their open question, (2) offer the possibility to ask questions at the informational screens, and (3) allow users to ask their questions while they are conversing with the virtual assistant.

While option one offers the most flexibility, a user might ask questions regarding subjects still to come. We want the open question answering feature to be an addition and not disturb the flow of the eCG agent and cause a user to receive the same information multiple times. In option two, the information page is often relatively short, and the subject a user might address is limited. Additionally, some of the questions might also be discussed in the conversation with Anna, the virtual assistant, at a later stage. Option three is the most logical since some frequently asked questions are explained well in a more closed answer system, the user has received all the information in text form, and the agent interface is already present. Thus the open question answering feature will be implemented at the end of the conversation with Anna to allow users to ask additional questions without discussing topics yet to come.



Deze vragen hebben we over HCM behandeld:

- ✓ Wat is HCM?
- ✓ Wanneer ben je ziek?
- ✓ Wat zijn de klachten?
- ✓ Hoe wordt HCM voorkomen of behandeld?

Hallo, ik ben Anna

Waar kan ik je mee helpen?

Wat gebeurt er in het hart bij HCM?
Maakt het uit welke sport je doet?
Wat betekent HCM voor mijn sporten?
Wat als ik hier wakker van lig?
Ik heb andere vragen
← Terug
Ga door →

Figure 4.4: Implementation of open question answering feature in the eCG Family Clinic. The user can ask open question by pressing "Ik heb andere vragen.", which can be translated to "I have other questions."

In the chat with Anna, users can click "I have other questions" to access the open question answering feature (cf. Figure 4.4). Different dialogue states are used to handle the dialogue with the user once the "I have other questions" button is pressed. These states are programmed using wool[11]. In Figure 4.5, the flow of this conversation is shown. Each time Anna says something, the user can choose their response from multiple

---

[11] https://www.woolplatform.eu/about.php

answers or type an open question. After an answer is retrieved from the QA system, the user is asked if it satisfies their question. If not, there is a possibility of retrieving the second-best answer from the QA system (if available), with a maximum of three options.

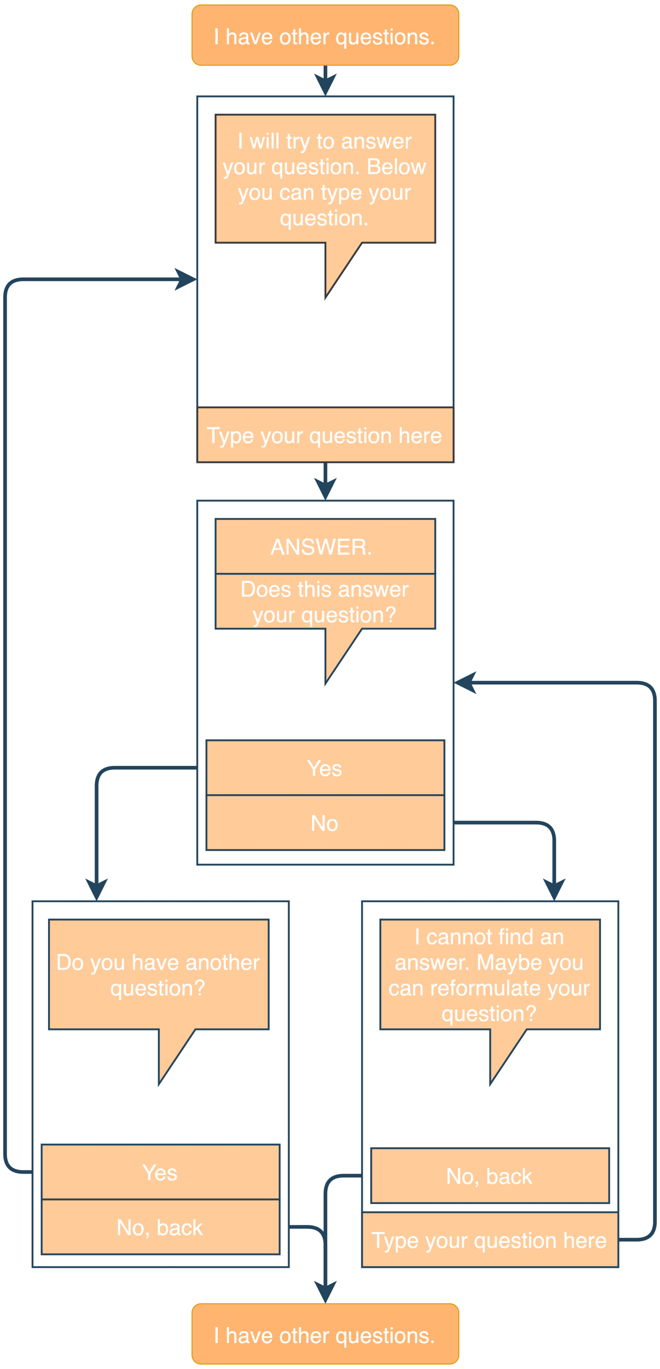

I have other questions.
I will try to answer your question. Below you can type your question.
Type your question here
ANSWER.
Does this answer your question?
Yes
No
Do you have another question?
Yes
No, back
I cannot find an answer. Maybe you can reformulate your question?
No, back
Type your question here
I have other questions.

Figure 4.5: System design of the open question answering feature.

Both the open question answering algorithm and the eCG Family Clinic platform are hosted on a server, where the eCG Family Clinic can request a JSON response with answers by sending the question a user asked through an HTTP GET request. An example is shown in table 4.2. Here it can be seen that a question asked is about the symptoms of HCM. An array with multiple possible answers is returned. These are always limited to a maximum of three, even if there would be more possibilities available. This ensures that lower-ranked answers are not offered to the user.

Question prompt	Answer

What is HCM?	{
	Answer: ["Hypertrophic cardiomyopathy is a heart muscle disease. The heart muscle, especially the septum between the heart chambers, is thickened.", "If you do not have the hereditary predisposition to HCM, you do not have an increased risk of developing HCM. You do not have to follow the rules of life, such as not exercising intensively."]
	}

What is on the McDonalds menu?	{
	Answer: [ ]
	}

Table 4.2: Overview JSON responses to query with question.

5. User Evaluation

User tests were conducted with two aims. Firstly, through user testing, we can determine whether the agent's answers align with what the user wants to know. Additionally, we want to find the perceived acceptability of users of the QA system. A semi-structured interview has been chosen where the questions are based on the theoretical framework of Technology Acceptance Model (TAM) [6]. To achieve these goals, we will report on user satisfaction with the answers received for their questions and conduct a thematic analysis, which will enable us to gain valuable insights into users' needs and preferences.

First, we will discuss the considerations for selecting the theoretical framework. This is followed by the chosen research procedure. Next, details of the data collection are discussed and lastly the planned data analysis is explained.

5.1 Theoretical Framework

For this research, we need to find a theoretical model that describes the application's acceptability and is focused on evaluating the question answering function and its implementation in the eCG Family Clinic. There are many models which could be used to measure the acceptance of technology, such as the Unified Theory of Acceptance and Use of Technology (UTAUT) [42], the Theory of Planned Behaviour (TPB) [2], Theoretical Framework of Acceptability (TFA) [33], and the Technology Acceptance Model (TAM) [6]. The TAM is one of the most widely used models to measure technology acceptance [23]. This model considers usefulness and ease of use to measure perceived acceptability [7]. When evaluating conversational agents, this model is applied in research focused on evaluating e-health applications [3, 17]. Since the TAM model aligns with the aim of evaluating the acceptability of the question answering function and its implementation, this model will be used within this user evaluation. From the many variations of TAM [6, 39, 41] it has been decided to work with the original TAM model (cf. Figure 5.1), since this model suffices to understand the factors influencing users' perceptions of usefulness and ease of use without going into depth on constructs such as social influence and cognitive instrumental processes.

TAM consists of perceived usefulness and ease of use as described by Davis [6] as: (1) perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance," [6] and (2) the ease of use is defined as "the degree to which a person believes that using a particular system would be free of an effort." [6]

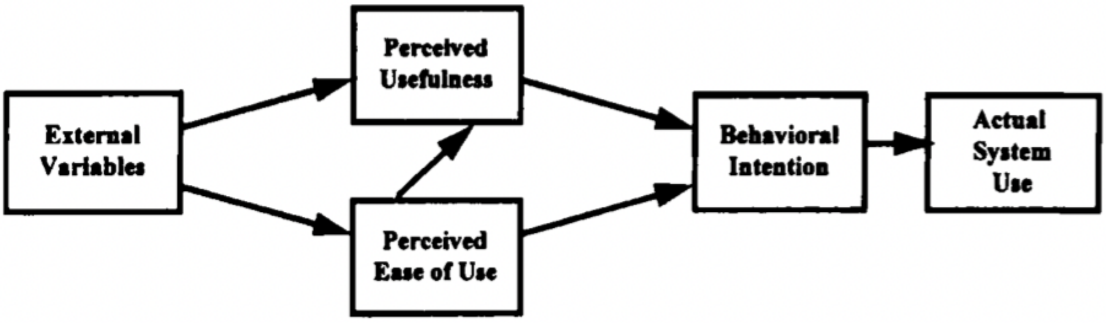


Figure 5.1: TAM by Venkatesh and Davis [40]

5.2 Methodology

The experiment consists of a session with a maximum duration of an hour. In Figure 5.2, an overview of activities within this experiment is shown. In Appendix A, the entire research protocol can be found (in Dutch). The user test procedure consists of (1) Setup, (2) Welcome and informed consent, (3) Briefing, (4) Task 1: Interaction with eCG Family Clinic based on a scenario, (5) Task 2: Asking Anna questions based on topics, (6) Semi-structured interview, and (7) Debriefing.

Each method step is explained in more detail below.

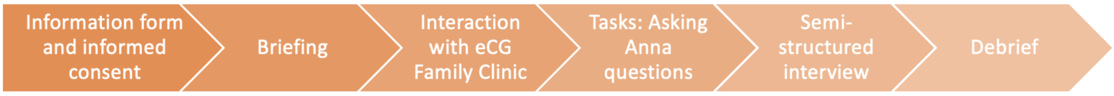

Information form and informed consent
Briefing
Interaction with eCG Family Clinic
Tasks: Asking Anna questions
Semi-structured interview
Debrief

Figure 5.2: Overview steps of the experiment

1) Setup

The experiment took place in an office at the University of Twente, where the room setup was arranged as shown in Figure 5.3. The room was enclosed to give the participant and researcher privacy and reduce disturbances. Here the participant was provided with (1) an external screen, (2) a keyboard, (3) a mouse, (4) blank A4 paper to take notes, and (5) a pen. For the researcher, there was available: (1) a laptop and (2) blank A4 paper to take notes. A small stand-alone audio recorder was placed in the middle of the table. The laptop of the researcher was a mirror of that of the participant. A web browser was open on the screen with two instances of the eCG Family Clinic (one for the interaction task and one for the question tasks) with the browser navigation bar hidden.

2) Welcome and Informed Consent

At the beginning of the user evaluation, the participant was asked to read the information sheet (cf. Appendix C) and sign the informed consent form (cf. Appendix B). The researcher could then be asked any question. The audio recording was started after the participant had given their consent.

It was explained to the participant that they could withdraw from the study at any time, either during the experiment or up to 48 hours after its completion. If the participant

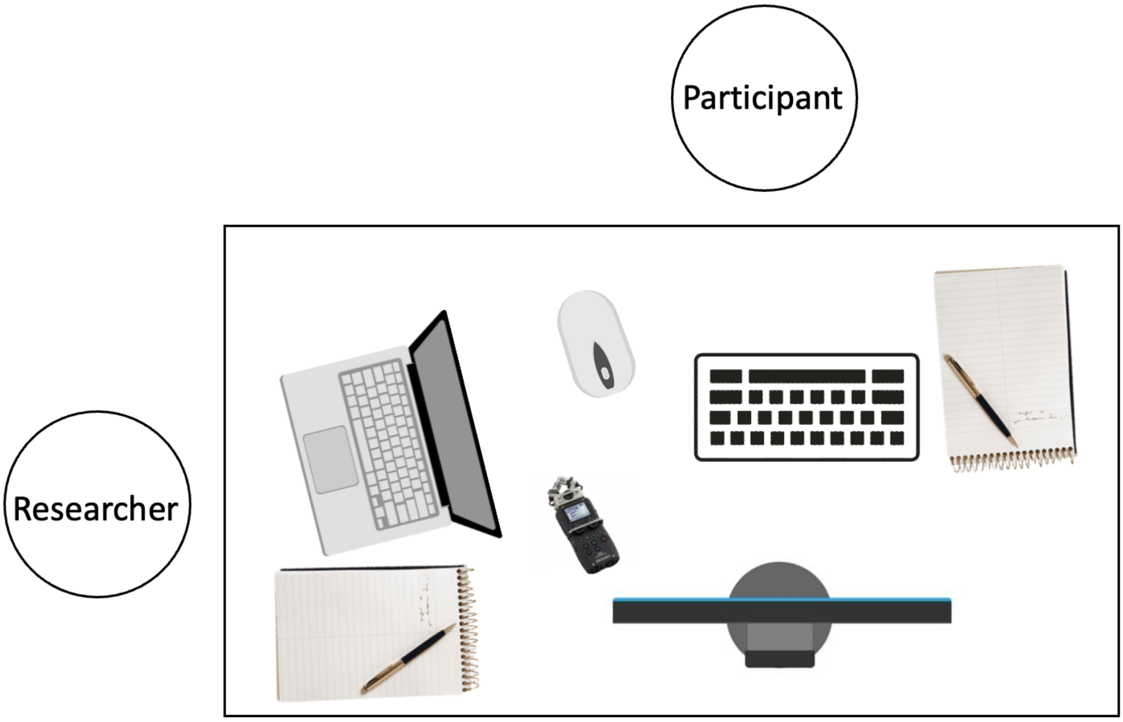


Figure 5.3: Experiment setup.

chose to withdraw, the audio recording would be deleted. This information was included in the information sheet given to the participant.

