Beyond **digital diagnosis:**

Visualizing decision making paths of different Self-Diagnosis Apps

Name: Zixun Huang
Student ID: 10648500

Tutor: Michele Mauri
Beatrice Gobbo

LM Communication Design 2020
Design School - Politecnico di Milano

Abstract

The Self-Diagnosis application, which use the chatbot system to help people check symptom are more widely used in recent years. While they promise faster, more convenient and accurate, little is known about how their inner algorithm works and what are the differences between different diagnostic paths for lay users, especially when they give unfaithful outputs. According to the user review on Google Play Store, negative reviews about the output account for not a small proportion of diagnosis applications, and through the subsequent experiments, the different decision-making paths can result in totally different outputs even with the same input. Therefore, there is a keen need to interpret the diagnostic process and build an explanation path for lay users or relevant stakeholders to help them understand how this kind of algorithm takes high-stake decisions in people's daily life.

The user reviews of the top 3 downloads diagnosis applications on Google Play Store are collected in this project, intending to define the most common pain point of symptom and disease from the user's perspective, and then use these symptoms as cases to build a visualization method to disclose how decision-making model works and show the comparison of different diagnostic paths in an understandable way.

Key word: Self-Diagnosis Application; AI chatbot; Decision making path; Question-answering system; Visualization model;

Index

1. Introduction

With the rapidly growing of Artificial Intelligence and machine learning technologies, using smart applications to help people make some important decisions is more and more common in our daily life, such as diagnostic applications for symptom checking. Although the digital diagnosis provides users with a fast, low-cost and highly portable health assessment through smart algorithms, there is still s a wide gap between the lay user's elementary knowledge and those sophisticated machine learning models.

1.1 Diagnosis Applications

In recent years, the development of the artificial intelligence industry results in the rise of advanced digital medical and health platforms, and seeking health information or help on the internet becomes more and more common in people's daily life. Until 2019, there are over 100,000 mobile phone software applications[1] have been designed for the medical and health field, including diagnosis, monitoring, providing information and so on according to the function. Among those categories, diagnosis application constitutes a major part and it has more than 200,000 downloads and over 100,000 user reviews on both Google Play store and App store.

Digital diagnosis provides users with a fast, low-cost and highly portable health assessment through complicated and complex algorithms, which usually is an AI Chatbot system. The AI Chatbot applied in diagnosis applications contains various machine learning models, including natural language processing and deep learning, which finally presents as a question-answering system for the user. They input the key word of their symptoms, and then the system will give possible causes of the symptom after a series of question and answer paths. Although sophisticated computer science allowed the digital diagnosis faster, smarter, more convenient and bigger database, the potential of negative consequences still exist such as inaccurate diagnosis and inappropriate treatment. Moreover, little is known about the inner work of this kind of application, since its final user is general public rather than medical experts and no adequate professional information is provided for them.Thus, we cannot deny that there is potential risk for lay users when they are excluded from relevant knowledge. For instance, an elderly person who without any medical knowledge wants to seek some information for his mild headache online, and the diagnosis application tells him that he may get cancer after asking a few simple questions. Leave aside if the result is unfaithful or not, just the mysterious algorithm procedure can make him panic-stricken, which seems like totally a black box for him. Unfortunately, this kind of case happens hundreds of times based on the user reviews on Google Play Store and App Store.

[1] Jutel, Annemarie & Lupton, Deborah. (2015). Digitizing diagnosis: a review of mobile applications in the diagnostic process. Diagnosis. 2. 10.1515/dx-2014-0068.

1.2 AI Chatbot used for healthcare

Chatbot is a kind of smart software tool used for online conversation between computer/mobile phone and human via text or text-to-speech. This kind of dialog system applied in various fields including e-commerce, education, finance, healthcare and so on, which is always equipped with various complicated algorithms such as natural language processing, dialogue management and database management.

AI chatbot has been widely used for the healthcare field in recent years like doctor appointment, medicine reminders, mental healthcare counseling and symptom checker. A study suggested that physicians in the United States believed that chatbots would be most beneficial for scheduling doctor appointments, locating health clinics, or providing medication information[1]. On the one hand, healthcare is seeing big cost savings with the adoption of chatbots. Annual cost savings are estimated to reach $3.6 billion globally by 2022, up from an estimated $2.8 million in 2017[2]. On the other hand, AI in the health field designed for the lay user is helping them take medium and high-stake decisions in their daily life, which is not a small risk for people although it promises reliability and smartness. For instance, some people who are not available or inconvenient to see a doctor, they will choose to use diagnosis applications to check their symptoms. In general, there are several procedures when chatbot works, the first step is the input from the user, and analyze the user's request through the key words of input, and then identify the intent and entities, and the last step is composing a reply. As for chatbots of diagnosis applications, the whole procedure is much more linear and goal-directed. First, users input the symptoms they are experiencing and select the best fit from application. Then, the chatbot will ask them a few more questions about the symptom, and most question-answering type is choice question including single choice question and multiple choice question. After gathering enough details, the chatbot will provide users with information about conditions, and even medication suggestions. However, the implementation of AI chatbots used for healthcare field is still a developing area with high risk because of the limited database and irregular language processing, especially when it is related to some complicated diseases such as cancer and syndrome.

[1] Palanica, Adam; Flaschner, Peter; Thommandram, Anirudh; Li, Michael; Fossat, Yan (January 3, 2019). "Physicians' Perceptions of Chatbots in Health Care: Cross-Sectional Web-Based Survey". Journal of Medical Internet Research. 21(4): e12887. doi:10.2196/12887. PMC 6473203. PMID 30950796.

[2] Swapnil. Dambe. Chatbots for healthcare. Retrieved from https://www.engati.com/chatbots-for-healthcare

1.3 Why is interpretation needed?

With the rapid development of computer technology and it's popularization to the public, there is growing demand among building plain access between professional experts and lay users who are influenced by these technologies. Until now, there are abundant AI-based medical visualization systems for various usages like 3D medical imaging for breast cancer detection, which allows radiologists to capture images from multiple angles and depths, and 3D computed tomography angiography for mapping vascular anomalies, in which doctors can visualize arterial and venous vessels. In other computer science fields, visualization for interpreting machine learning is also experienced and mature. In supervised learning, there is Baobab View system[1] for interactively visualizing and analysis of decision tree, and in neural network models there is directed acyclic graph[2] to help understand convolutional neural networks, which is a node-link visualization and the color, shape and symbol can represent different elements of the network. For image classifiers, experts use gradient maps to help them label the factors that have influence on prediction. Due to the complexity of machine learning models, visualization or explanation of artificial intelligence enable people to reveal the inner work of operations that how a given input turns to a certain output and further develop the performance of sophisticated models.

However, most of these visualizations in the computer science field are developed for experts to better understand machine learning models, rather than for lay users. There is still a big gap between the operation of algorithms and stakeholders' understanding of related professional knowledge, which can result in severe consequences affecting ordinary people even give rise to ethical problems in society. Obviously, effective visualizations provide new insights for computer engineers and make machine learning in the healthcare field meaningful for clinicians, but the patients are also part of the user or stakeholder, even the most affected one, and they should not be excluded from the so-called mystic "black box" which they have the right to know. For instance, a patient can fully trust his doctor but it's hard for him to unconditionally believe an untouchable AI, which is unfamiliar and never explains to him the diagnosis process and result.

So far we cannot find many visualizations in the healthcare field, especially the AI symptom checker targeted at lay users, so there is a keen need to build a bridge between obscure algorithms and normal people. Aside from those professional computer science images, interpreting machine learning models can have many other aspects, while a complicated algorithm is hard to build explanation path, we still can find some effective ways to translate it to understandable visualizations that people who don't have any computer related knowledge can even understand without boundaries, because the lay user should be allowed to know how this kind of algorithm works and takes high-stake decision in their daily life. As a communication designer, the visualization from communication and design perspective can provide helpful insights and methods to translate the complicated machine learning models in understandable ways, which can give lay users the full view of how it works and highlight the comparison of different decision making paths.

[1] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker. Interacting meaningfully with machine learning systems: Three experiments. International Journal of Human- Computer Studies, 67(8):639-662, 2009.

[2] A. W. Harley. An interactive node-link visualization of convolutional neural networks. In Int. Symp. on Visual Computing, pp.867-877. Springer, 2015.

2. Methods and experiments

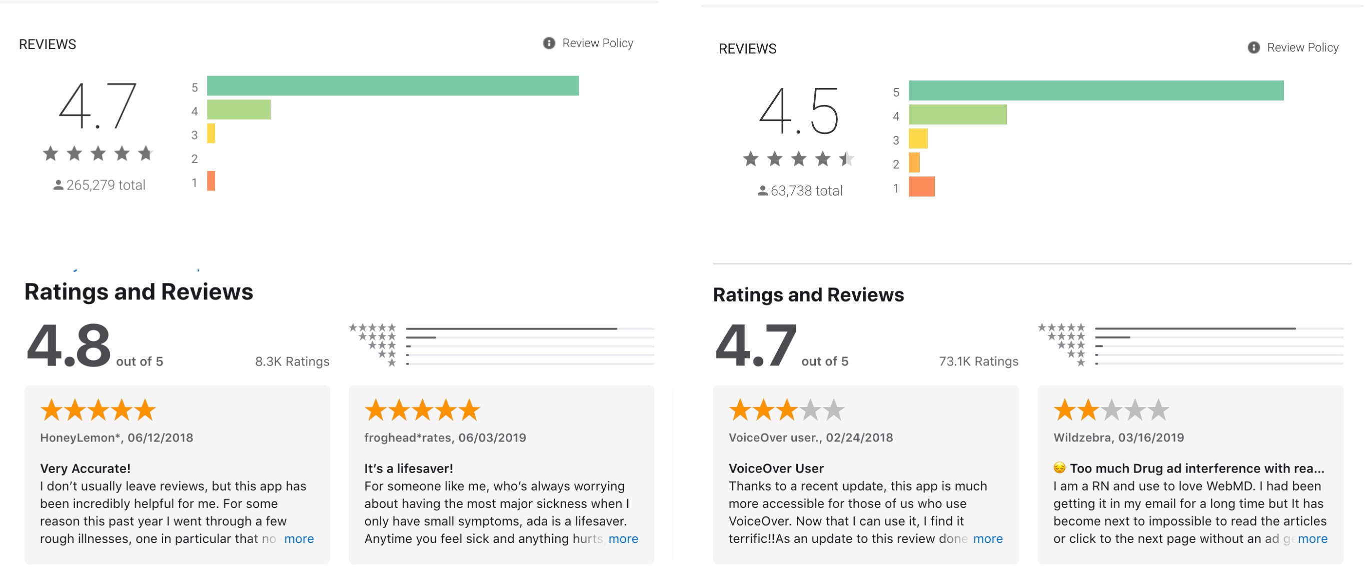
In the whole process of determining the experimental methods and doing experiments, the research is based on the data and the findings after data collection and analysis. All the data is from the user reviews of Google Play Store, and in the process of data collection, these user reviews are screened and sorted out according to the research needs. Subsequently, the experiment of this project is conducted according to the results of data analysis, which are the seven chosen symptoms of pain and top three downloaded applications on Google Play Store. After selecting the symptoms, they are tested in three different applications by controlling variables, and then the whole process and results of the experiments will be contrastively analyzed and visualized.

The main research question is how to collect and classify the user review to define the most mentioned pain point about symptoms of the user? how to map and visualize the decision making path of specific symptoms in Self-Diagnosis Apps? In order to bring experiments in three carefully selected applications into correspondence with the same standard, in the following the methods of choosing and controlling parameters will be presented detailedly.

One thing needs to be emphasized here is that this experiment didn’t contain all the symptoms of each application, and those parameters in the experiment could also be changed according to different profiles or users. The main aim of the experiment is to provide a visualization method to explore how decision making model works and disclose the behavior or surface of the question-answering system of AI Chatbot applied in different applications, which can enable users to have general idea of how chatbot algorithm works in understandable way and make them be aware of what is behind the unfaithful outputs.

2.1 Data source

There are more than 1,000 diagnosis applications in both Google Play Store and App Store if you search the key word of ‘symptom checker’, and the user reviews of some diagnosis apps are more than 50,000 such as Ada health and WebMD(Figure.1). Although most of those applications have a high score over 4.5, there are still a large number of negative comments about the unfaithful or inaccurate diagnosis results. Those negative comments are from various aspects, which could be classified as registration problem, update and subscription problem, limited options of symptom and unfaithful outputs. Among these classifications, the unfaithful output accounts for not a small amount and it includes “inaccurate results”, “extreme results”, “fail to recognize” and so on. Even the application with the highest score like Ada health has more than 5,000 one-star negative user reviews, so we cannot deny that this kind of self-diagnosis applications who promise smarter and more precise still have unfaithful diagnoses which can lead to serious influence on their lay users.