3) Briefing

During the briefing, the researcher provided the participant with information regarding the study and two tasks. Specifically, the participant was briefed on: (1) what to expect from the user experiment, e.g., tasks and interview, (2) limitations of the current prototype, e.g., buttons that are not working and a limited version of the prototype, and (3) the request to think and read aloud accompanied by an example.

4) Task 1: Interaction with eCG Family Clinic Based on Scenario

Participants were first given a scenario of the target user of the application: a family member has HCM and the participant needs to make an informed decision on DNA testing or regular check-ups. Next, the participant was asked to use the eCG Family Clinic with this scenario in mind. This means the participant would read the information from the informative screens, answer several questions regarding sports, and get the opportunity to ask Anna questions via predetermined or open options. While predetermined options entail buttons on which a user can click in order to ask questions to Anna, the open options entail questions which the user can type themselves. The last method of asking questions is our contribution to the eCG Family Clinic.

5) Task 2: Asking Anna Questions

Next, the participant was given a set of eight topics extracted from information in the eCG Family Clinic. For each topic, the participant was asked to ask an open question or click on a predetermined option. If the participant already asked questions regarding specific topics in task 1, that specific topic was skipped. Of the eight required questions, at least one needed to be an open question, and one needed to be from the predetermined options. The topics were carefully chosen by considering the information in the knowledge source and with the aim of facilitating users in asking questions, as it can be challenging for them to ask a series of questions without any guidance. The eight topics are (1) Sport in HCM, (2) Tilting weight in HCM, (3) Symptoms, (4) Treatment, (5) Occurrence, (6) Meaning, (7) Heart, and (8) Hereditary. While giving predefined tasks to the participant does not stimulate a natural interaction with Anna, it ensures the participant has asked enough questions to form a grounded opinion of Anna's capabilities and limitations. Additionally, the topic formulation is kept general to try and not influence wording used by participants.

6) Semi-Structured Interview

After completing both tasks, the participant was interviewed using a semi-structured format. The questions were based on the TAM questionnaire to cover both the constructs of ease of use and perceived usefulness. In this interview, the conversation was steered to those specific topics. The researcher tried to get opinions specifically on the open question answering and the predetermined click options. However the interview did cover the entire interaction with the eCG Family Clinic to decrease the chance of bias by participants knowing the focus of the research. The questions used as the basis for this semi-structured interview can be found in Appendix A.

7) Debriefing

During the debriefing, the participant was allowed to ask any last questions and was informed of the researcher's goal of developing and evaluating the open question answering feature within the eCG Family Clinic. The participant was given a thank-you snack (home-baked apple pie).

5.3 Data Collection

Students and employees of the University of Twente were asked to participate in this user evaluation. The inclusion criteria were: (1) being able to speak, read, and write a native level of Dutch due to the eCG application being in Dutch, and (2) being active in the form of sports since themes regarding sports are discussed in the experiment. Participants were recruited through WhatsApp or in person. A convenience sample of 8 participants was recruited. All participants were interviewed in April 2023, each in an office at the University of Twente. The data collected from participants consists of log files of questions asked by the user to Anna and transcribed audio recordings of the entire experiment.

This study was approved by the ethical committee of Computer & Information Science of the University of Twente with application number 230138. All participants were provided with an information form and provided written informed consent prior to participation. The informed consent and accompanying information forms can be found in Appendices B and C.

5.4 Data Analysis

After data collection, the audio recordings were transcribed. This transcription was done using the Microsoft Word transcription tool. After this automatic transcription step, the researcher did a manual check. During this manual check, text spoken by the participant (except for the read-out-loud eCG text) is corrected. Additionally, the researcher took initial notes and familiarized themselves with the data.

In the thematic analysis, we used a research-driven deductive approach, where based upon the theoretical framework TAM, we can expect specific themes to emerge from the data. The themes we expect to see in the data are ease of use and perceived usefulness. Both themes contribute to the acceptability of the theoretical TAM framework. For this analysis, Delve[1] was used. This qualitative coding tool can be used for the analysis of qualitative data.

In addition to the thematic analysis, the log files were analyzed to gain a deeper understanding of the types of questions participants were asking and how accurate the answers provided by the system were. The researcher examined the questions asked by each participant and formulated conditions for determining whether an answer was satisfactory, based on the participant's behavior during the user evaluation. Once the conditions were established, the accuracy of the answers provided by the system was measured.

---

[1] https://delvetool.com

6. Results

This chapter discusses the results of the user testing and the implementation of requirements. The user testing involved a total of 8 participant, all master's students. The results will first discuss the thematic analysis based upon the participants' interaction with the eCG Family Clinic and the interview. Then, we will discuss the accuracy of the open question algorithm on the questions the participants have asked Anna.

6.1 Thematic Analysis

A total of 8 participants resulted in 197 snippets with initial codes such as completeness, layout, preference, incompleteness, unlabeled but interesting, trust, personalization, ease of use, usefulness, and navigation. Coding was done using a hybrid approach. After iterative steps, these codes were transformed into conversation management, ease of use, and perceived usefulness themes (cf. Figure 6.1). Ease of use consists of the sub-themes: (1) Aesthetics and personality, (2) Learnability, (3) Flexibility, and (4) Navigation. The perceived usefulness theme consists of four sub-themes as well: (1) Convenience, (2) Accuracy, (3) Trustworthiness, and (4) Personalization. The definition used for these themes and sub-themes can be found in table 6.1.

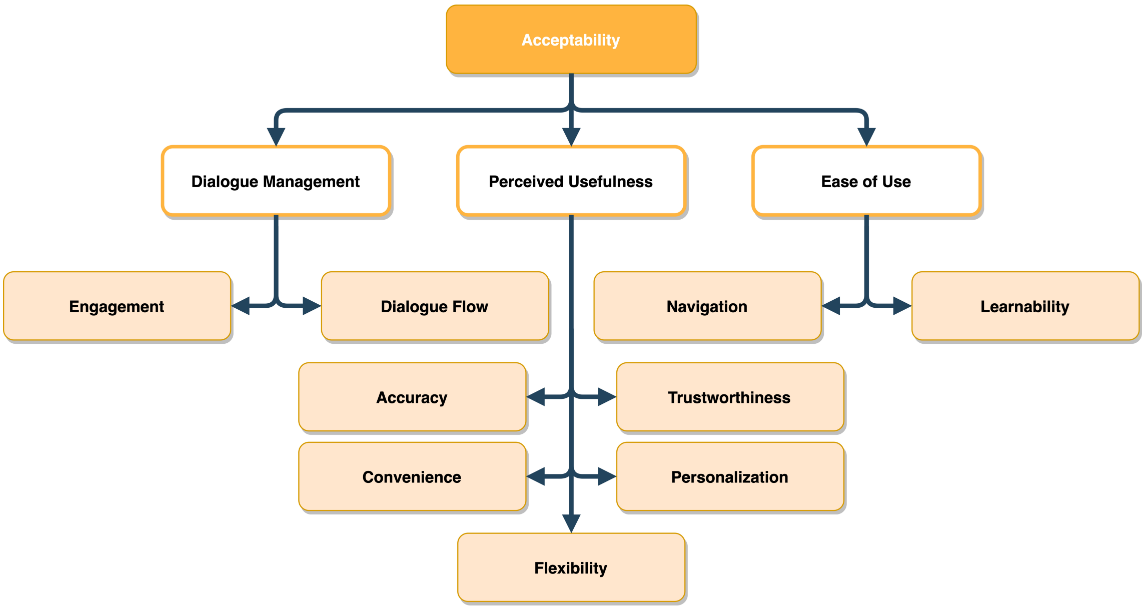


Figure 6.1: Overview of themes used for thematic analysis.

Main theme	Description	Sub-theme	Description
Perceived Usefulness	Users' perception of the expected benefits of using Anna to receive information.	Accuracy	User satisfaction of responses received from the virtual assistant.
		Personalization	To what extent does the virtual assistant satisfy the need for personalized advice, and does the assistant consider user unique preferences?
		Flexibility	To what extent does the virtual assistant conform to the user's needs and preferences: sufficient interaction and sufficient functionalities.
		Trustworthiness	Thoughts and preconceived notions about the assistant's overall reliability or accuracy: future use, frustrations, willingness to use.
		Convenience	How easy and convenient is it to use the virtual assistant or functions of the virtual assistant compared to alternatives: user preferences and priorities.
Dialogue Management	Process of controlling and guiding the interaction between the participant and virtual assistant to ensure a smooth and engaging conversation.	Engagement	Engagement of the user in the conversation with Anna: frequency and relevance of prompts, use of humor or character, and overall tone and style of communication.
		Dialogue Flow	Structure and organization of the conversation: sequence of topics, transitions, the effectiveness of prompts in guiding the conversation, and potential interruptions and/or errors.
Ease of Use	The degree to which a user believes that using Anna to receive information would be free of effort.	Learnability	How much effort does it take users to learn how to work with the virtual assistant: clarity of instructions, ease of understanding the virtual assistant capabilities, and learning curve.
		Navigation	How user-friendly is the navigation through the user interface: simplicity, clarity, and ease of navigation.

Table 6.1: Overview definitions themes and sub-themes

Snippets are only identified if they can be related to the functionality of, attitude towards, or suggestions regarding the predetermined or open questions. With snippets we mean small excerpts or sections of text, such as a sentence or paragraph, which are extracted from the interview with participants. The contributions to each theme and sub-theme for each participant can be seen in table 6.2. Some extracted excerpts can have multiple labels. Although this thesis did not focus on the chat interface and the already existing prototype of the eCG Family Clinic, but rather the development of the open question function, we still believe that it is valuable to present the full picture. Thus in the reporting of findings, there is a prioritization of themes that are the most relevant to this thesis. The same prioritization is done within the sub-themes. Figure 6.3 gives an overview of which components the themes mainly relate too.

Name (sub)theme	P1	P2	P3	P4	P5	P6	P7	P8
Perceived Usefulness	18	16	8	17	13	14	12	7
-> Accuracy	8	7	2	13	7	5	8	5
-> Personalization	1	0	1	5	2	3	1	1
-> Flexibility	6	2	5	0	1	3	2	0
-> Trustworthiness	2	3	1	3	6	1	1	1
-> Convenience	3	5	1	3	3	2	1	1
Dialogue Management	6	1	5	4	2	6	5	3
-> Engagement	1	1	2	0	1	2	3	0
-> Dialogue Flow	5	0	3	4	1	5	2	3
Ease of Use	3	13	2	9	6	7	2	3
-> Learnability	1	4	1	2	2	0	2	1
-> Navigation	2	9	1	7	4	7	0	2
Total	27	30	15	30	21	27	19	13

Table 6.2: Overview contributed snippets per participant

Name (sub)theme	Question answering	Chat interface	eCG Family Clinic
Perceived Usefulness	X	X	X
-> Accuracy	X		
-> Personalization	X		
-> Flexibility	X	X	X
-> Trustworthiness	X		X
-> Convenience			
Dialogue Management	X	X	
-> Engagement	X		
-> Dialogue Flow	X	X	
Ease of Use		X	
-> Learnability		X	
-> Navigation		X	

Table 6.3: Overview relation theme and components

6.1.1 Perceived Usefulness

The theme 'Perceived Usefulness' describes the users' perception of the expected benefits of using Anna to receive information and consists of (1) Flexibility, (2) Convenience, (3) Accuracy, (4) Trustworthiness, and (5) Personalization.

Accuracy

The user satisfaction with received answers from Anna and, thus, the usefulness of those answers can be analyzed through the sub-theme of 'Accuracy'. Here participants indicate that the open question answer is often not in line with the questions asked:

> "Only, Anna isn't very good at listening. Or well, can't really find things on the basis of the questions when I type them myself, actually she always

> comes up with the same answer always, yes." (P8)

> "I found the chatbot itself still-, often did not really answer my question, but gave an answer that kind of had to look like it or was close to, but was kinda related to [the question]. I have also seen the same answer pass by several times, while in my eyes the questions were really looking for something else." (P2)

> "The open questions is just not production ready." (P5)

Furthermore, participants indicated that retrieved answers to open questions were not specific enough:

> "But very often I did get, yes, answers that were still very general." (P4)

> "Some of the answers were just a little vague, but with the context of what I'd learned about the disease before, it was clear enough, but maybe without it, it wasn't." (P6)

Some participates do indicate that several open questions were answered to satisfaction:

> "When I asked about the symptoms, they were very clear, so that was actually the first open question I asked and it was answered very clearly." (P1)

> "But the open question: I was especially surprised by the first open question I asked that it was immediately answered completely right." (P7)

The predetermined question option received more positive reactions from the participants, where participants stated that the predetermined answers were complete:

> "And, I also thought that the predefined questions were very complete, so I thought that went pretty quickly." (P2)

> "But I have to say that the closed questions were quite nice, so there I recognized questions I would have, I recognized them in the suggestions, so I could easily click them immediately." (P5)

However, participants noticed that a sometimes lacking conversational flow did not allow them to receive all the information they wanted in the predetermined question options. One example: the predetermined option 'Does it matter which type of sport I practice?' brings the user to a point where they can choose between 'Which sport asks much of the heart?' and 'Which sport requires much oxygen?'. Once one is picked, the other option is unavailable later in the conversational flow of that specific predetermined question. Participants indicate:

> "Are you actually missing the option: 'Which requires much oxygen?', because it was with the previous one." (P1)

> "Actually, I would also like to know which sports require a lot of oxygen, but I can imagine it a bit." (P4)

Moreover, participants would have preferred to have more predetermined options in order to receive a more accurate set of information:

> "The most annoying for the click options was that they were consequently the same, so the first time you do not see the answer to your question, they will not have it. [...] You return to the same questions each time, while I have already seen them. So I expected to see other questions, not the same each time." (P2)

> "But that [the predetermined options] could be more extensive, because what I say every time-, now she goes back to that overview every time, but that overview is very based on whether or not you exercise." (P6)

For the 'Accuracy' theme, several trends can be found: (1) the open question answers are often not in line with the expected answers, (2) some answers to open questions do match, (3) participants viewed the predetermined questions positively, (4) participants missed information during the predetermined questions due to errors in the conversation tree, and (5) participants prefer more predetermined options to give a more accurate overview of desired information.