HoneyLemon*, 06/12/2018

Very Accurate!
I don’t usually leave reviews, but this app has been incredibly helpful for me. For some reason this past year I went through a few rough illnesses, one in particular that no

froghead*rates, 06/03/2019

It’s a lifesaver!
For someone like me, who’s always worrying about having the most major sickness when I only have small symptoms, ada is a lifesaver. Anytime you feel sick and anything hurts,

VoiceOver user., 02/24/2018

VoiceOver User
Thanks to a recent update, this app is much more accessible for those of us who use VoiceOver. Now that I can use it, I find it terrific!!As an update to this review done

Wildzebra, 03/16/2019

😔 Too much Drug ad interference with rea...
I am a RN and use to love WebMD. I had been getting it in my email for a long time but It has become next to impossible to read the articles or click to the next page without an ad ge

Figure.1 The score of Ada health in Google Play Store and App Store
The score of WebMD in Google Play Store and App Store

This finding from the user reviews generated three research questions: *what disease or symptom has the highest error rate from the user’s perspective and they complain about most? Why does this kind of application result in inaccurate or unfaithful outputs sometimes? How does the algorithm system work to make decisions?* With the intentions of researching these meaningful questions, two experimental tests have been carried out, which help to provide research evidence and guidance for the following data collection and experiments procedure. The input or the symptom of the tests respectively are “cough” and “blurred vision”, and then these two inputs are tested in three different applications. In the process of test, to simulate a real scenario, the profile background and testing environment are the same for different inputs and apps:

- *Profile Background:* 25-year-old, female; around 1 week duration; without other medical histories

- *Environment:* i phone 8; App version: Ada health - 4.2.0, WebMD - 3.4.0, Healthily - 4.1.3

The three different apps respectively are Ada Health, WebMD and Healthily[1] based on the number of user reviews on Google Play Store. The parameters of the first test(Figure.2) with the input of "cough" are *cough, runny nose and sneezing*, which are always the same in different apps by controlling the answers. For example, in Ada health there is a question asking "Do you have the symptom of runny nose", the answer should be "yes" according to the determined parameter, and if it asks about symptom without mentioning about above-mentioned parameters, the answer should be "no" or "not sure". The same procedure and routine carried out in other two apps. The parameters of the second test(Figure.3) are *blurred vision, thirst and tiredness*, which should be the symptom of diabetes on the basis of disease descriptions page in Ada health website[2]. For the outputs, the diagnoses of the first test with the input of "cough" are almost the same, which are some mild cases like common cold or flu. However, the outputs of the second test are vastly different-two apps are cataract and another one is diabetes. From the various, it could be acknowledged that although the input and parameter are the same, the output from different apps can be totally different and some of them must be inaccurate sometimes, especially for those complicated or chronic diseases.

[1] https://play.google.com/store/apps/details?id=com.ada.app&hl=en
https://play.google.com/store/apps/details?id=com.webmd.android&hl=en
https://play.google.com/store/apps/details?id=md.your&hl=en

[2] https://ada.com/conditions/diabetes/

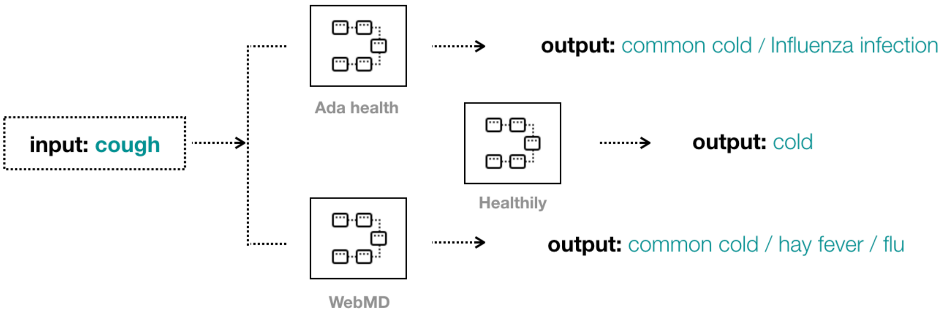

input: cough
Ada health
output: common cold / Influenza infection
Healthily
output: cold
WebMD
output: common cold / hay fever / flu

Figure.2 The first experimental test with the input of "cough"

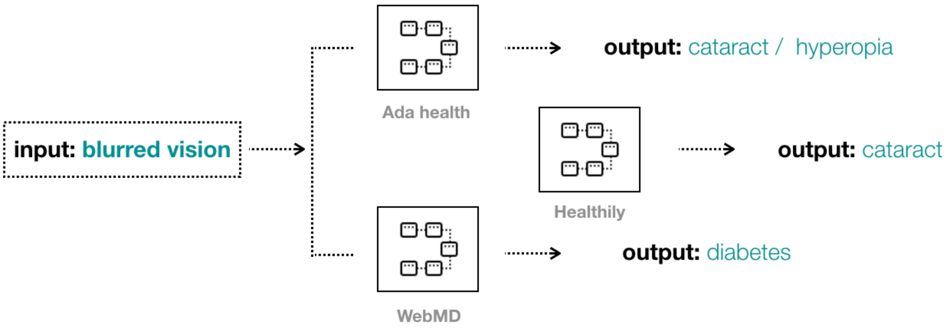
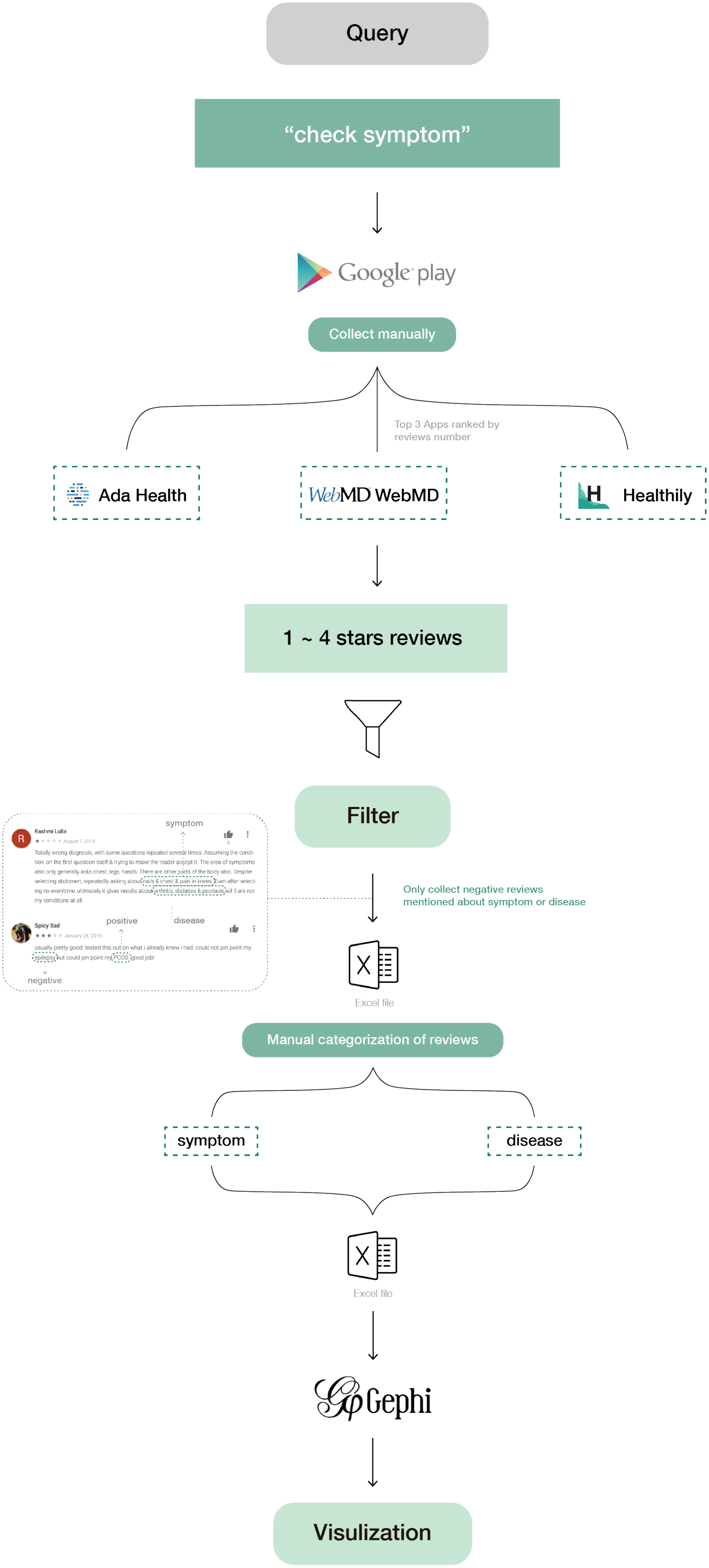


Figure.3 The second experimental test with the input of "blurred vision"

The experimental tests prove the possibility of inconsistencies, mistakes or error diagnoses in diagnostic apps, even those apps who have higher scores in application store, which has potential risk for people's daily life for the reason that the target user of this kind of apps is lay user, especially who don't want or inconvenient to seek advice from professional doctors. Hence if they are utterly ignorant of how it works, they cannot build independent justifications and critical thinking of those unfaithful outputs. After the experimental tests, the clue of finding and collecting data according to the previously mentioned research questions is more clear.

The aim of data collection is to figure out what are the user complaining about and what is the most common pain point of users based on their reviews on application store, and the protocol of data collection(Figure.4) includes four steps:

1. defining the applications
2. collecting user reviews
3. filtering the reviews
4. visualization



Rashmi Lulla
★☆☆☆☆ August 7, 2019
Totally wrong diagnosis, with same questions repeated several times. Assuming the condition, on the first question itself & trying to make the reader accept it. The area of symptoms also only generally asks chest, legs, hands. There are other parts of the body also. Despite selecting abdomen, repeatedly asking about nails & chest & pain in knees. Even after selecting no everytime, ultimately it gives results about arthritis, diabetes & psoriasis, all 3 are not my conditions at all.

Spicy Sad
★★★☆☆ January 28, 2019
usually pretty good. tested this out on what i already knew i had. could not pin point my epilepsy but could pin point my PCOS. good job!

Figure.4 Protocol of data collection

- *Define apps:* In consideration the user of Android System is more extensive than Apple iOS System, so the Google Play Store is chosen as the appropriate platform to collect user reviews. Since there are more than 1,000 diagnosis apps in each platform and it's impossible to browse through all user reviews, the standard of numbers of reviews can help filter the apps. Therefore, the first step of data collection is to input the query of "check symptom" in Google Play Store and choose the top three apps based on the ranking of the user review numbers, which respectively are Ada health, WebMD and Healthily.

- *Collect reviews:* The aim of data collection is to define what is the user complaining about most, which is rooted in their negative comments, so the collected reviews are all from one to four-star reviews, except five-star reviews.

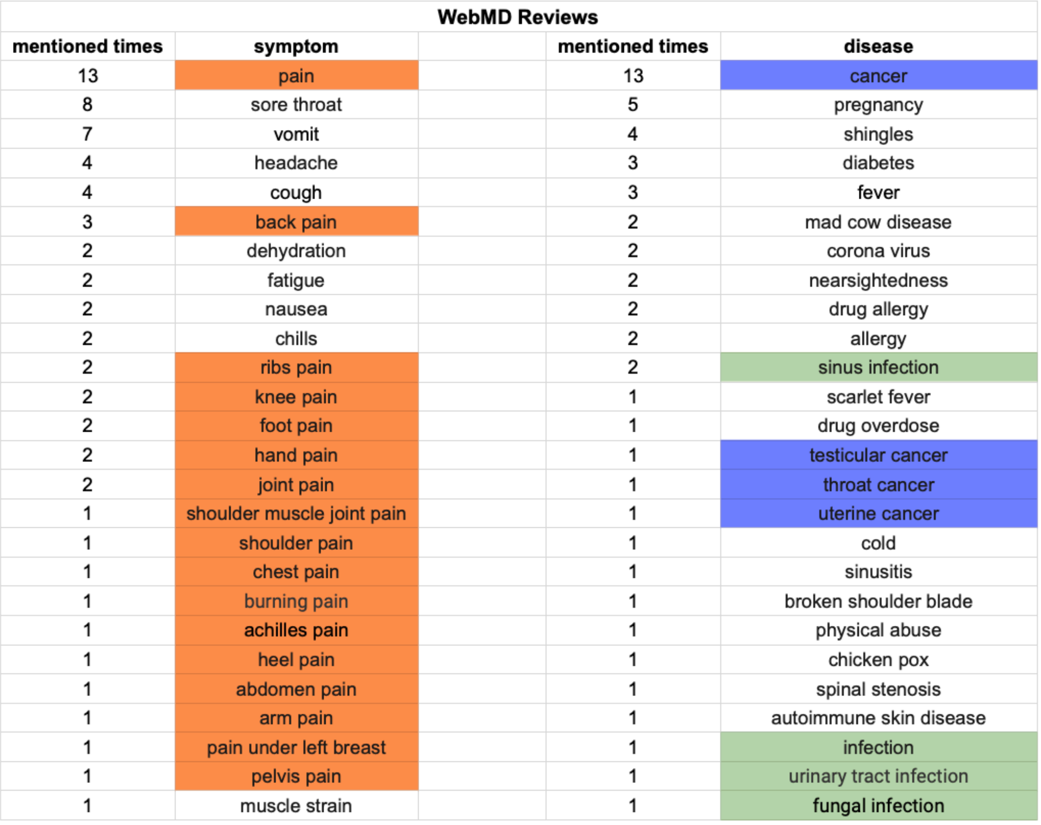
- *Filtrate the review:* There are around 10,000 negative comments from 1-4 star user reviews, and those negative comments can be classified into various types, such as registration problem, update and subscription problem, limited options of symptom and unfaithful outputs. Since the research question is focusing on the user's common pain point from those unfaithful or inaccurate outputs, in this step only those negative comments mentioned about symptoms or disease can be collected into the database. For example, a user review said: "it repeatedly asks me about chest pain and pain in knees, but it gives results about arthritis and diabetes, those diseases are not my conditions at all." In this review, "not my conditions at all" means it's a negative review, and the symptom is "chest pain and pain in knees"; the disease is "arthritis and diabetes". By using this rule, all these negative comments can be classified into disease and symptom, and then ranked by their mentioned times finally, which is the common pain points that users complain about most.

- *Visualization:* The reviews are from three different apps, so in order to effectively compare their commons and differences, it's better to use visualization to make the data visible and clear to compare. The visualization tool used in this case in Gephi, which can present both the mentions by different sizes and the network between the corresponding symptom and disease.

The data collection is a heavy workload, but it is the cornerstone of this research and provides guidance for the following experiments and final visualizations.

2.2 Data Analysis

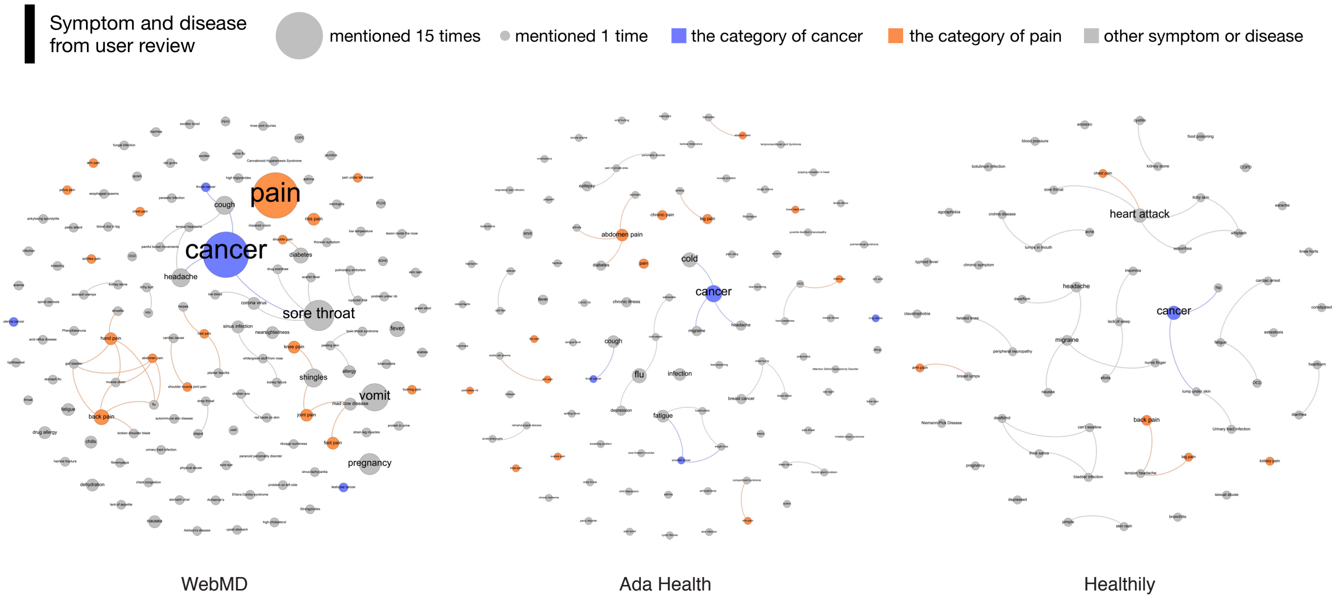
After collecting and classifying those reviews, arranging reviews by their mentioned times and types of disease and symptom can help us understand data and generate useful findings. Since the types of disease and symptom are limited, the reviews are arranged manually and ranked by mentions orderly. For example, Figure.5 shows the processed data from the negative user reviews of WebMD on Google Play Store, and there are more than 60 types of symptoms and diseases, but the symptom of “pain” and the disease of “cancer” both have 13 mentions which are the most mentioned ones so rank the first. As for the classification of disease and symptom, different colors represent different types of diseases or symptoms, and only the same type of disease or symptom mentioned more than four times will be colored because it can be classified into a large category to be compared and analyzed. In this figure, all the symptoms with the same key word of “pain” are colored by orange, and all the diseases with the key word of “cancer” are colored by bluish violet, which can visually show their big proportion in the whole types of diseases and symptoms.



WebMD Reviews				
mentioned times	**symptom**		**mentioned times**	**disease**
13	pain		13	cancer
8	sore throat		5	pregnancy
7	vomit		4	shingles
4	headache		3	diabetes
4	cough		3	fever
3	back pain		2	mad cow disease
2	dehydration		2	corona virus
2	fatigue		2	nearsightedness
2	nausea		2	drug allergy
2	chills		2	allergy
2	ribs pain		2	sinus infection
2	knee pain		1	scarlet fever
2	foot pain		1	drug overdose
2	hand pain		1	testicular cancer
2	joint pain		1	throat cancer
1	shoulder muscle joint pain		1	uterine cancer
1	shoulder pain		1	cold
1	chest pain		1	sinusitis
1	burning pain		1	broken shoulder blade
1	achilles pain		1	physical abuse
1	heel pain		1	chicken pox
1	abdomen pain		1	spinal stenosis
1	arm pain		1	autoimmune skin disease
1	pain under left breast		1	infection
1	pelvis pain		1	urinary tract infection
1	muscle strain		1	fungal infection

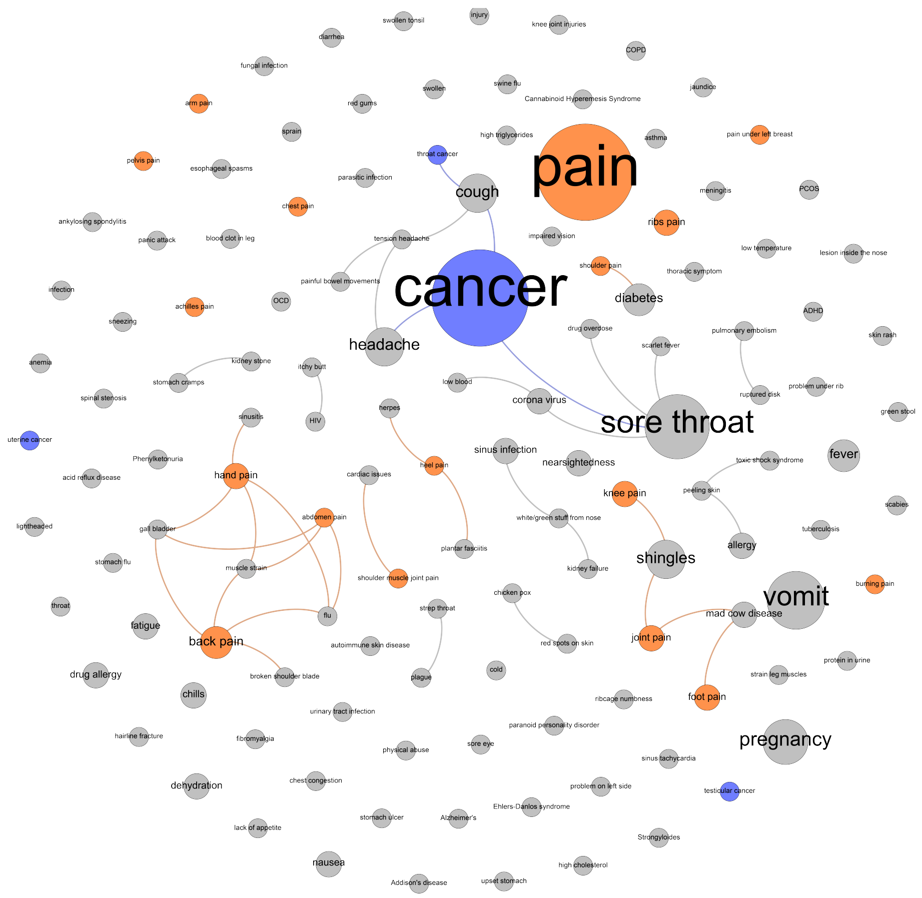
Figure.5 Processed data of the user reviews of WebMD

Compare the three different apps, WebMD has 67 types of symptoms and 64 types of disease in all, in which “pain” is the most mentioned symptom that has 13 mentions and “cancer” is the most mentioned disease that also has 13 mentions; Ada health has 44 types of symptoms and 61 types of disease in all, in which “fatigue”, “abdomen pain” and “cough" are the most mentioned symptoms that all have 3 mentions and “cancer” is the most mentioned disease that has 5 mentions; Healthily has 28 types of symptoms and 37 types of disease in all, in which “headache” and “back pain” are the most mentioned symptoms that both have 2 mentions and “heart attack” is the most mentioned disease that also has 4 mentions. To compare those symptoms and disease in an effective and visible way, a visualization made by Gephi is a suitable tool to display those data which can present both the mentions by different sizes and the network between the corresponding symptom and disease, which is shown as figure.16.

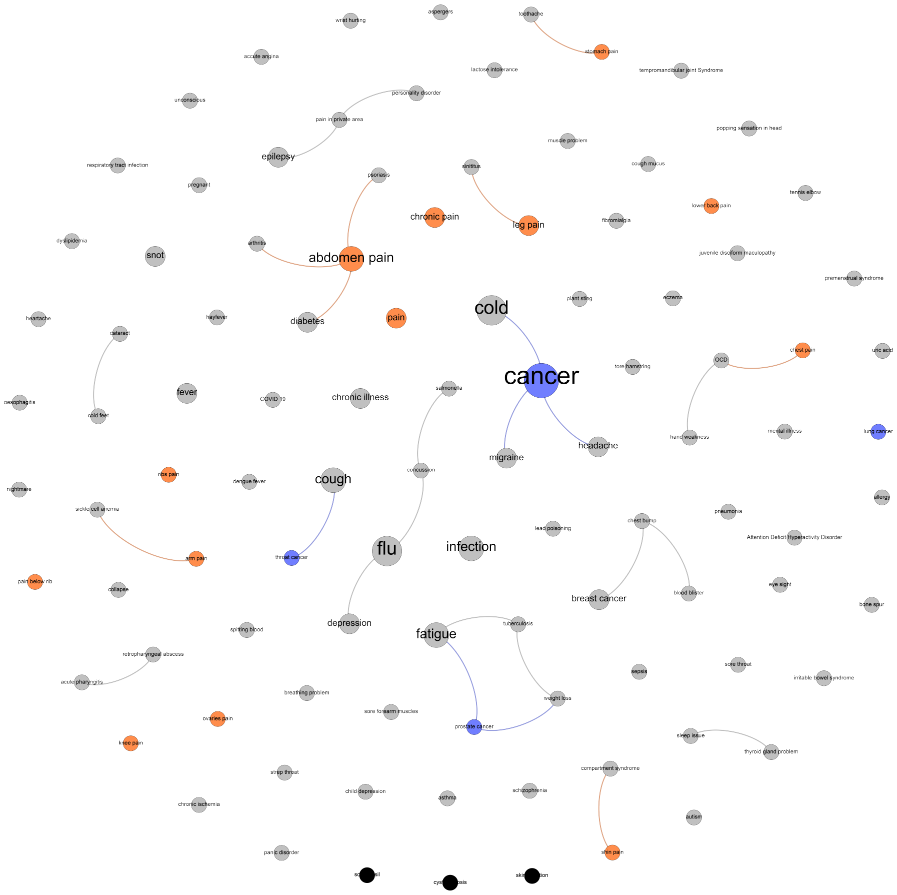


(*the size is in the same proportion)