Personalization

The sub-theme 'Personalization' aims to describe the extent to which the usage of Anna gives participants personalized information when answering questions and takes their user preferences into account.

Participants expressed the desire to receive more personalized information from Anna when asking questions (both predetermined as open):

> "What does that mean for me then? Because I feel that here that it is, with this application, that-, that is still a bit difficult, Because, well, everyone has a completely different situation of course." (P4)

> "Well the role is to provide insight into the information, so precisely the edge cases [...] But it is really that you ask can ask those edge questions [open questions] to the assistant." (P5)

> "Yes, you probably want to relieve that cardiologist by doing this [interacting with Anna], but this [the answers to questions] remains so general." (P6)

Through this sub-theme, we can see that the overall consensus among participants is that more personalized answers from Anna are desired to increase their usefulness.

Flexibility

The sub-theme 'Flexibility' described the extent to which the virtual assistant conforms to the users' needs and preferences for asking questions and to the chat-interface. One reoccurring user preference which was not offered is the possibility to view the information from the information screens while asking Anna questions:

> "Then [with a certain topic] I had a bit of a tendency to go back again in the informative text I got before Anna, because you want to keep track of that. But that's also because during studying and stuff-. You quickly look at books and stuff." (P1)

> "Perhaps that can be solved by making this [the informative topics from the informative screens above the Anna dialogue] fold-out, because it is not like

> the [informative] texts were that big. Or maybe you can make that fold-out and then shift Anne to the left or to the right, because there's quite a lot of space left there [next to the conversation window]." (P6)
>
> "Maybe put those buttons [the four topics discussed] on the side, which are of course not in operation now. So that you can also do pure reading yourself for people who do that. Like I like to do." (P3)

Additionally, participants expressed a need to archive things they asked or answers they received from Anna to view them later for both the predetermined and open questions. Example:

> "A bit of archiving so that you can read back: "oh God, haven't I seen this before", or whatever the story may be, or "what was that again?"" (P2)
>
> "Well, I do notice that I can't look back now to what I asked before [earlier in the interaction with Anna]. What are the important things that I just want to remember or something. If this [the eCG Family Clinic] is a thing, where you are in your own bubble. So something that you want to remember, that you have a code, some sort of log-in. A personal thing, so to speak." (P3)

Furthermore, participants would prefer the predetermined option and open questions to be more integrated:

> "Maybe it would help if you are able to ask open questions in the beginning [when you start asking questions to Anna] to really exactly come up with a problem and then you get points which offer you further suggestions that you can click on to find out more." (P7)
>
> "Maybe there should always be some kind of option to ask those predefined questions again, you could even combine it in such a way that it's always a kind of search bar [in the chat-interface] where you can type your own texts, but that you get some kind of recommendations. So those are closed questions. [...] then it's a search engine with a few predefined options." (P1)

These user experiences show that users are missing some functionalities, which are: (1) viewing the information of the information screen during the conversation with Anna, (2) an opportunity to archive questions or information to view later, and (3) integration of open and closed questions.

Trustworthiness

The theme of 'Trustworthiness' describes the thoughts about Anna's reliability or accuracy for future use or preconceived trustworthiness based on past experiences. Here we can see that participants already have a lack of trustworthiness for Anna based on past interactions with virtual assistants. Exactly what these negative interactions entail is not clear:

> "'Unfortunately, because I have to retrieve the answers, I can sometimes make a mistake.' [said by Anna when someone asks an open question] I was already afraid of that." (P2)
>
> "I would say, as a user I have had perhaps somewhat less reliable experience

> with chatbots, so that I will immediately doubt Anna's correctness. Even though she is in this fairly reliable system, as I experience it. But still, because it's a chatbot, I would immediately think: 'oh, would it be correct?'" (P5)

Other participants express trust in the accuracy of the information given by Anna due to the companies which collaborated on creating the eCG Family Clinic:

> "It's just a lot of reliable information. I would rather trust this system than Google." (P5)
>
> "I assume that it is from a reliable source and that what is written here [the logo's on the bottom of the page with companies who worked on the eCG Family Clinic] is correct. And if you google your question that can be questioned, especially if you google with symptoms." (P8)

Additionally, participants find that their trust decreases if Anna does not give accurate answers to open questions:

> "But I think at some point I would think: 'well, It's okay with you, Anne, or Anna, I'll call a counselor in a bit and then they can talk me through it and they can also respond more actively to the questions I ask.'" (P6)
>
> "That's mainly for that option of: 'can you try again?'. 'Can you give me one more answer?': that immediately feels like: just keep trying until something works. And then I doubt the correctness of the information, because if it the first answer is not correct, why should I trust that the second answer is correct. But there is of course a difference in the correctness of the information, and whether the question is correctly answered. Of course, but it still lowers confidence a bit if you [Anna] don't understand what I'm saying." (P5)

These transcriptions of the trustworthiness of Anna that influences the usefulness can be summarized as (1) participants might have issues trusting Anna due to past experiences, (2) participants trust the information Anna provides since it is from reputable sources, and (3) trust in Anna might decrease if inaccurate answers are given.

Convenience

The sub-theme of 'Convenience' can be described as the convenience of using Anna to ask questions compared to alternatives. Alternatives stated by participants were the doctor and Google. Participants indicated that using the eCG Family Clinic to get answers to initial problems might improve the visit they would have with a genetic counselor at a later stage:

> "I think that's why the platform, so that you don't have to go to a cardiologist for your first orientation, because that naturally takes a lot of time and maybe that also provokes you to go deeper into the material than necessary." (P6)
>
> "Just to find information, then I don't have to ask a doctor for certain things. Then I think that it is faster via Anna and then I also know that I can prepare. If you then speak to a doctor, for example-, then you can ask different kinds of questions." (P3)

One participant would use the platform but would visit the doctor when getting nervous:

> "So I think I'll just click around on the couch at home in the platform for information, but if I'm nervous then I'll still go to the doctor. Which is very accessible anyway, so I can discuss it there and then I can also take a look at the rules of life are also applicable for me to prevent me from getting HCM complaints. But I would use it [the platform and Anna], I think." (P1)

Other participants found that using Anna would be more convenient since getting in contact with a specialist might take time:

> "Especially nowadays. I mean, It's still a bit with corona. Hospitals are just very taxed in that respect, so before you have such an appointment it will take a while." (P4)

> "A specialist. That is dependent on in how much of a hurry you are. In principle is this [asking questions to Anna] immediate, and making an appointment with the doctor, it can take, so to say, a week or two if it is not an emergency." (P8)

Participants differed in opinions when it came to the convenience of using Anna over an alternative such as Google:

> "I think it's faster, because if you forward this [the log-in/platform] to relatives from which someone has that gene, they can start asking the questions at that point, instead of thinking, okay, [swear word] and that they will then Google it later. So the information is already in the right place." (P1)

> "I think that a verified medical chatbot at all is a lot more effective than just random terms in Google search. Mainly that verification step." (P2)

> "The question is how much would I use this more than, say, Google? I don't know that very well." (P5)

From the 'Convenience' theme, we can identify four main findings: (1) participants would use Anna to have a better-informed appointment with a specialist at a later stage, (2) participants might initially use the platform but make an appointment with a specialist when getting nervous, (3) Anna would be a more immediate alternative to an appointment with a professional, (4) participants do not agree on their preference in Google vs Anna.

6.2 Dialogue Management

The theme 'Dialogue Management' describes the process of controlling and guiding the interaction between the user and Anna to ensure a smooth and engaging conversation. This theme consists of two sub-themes: (1) Engagement and (2) Dialogue Flow.

6.2.1 Engagement

The sub-theme 'Engagement' describes the engagement of the user while interacting with Anna. This might include the frequency and relevance of prompts, the use of humor or character, and the overall tone and style of communication. Participants disagreed on the relevance of Anna, saying, "Thank you for your question, I found this answer. I hope this helps." when a participant asked an open question:

> "Yes, I thought it became a bit redundant at some point that she [Anna] kept saying: thanks for your question, I found the answer, hopefully this helps. I found that a bit redundant after a few times." (P7)

> "She asks every time: 'does this help you further?' I like that." (P3)

Participants also noted that Anna is reacting neatly overall:

> "I think she just responds nicely." (P3)

> "The type of language Anna uses, i think, is very nice. [...] It's kind of personal, but also as if Anna knows a lot about it. [...] And formal, a personal and formal combination." (P7)

Through these transcriptions, we find that participants (1) have different opinions regarding the response Anna gives when a participant has asked an open question, and (3) the overall reactions Anna gives are received positively by participants.

6.2.2 Dialogue Flow

Through the sub-theme of 'Dialogue Flow', we examine the structure and organization of the conversation, including the sequence of topics, transitions, the effectiveness of prompts in guiding the conversation, and potential interruptions and/or errors. Participants express a lack of effectiveness in guiding the conversation:

> "I am glad that there is some understanding [when participant asked the predetermined option: 'What if I lay awake at night.'], but sometimes the conversation is to quickly steered to the next topic." (P4)

> "'I want a different answer' [one of the options when the answer Anna gives to an open question is not satisfactory] also feels so stupid, like: 'no, you didn't give me the answer I wanted to hear, say something else.'" (P6)

Participants found that there were some mistakes and some redundancy in the conversational flow of the predetermined question options:

> "Shall I first tell you how the heart is under normal -? [option in the conversational flow of one of the predetermined options, second time it is mentioned in the same flow] No. No, I do know that. I don't need that." (P3)

> "Yes, I wouldn't know why I would want to go back to a normal heart now, but anyway, because I also said earlier that I knew that and had ignored that option before." (P6)

Additionally, participants expressed that it was for them unclear that pressing the button 'I have other questions.' would bring them to a place where they could type their own questions:

> "Because now it's kind of like a fifth button and you don't know what happens when you push that button. Is it an open question or do you get a new list of 4? So that wasn't necessarily very intuitive." (P8)

> "For example the button: 'I have other questions', am I supposed to ask an open question or will there be another list of questions?" (P7)

A few participants made suggestions on how the dialogue flow could be improved further:

> "Perhaps it would be useful if these questions [the predetermined options] were also ticked at some point. Which ones have you already had and which haven't?" (P4)

> "So I would like another question like that among the suggestions [the predetermined options] of 'I'm done talking'. Especially for users who are experienced with technology, because it is a natural conversational way to end a conversation. It feels weird to press go in the middle of the conversation." (P5)

The quotes show that the trends between participants within the sub-theme 'Dialogue Flow' can be summarized as (1) improvements could be made in steering the conversation, (2) there were mistakes and redundancy in the conversational flow of predetermined question options, (3) it was not clear for users that the 'I have other questions' button would bring them to a spot where they could ask open questions, and (4) participants have offered suggestions on how elements of the conversation could be improved.

6.2.3 Ease of Use

The theme 'Ease of Use' describes the degree to which a user believes that using Anna to receive information would be free of effort. The defined sub-themes will be discussed: (1) Learnability and (2) Navigation.

Learnability

'Learnability' will show the effort it will take for a user to learn how to work with the virtual assistant, which for instance, includes clarity of instructions, ease of understanding the virtual assistant's capabilities, and the learning curve. The overall chat interface seems easy to understand for participants due to the similarities with already existing chat interfaces:

> "It is just a chat-interface where most of the users are familiar with, so that feels familiar." (P5)

> "I don't think there is a learning curve. Because it [the interface] is used quite often by other websites." (P8)

Asking questions for participant did not seem to be difficult:

> "I don't have the idea that asking questions [both open and closed] is hard at this moment." (P2)

> "I think it [finding out how you can ask questions] is very clear." (P5)

The participants' experience with Anna shows that there is little effort needed to learn how to use Anna within the eCG Family Clinic and that asking questions is experienced as easy and straightforward.