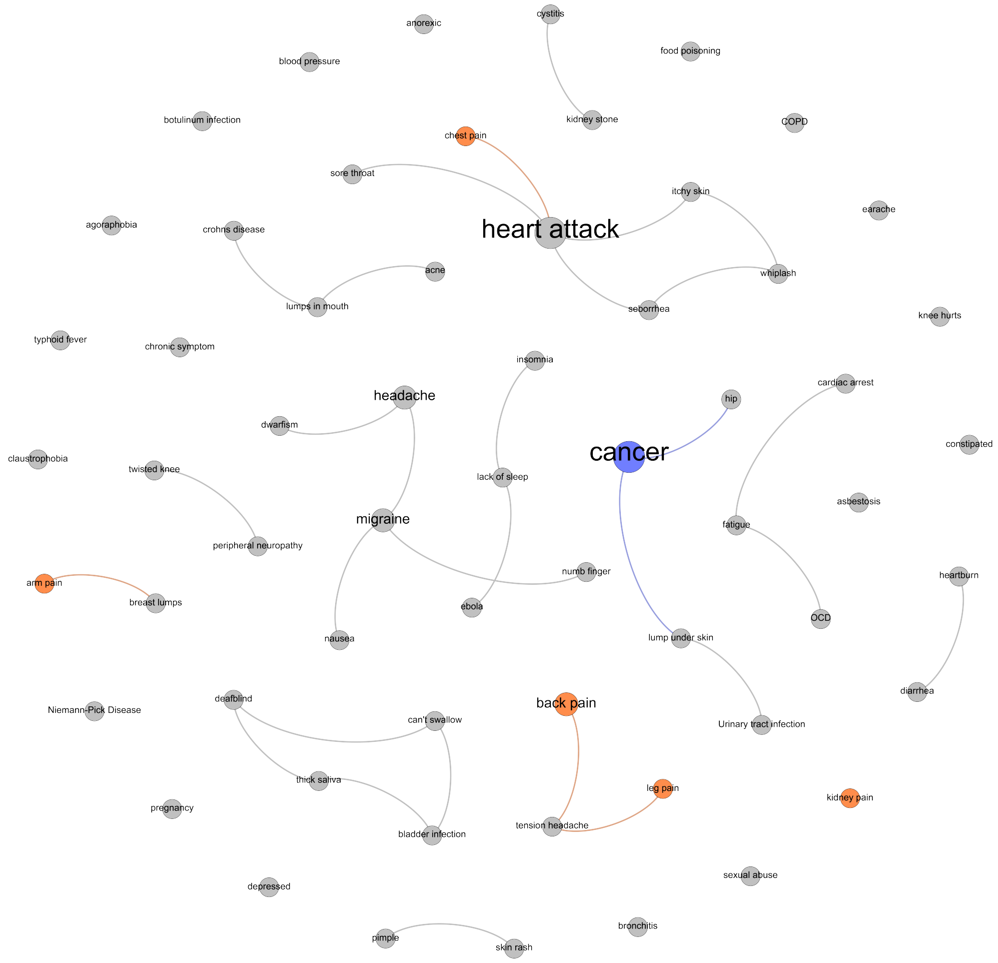
Figure.6 Symptom and disease from user review



WebMD



Ada Health


cystitis
anorexic
food poisoning
blood pressure
botulinum infection
kidney stone
COPD
chest pain
sore throat
itchy skin
earache
agoraphobia
crohns disease
heart attack
acne
whiplash
lumps in mouth
seborrhea
knee hurts
typhoid fever
chronic symptom
insomnia
cardiac arrest
headache
hip
dwarfism
cancer
constipated
claustrophobia
twisted knee
lack of sleep
asbestosis
migraine
fatigue
peripheral neuropathy
numb finger
heartburn
arm pain
breast lumps
ebola
OCD
nausea
lump under skin
diarrhea
deafblind
Niemann-Pick Disease
back pain
can't swallow
Urinary tract infection
thick saliva
leg pain
kidney pain
pregnancy
tension headache
bladder infection
sexual abuse
depressed
bronchitis
pimple
skin rash

Healthily

In the data visualization (Figure.6), all the diseases and symptoms of three different apps are shown in one image in the same size, clearly displaying the comparison of different apps and which disease or symptom accounts for the biggest proportion. Since the visualization is to define the user’s most pain points, it does not distinguish the symptom or disease by visual language, but only highlights the most mentioned key categories by different colors which respectively are the category of “pain” and the category of “cancer”, and other less mentioned diseases and symptoms are all colored by inconspicuous green. From the visualization, some findings can be generated: The disease of cancer has the most mentions in all the apps but not shown in many specific cancers; The category of pain is the most mentioned symptom, and it also has various detailed description such as back pain and abdomen pain; Some common diseases and symptoms also have many mentions, such as sore throat, cough and infection.

After data arranging and visualizing, the most common pain points of the user are already defined which are the category of “pain”, mentioned 35 times in three different apps in total; and the category of “cancer”, mentioned 10 times in three different apps in total. The research question of “what disease or symptom has the highest error rate from the user’s perspective and they complain about most?” is solved, and the next step is carrying out the experiments in different apps based on the findings of data visualization to research why diagnosis apps result in unfaithful outputs sometimes? And how does the algorithm system work to make decisions? In order to conduct the experiments effectively, the experimental object should be selected carefully. From the user reviews, cancer and pain are the most mentioned disease and symptom, but cancer as a disease, which should be the output of the diagnosis path, is futile and difficult to satisfy the research aim. Therefore, the experimental object is targeted at the symptom of pain. Since the “pain” is too general and cannot be input in those diagnosis apps, the more specific pain should be chosen. Ranked by the mentioned times, there are 33 kinds of specific pain in all, and 7 of them are mentioned two or more than 2 times, in which back pain is the most mentioned one that has 3 mentions. As a result, the 7 specific symptoms are selected as experimental objects, which are: back pain, abdomen pain, leg pain, chest pain, knee pain, foot pain and joint pain.

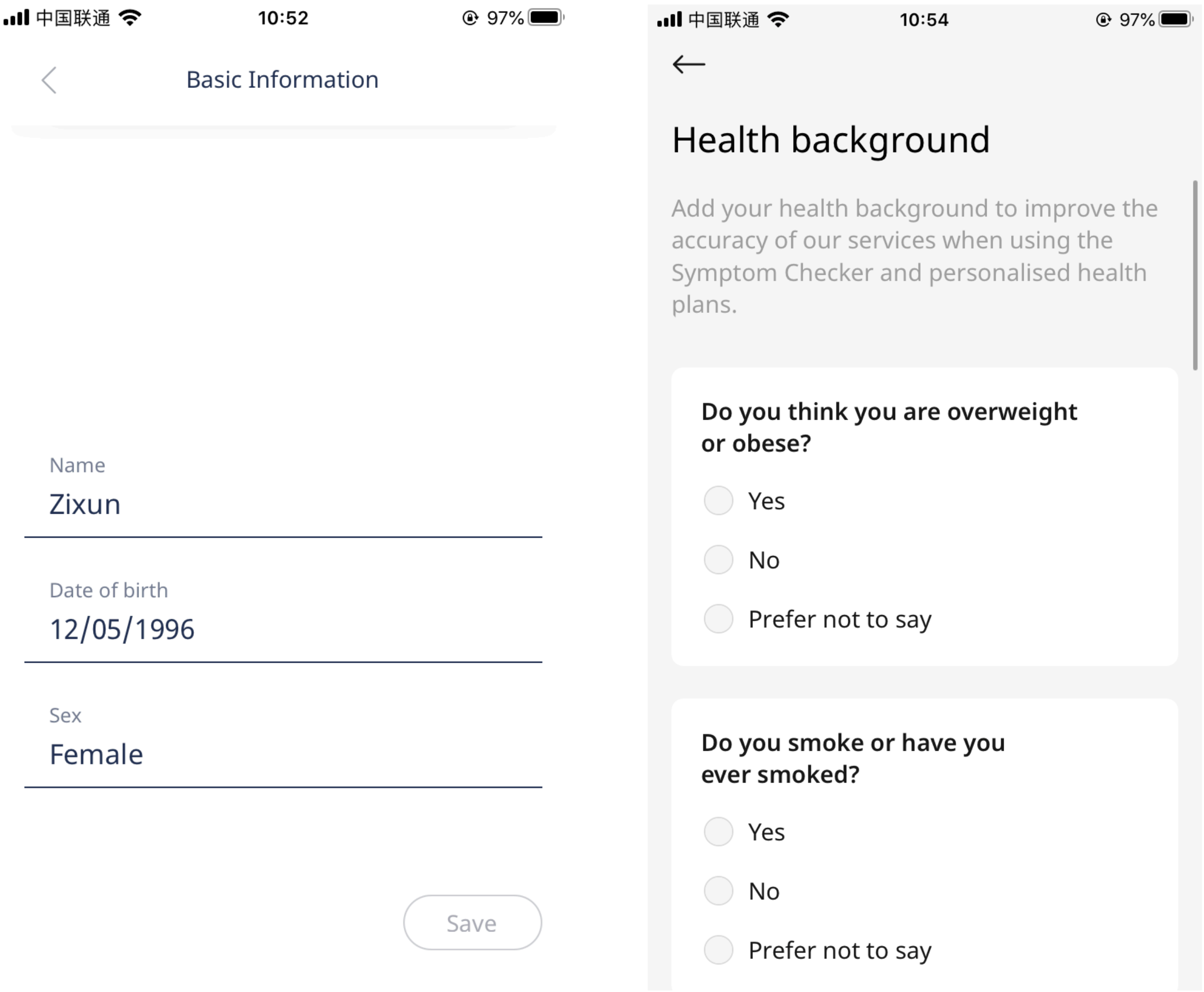
2.3 Experiments

Following the initial research questions and data visualization, the aim of the experiments is to figure out how the algorithms of diagnosis apps work to make decisions and how to visualize its decision path for lay users to understand. The experimental objects have been defined and the flow path of experiments is carried out on the grounds of previous research. The two experimental tests have been done before the data collecting step, which provides the guidance and method for next experiments. The referable part from previous experimental tests are using the same profile background and controlling the parameters to ensure all the elements are the same in different apps, and they are still suitable for the experiments.

The seven specific symptoms and three different apps result in twenty-one total experiments and all of them have the same experimental background and follow the same experimental procedure, which is constituted by following four steps: Define profile and parameter, input, question-answering path, record the path. It should be noted that WebMd does not have the diagnostic process and it only provides the user outputs at once when input the symptom, so the step of question-answering path and the subsequent data statistical analysis will not include this app.

- **Define profile and parameter:** In order to ensure the outputs are not influenced by the experimental process, it is necessary to define the same profile and parameter for all the different symptoms and apps. All diagnosis applications will ask the user's age, gender and health background, which is the profile of users. To make the experimental process concise and comparable, the profile background is the same for all the seven different symptoms and three different apps in the experiments: female, around 30-year-old, no smoker and without other medical history. In different apps, the question asking about profile background is in different scenes, but no matter where it is, these elements of profile should always be the same in different symptoms, like the Figure.7 shows.

  Besides the profile, another parameter needs to be the controlled as the same is the relevant symptoms, which are those symptoms mentioned by various kinds of questions from the diagnostic path, and once the relevant symptoms of one input are defined, they should be always the same in that input tested in different apps. For instance, supposing the input is back pain, the relevant symptoms could be lower back pain and worse with activities. If the questions from different apps asking about "Do you have lower back pain?" or "Is your symptom becoming worse when you are doing activities?", the answer should always be "yes" according to the defined parameters, and the questions mentioned about other symptoms such as "Do you have abdomen pain?" should be answered with "no" or "not sure".



Basic Information

Name

Zixun

Date of birth

12/05/1996

Sex

Female

Health background

Add your health background to improve the accuracy of our services when using the Symptom Checker and personalised health plans.

Do you think you are overweight or obese?

○ Yes

○ No

○ Prefer not to say

Do you smoke or have you ever smoked?

○ Yes

○ No

○ Prefer not to say

Figure.7 Profile and background page

In the experiments, the parameters of seven different inputs are as following:

Abdomen pain: pain in both sides;
worse with activities;

Back pain: lower back pain;
menstrual pain;
worse with activities;

Chest pain: heavy and tight feeling;
worse with activity;
feel exhausted;

Foot pain: worse with activity;
pain in heel;
difficulty walking;

Joint pain: difficult to move joints;
stiff joints;
feel exhausted;

Knee pain: pain in both sides;
pain in outer side of knee;
difficult to bend

Leg pain: tender thigh muscles;
pain in thigh;
pain in back thigh

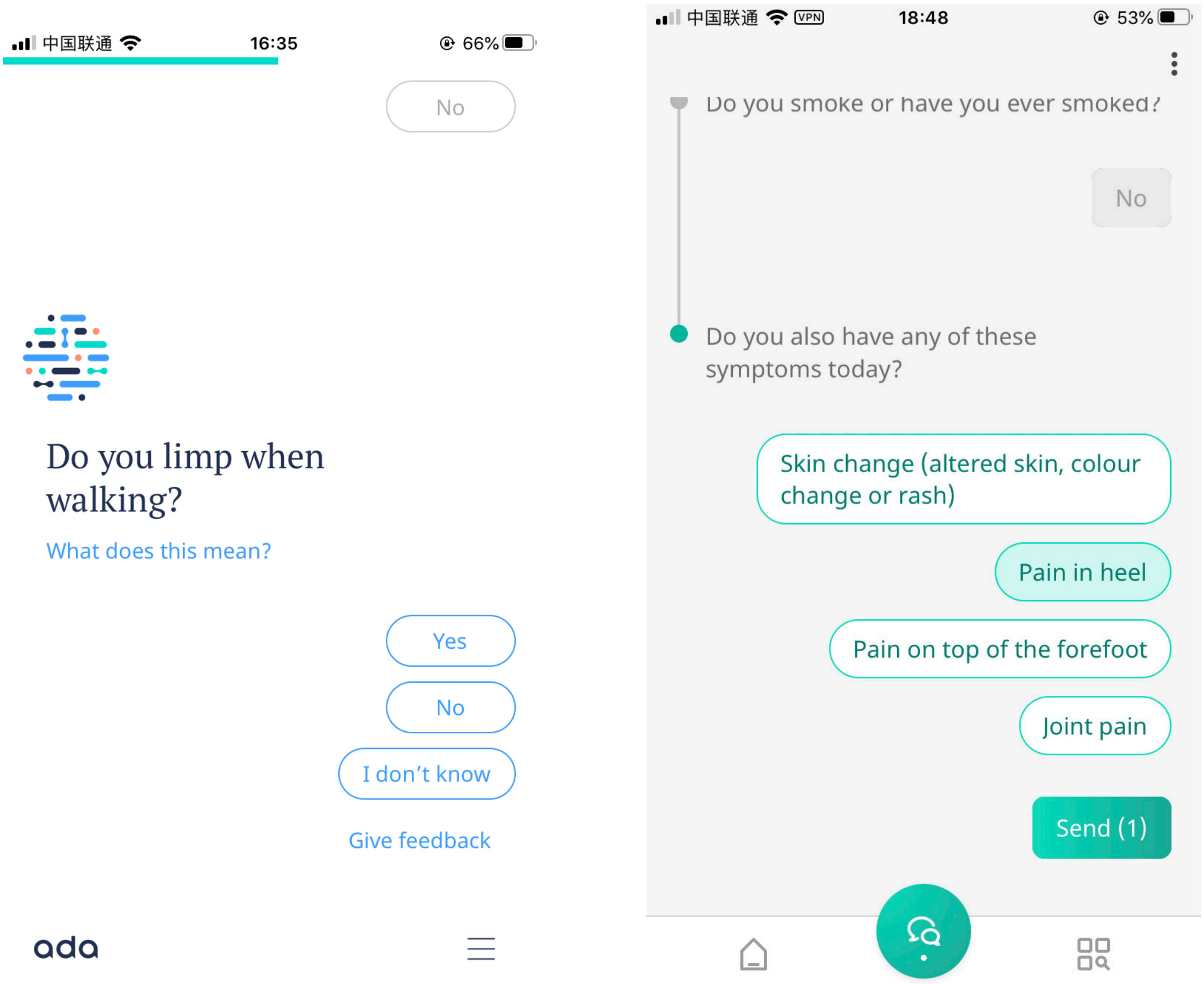
In order to avoid the complexity and influence on experimental results, the parameters should not be excessive or uncommon, such as some rare symptoms like phenylketonuria, which even does not show in some of these diagnosis apps. In the experiments, the parameters for each input are not more than 3 elements, and all of them are common symptoms such as feeling exhausted and pain being worse with activities. However, one important thing needs to be clarified here is that all the profile background and parameters in the experiment are not unmodifiable, and they could be changed according to different profiles or users based on different needs. Also, this kind of method can be used for other similar projects or apps to carry out experiments.

- **Input:** As the data visualization shows, there are 32 kinds of pains and 7 kinds of cancer in total, which respectively are the most mentioned symptom and disease from the user reviews. Since cancer as a disease should be the output of experiments, it can not be used as input even if there are some specific types of cancer. “Pain” is also too general for the input, which needs to be more specific, so the defined inputs are: back pain, abdomen pain, leg pain, chest pain, knee pain, foot pain and joint pain, that are all the specific symptoms of “pain” mentioned more or equal to two times in user reviews.

  These inputs are tested in three apps successively, and the user profile and parameters for each input should be the same in different apps. Every input has three different diagnostic paths based on three different apps which are Ada health, Healthily and WenMD, and all the paths will be compared and analyzed together finally.

- **Question-answering path:** Question-answering path is the most important and complicated part of the diagnostic process for the reason that it is the procedure to ensure all the profile background and parameters are the same by controlling the answer, and it is also the factor that has the influence on the outputs. When start one experiment in a time, the input is same for three apps, and in order to make the outputs are not influenced by the different questions in the diagnostic process of different apps, it’s necessary to comply the standard, which is the profile background and parameters, by replying the same answers in different apps.

  In general, the question number is different according to different inputs and apps, and the types of questions are also various(Figure,8) such as there are yes-no questions and multiple choices questions in the same app, but the key is focusing on the specific symptoms or elements mentioned in those questions. For example, when testing the input of “foot pain”, there is a multiple choice question in Healthily asking “Do you have any of these symptoms?” and the answers include “skin change; pain in heel; pain on the top of the forefoot; joints pain”. On the basis of defining profile background and parameter step, “pain in heel” is one of the define parameters of “foot pain”, so the chosen answer should only be “pain in the heel”. Nonetheless, there are no questions mentioned about “pain in heel” in Ada health, so the proper way is choosing the answers of “no” or “not sure” when encountering the questions of other irrelevant symptoms to defined parameters.



Do you limp when walking?

What does this mean?

Do you smoke or have you ever smoked?

Do you also have any of these symptoms today?

Figure.8 Screenshots of questions in Ada health and Healthily

On the whole, most questions in Healthily are multiple choice questions; in Ada health most questions are yes-no questions, and WebMd doesn’t have any question-answering path. Since the parameters are same for each input, so output can only be influenced by the decision making algorithms of different apps, which also conforms to the aim of the experiments that how the algorithms of diagnosis apps work to make decisions and why their outputs are unfaithful or various, and the answer is from those different question-answering paths.

- **Record the path:** The question-answering path is already completed during the experimental process, and in order to better analyze the paths for figuring out the research questions in previous, it’s necessary to use an appropriate and effective method to record those paths and make them easy to compare and statistically analyze. The whole experimental process is a linear flow, and it contains inputs, questions, answers and outputs. Figure.9 completely restores the whole process of one input experimented in one app, and in this process there are 30 questions and 2 outputs in all.

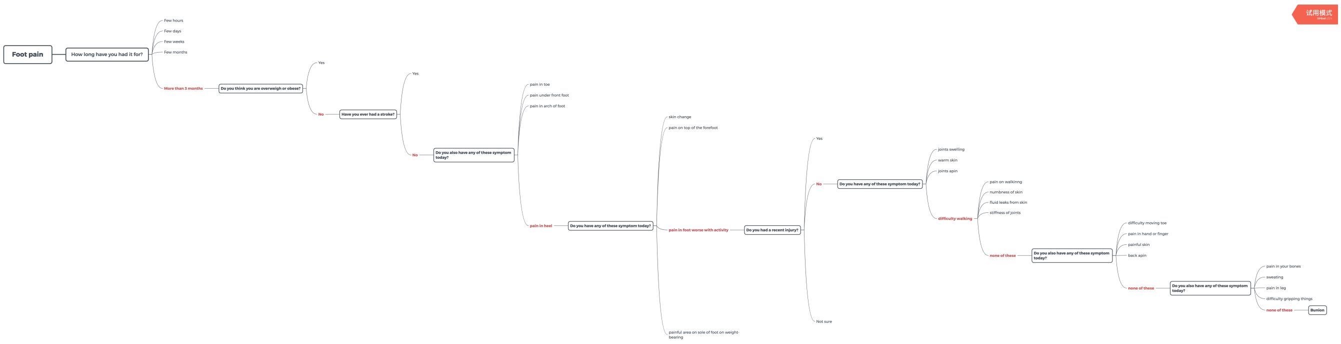
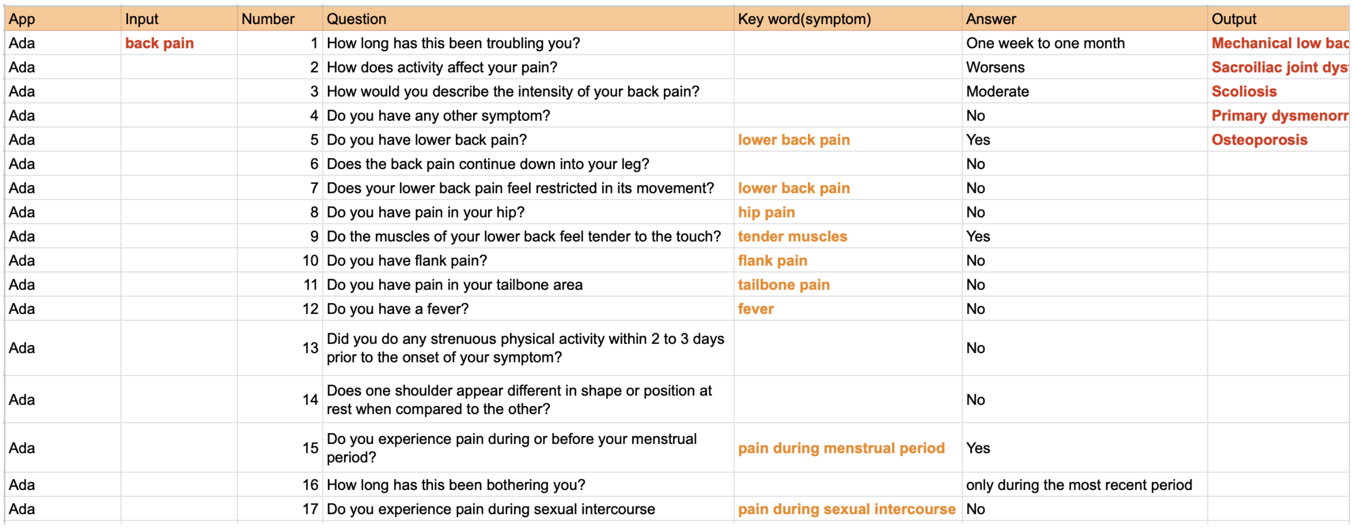

Foot pain
How long have you had it for?
Few hours
Few days
Few weeks
Few months
More than 3 months
Do you think you are overweigh or obese?
Yes
No
Have you ever had a stroke?
Yes
No
Do you also have any of these symptom today?
pain in toe
pain under front foot
pain in arch of foot
pain in heel
Do you have any of these symptom today?
skin change
pain on top of the forefoot
pain in foot worse with activity
painful area on sole of foot on weight-bearing
Do you had a recent injury?
Yes
No
Not sure
Do you have any of these symptom today?
joints swelling
warm skin
joints apin
difficulty walking
pain on walkinng
numbness of skin
fluid leaks from skin
stiffness of joints
none of these
Do you also have any of these symptom today?
difficulty moving toe
pain in hand or finger
painful skin
back apin
none of these
Do you also have any of these symptom today?
pain in your bones
sweating
pain in leg
difficulty gripping things
none of these
Bunion

Figure.9 The process of foot pain experimented in Healthily

However, only the linear flow image is difficult to display the numbers and the types of question visually, so a form is used to record and manage the diagnostic paths. As Figure.10 shows, there are seven main columns in the form: App name, input, question, key symptom, answer and output. Among them, the key symptom is those specific symptoms mentioned by the questions, which is an important factor to influence the outputs of different apps because different apps have totally different types of questions and mentioned symptoms on account of the difference of algorithm and data base. For example, when testing the input of “back pain”, there is a question on Ada health asking “Do you have pain in your tailbone area?” but does not show in Healthily, so “tailbone area” is a key symptom for Ada health. Also, some symptoms could be the same of different apps, and they are the commonalities of those apps. Comparing the question numbers and key symptoms of different apps can help us analyze how the algorithms work and why they result in totally different results.