Navigation

The 'Navigation' sub-theme aims to describe the user-friendliness of the virtual assistant concerning the simplicity, clarity, and ease of navigation. Here many participants shared experiences of what could be improved, such as the scrolling within the chat interface:

> "It's nice that you can scroll back. Maybe a scroll bar will indicate that, because now you don't see that and sometimes that's not an option with this kind of agent." (P6)
>
> "The 3 blocks actually didn't fit on the screen so I had to scroll every time [...]" (P2)

Additionally, participants found out that the enter key did not work for sending open questions to Anna:

> "Enter, that doesn't work. So click the arrow. Arrow does work. [During think-out-loud]" (P2)
>
> "So unclear, the enter does not work. [During think-out-loud]" (P8)

Moreover, participants found that the arrow used to send open questions in the dialogue with Anna was unclear.

> "Well, I found that the arrow, I don't think that it is a really good arrow to use. [...] such an airplane is really for sharing on social media with others." (P6)
>
> "I know the logo [the arrow], that you can see, from Instagram were you can share things, I think, so maybe just a normal arrow like this 'ga door' button [continue button on the bottom of the page] would have been better." (P4)

Further improvements can be made by adapting the input field of the open questions according to the participants:

> "Quite logical that it [the text box where a user can type] is a text message, the text bar could have been larger in my opinion, because now you cannot see your entire question. Which means that you have to remember very well what you said. Or indeed scroll trough it with the arrow buttons to reread it again. Or just accept that what you said is probably correct." (P2)
>
> "I expected it to be in a WhatsApp style. That is the first you think of, so have it broader in the bottom of the screen instead of it already being a text bubble, so to say." (P5)
>
> "If the bar [text box where the user can type] is a bit longer it invites more complete answers or questions." (P6)

There was a clash of opinions when it came to the way of asking open questions, where it is currently only possible to type:

> "Ah, it is nice you can type yourself." (P5)
>
> "I am never happy if I need to type myself. But there is not an easier way and I can say very nicely 'oh, you might speak your question', but that is

effectively a point which will make the application work differently." (P2)

Through the sub-theme of 'Navigation', several improvements are identified which would potentially increase the ease of use: (1) improve the scrolling, (2) improve the timing of when the messages appear, (3) change the arrow with which users can send their open questions, and (4) adapting and making the input text field for open questions larger. Participants did not share an opinion on the method of inputting their open questions, letting participants type their questions.

6.3 Algorithm Accuracy

Each open question asked during the user testing was logged. During the user testing, the researcher wrote down whether an answer satisfied the user's question. An answer is deemed satisfactory if (1) the participant explicitly stated during the user test that it satisfied their question, (2) the participant answered "yes" to the question from Anna whether the answer is satisfactory (only if the participant did not contradict this during the user testing), and (3) the participant explicitly mentioned that the answer did satisfy their question but wanted to press "no" on the question from Anna whether the answer is satisfactory to explore the interface. All answers which did not satisfy one of these three requirements were deemed unsatisfactory, including answers of which the participants felt the answer was only partially satisfactory. Participants were not instructed to indicate if an answer was satisfactory, but the researcher made notes during the user evaluation with the think out loud method.

Participant	Nr of questions	No answer	Correct@1	Correct@2	Correct@3	Top-1 Accuracy	Top-2 Accuracy	Top-3 Accuracy
P1	5	0/5	2/5	0/2	0/0	40,0%	40,0%	40,0%
P2	7	1/7	2/6	1/2	0/0	28,6%	42,9%	42,9%
P3	13	3/13	2/10	0/4	0/0	15,3%	15,3%	15,3%
P4	10	2/10	2/8	1/4	0/0	20,0%	30,0%	30,0%
P5	7	0/7	1/7	0/2	0/0	14,3%	14,3%	14,3%
P6	13	2/13	1/11	1/4	1/1	7,7%	15,4%	23,1%
P7	6	1/6	1/5	1/4	0/0	16,7%	33,3%	33,3%
P8	9	1/9	1/8	0/4	0/0	11,1%	11,1%	11,1%
Average	8,75					19,2%	25,3%	26,25%

Table 6.4: Overview accuracy open question answering

An overview of satisfaction with the accuracy of the open question feature by participants can be found in table 6.4. The table shows how many questions were asked and how many correct answers were given. After an incorrect answer, a participant could make the decision to let Anna retrieve up to two new answers. Participants did not always do this. Participants were also given the opportunity to rephrase a question if no answer was available. These newly rephrased questions count as separate questions. The 'No Answer' column indicates for how many questions no potential answer could be found. Of those it could retrieve potential answers, the table shows how many correct answers are retrieved, where correct@1 is the first answer retrieved, and correct@3 is the third answer retrieved. A second or third answer is only retrieved if participants request Anna to do so. Lastly, the top-1, top-2, and top-3 accuracy shows us the accuracy of the correct answer being in the top 1, top 2, and top 3 potential answers respectively. Do note: the top-2 and top-3 accuracy measures do not describe the full objective accuracy, since a second or third best answer could only be found satisfactory if the participant actively choose to receive another answer.

Each participant had a different way of formulating their question. The type of formulation was consistent per participant throughout the entire interaction with Anna. For example, P6 mainly used keywords, while P7 formulated relatively long sentences. See:

> "Prevention" (Question by P6 based upon 'prevention' topic)
>
> "You were talking about lifestyle changes to treat HCM, what can I think about?" (Question by P7 based upon 'treatment' topic)

Examples of some answers found satisfactory by participants are:

> Question by P7: "What is the chance that if I have HCM that one of my children has it as well?"
> Answer from Anna: "HCM is hereditary. If you have a predisposition to HCM, there is a 50 percent chance that you will pass it on to your children."
>
> Question by P4: "What can I do to prevent HCM?"
> Answer from Anna: "If HCM has been diagnosed, a treatment plan can be drawn up in consultation with your cardiologist to prevent or reduce the risk of serious symptoms. This may, for example, consist of medication, rules of life, and (in the case of serious complaints) surgery."
>
> Question by P3: "What does it mean to me if I have HCM?"
> Answer from Anna: "If you appear to have a predisposition to HCM, you will have to consult a cardiologist to see what is or is not wise. For anyone with a predisposition to HCM, the lifestyle changes are important, but if you already have symptoms, it becomes even more important to follow the lifestyle changes."

Examples of answers found unsatisfactory by participants are:

> Question by P2: "Possibily I have the gene for HCM, how do I find out?"
> Answer from Anna: "HCM is a hereditary disease. DNA testing can be done to determine a hereditary cause. In about 50% of people with HCM, DNA testing finds the cause in the DNA. Because not all genes involved in HCM are known yet, a mutation (change in the DNA) is not found in everyone. It is still possible that the HCM is hereditary. Sometimes a mutation is also found of which it is not (yet) clear whether or not it can cause the disease."
>
> Question by P4: "How much kilo am I allowed to lift?"
> Answer from Anna: "People with HCM do not always have symptoms. Complaints that can occur are, for example: chest pain, fast or irregular beating of the heart, reduced fitness and endurance, shortness of breath, and cardiac arrhythmias (sometimes cardiac arrhythmias can lead to fainting, dizziness or even sudden death)."
>
> Question by P8: "How big is the chance that a person has HCM?"
> Answer from Anna: "Through treatment with medication, lifestyle changes, and (in severe cases) surgery, many people with HCM have a normal life expectancy."

Additionally, a few participants found that an answer could only partially answer their question. For the results, these answers were deemed unsatisfactory. Examples are:

> Question by P3: "What are the most often occurring symptoms?"
> Answer from Anna: "People with HCM do not always have symptoms. Complaints that can occur are, for example: chest pain, fast or irregular beating of the heart, reduced fitness and endurance, shortness of breath, and cardiac arrhythmias (sometimes cardiac arrhythmias can lead to fainting, dizziness or even sudden death)."
>
> Question by P4: "Can dancing be harmful if you have HCM?"
> Answer from Anna: "You may experience chest pain. The pain is usually most pronounced during exercise. The pain may be caused by the blood not flowing properly from the left ventricle. The pain can also be caused by the thicker heart muscle not getting enough oxygen."
>
> Question by P6: "Prevention"
> Answer from Anna: "HCM is a chronic disease. This is a disease that cannot be cured. The complaints can be more severe in one period than the other. That also depends on the treatment and how you respond to it."

6.4 Discussion of the Results

The explored method of building a retrieval-based open question feature was evaluated through user testing. This qualitative evaluation was analyzed through thematic analysis and investigation of the algorithm's accuracy. The key findings of this evaluation are: (1) the platform and chat interface, in which the open question function is implemented, are easy to use, (2) users' perceptions and experiences show that participants see the value of the addition of the open question feature, and (3) the usefulness of the feature is limited due to a lack of accuracy. We will first discuss the thematic analysis followed by the algorithm's accuracy found through logging the questions asked by participants.

6.4.1 Thematic Analysis

The thematic analysis of the user evaluation can help us determine the users' attitude towards the open question answering feature (RQ2). While not every theme within the thematic analysis is directly related to this specific research question and the open question answering feature, we will still discuss the other trends found since they might be valuable for the further development of the chat-interface, Anna, or the eCG Family Clinic. Below the themes of 'Perceived Usefulness', 'Dialogue Management', and 'Ease of Use' are discussed in further detail.

The first theme, 'Perceived Usefulness,' gave us valuable insight into the users' perception of the expected benefits of using Anna to receive information through the sub-themes of 'Accuracy', 'Personalization', 'Flexibility', 'Trustworthiness', and 'Convenience'. Especially the sub-themes of 'Accuracy' and 'Personalization' are focused on implementing the open question feature and the predetermined questions. The first sub-theme shows that the open question feature could have been more useful if it had returned the desired answer more often. Here, participants described that the answers given often did not match the questions asked and that answers given were often very general. These questions were too specific for the algorithm. This preference for less generalized answers from Anna was also found in the 'Personalization' sub-theme. This need was indicated for both the open and the predetermined questions. However, users found the predetermined options more often answered their questions. Two points of improvement participants mentioned were: (1) fixing errors in the conversation flow and (2) having more predetermined questions. From the 'Perceived Usefulness' theme we can thus conclude that more accurate and personalized answers are needed to enhance the usefulness of the open and predetermined questions.

Moreover, the sub-themes of 'Flexibility' and 'Trustworthiness' also partially cover the predetermined and open questions. Here, the lack of accuracy is described as harming the trustworthiness of the open question feature. Additionally, users indicate that the integration of the open and closed questions could be combined in a manner where they complement each other. For instance: predetermined options could be more related to the users' needs based on the open questions asked. This could for instance be done by displaying related predetermined options while a user is still typing.

The theme 'Perceived Usefulness' also left us with some trends which could be valuable for future development but do not directly relate to the goal of this research. The sub-themes of 'Flexibility', 'Trustworthiness', and 'Convenience' included trends that could be related to Anna or the chat interface but not specifically to the question and answering feature. Firstly, the sub-theme 'Flexibility' described functionalities that would increase the benefits of using Anna for the users. These included being able to view the information from the earlier information screens during the interaction with Anna and an opportunity to archive answers or information given by Anna for later reference. Secondly, participants described in the sub-theme 'Convenience' when they would and would not use Anna for the purpose of gathering information. Thirdly, one danger that might need to be taken into account is that there is a lack of trustworthiness in Anna based on previous interactions with chatbots. However, we also found participants indicating they trust the information from Anna since it is from a reputable source. It is not known if this effect of trusting reputable sources is enough to counter the lack of trustworthiness of participants based upon previous experiences.

The second theme, 'Dialogue Management', gives us trends regarding controlling and guiding of the conversation with the user. This theme gives us valuable insights into the dialogue management used for open and predetermined questions. The most important key takeaway of this theme is that it is unclear to participants that they are able to ask open questions when pressing the 'I have other questions' button. Further, the sub-themes of 'Dialogue Flow' and 'Engagement' show that the overall reactions Anna gives are positively received, with the exception of a few specific sentences, which can seem redundant or too steering. Take for example Anna giving the same predetermined option multiple times and Anna moving on too quickly from specific topics.

The third theme, 'Ease of Use', shows a few final minor concerns about the effort needed to receive information. Here the 'Learnability' sub-theme shows that there is little to no learning curve to using the chat interface. The 'Navigation' sub-theme mainly highlights technical elements of the chat interface, which could be improved. While this theme does not contribute to the main takeaways of this research, improvements could be made to a future prototype of the chat interface.

6.4.2 Algorithm Accuracy

The log data quantitatively reflected the users' perception that the open question feature was not accurate. The algorithm's accuracy is limited to 19.2%, 25.3%, and 26.25% for top-1, top-2, and top-3 accuracy. Multiple factors could have influenced the accuracy. Firstly, the top-2 and top-3 accuracy measures do not describe the complete picture of the accuracy at that level since not every participant explored the second or third-best answer and did not get the opportunity to judge whether it was satisfactory. If all the participants had seen all the top three options, the top-2, and top-3 accuracy could potentially have been higher. Secondly, participants tend to reformulate questions for which no answer can be found. It could be that for that type of question there is no correct answer in the knowledge base. In these cases rephrasing will not result in finding a correct answer and will continue to lower the accuracy. Thirdly, we were strict with excluding the answers which only partially answered a user's question. Lastly, it could be argued that participants did not always carefully read the answer to the question asked before deeming it satisfactory or unsatisfactory. A participant, for example, asked about the prevention of HCM but got an answer regarding the treatment plan. Objectively, the answer does not align with the question, but the participant deemed the answer satisfactory. This means that the accuracy score does not always correctly reflect the objective satisfaction of a question by retrieved answers. The results of the algorithm accuracy suggest that the open question feature did not perform well, since many questions were answered unsatisfactorily or were not answered at all.