The column of answer is recording the answer replying to each question, and it is the step to control all the chosen symptoms are the same for different apps with one same input, by defining the profile background and parameters as mentioned before. The column of output shows that although the input and parameters are the same, the results of different apps are almost completely different. With the seven inputs, there are 79 outputs in all after experimenting in three different apps, and only 4 of them are the same.



App	Input	Number	Question	Key word(symptom)	Answer	Output
Ada	**back pain**	1	How long has this been troubling you?		One week to one month	**Mechanical low bac**
Ada		2	How does activity affect your pain?		Worsens	**Sacroiliac joint dys**
Ada		3	How would you describe the intensity of your back pain?		Moderate	**Scoliosis**
Ada		4	Do you have any other symptom?		No	**Primary dysmenorr**
Ada		5	Do you have lower back pain?	**lower back pain**	Yes	**Osteoporosis**
Ada		6	Does the back pain continue down into your leg?		No	
Ada		7	Does your lower back pain feel restricted in its movement?	**lower back pain**	No	
Ada		8	Do you have pain in your hip?	**hip pain**	No	
Ada		9	Do the muscles of your lower back feel tender to the touch?	**tender muscles**	Yes	
Ada		10	Do you have flank pain?	**flank pain**	No	
Ada		11	Do you have pain in your tailbone area	**tailbone pain**	No	
Ada		12	Do you have a fever?	**fever**	No	
Ada		13	Did you do any strenuous physical activity within 2 to 3 days prior to the onset of your symptom?		No	
Ada		14	Does one shoulder appear different in shape or position at rest when compared to the other?		No	
Ada		15	Do you experience pain during or before your menstrual period?	**pain during menstrual period**	Yes	
Ada		16	How long has this been bothering you?		only during the most recent period	
Ada		17	Do you experience pain during sexual intercourse	**pain during sexual intercourse**	No	

Figure.10 The form recording the process of back pain experimented in Ada health

2.4 Data Design Process

To better compare and analyze the diagnostic paths of different apps, it’s necessary to go deeper for the data based on the record form because the different key symptoms and outputs are not enough for representing the whole decision making process. Since the target of this project is lay users who don’t have knowledge about complicated machine learning models or algorithms, the analysis should focus more on the behavior and surface or the diagnostic process rather than obscure inner programs, which also can demonstrate the algorithms of different apps in an effective way. As mentioned above, the most important procedure of the diagnostic process is the question-answering path that has a big influence on the outputs. The question-answering path contains questions from the applications and answers from the user, and the different types and key symptoms of questions from different apps result in various outputs, so further classifying and contrastively analyzing the questions can provide a clue to why there are different results with the same input.

Throughout all the questions from three apps, the types of question can be classified into three categories:

- yes-no question
- single choice question
- multiple choice question

In Ada health, the type of yes-no question accounts for the largest proportion, and in Healthily the type of multiple choice question has the largest number, which doesn’t show any in Ada health. Besides the types of questions, those questions that are the same from different inputs and apps should also be emphasized. For instance, the question “How long has this been troubling you” is always asked in Ada health for all the seven inputs, and when experimented the “back pain”, the question “Do you have lower back pain” shows in both Ada health and Healthily, although the the way of asking is not identical, they both mention about “lower pain”. The proportion that those same questions account for in different apps can demonstrate how many differences of the apps’ algorithms and database.

On the whole, the types and numbers of questions are different in different apps with the same input. For “abdomen pain”, the total question numbers in Ada health is 31 and in Healthily is 24, and among them there are 9 same questions for both the two apps; for “back pain”, the total question numbers in Ada health is 30 and in Healthily is 21, and among them there are 13 same questions for both the two apps; for “chest pain”, the total question numbers in Ada health is 32 and in Healthily is 23, and among them there are 11 same questions for both the two apps; for “foot pain”, the total question numbers in Ada health is 25 and in Healthily is 11, and among them there are 9 same questions for both the two apps; for “joint pain”, the total question numbers in Ada health is 26 and in Healthily is 25, and among them there are 13 same questions for both the two apps; for “knee pain”, the total question numbers in Ada health is 13 and in Healthily is 7, and among them there are 4 same questions for both the two apps; for “leg pain”, the total question numbers in Ada health is 18 and in Healthily is 12, and among them there are 6 same questions for both the two apps.

Since the inputs are all about “pain” which is the pivotal symptom to determine a disease, the questions mentioned about the key word of pain should also be emphasized, such as questions about “lower back pain” from the input of “back pain”. Looking over the questions of different inputs and apps, the times they mention about pain are different. For example, the input of “back pain” has most questions mentioned about pain in both Ada health and Healthily, which respectively are 14 times and 9 times. The questions mentioned about pain with the same input in different apps can also demonstrate how the different algorithms work to make decisions. On the whole, for “abdomen pain”, the question mentioned about pain in Ada health occurs 5 times, and in Healthily occurs 7 times; for “back pain”, the question mentioned about pain in Ada health occurs 14 times, and in Healthily occurs 8 times; for “chest pain”, the question mentioned about pain in Ada health occurs 5 times, and in Healthily occurs 6 times; for “foot pain”, the question mentioned about pain in Ada health occurs 6 times, and in Healthily occurs 5 times; for “joint pain”, the question mentioned about pain in Ada health occurs 4 times, and in Healthily occurs 6 times; for “knee pain”, the question mentioned about pain occurs 5 times in both Ada health and Healthily; for “leg pain”, the question mentioned about pain in Ada health occurs 6 times, and in Healthily occurs 5 times; Except the input of “back pain”, the times that questions mentioned pain of other inputs are very similar which is around five to seven in different apps. To better compare those different types of questions and the same questions of different apps, Figure.11 visually shows the comparison of three different apps.

Some findings can be generated from the data statistical analysis, the first finding obviously is that the same inputs and parameters can result in totally different outputs because of the disparate decision making paths. Second, some questions are the same of different apps, but most questions including the types and key symptoms mentioned are different from different apps and inputs, like Ada health has more yes-no questions and most questions in Healthily are multiple choice questions. Another finding is about the output, although the WebMD doesn’t show any decision making path, it always gives more outputs than other two apps and some outputs are even the same with others. The step of data analysis figures out the research questions of how the algorithms of diagnosis apps work to make decisions and why their outputs are various sometimes with the same input and parameters, and it also provides the basics and methods for the following visualization.

comparison of the question of 3 apps based on the decision path

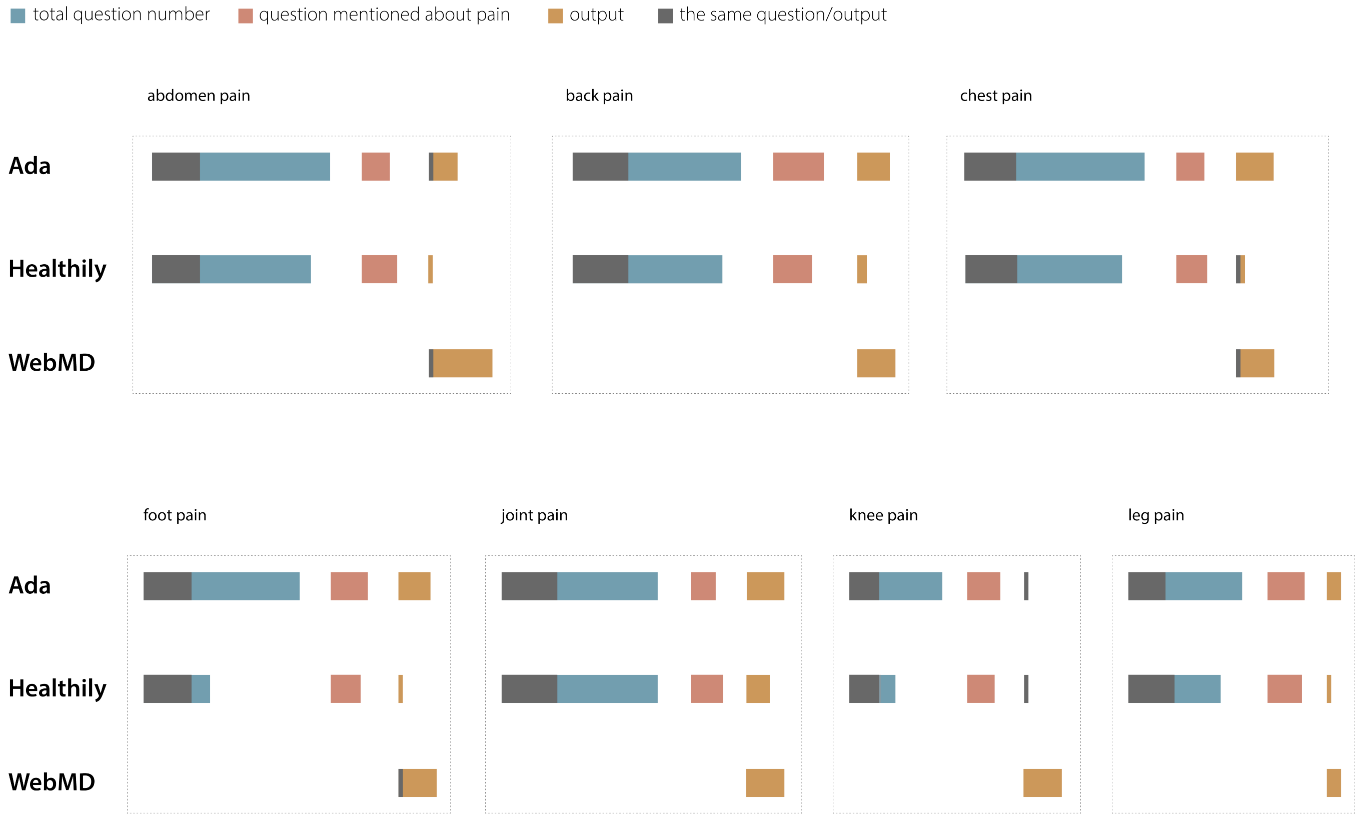


Figure.11 The comparison of questions of 3 different apps

2.5 Visualization Methods

The data statistical analysis can explain how different the algorithms from different apps, but it is not an effective way to demonstrate for the people. Since the target user of this project is lay users and apparently those intricate professional computer science visualizations are difficult for them to understand, so the way of visualizing the diagnostic path is supposed to be understandable and lucid for common people because they should not be excluded from knowing how this kind of algorithm works and takes high-stake decision in their daily life. Therefore, as the same with the data analysis step, the visualization will focus more on the behavior and surface of those algorithms which are the question-answering path. Through the visualization, the aim is to provide a method to explore how decision making models work and disclose the inner work of this kind of algorithm by visualizing the decision making path, and the visualization method can also be used for other similar apps and AI chatbot systems except the diagnosis apps.

From the data statistical analysis, we know that the different decision making paths can result in diverse outputs even with the same input, so the visualization intends to enables lay users to have general idea of how chatbot algorithm works in understandable way and make them be aware of what is behind the unfaithful or diverse outputs. As a result, the user can develop justification and critical thinking of this kind of algorithm/app by themselves. In order to make the visualization not difficult to comprehend for lay users, the perspective from a communication designer is essential and effective to be used for translating the decision making process. Also, all the paths and outputs are orderly shown in one integrated visualization by the same way, that can give lay users the full view of how it works and highlight the comparison of different decision making paths.

In the section of experiment, the method of recording data is already introduced and the process of the whole diagnostic path is also restored by mind mapping software, which completely retains the question-answering path and is the substantial content of visualizations. The diagnostic path is a linear flow and with a lot of branches that are the different answers, so the alluvial diagram and dendrogram can be used as references for the basic structure of visualization. To maintain the readability for lay users, the first step of visualizing is simplifying the restored image and extracting the essential elements by visual language. Looking through the restored image of experiment in Ada health(Figure.13), it stars from an input which is “back pain” in this case, and then made up by a series of questions from the app and answers from the user that account for the the largest proportion of the flow, and finally the output is generated from the question-answering path. Hence, the questions, the chosen answers by users and the connection between them should be highlighted and linked clearly. As the simplified version of visualization shows(Figure.13), the black dot represents the questions and those green lines mean the answers, in which the line connecting one dot to another is the answer chosen by the user. This image can visually show how long the diagnostic path is and how the questions and answers generate the outputs step by step.

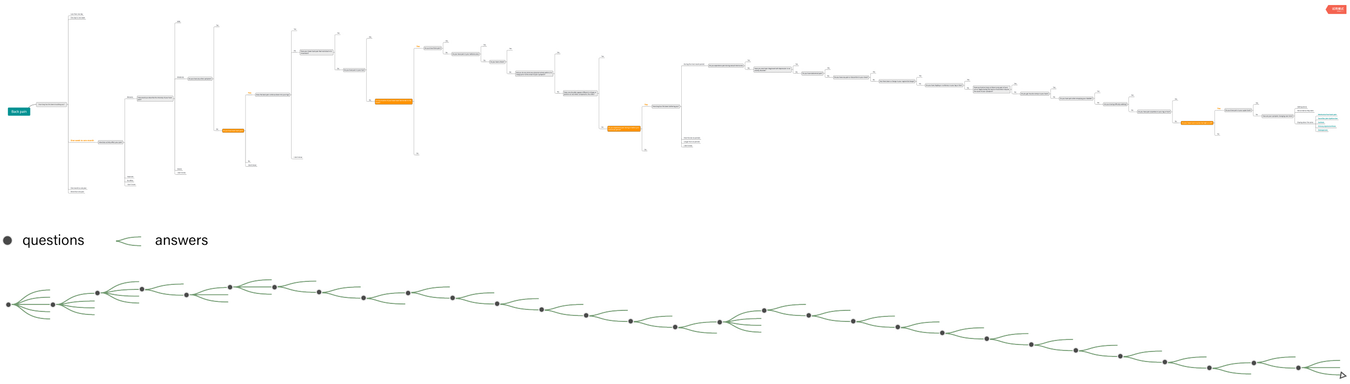


Figure.12 The restored image and simplified visualization of “back pain” in Ada health
(The first image only shows the flow of diagnostic path, so it doesn't need to see the texts clearly)

As I mentioned in the section of data design process, the types of questions and key symptoms mentioned by questions of different algorithms are also the essential elements to result in the different outputs, so highlighting those various kinds of questions is important to help lay users understand what are the differences between the different algorithms of three apps. Besides the questions, answers are the step to control the profile and parameters, thus displaying what answers are chosen can show how one question is connected to another. For most questions the user only needs to choose “yes”, “no”, or “none of them” since the defined parameters are limited and the multiple choice questions only occur in one app that is Healthily. In general, combined with the data statistical analysis, in visualization the questions can be classified into: single choice question, yes-no question, multiple choice question and question mentioned about pain, while the answers can be classified into answer with yes and answer with no/none of them. As for the outputs, although most outputs are totally different, a few same outputs should also be emphasized because it will give a clue of how various those outputs are. In order to compare the diagnostic paths of three apps effectively, the visualizations of different flows could be put together to enhance the comparison and visual impact, and the curving flow is regulated to a straight path with different dots. Since the categories of questions and answers are numerous, it’s a smart and efficient way to distinguish by different shapes, sizes and colors. For example, the different types of questions are highlighted by the same circle but in different colors and the questions mentioned about pain are highlighted by triangles.

All these regulations and elements generate the visualization of diagnostic decision making paths together(Figure.13), and this method can also be used for other similar question-answering systems and apps, such as the AI customer service chatbot in the e-commerce field.

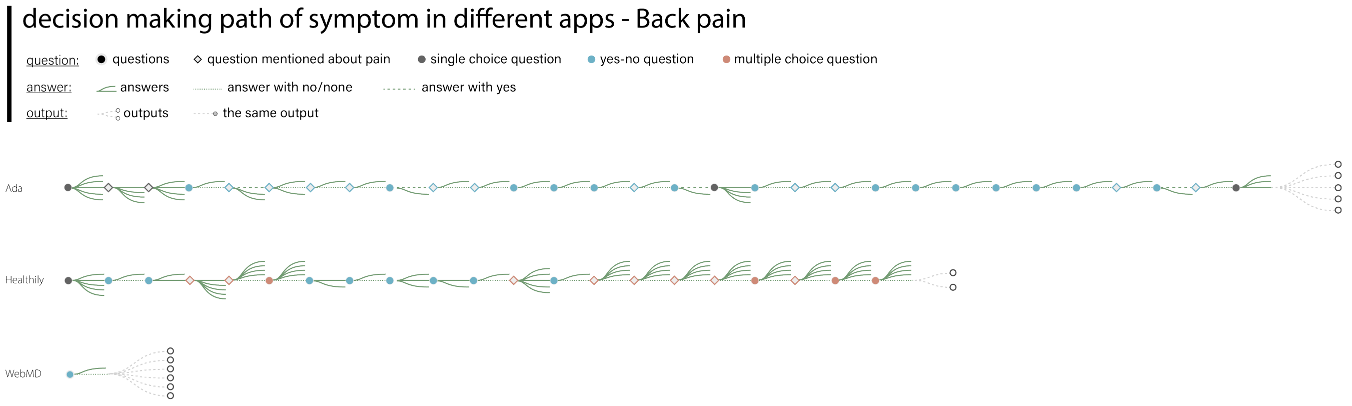


Figure.13 The decision making path of “back pain” in different apps

3. Results

Seven inputs about pain have been experimented in three different apps, for a total of 21 experiments all together. All the experiments are following the same regulations and completed the full path, in which the only variables are different inputs and diagnosis apps. The rigorous experimental methods and process lead to a comprehensive comparison and analysis that is beneficial for preliminary visualization and the visualization also provides a helpful perspective to understand better those experiments and results correspondingly. These two procedures combined together to enable lay users to have an overview of how diagnostic apps take high-stake decisions in their daily life and why the diagnoses are inaccurate and diverse sometimes.

3.1 Overview of experimental results

To evaluate the 21 experiments comprehensively, there are several aspects that need to be considered: input, question-answering path, the characteristic of questions and answers ,outputs. The inputs are determined which are the seven types of pain. The question-answering path mainly refers to the length of diagnostic path, for example, for the same input, Ada health always has more questions and longer decision making process while WebMD only gives one question every time. Through the analysis in the section of data design process and visualization method, the questions of three apps can be be classified into four categories: yes-no question, single choice question, multiple choice question and question mentioned about pain; while for the answers only the answer with “yes” and the answer with “no/none” are emphasized. Throughout all the experiments, the input of “abdomen pain” has the longest question-answering path which is 31 questions for Ada health and 25 questions for Healthily in total, and Ada health always has more questions than Healthily and WebMD in all the inputs. As for the characteristic of questions, most questions in Ada health are yes-no questions, while Healthily has most multiple questions and WebMD doesn’t have any diagnostic path. Comparing all the questions of two different apps, only less than half of them asked the same questions or mentioned the same symptoms, and most questions including the types and the mentioned key symptoms are totally different. Also, for different inputs there are different numbers of questions mentioned about pain, but always they are not more than half for both the two apps.

For the outputs, 79 different outputs are generated by the the seven inputs experimented in three apps, and only 4 of them are the same, which are “Gallstones” in the input of “abdomen pain”, “Costochondritis” in the input of “chest pain”, “Plantar fasciitis” in the input of “foot pain”, and “Iliotibial band syndrome” in the input of “knee pain”. Although WebMD doesn’t show any decision making path, it always has more outputs and provides 3 same outputs with other apps. This result shows how different these algorithms of diagnostic apps are even the same inputs and parameters can lead to totally various outputs. Figure.14 demonstrates the general comparison of different diagnostic paths and outputs from the seven inputs. Through the experiments, some significant findings can be generated: different diagnostic apps produce diverse outputs because of different decision making paths, which is influenced by the key symptoms or elements mentioned by questions and the way of asking. Therefore, it’s unadvisable for lay users to fully trust one app and algorithm, and if the user wants to gain more accurate diagnoses without seeking help from a doctor it’s better to check his symptoms in more than one diagnosis apps; another finding is that the diagnoses are not only determined by the diagnostic path, for example, WebMD has no question-answering path but it also gives diagnoses and some of them even are the same with other apps. Therefore, the diagnostic path can only be a reference for the outputs and it doesn’t mean the longer diagnostic path will generate the more accurate diagnoses.

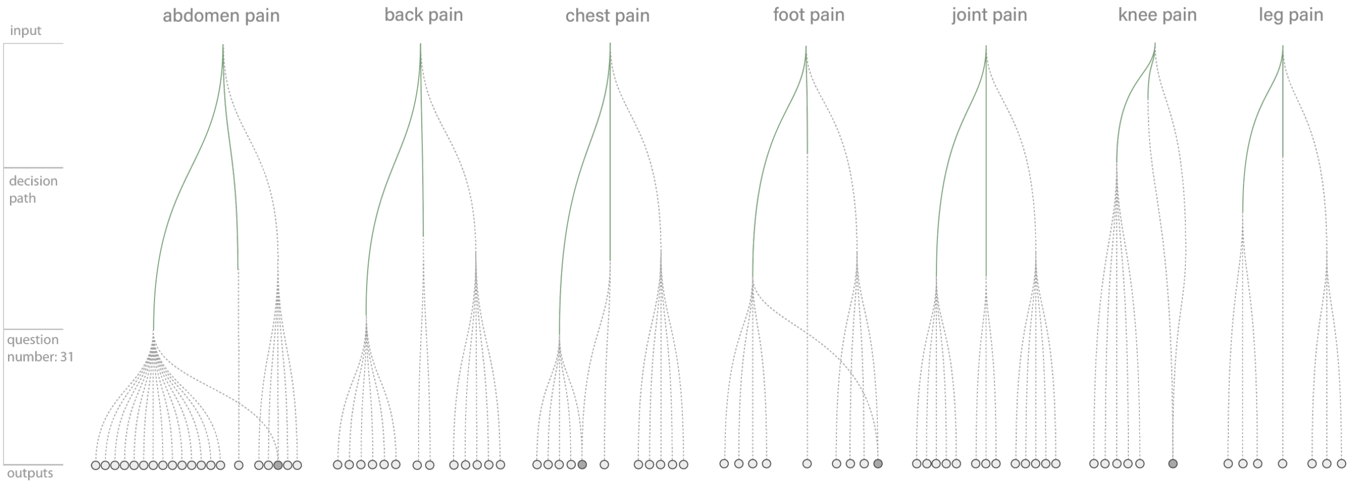


Figure.14 The comparison of the decision making paths and outputs of seven inputs

The statistical analysis of experiments figures out the research questions of how the different inner decision making processes of diagnostic apps work and why they can result in diverse outputs sometimes, while the visualization displays the reasons and factors in an understandable way by showing the algorithms' surface and behavior for lay users. The experiments and preliminary visualization enables lay users to be aware of what behind those unfaithful outputs is different AI chatbot models and databases, thus they can develop their own justification and critical thinking of this kind of app and algorithm.

3.2 Visual analytics in interpretation

With the rapidly growing of Artificial Intelligence and machine learning technologies, there are abundant visualizations for complicated algorithms and the user group is gradually expanding to normal users, such as Explainable Artificial Intelligence (XAI), which aims to make the algorithm and machine learning more understandable to users rather than computer experts. In recent years, good progress has been made in XAI and these relevant research can be classified by the types of techniques being illuminated like black-box techniques and white-box techniques[1]. However, these XAI works are always based on an algorithm-centric view and somehow ignore the real needs of lay users, relying on “researcher' intuition of what constitutes a 'good' explanation”[2]. For example, one of the most popular methods to explain a prediction machine learning model[3] is by listing the features of algorithm data with the highest weights contributing to a model's outputs, which is not clear that whether such an explanation satisfies the expert's needs or builds an understandable path for lay users.

To make the lay users who don't have a deep technical understanding of machine learning not neglected by researchers from the Artificial Intelligence field, it's necessary to consider the needs from the users and visualize those complicated machine learning models from a designer perspective who has the ability of communicating with them. For lay users, they do not need to understand deeply the inner work of algorithms such as how to transform the natural language input with potential ambiguity into a kind of unambiguous computer internal language, and they also do not need to know what are the features of different databases. The thing that they are interested in and feel relevant to their daily life is how these models work in practical aspects, like the different question-answering paths of diagnostic application, which they can actually feel and contact with. Therefore, the visualization will focus on the behavior and surface of those machine learning models, which mostly are the question-answering paths in diagnostic applications.

Nevertheless, visualizing the surface and behavior does not mean translating those sophisticated algorithms in a superficial way. In diagnostic applications, question-answering path contains the basic logic of machine learning models, information of database and data generation process, which is the exterior performance of the intricate inner algorithm and it can represent the inner work of this system to a large degree. Moreover, focusing on the surface can make the complicated machine learning models furthest transparent and understandable for lay users and they can compare and analyze different AI chatbot systems even without any relevant knowledge.

[1] J. Zhu, A. Liapis, S. Risi, R. Bidarra and G. M. Youngblood, "Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation," 2018 IEEE Conference on Computational Intelligence and Games (CIG), Maastricht, 2018, pp. 1-8, doi: 10.1109/CIG.2018.8490433.

[2] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018).