Additionally, the type of questions asked and the phrasing of questions differed per participant. This caused a high variation in accuracy between participants and could cause participants with the same question to receive a different set of answers. Thus, this shows that the algorithm, even though it uses semantic similarities to check for word similarity, is sensitive to the phrasing of the question.

Overall, it can be concluded that the open question feature is not accurate enough to be used in its current state. The lack of accuracy causes low acceptability since it harms the trustworthiness of the open question feature. However, participants do express that they see the value in the addition of this feature as it offers more flexibility and convenience than the predetermined question. We expect that the perceived usefulness and acceptability can be increased if future research focuses on improving the accuracy of the open question algorithm.

6.5 Implementation of Requirements

In Chapter 3, we formulated system requirements according to the MoSCoW method; see table 6.5 for the implementation status within this thesis. We can see that all 'must have' requirements have been implemented. From the 'should have' requirements only the expected answer type has not been implemented due to an outdated library.

Label	Requirement	Priority	Implementation
1	The user can type their open question.	Must have	Implemented - The user can type their question in a textbox and press the arrow to send it.
2	The algorithm is able to retrieve an appropriate answer from a list of website passages related to the subject just discussed by the virtual assistant.	Must have	Implemented - The algorithm returns an array of three relevant answers.
3	The algorithm makes use of keyword matching.	Must have	Implemented - The algorithm uses keyword matching with a bigram model and semantic similarity.
4	The algorithm makes use of sentence structure.	Must have	Implemented - The algorithm prioritizes sentences with similar structures.
5	The agent should be able to display chosen the answer.	Must have	Implemented - The chosen answer is displayed.
6	The open question answer function should be hosted in such a way that it will be able to communicate with the eCG virtual assistant.	Must have	Implemented - Both the eCG virtual assistant and the algorithm are hosted and communicate through an HTTP get request.
7	The eCG virtual assistant should have an option to switch between the predetermined question options and open questions.	Must have	Implemented - A user can navigate towards the open question feature by pressing 'I have other questions.'
8	A communication protocol should be specified between the assistant and hosted function.	Must have	Implemented - There is a communication protocol.
9	The algorithm makes use of expected answer type.	Should have	Not implemented - Due to time constraints and an outdated library, this was not implemented.
10	The algorithm should be able to indicate if an appropriate answer is not found.	Should have	Implemented - Anna will indicate when the answer is not found and will allow the user to reformulate their question.
11	The algorithm should give another response if appropriate answer is not found.	Should have	Implemented - The user can request the second and third best answer.
12	The algorithm should adapt responses bases upon the users situation.	Won't have	Not implemented.
13	The algorithm is able to translate a webpage into a list of passages.	Won't have	Not implemented.

Table 6.5: Overview design requirements implementation

7. Discussion

This research explored how an open question answering feature could be built for genetic counseling. This was achieved by looking at methods used within e-health conversational agents, implementing a prototype based upon those methods, evaluating the accuracy of the developed open question feature, and analyzing users' attitudes towards the open question feature.

To answer our main research questions, two subquestions were formulated. This discussion will focus on outlining the main findings, limitations, and future recommendations related to these research questions.

> RQ: How to design an open question answering feature for a virtual assistant in a genetic counseling platform?
>
> RQ1: What techniques for question answering systems can be used to develop a retrieval-based agent for open question answering in the context of a genetic counseling virtual assistant?
>
> RQ2: What is the users' attitude towards the open question answer feature in a genetic counseling virtual assistant?

7.1 Techniques for QA Systems

Through literature, it was determined that a retrieval-based method would best suit this research. Within a genetic counseling virtual assistant, answers given should be explainable. Thus, generative-based methods are problematic and not the best option. Additionally, using generative-based methods might cause concerns regarding the accuracy and validity of responses [27]. Through RQ1 we aim to find techniques that can be used to develop a retrieval-based agent for an open question answering function, specifically within genetic counseling. The developed retrieval-based algorithm is based on related literature and uses NER, Bigrams, Semantic Similarity, and POS tagging to detect similarities between user input and a potential answer in the knowledge base. This resulted in a working prototype with an accuracy of 19.2%, 25.3%, and 26.3% for the top-1, top-2, and top-3 accuracy respectively, with top-2 and top-3 accuracy possibly not describing the full objective accuracy. Looking purely at the top-1 accuracy, this is much lower than those of related open question features found in literature [5, 8, 18, 20] and under-performs the lowest accuracy of 33% percent found in related systems [20].

Reflecting back on the development of the algorithm, there are several limitations, such as the selection of technical components for the algorithm, the adaption of thresholds of the algorithm components, the addition of expected answer type as a component, and the knowledge source. These limitations are further discussed in Section 7.1.1. Several solutions to these limitations and suggestions for future work are discussed in Section 7.1.2.

7.1.1 Limitations

Firstly, selecting different techniques for the open question feature could potentially influence the algorithm's accuracy. The techniques used to build this retrieval-based method were based upon the research by Vargas-Vera and Lytras [38] and Damiano et al. [5], for example POS tagging, stop words removal, stemming, named entity recognition, and dependency parsing. In literature no methods were highlighted as best techniques to use, thus we selected and added components through an iterative design process. While the selected components resulted in a working prototype, no quantitative research was done into the best components to select.

Secondly, the thresholds of the individual technical components used in the algorithm could potentially be adapted as improvements in the open question algorithm. In this research they were not determined through a systematic and unbiased method, but were adapted through manual testing by adjusting the weights and thresholds to better answer questions we asked. This could have induced bias.

Thirdly, one of the components which were attempted to be implemented based on literature was the expected answer type. Kia et al. [18], for example, used this method through the Giveme5W1H tool[1] and Damiano et al. [5] implemented a deep learning approach to detect the answer type depending on the asked question. The Giveme5W1H tool seemed promising for this thesis and was thus attempted to be implemented in the open question answering algorithm. Hamborg, Breitinger, and Gipp [10] state that their tool can label pieces of text with what type of question it might answer. The possible labels are: what, why, where, when, and how. While implementing this tool, we discovered that the library was outdated and did not work sufficiently to be easily implemented. Thus it was decided to not implement this component in our algorithm. While this tool might positively affect the accuracy of the open question answering algorithm, this would need to be tested in future research.

Another limitation was that the knowledge base needed to include more (specific) information. For some questions participants in the user evaluation could not get a correct answer to their questions since the information in the knowledge base was too general. This could have had an influence on the accuracy of the algorithm. Firstly, this partly could be solved by broadening the knowledge source, researching the type of questions the participant will ask, and ensuring that information is included in the knowledge source. Secondly, in the user evaluation, we saw that participants preferred to receive more personalized answers instead of generalized answers retrieved from web pages. This need for personalization could be a broader problem and changes in the knowledge base might only partially solve this. Alternatives could be to implement a different type of knowledge source. While this would be possibly less flexible and take more effort for the developers/genetic counselors, it could potentially return more specific answers. For example: the answer targeting the specific sport the user practices instead

---

[1] https://github.com/fhamborg/Giveme5W1H

of just sports overall.

Moverover, the knowledge source was created by us through gathering paragraphs from approved websites. This method of setting up a knowledge source could have induced a bias in the information present. Alternative methods should have been considered to reduce bias and ensure a better knowledge source which could have functioned as ground truth.

Next to the main limitations and potential improvements, there are some more minor limitations that are worth mentioning. The eCG Family Clinic is in Dutch, while the algorithm handles English texts. To solve this, manual translation of the knowledge source and automatic translation of user questions have been done using Googletrans [2]. This could have slightly influenced the accuracy due to mistakes in automatic translation.

7.1.2 Future Work

Based upon the identified limitations of the open question answering feature, we suggest several areas of future research to further improve the algorithm's performance.

Firstly, future research could focus on examining which set of components would perform best. The selection in this research is based upon a limited search in literature and an iterative design process. A different set of components could potentially increase the accuracy achieved in this research. However, there is no clear overview of which components would ensure this. It is recommended to conduct a systematic literature review to create an overview of all retrieval-based question and answering research and the methods used together with their strengths and weaknesses. Moreover, this literature review should not be limited to counseling question and answering systems, since a broader overview would give us the opportunity to identify trends in the selection of components between research based upon performance. These trends, the performance, and component strengths and weaknesses would allow us to determine which components would need to be selected for an algorithm with a higher accuracy.

Secondly, for future research, a better knowledge base should be built. One way to achieve this improved knowledge base is to gather realistic questions from potential users of the system. This would hopefully give a better overview of which types of questions the virtual counselor should be able to answer. Next, health care professionals can be asked to write appropriate answers to these questions and we can test with users if the answers sufficiently answer their questions. Having a knowledge base with both questions and answers regarding HCM would allow us to have a ground truth dataset for future research. For example, within this thesis, this adaption of technical component thresholds was done through manual testing. A ground truth dataset, however, would allow us to adapt these thresholds using the grid search approach. Through this grid search, the optimum weights can be determined, which may increase the accuracy.

Another avenue for future research is the implementation of the expected answer type. More research should first be done into the best implementation method with, as examples, the Giveme5W1H tool [18] and the deep learning approach by Damiano et al. [5]. While it is not yet known what the effect of implementing such a method can be, it could drastically change the accuracy since the set of potential answers for a specific question

---

[2] https://pypi.org/project/googletrans/

could be drastically reduced due to a specific type of expected answer. We recommend exploring different methods of determining the expected answer type since we could then determine which method would perform best in a question-answering feature.

Lastly, future research could include investigating the best type of knowledge base to use in the context of a genetic counseling virtual assistant. The requirement of minimal effort for the experts to develop the knowledge base led us to choose a CSV file with excerpts and paragraphs of websites approved by those experts. However, the users express the desire to have answers which are more specific to their circumstances. We must ask ourselves if the current data format would suffice if this additional information were added. This information could potentially not be extracted from websites and would thus require more manual input from experts to write additional texts. If this extra manual input is needed, we have to determine if other types of knowledge bases might require less or similar effort from experts and might increase the algorithm's capability to answer with more precise answers. This could then possibly give users the feeling of receiving more personalized answers. While this research already discussed multiple types of knowledge bases such as ontologies, structured tables, or templates, more research could be done on other available approaches. Our recommendation is to explore ontologies and knowledge graphs to examine the effort and effects of using those knowledge bases in comparison to the approach chosen in this thesis.

7.2 Users' Attitude Towards the QA Feature

To answer RQ2, the users' attitude towards the open question answer feature was measured through thematic analysis guided by the Technology Acceptance Model (TAM). This user evaluation shows that participants express the value of adding the open question feature to the eCG Family Clinic due to its flexibility and convenience. However, the participants did express the necessity to increase the accuracy of the feature in order to receive a higher acceptability. Several limiting factors could be improved: a bias in participant selection, the extent of personalization, and the accuracy. Let us now discuss these subjects in more detail.

7.2.1 Limitations

Firstly, for this research, 8 participants were invited, who were all doing a master's. Six out of eight participants are studying in the field of interaction technology. Using this convenience sampling, the results cannot be generalized beyond the sample used [1]. We can assume that our sample of participants are all relatively well-versed with technology and would thus have fewer troubles interacting with the eCG Family Clinic. Additionally, we know that participants within our sample, highly educated and young people, are less likely to suffer from low health literacy [29]. This means that we exclude a big group of potential users. However, since this study focused on an initial exploration of the open question feature within the eCG Family Clinic, we argue that using a sample size where most participants are familiar with technology and similar systems would result in a more in-depth technical evaluation of the open question feature, Anna, and the eCG Family clinic.

Moreover, during the recruitment of participants, we assumed that they were non familiar with the eCG Family Clinic since it was only in its prototype stage and not publicly available. However, after concluding most of our user evaluations, we discovered that a similar user evaluation was being conducted for another research project regarding the eCG Family Clinic. For this user evaluation, the same pool of students was used, which caused participants of our user evaluation to have prior knowledge of the eCG Family Clinic. It is unclear what number of participants already had this prior knowledge due to participation in the other user evaluation prior to ours. Despite the potential cause of bias, the decision was made to keep the results and not repeat the experiments. Potential effects could be that participants knew the experiment targeted the open question feature, had less questions regarding the HCM information since they had read it twice, and had preconceived opinions regarding the eCG Family Clinic. After careful consideration, we estimated that the bias was likely limited since the other user test did not give them prior knowledge of the open question feature. Additionally, the experiments still provided us with valuable information.

Secondly, participants expressed a lack of personalization in the answers retrieved by the open question feature. Participants want to know the impact of HCM on their personal situation if they were to be tested positively for HCM, e.g., their sport, the impact on their family members, and how much chance they have of HCM. Literature shows that tailoring conversations to users' medical history and specifics can increase user satisfaction within e-health chatbots [30, 46]. This entails that future research might focus on researching how the retrieval-based method might be combined with more personalized information in order to increase the acceptability of the open question agent.

Thirdly, participants found that the open question answer feature needed to be more accurate. This heavily influences the perceived usefulness of this feature. Additionally, participants indicated that the trustworthiness of the virtual assistant was negatively influenced. Toader et al. [36] show that erroneous virtual assistants do have a negative impact on perceived competence and trust. This is thus in line with the observations made in our user evaluation. This decreased trust shows the necessity to increase the accuracy of our algorithm in the future.

Additionally, participants suggested that a combination of the open and predetermined questions might be useful for them. Suggestions included: having the predetermined questions adapt to the open questions asked and giving suggestions regarding frequently asked questions related to the open question while it is being typed. While this method of implementation was not considered prior to the start of this research, it might be valuable to explore the potential benefits of this implementation in future research.

Next to these main limitations and potential improvements, there are some limitations not directly related to RQ2, which are worth mentioning. Firstly, the eCG Family Clinic was still limited in its functionality since it was still a prototype. This caused some buttons not to work and some navigation issues, which the participants noticed. Moreover, not all topics which will be discussed in the eCG Family Clinic in the future were implemented. Secondly, for participants, there was unclarity as to where they could ask open questions since the text on the 'I have other questions' button suggested to some participants that more predetermined options would be presented when clicked. This unclarity should be considered in further research. Lastly, within conversational flow, several statements from Anna were found too steering and redundant; these could be adapted to increase user satisfaction towards the open questions feature and Anna overall.

7.2.2 Future Work

Firstly, future work can target the personalization of answers given to the user. Personalization in this context is targeted toward giving answers more specific to the user's circumstances. For example, the sport the user practices or the user's relation to the index patient. The best methods for adapting answers to users' circumstances should be researched. Additionally, the extent to which participants desire this personalization should be determined. This could be done by testing multiple types of personalization based upon literature.

Secondly, more research needs to be done into the best manner to combine open questions with predetermined question options. The implementation of this research did not focus on the UX in regards to the open question answering feature implementation; however, between participants, a trend could be seen that they did desire a more logical implementation. Participants also gave suggestions of possible implementations, such as adapting the predetermined question options based on what the user is typing and changing the predetermined options based on the last open question asked by the user. Further research in the literature could also result in additional options for better implementation. Both the research gathered implementation options, and the user suggestions could be explored further by implementing them and comparing these implementations through user evaluations.

Another avenue for future research would be to focus on the best chat interface for Anna. While this research did not focus on the chat interface, many participants in the user evaluation commented on the interface. Further exploring the chat interface could increase the users' overall satisfaction with chatting with Anna. An approach could be to identify common design features in textual conversational agents and messaging platform through literature. These could then be implemented to create a user interface which could be tested with potential users.

Lastly, more research could be done on the conversational flow within the chat with Anna. Participants indicated that several statements were found redundant or too steering by participants. More research could determine the consequences of these kinds of statements, how to avoid these kinds of statements, and how to best adapt the conversational flow to increase user satisfaction in the overall conversation. Since some research is already done on the design of conversational flow [16] we believe guidelines for a better conversational flow can be determined through a literature search.

7.3 Relevance

User tests shows that users see value and potential in the open question answering feature. However, we need to be aware of what such a feature could and could not do within genetic counseling. To ensure users get good care, there are several steps within the eCG Family Clinic that users have to complete. This can include reading specific information, answering certain questions, or conversing with Anna regarding a specific subject. Users of the system may not always want to go through these steps. However, the users are required to decide on genetic testing fully informed. This research's open question answering feature is developed as an addition to the eCG Family Clinic. This means that users can ask any additional questions they might have, but the question answering feature will not replace any of the mandatory steps the user has to go through within the system. While it could be explored in the future how to offer good care with solely an open conversation or how to replace several mandatory steps with an open question answering feature, this was not the focus of this research. However, reflecting upon the fast change within the natural language processing field, examining how such features and open conversations can offer users reliable and good care can be recommended.

Additionally, within the developed open question answering feature, more awareness should be created of when a conversation would need to be handed over to a genetic counselor. Take, for instance, a user who panics while asking open questions to Anna. In that case, the user should be connected with a genetic counselor via chat or for an appointment. The exact flags and points where this conversation should be handed over should be determined in future research.

Overall this research shows how such an open question answering feature can be implemented and provides the code used within this research. While the accuracy is not yet up to par, this research can contribute to other studies by forming a foundation on which can be built further. Additionally, this study offers insights into the design consideration and method which was used. Other studies can adapt this for further research or application within other domains.

8. Conclusion

This research set out to answer the main research question of *"How can an open question answering feature be designed for an virtual assistant in a genetic counseling platform?"*. To answer this question, we both looked at the techniques that can be used to develop this feature and measured users' attitudes toward this feature. Literature research determined that a retrieval-based open question feature would best fit the context of HCM genetic counseling offered within the eCG Family Clinic. The selected method used NER, Bigrams, Semantic Similarity, and POS tagging to select an answer from a set of paragraphs extracted from approved medical websites. While the accuracy could be further improved to compete with related systems, this method is a good foundation to build on for future research. User testing shows us that users see the value and potential of such an open question feature within the eCG Family Clinic since it offers flexibility and convenience. However, users did express the need for improvements and personalization.

Future research could focus on increasing the algorithm accuracy. Several avenues could be the selection of different question answering components, adapting the weights of the algorithm components, adding expected answer type as component and extending or adapting the knowledge sources. Additionally, more research could be done to increase users' attitude towards the open question answering feature: a personalization of answers given by the algorithm, a better integration between predetermined question options and the open questions, find the best way to design the chat-interface, and improve upon the conversational flow.

Overall, this research thoroughly explored which method would work best for an open question feature within the eCG Family Clinic and lays a good foundation for future work to build upon and further improve the open question feature.

Bibliography

[1] Anita S Acharya et al. “Sampling: Why and how of it”. In: *Indian Journal of Medical Specialties* 4.2 (2013), pages 330–333 (cited on page 64).

[2] Icek Ajzen. “The theory of planned behavior”. In: *Organizational behavior and human decision processes* 50.2 (1991), pages 179–211 (cited on page 37).

[3] Raquel Chocarro, Monica Cortinas, and Gustavo Marcos-Matás. “Teachers’ attitudes towards chatbots in education: a technology acceptance model approach considering the effect of social language, bot proactiveness, and users’ characteristics”. In: *Educational Studies* 49.2 (2023), pages 295–313 (cited on page 37).

[4] Bernardo Cuteri, Kristian Reale, and Francesco Ricca. “A logic-based question answering system for cultural heritage”. In: *European Conference on Logics in Artificial Intelligence*. Springer. 2019, pages 526–541 (cited on pages 19–24, 28).

[5] Emanuele Damiano et al. “Towards a framework for closed-domain question answering in Italian”. In: *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2016, pages 604–611 (cited on pages 19–24, 26, 61–63).

[6] Fred D Davis. “Perceived usefulness, perceived ease of use, and user acceptance of information technology”. In: *MIS quarterly* (1989), pages 319–340 (cited on page 37).

[7] Fred D Davis et al. “Technology acceptance model: TAM”. In: *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption* (1989), pages 205–219 (cited on page 37).

[8] Caner Derici et al. “A closed-domain question answering framework using reliable resources to assist students”. In: *Natural Language Engineering* 24.5 (2018), pages 725–762 (cited on pages 19–21, 23, 24, 26, 61).

[9] Paula Gardiner et al. “Reaching women through health information technology: the Gabby preconception care system”. In: *American Journal of Health Promotion* 27.3_suppl (2013), eS11–eS20 (cited on page 17).

[10] Felix Hamborg, Corinna Breitinger, and Bela Gipp. “Giveme5w1h: A universal system for extracting main events from news articles”. In: *arXiv preprint arXiv:1909.02766* (2019) (cited on page 62).

[11] Harteraad. *Erfelijke hartaandoeningen*. 2018. URL: https://www.harteraad.nl/wp-content/uploads/2018/06/brochure-erfelijke-hartaandoeningen.pdf (cited on page 16).

[12] *HCM*. Dec. 2019. URL: https://www.erfelijkehartziekten.nl/patienten/hcm/ (cited on pages 9, 16).

[13] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. "A survey on conversational agents/chatbots classification and design techniques". In: *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer. 2019, pages 946–956 (cited on page 9).

[14] *Hypertrofische Cardiomyopathie (HCM)*. URL: `https://www.umcutrecht.nl/nl/ziekenhuis/ziekte/hypertrofische-cardiomyopathie-hcm` (cited on pages 9, 16).

[15] *Hypertrophic cardiomyopathy*. May 2022. URL: `https://www.mayoclinic.org/diseases-conditions/hypertrophic-cardiomyopathy/symptoms-causes/syc-20350198` (cited on pages 9, 16).

[16] Mohit Jain et al. "Evaluating and informing the design of chatbots". In: *Proceedings of the 2018 designing interactive systems conference*. 2018, pages 895–906 (cited on page 66).

[17] Mary-Louise Jung and Karla Loria. "Acceptance of Swedish e-health services". In: *Journal of multidisciplinary healthcare* (2010), pages 55–63 (cited on page 37).

[18] Mahsa Abazari Kia et al. "Adaptable Closed-Domain Question Answering Using Contextualized CNN-Attention Models and Question Expansion". In: *IEEE Access* 10 (2022), pages 45080–45092 (cited on pages 19–24, 26, 61–63).

[19] Everlyne Kimani et al. "A smartphone-based virtual agent for atrial fibrillation education and counseling". In: *International conference on intelligent virtual agents*. Springer. 2016, pages 120–127 (cited on page 17).

[20] Majid Latifi, Horacio Rodriguez Hontoria, and Miquel Sànchez-Marrè. "ScoQAS: A semantic-based closed and open domain question answering system". In: *Procesamiento de lenguaje natural* 59 (2017), pages 73–80 (cited on pages 19–21, 23, 24, 26, 28, 61).

[21] Sweta P Lende and MM Raghuwanshi. "Question answering system on education acts using NLP techniques". In: *2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave)*. IEEE. 2016, pages 1–6 (cited on pages 19–24, 26).

[22] Christine Lisetti et al. "I can help you change! an empathic virtual agent delivers behavior change health interventions". In: *ACM Transactions on Management Information Systems (TMIS)* 4.4 (2013), pages 1–28 (cited on page 16).

[23] Nikola Marangunić and Andrina Granić. "Technology acceptance model: a literature review from 1986 to 2013". In: *Universal access in the information society* 14 (2015), pages 81–95 (cited on page 37).

[24] Mónica Marrero et al. "Named entity recognition: fallacies, challenges and opportunities". In: *Computer Standards & Interfaces* 35.5 (2013), pages 482–489 (cited on page 29).

[25] Jennifer Gladys Mulle and Viola Vaccarino. "Cardiovascular disease, psychosocial factors, and genetics: the case of depression". In: *Progress in cardiovascular diseases* 55.6 (2013), pages 557–562 (cited on page 9).

[26] Prasanth Murali et al. "Health Counseling by Robots: Modalities for Breastfeeding Promotion". In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2019, pages 1–6 (cited on pages 16, 17).

[27] San Murugesan and Aswani Kumar Cherukuri. "The Rise of Generative Artificial Intelligence and Its Impact on Education: The Promises and Perils". In: *Computer* 56.5 (2023), pages 116–121 (cited on page 61).

[28] Elizabeth G Nabel. "Cardiovascular disease". In: *New England Journal of Medicine* 349.1 (2003), pages 60–72 (cited on page 9).

[29] Michael K Paasche-Orlow et al. "The prevalence of limited health literacy". In: *Journal of general internal medicine* 20.2 (2005), pages 175–184 (cited on page 64).

[30] Juanan Pereira and Óscar Dıaz. "Using health chatbots for behavior change: a mapping study". In: *Journal of medical systems* 43 (2019), pages 1–13 (cited on page 65).

[31] Deborah Richards and Patrina Caldwell. "Improving health outcomes sooner rather than later via an interactive website and virtual specialist". In: *IEEE journal of biomedical and health informatics* 22.5 (2017), pages 1699–1706 (cited on page 17).

[32] Lazlo Ring, Timothy Bickmore, and Paola Pedrelli. "An affectively aware virtual therapist for depression counseling". In: *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) workshop on Computing and Mental Health*. 2016, pages 01951–12 (cited on pages 16, 17).

[33] Mandeep Sekhon. "Acceptability of healthcare interventions". PhD thesis. City, University of London, 2017 (cited on page 37).

[34] Satwinder Singh and Himanshu Beniwal. "A survey on near-human conversational agents". In: *Journal of King Saud University-Computer and Information Sciences* (2021) (cited on pages 17, 24).

[35] Centraal Bureau voor de Statistiek. *54 procent sterfgevallen in 2019 door kanker of hart- en vaatziekten*. June 2020. URL: `https://www.cbs.nl/nl-nl/nieuws/2020/27/54-procent-sterfgevallen-in-2019-door-kanker-of-hart-en-vaatziekten#:~:text=Het%20aantal%20mensen%20dat%20overlijdt,bij%20vrouwen%20(15%20procent).` (cited on page 9).

[36] Diana-Cezara Toader et al. "The effect of social presence and chatbot errors on trust". In: *Sustainability* 12.1 (2019), page 256 (cited on page 65).

[37] Universitair Medisch Centrum Utrecht. *Door erfelijke ziekte geen verzekering*. 2019. URL: `https://www.umcutrecht.nl/nieuws/door-erfelijke-ziekte-geen-verzekering` (cited on page 16).

[38] Maria Vargas-Vera and Miltiadis D Lytras. "Aqua: A closed-domain question answering system". In: *Information systems management* 27.3 (2010), pages 217–225 (cited on pages 19–21, 23, 24, 26, 28, 62).

[39] Viswanath Venkatesh and Hillol Bala. "Technology acceptance model 3 and a research agenda on interventions". In: *Decision sciences* 39.2 (2008), pages 273–315 (cited on page 37).

[40] Viswanath Venkatesh and Fred D Davis. "A model of the antecedents of perceived ease of use: Development and test". In: *Decision sciences* 27.3 (1996), pages 451–481 (cited on page 38).

[41] Viswanath Venkatesh and Fred D Davis. "A theoretical extension of the technology acceptance model: Four longitudinal field studies". In: *Management science* 46.2 (2000), pages 186–204 (cited on page 37).

[42] Viswanath Venkatesh et al. "User acceptance of information technology: Toward a unified view". In: *MIS quarterly* (2003), pages 425–478 (cited on page 37).

[43] Kelly Waters. "Prioritization using moscow". In: *Agile Planning* 12 (2009), page 31 (cited on page 25).

[44] Munazza Zaib et al. "Conversational question answering: A survey". In: *arXiv preprint arXiv:2106.00874* (2021) (cited on pages 17, 19, 24, 26).

[45] Shuo Zhou and Timothy Bickmore. "Automating Cancer Genetic Counseling with an Adaptive Pedagogical Agent". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pages 1–7 (cited on pages 16, 17).

[46] Yonghan Zhu, Rui Wang, and Chengyan Pu. ""I am chatbot, your virtual mental health adviser." What drives citizens' satisfaction and continuance intention toward mental health chatbots during the COVID-19 pandemic? An empirical study in China". In: *Digital Health* 8 (2022), page 20552076221090031 (cited on page 65).

A. Research Protocol

Onderzoeksprotocol eCG Family Clinic

In dit protocol staat de taak van de onderzoeker gedefinieerd en daarbij de stappen die de onderzoeker moet nemen om het onderzoek consistent uit te voeren. Dit stappenplan is onderverdeeld in meerdere secties:

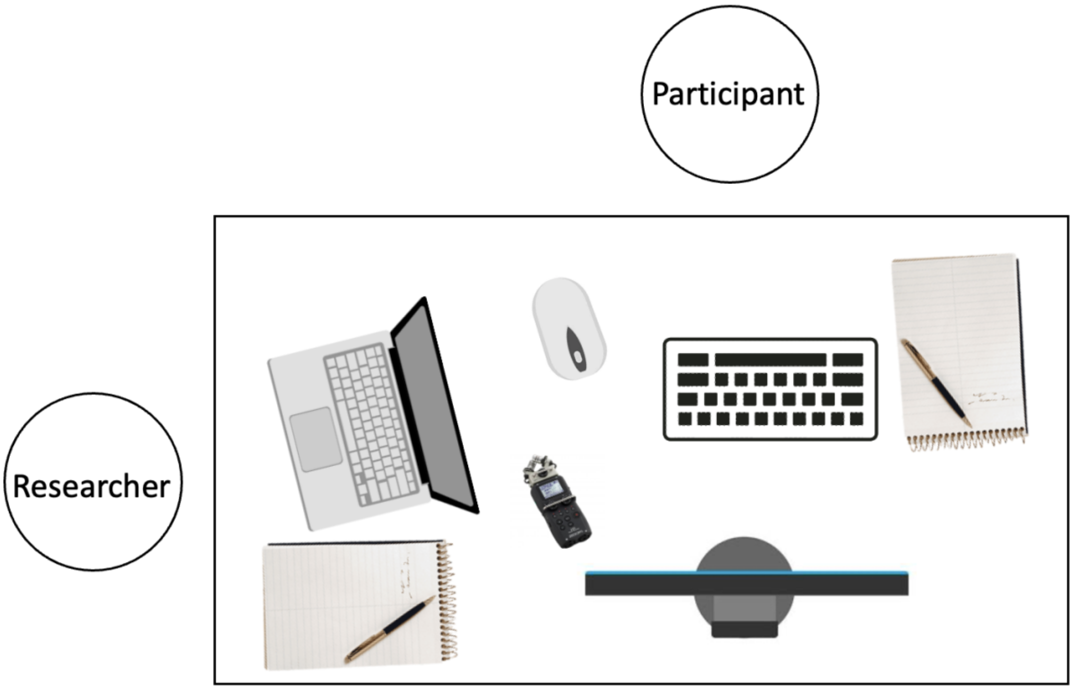
1. De onderzoeker bereidt het experiment voor.
2. De onderzoeker verwelkomt en brieft de participant.
3. De onderzoeker laat de participant de eCG Family clinic laten gebruiken.
4. De onderzoeker zal een semigestructureerd interview afnemen bij de participant.
5. De onderzoeker sluit de sessie af.

Elke sectie is hieronder toegelicht. Langere teksten die de onderzoeker moet vertellen staan in een blauw vak waarbij acties die de onderzoeker moet doen dikgedrukt zijn.

Voorbereiding

De onderzoeker zorgt dat de ruimte en materialen klaar zijn voordat de participant binnen komt. Dit houdt in dat de volgende dingen klaarstaan en geregeld worden:

- De onderzoeker zet de ruimte op zoals geïllustreerd in figuur 1. Dit houdt in dat de participant en onderzoeker schuin tegenover elkaar zitten op een manier waardoor de participant en onderzoeker oogcontact hebben tijdens het interview. De participant heeft een laptop voor zich met een losse muis. De onderzoeker heeft een extern scherm voor zich die niet zichtbaar is voor de participant met een mirror van het laptopscherm. De participant heeft een leeg A4 blad met een pen naast de laptop liggen voor eventuele aantekeningen. De onderzoeker heeft een notitie blad voor zich voor eventuele aantekeningen.



Figuur 1: De setup

- De onderzoeker heeft een uitgeprinte versie van dit protocol voor zich.
- De audiorecorder staat klaar om te gebruiken om audio mee op te nemen.
- De laptop staat open met twee instanties van de eCG Family Clinic.
    - De eerste instantie van de eCG Family Clinic staat open op de start pagina van de demo, getiteld: ‘Welkom!’. Deze instantie staat fullscreen op de laptop.
    - De tweede instantie van de eCG Family Clinic staat open op de dialoog met Anna. Om deze instantie klaar te kunnen zetten zal de onderzoeker door de applicatie moeten klikken totdat dit dialoog is bereikt. Op de pagina ‘Hoe wordt HCM voorkomen of behandeld?’ staan twee vragen die beantwoord moeten worden voordat je op de dialoog uitkomt. Deze vragen beantwoord de onderzoeker met ‘nee’.

Verwelkomen en briefing

Als de participant binnen komt is deze verwelkomt door de onderzoeker en word de stoel aangewezen waar de participant plaats mag nemen. De onderzoeker stelt de participant op zijn gemak en begint daarna met de briefing. De onderzoeker vertelt het volgende en beantwoord vragen als de participant deze heeft:

Bedankt voor het deelnemen in dit onderzoek. Voordat we beginnen wil ik je graag dit informatie formulier laten doorlezen.

Geef informatie formulier

Deelname aan dit onderzoek is geheel vrijwillig en kan op elk moment worden ingetrokken. Let op: het gaat over een ziekte en hier worden onder andere symptomen besproken, neem bijvoorbeeld plotseling overlijden. Mocht je je niet comfortabel hierbij voelen, dan kun je op ieder moment je deelname terugtrekken. Dit kan tot 48 uur na afloop van het onderzoek. Ook kan je aangeven als je vragen niet wil beantwoorden. Ik ben niet medisch opgeleid en kan geen vragen beantwoorden op medisch vlak. Mocht je na afloop vragen hebben, dan verwijs ik je door naar je huisarts. Dit onderzoek omvat één sessie met een duur van ongeveer 45-60 minuten. We zullen straks enkele dingen gaan doen. Ten eerste zal ik je straks een introductie geven, daarna mag je gebruik maken van de eCG Family Clinic, daarna zal ik je vragen een vragenlijst met 24 vragen in de vullen en als afsluiting zal ik je interviewen. Je mag het informatieformulier doorlezen.

Kijk of de participant vragen heeft en beantwoord deze.

Als er geen vragen meer zijn geef het consentformulier aan de participant.

Ik heb hier het consentformulier. Lees het formulier door en vink het vakje aan die voor jou van toepassing is. Ik geef je twee formulieren, waarvan je er 1 mee naar huis mag nemen.

Check na het tekenen of de participant overall ‘Ja’ of heeft geantwoord. Als dit het geval is, ga dan door. Als er een nee staat of als de participant niet wil tekenen dan sluit de onderzoeker op een vriendelijke manier het onderzoek af en bedankt de onderzoeker de participant voor zijn tijd.

De onderzoeker geeft aan dat als er geen objecties zijn, de audio opname nu word gestart en start vervolgens de audio opname.

Oke, dan gaan we beginnen. Ik zal je straks vragen om de eCG Family clinic te gebruiken. Voordat je het gebruikt wil ik je het scenario schetsen van een algemene gebruiker van dit systeem, zodat je je kan inbeelden vanuit de gebruiker die normaal in aanraking komt met de eCG Family Clinic.

Stel voor dat je een bericht hebt gekregen van een familielid dat bij hun HCM is geconstateerd. Hierdoor blijkt het dat jij, omdat je familie bent, ook het risico loopt om HCM te hebben. Jij hebt via dat familielid inlog gegevens gekregen voor de eCG Family Clinic. Dit is jouw eerste interactie met de eCG Family Clinic.

Je zult eerst wat informatie krijgen van de eCG Family Clinic en daarna de mogelijkheid hebben om vragen te stellen aan het platform. Eventuele vragen die je hebt kun je opschrijven op het kladpapier voor je.

We zullen vandaag werken met een prototype. Dat houdt in dat nog niet alle knopjes werken. Deze knopjes boven aan werken niet en deze knopjes aan de zijkant werken niet. Ook wordt maar een deel van de informatie behandeld die een normale gebruiker zou krijgen en is er geen toegang tot functies zoals het contact met een arts.

Probeer zo hardop te denken en lezen. Bij deze eerste pagina die je voor je ziet zal ik een voorbeeld geven van hardop denken:
Ik zie een scherm waarop ik verwelkomt wordt en er staat ga door als je klaar bent om te beginnen. Ik zie hier een knop ‘ga door’ en klik daarop.

Participant interactie met eCG Family Clinic

De onderzoeker laat de gebruiker zelf door de applicatie lopen. Eerst zal de participant dit doen op een natuurlijke manier met aan het eind de kans om zelf open vragen te stellen. Als de participant hier klaar mee is zal er enkele opdrachten worden gegeven aan de participant. Deze twee stappen en voorbeelden van wanneer er ingegrepen dient te worden staan hieronder verder uitgewerkt.

Applicatie gebruiken

De participant mag via de laptop die voor zich staat de applicatie gebruiken. Bij het punt ‘Wanneer ingrijpen’ staat beschreven welke acties de onderzoeker moet nemen. Wanneer de participant aan het eind van de applicatie is, mag de onderzoeker naar de volgende stap gaan. Het eind van de applicatie houdt in dat de participant vragen heeft kunnen stellen aan Anna en verder geen vragen meer heeft. De onderzoeker houd bij welke van de thema’s in de opdrachten al geadresseerd zijn door de participant en of de participant gebruik heeft gemaakt van de gesloten opties.

Opdrachten

Wanneer de participant klaar is, geeft de onderzoeker aan dat er nu enkele opdrachten zijn. Wanneer nodig kan de onderzoeker naar de andere instantie van de eCG Family Clinic gaan als de participant op ‘ga door’ heeft gedrukt.

Bedankt. Dat was de eerste stap, nu heb ik enkele opdrachten voor je. Je mag zo vragen gaan stellen aan Anna. Ik heb hier een lijst met thema’s waarover je vragen mag stellen.

Geef het overzicht aan de participant en streep thema’s af die de participant al heeft behandeld. Bij twijfel over een thema moet het thema niet worden afgestreept.

Deze thema’s heb je net al een vraag over gesteld en hoef je niet meer mee te nemen. Dat houdt in dat je over elk resterende thema een vraag mag stellen. Dus in totaal … vragen. Zorg dat je minimaal 1 vraag stelt via de gesloten optie en minimaal 1 vraag via de open optie. De volgorden en welke vragen dat zijn, mag je zelf beslissen. Heb je nog vragen?

De onderzoeker geeft het uitgeprinte overzicht van de volgende onderwerpen:
Sport & HCM
Gewichten tillen & HCM
Symptomen
Behandeling
Voorkomen
Betekenis
Hart
Erfelijkheid

De onderzoeker ligt de opdrachten toe waar nodig en beantwoord vragen. Tijdens deze opdrachten laat de onderzoeker de participant deze opdrachten zelf afhandelen en grijpt alleen in wanneer het gaat over een situatie zoals hieronder beschreven.

Wanneer ingrijpen

De onderzoeker mag ingrijpen in situaties die niet direct gerelateerd zijn met het vragen stellen. Hier mag de onderzoeker de participant verder helpen. Bij situaties die wel gerelateerd zijn aan het vragen stellen moet de onderzoeker eerst afwachten en inschatten of de participant het zelf kan oplossen door trial and error. Als het blijkt dat de participant echt vastzit kan alleen dan de onderzoeker ingrijpen. De onderzoeker mag nooit medisch advies geven of medische vragen beantwoorden. Hieronder zijn enkele situaties geschetst en daarbij de gewenste reactie van de onderzoeker.

Situatie: Er is onduidelijkheid over hoe de ecg-applicatie werkt en deze onduidelijkheid is niet verbonden aan het stellen van vragen. Hieronder valt: 3 keer klikken op een knop, actief

een vraag krijgen van de participant over deze functionaliteit, op zoek zijn naar een functionaliteit die de applicatie niet bevat (na 15 seconden ingrijpen).
Reactie: In dit geval legt de onderzoeker uit hoe de participant verder moet.

Situatie: Er is onduidelijk over hoe de ecg-applicatie werkt en deze onduidelijkheid is verbonden aan het stellen van vragen.
Reactie: De onderzoeker wacht af om te kijken of de participant er zelf uit komt. Mocht het blijken dat de participant echt vastzit (10 keer klikken, actief vragen, of 2x dezelfde vraag aan Anna stellen, en op zoek naar een functionaliteit die er niet is waarbij de participant elke optie heeft geprobeerd en na 1 minuut er nog steeds niet uit is), dan grijpt de onderzoeker in door een aanwijzing te geven (aangeven dat een functie er niet is, dat Anna het antwoord waarschijnlijk niet weet, dat de participant door kan gaan met de volgende vraag)

Situatie: De participant stelt de onderzoeker een medische vraag aan de onderzoeker waarover de participant zich duidelijk geen zorgen maakt.
Reactie: De onderzoeker geeft aan dat hij/zij geen medische ervaring heeft, deze vraag niet kan beantwoorden en dat de participant met vragen en zorgen bij een arts te recht kan.

Situatie: De participant stelt de onderzoeker een medische vraag aan de onderzoeker waarover de participant zich duidelijk zorgen maakt.
Reactie: De onderzoeker geeft aan dat hij/zij geen medische ervaring heeft, deze vraag niet kan beantwoorden en dat de participant met vragen en zorgen bij een arts te recht kan. Ook geeft de onderzoeker aan dat als de participant zich oncomfortabel voelt of niet verder wil, dat een optie is.

Situatie: De participant stelt de onderzoeker een medische vraag aan de onderzoeker en de onderzoeker weet niet of de participant zich er zorgen over maakt.
Reactie: De onderzoeker geeft aan dat hij/zij geen medische ervaring heeft, deze vraag niet kan beantwoorden en dat de participant met vragen en zorgen bij een arts te recht kan. Ook geeft de onderzoeker aan dat als de participant zich oncomfortabel voelt of niet verder wil, dat een optie is.

Situatie: De participant blijft proberen dezelfde vraag te stellen aan Anna, maar er is duidelijk dat er geen antwoord terug gaat komen.
Reactie: Als de participant na meerdere malen proberen niet het antwoord krijgt wat hij/zij wil, dan mag de onderzoeker ingrijpen en aangeven dat Anna dan waarschijnlijk geen antwoord heeft.

Situatie: De participant denkt en leest niet hardop. Na 15 seconde stilte ingrijpen.
Reactie: Wijs de participant hierop.

Interview

De onderzoeker geeft aan dat er nu enkele vragen komen en dat er geen goede of foute antwoorden zijn. De vragen die hieronder genoemd zijn, zijn bedoeld als houvast. Dit houdt in dat de onderzoeker zelf zal doorvragen op de onderwerpen en vooral een duidelijk beeld probeert te krijgen van de participant over de open en gesloten vragen.

De vragen die de onderzoeker stelt zijn:

- Voordat je met Anna kennismaakte, wat verwachtte je van Anna?
    - Kun je voorbeelden geven op welke manier Anna hier wel of niet aan voldeed?
- Zou je kunnen omschrijven hoe het gebruik van Anna het verkrijgen van informatie voor jou zou beïnvloeden?
    - Als je vragen hebt rondom een medisch onderwerp, op welke manier vind je normaal je antwoord op deze vraag?
        - Kun je uitleggen hoe je denkt dat het vragen stellen aan Anna sneller of langzamer zou kunnen zijn dan andere methodes?
            - Zit hierbij verschil tussen het open vragen stellen en het gesloten vragen stellen?
        - Hoe zou je de impact van Anna op de productiviteit van het krijgen van antwoorden beschrijven?
            - Zit hierbij verschil tussen het open vragen stellen en het gesloten vragen stellen?
        - Kun je omschrijven hoe het gebruik van Anna de effectiviteit van het krijgen van antwoorden zou veranderen ten opzichte van andere methodes?
            - Zit hierbij verschil tussen het open vragen stellen en het gesloten vragen stellen?
        - Op welke manieren zou, door het gebruik van Anna, het krijgen van antwoord op de vragen makkelijker of moeilijker zijn?
            - Zit hierbij verschil tussen het open vragen stellen en het gesloten vragen stellen?
        - Als jij als gebruiker vragen zou hebben na het lezen van informatie binnen de eCG Family Clinic, welke manier zou jij dan het liefst antwoorden op deze vragen gaan zoeken?
            - Waarom zou je die methode kiezen boven andere methodes?
            - Zit hierbij verschil tussen het open vragen stellen en het gesloten vragen stellen?
    - Zou je kunnen omschrijven van je vond van het communiceren met Anna?
        - Kun je je ervaring omschrijven rondom het uitvinden hoe je open en gesloten vragen kon stellen?
        - Hoe veel moeite koste het jou om Anna te laten doen wat je wou?
        - Op welke manier was je interactie met Anna duidelijk of onduidelijk?
        - Zou je kunnen omschrijven in welke opzichten Anna flexibel is in het gebruik of waar Anna flexibiliteit mist?
        - Zou je kunnen omschrijven hoe lang en hoeveel moeite het jou zou kosten om behendig te worden in het gebruik van Anna?
        - Op welke manieren zou Anna moeilijk of makkelijk zijn om te gebruiken?

- Heb je het gevoel dat jouw vragen volledig zijn beantwoord?
    - Kun je voorbeelden geven van welke vragen wel of niet? (Alleen vragen als de pagina met het gesprek met Anna nog open staat)

- Heb je andere onderwerpen die wij nog niet hebben besproken waar je nog wat over kwijt wil?

Afsluiten sessie

Bedankt voor je deelname, dit was het onderzoek. Zoals je misschien al doorhad ging dit experiment voornamelijk over de evaluatie over het vragen stellen richting Anna. Het ging ook specifiek over het stellen van open vragen en de evaluatie daarvan.

Mocht je medische vragen of zorgen hebben, dan vraag ik je om je dokter op te zoeken. Als je nog vragen hebt rondom het onderzoek dan kun je het zeker laten weten.

Mocht je je deelname nog willen intrekken, dan kan dit nog 48 uur door mij een bericht te sturen.

Voor nu: heb ik nog een klein bedankje. **Geef snack.**

Mocht je nog vragen hebben. Stel ze vooral!

De onderzoeker vraagt aan de participant of de participant nog vragen heeft. Als niet, dan worden deze bedankt voor hun deelname, de audio opname wordt gestopt en de participant wordt een snack aangeboden.

B. Informed Consent Form

Toestemmingsformulier voor gebruikerstest met eCG Family Clinic

U KRIJGT EEN KOPIE VAN DIT GEÏNFORMEERDE TOESTEMMINGSFORMULIER

Vink de vakjes aan die van toepassing zijn.	**Ja**	**Nee**
Deelname in het experiment		
Ik heb het informatieformulier d.d. [12/04/2023] gelezen en begrepen, of is mij voorgelezen. Ik heb vragen kunnen stellen over het onderzoek en deze vragen zijn naar mijn tevredenheid beantwoord.	☐	☐
Ik neem vrijwillig deel aan dit onderzoek en begrijp dat ik kan weigeren vragen te beantwoorden en ik me op elk moment tot 48 uur na afloop van het onderzoek me kan terugtrekken uit dit onderzoek, zonder dat ik een reden hoef op te geven.	☐	☐
Ik begrijp dat deelnemen aan dit onderzoek inhoud dat ik gebruik maak van de eCG Family Clinic, er mij wordt gevraagd een vragenlijst in te vullen, ik word gevraagd om deel te nemen in een interview, en dat de audio van deze hele sessie zal worden opgenomen.	☐	☐
Gebruik van informatie in het onderzoek		
Ik begrijp dat de door mij verstrekte informatie zal worden gebruikt voor het masteronderzoek van de onderzoeker.	☐	☐
Ik begrijp dat de verzamelde informatie wordt geanonimiseerd zodat deze niet naar mij kan worden herleid en geen persoonlijke informatie zal bevatten. Mijn naam zal niet buiten het onderzoeksteam gedeeld worden.	☐	☐
Ik ga ermee akkoord dat mijn informatie kan worden geciteerd in onderzoeksresultaten en begrijp dat deze citaten worden geanonimiseerd.	☐	☐
Ik ga ermee akkoord dat de audio van deze sessie wordt opgenomen en dat deze zal worden vernietigd zodra deze is getranscribeerd.	☐	☐

Handtekeningen

______________________ ______________________ ________

Naam van de participant

Handtekening Datum

Ik heb het informatieformulier nauwkeurig voorgelezen aan de potentiële deelnemer of deze door de deelnemer laten lezen en, naar mijn beste kunnen, ervoor gezorgd dat de deelnemer begrijpt waar hij vrijwillig mee instemt.

_________________________ ___________________ ________

Maaike Keurhorst

Handtekening Datum

Contact gegevens voor meer informatie:

Maaike Keurhorst (m.keurhorst@student.utwente.nl)

Contactgegevens voor vragen rondom rechten als deelnemer van dit onderzoek

Als u vragen heeft over uw rechten als deelnemer aan dit onderzoek, informatie wilt, vragen wilt stellen of zorgen over dit onderzoek wilt bespreken met iemand anders dan de onderzoeker, kunt u contact opnemen met de secretaris van de Ethische Commissie Information & Computer Science: ethicscommittee-CIS@utwente.nl.

C. Information Form

Informatie formulier: Genetic counsellor in de eCG Family Clinic

Onderzoeker: Maaike Keurhorst (m.keurhorst@student.utwente.nl)

Versie: 12 april 2023

Samenvatting van het project:

De eCG Family Clinic is een online omgeving die wordt ontwikkeld om geïnformeerde besluitvorming van familieleden die risico lopen op genetische hypertrofische cardiomyopathie (HCM) te stimuleren. Dit project is een samenwerking tussen het Universitair Medisch Centrum Utrecht (UMCU), KAESS en Human Media Interaction (HMI) van de Universiteit Twente. Informatie wordt aangeboden door teksten en de mogelijkheid tot het stellen van vragen. Het doel van dit onderzoek is om de eCG Family Clinic te evalueren.

Procedure:

Deelname aan dit onderzoek is geheel vrijwillig en kan op elk moment worden ingetrokken. Dit onderzoek omvat één sessie met een duur van ongeveer 45-60 minuten. De werkwijze is als volgt:

- Ontvang een briefing van wat de eCG Family Clinic is.
- Maak gebruik van de eCG Family Clinic.
- Neem deel in een interview.

Tijdens het testen wordt u gevraagd hardop te denken. Dit betekent dat u wordt gevraagd om eventuele gedachten over de eCG Family Clinic te verwoorden. De hele sessie wordt opgenomen en de onderzoeker maakt aantekeningen van uw antwoorden. U kunt op elk gewenst moment weigeren op vragen te reageren en/of uw deelname intrekken zonder gevolgen. Intrekken van deelname kan tot 48 uur na afloop van het experiment. Hierna zal de opname wordt getranscribeerd met behulp van de Microsoft Word-transcriptietool, waarna de audio-opname wordt verwijderd en daarbij ook informatie en data die naar u te herleiden is.

Vertrouwelijkheid en beveiliging van gegevens:

De gegevens die in dit onderzoek worden verzameld, zijn anoniem en worden opgeslagen op UT-systemen. Uw identiteit zal vertrouwelijk zijn en uw naam zal niet worden gebruikt bij publicatie van resultaten. In dit onderzoek kunnen citaten uit uw interview worden gebruikt. In dat geval worden de quotes geanonimiseerd; citaten die mogelijk uw identiteit zouden kunnen onthullen, worden niet gebruikt.

De gegevens die in dit onderzoek zijn verzameld, omvatten:

- Automatisch getranscribeerde (en geanonimiseerde) tekst van het interview en testen.
- Aantekeningen door de onderzoeker.
- Vragen die door de gebruiker aan de agent worden gesteld. Deze worden opgeslagen in een computerlogbestand.

U wordt om toestemming gevraagd voor het opnemen van audio van het experiment. Deze audio-opname wordt getranscribeerd in een geanonimiseerde tekst, waarbij persoonlijke gegevens zoals namen worden verwijderd. De audio-opnamen worden na transcriberen vernietigd. Er zal geen verder gebruik worden gemaakt van verzamelde gegevens zonder extra toestemming te vragen.

Risico's en voordelen:

Tijdens dit experiment behandelen we informatie over HCM (Erfelijke Cardiovasculaire Myopathie). Al deze informatie is afkomstig van gevalideerde websites, maar de virtuele assistent kan een antwoord geven dat niet relevant is voor uw vraag. De onderzoeker is geen medische professional en kan eventuele vragen over dit domein niet beantwoorden. Bovendien kan de onderzoeker geen medisch advies geven. Heeft u toch vragen over HCM of medisch domein, neem dan contact op met een arts. Als u zich niet op uw gemak voelt bij een besproken onderwerp, kunt u zich op elk moment tijdens het experiment terugtrekken of vragen niet beantwoorden.

Voor dit experiment wordt geen vergoeding aangeboden, maar de onderzoeker zal je heel erg dankbaar zijn en een versnapering geven.

Ethiek en zorgen

Als u vragen heeft over uw rechten als deelnemer aan dit onderzoek, informatie wilt, vragen wilt stellen of zorgen over dit onderzoek wilt bespreken met iemand anders dan de onderzoeker, kunt u contact opnemen met de secretaris van de Ethische Commissie Information & Computer Science: ethicscommittee-CIS@utwente.nl.