[3] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51, 5 (2019), 93

As I mentioned in the section of visualization methods, the question-answering path is made up of questions from the applications and the answers from the user. In general, the questions can be classified into: single choice question, yes-no question, multiple choice question and question mentioned about pain; while the answers can be classified into answer with yes and answer with no/none of them. Hence, the key point is how to interpret these different categories of questions and answers in a lucid and visual way. Figure.13 already displayed the method of using different colors, shapes and sizes to highlight the different types of question, and using lines and dotted lines to represent different types of answer as well as their connection with questions, but it was only shown as one single case that is the input of "back pain". On the contrary, figure.14 showed the general view of decision making paths of all inputs and applications, but it did not develop the detailed question-answering path. Therefore, to make lay users have a overview of all the decision making path of different inputs and applications in detail, it's necessary to put the visualizations of detailed question-answering path of all seven inputs together and show them orderly in one image, so that lay users can compare and analyze all the paths effectively. In addition, the previous visualizations never mentioned the profile background and parameter, which is indispensable for the experiments and can give the user a clue of how the experiment is carried out and why choosing the specific answer rather than a random one during the question-answering path. Hence, the final visualization(Figure.15) should contain all the elements that are useful for lay users as well as follow the same visual regulations as the previous images.

Through figure.15, it's clear for lay users to understand and compare different diagnostic paths even if they do not know any computer knowledge, and this method of visualizing complicated machine learning models for lay users can also be used for other similar applications or AI chatbot systems except the diagnosis apps. By comparing these decision making paths they can have a general idea of how this kind of algorithm works and in what way the different paths result in totally different output, and finally they can develop their own justifications.

decision path of symptom in different apps

control the parameter to test 7 different pains respectively in Ada, Healthily and WebMD

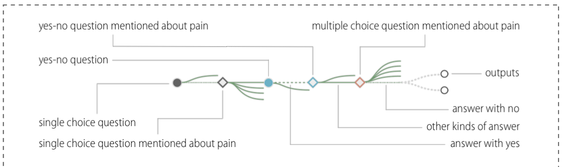
HOW TO READ?

QUESTION: ● **questions** ◇ **question mentioned about pain**

● **yes-no question** ● **single choice question** ● **multiple choice question**

ANSWER: **answers** **answer with no/none** **answer with yes**

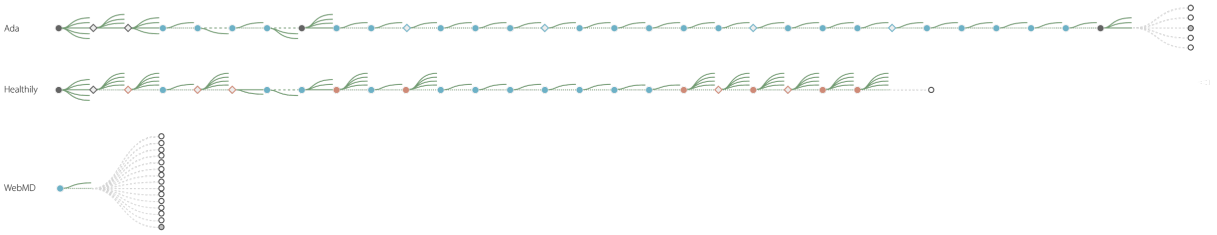
OUTPUT: ○ **outputs** ● **the same output**


yes-no question mentioned about pain
multiple choice question mentioned about pain
yes-no question
outputs
single choice question
answer with no
other kinds of answer
single choice question mentioned about pain
answer with yes

***** all the question and answer are arranged by original order*

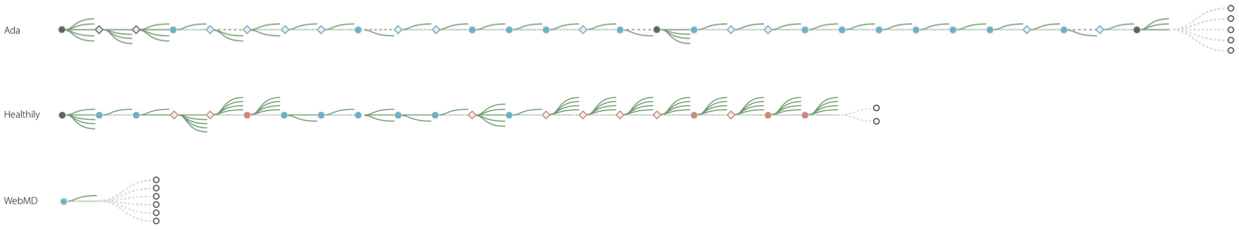
Abdomen Pain

parameter:
pain in both sides
worse with activities



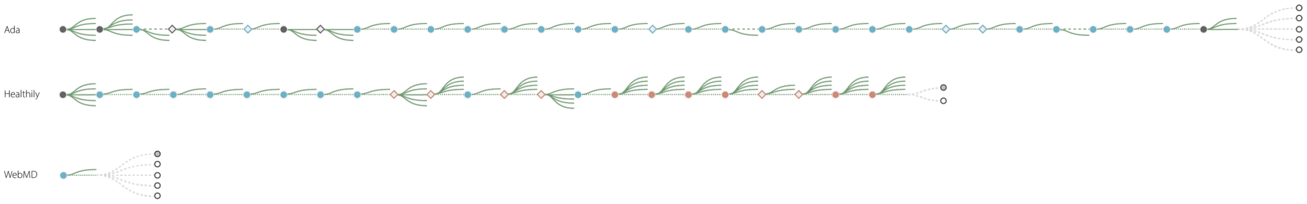
Back Pain

parameter:
lower back pain
menstrual pain
worse with activities



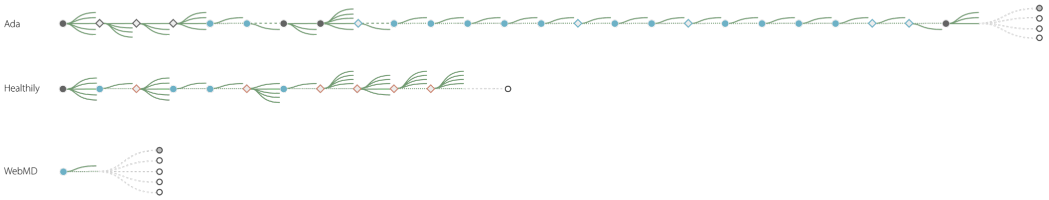
Chest Pain

parameter:
heavy and tight feeling
worse with activity
feel exhausted



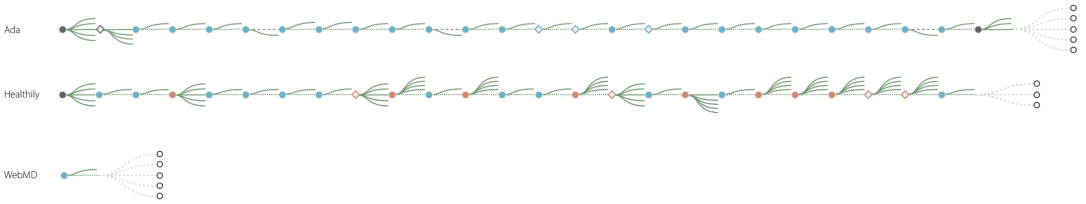
Foot Pain

parameter:
worse with activity
pain in heel
difficulty walking



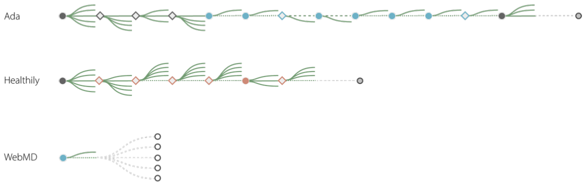
Joint Pain

parameter:
difficult to move joints
stiff joints
feel exhausted



Knee Pain

parameter:
pain in both sides
pain in outer side of knee
difficult to bend



Leg Pain

parameter:
tender thigh muscles
pain in thigh
pain in back thigh

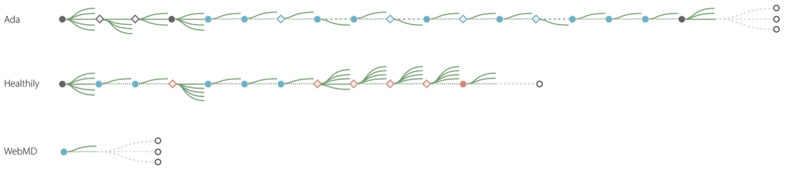


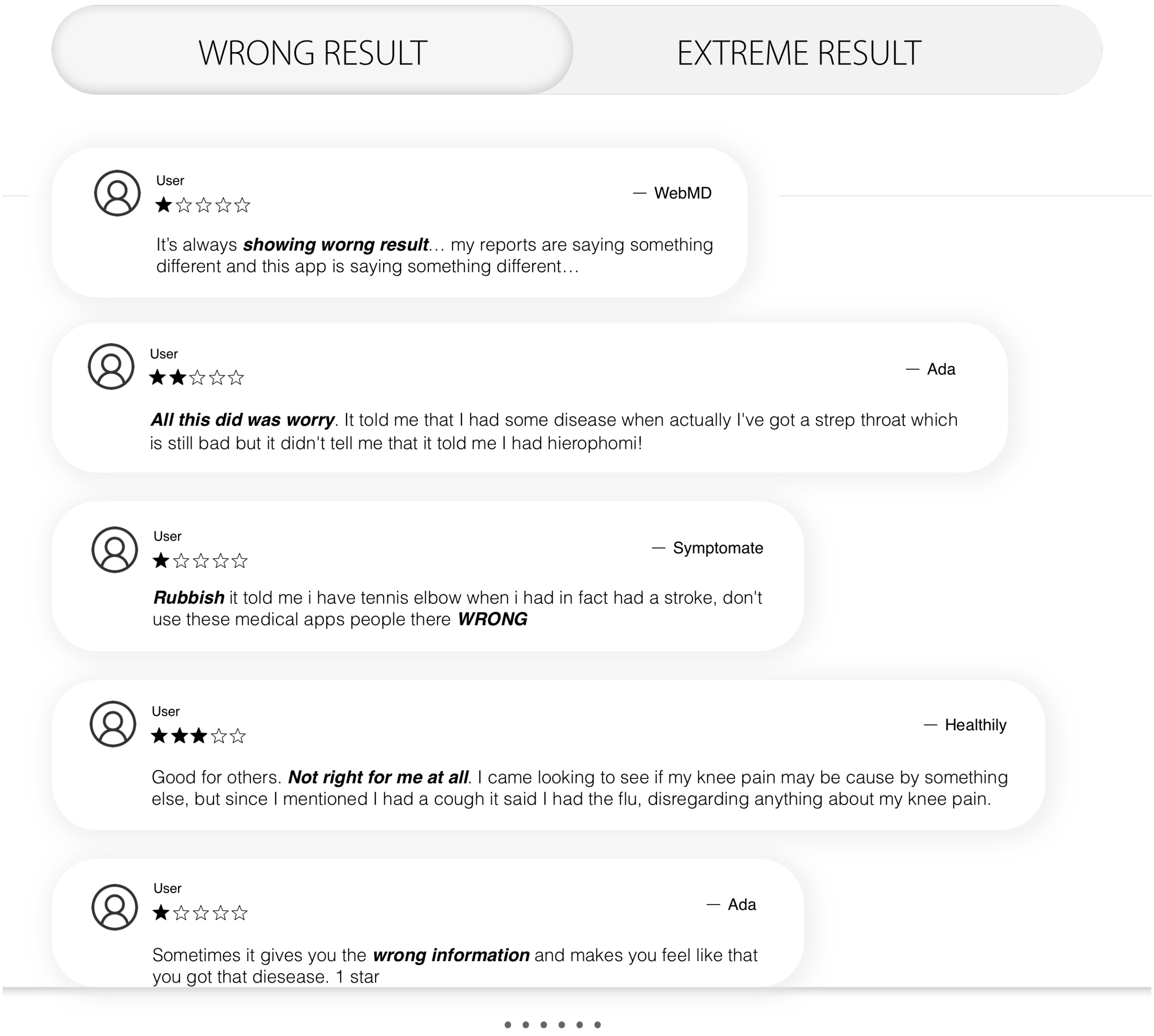
Figure.15 Visualization of all diagnostic paths with seven inputs

3.3 Visualization result

The final visualization contains all the 21 experiments as well as provides an understandable perspective of interpreting the sophisticated AI chatbot systems. However, more progress needs to be made to let lay users have the opportunity to come into contact with and then have a comprehensive understanding of this project. The single visualization is difficult to explain clearly the whole process of the project and attract users, so there should be a platform to involve all the information in a well structured and concise way. Thinking about the stability and convenience, an interactive website could be a suitable choice to present the whole project and final visualizations, which is convenient to communicate with lay users and is also easy to widely spread.

The website is not only a platform to transmit the project to lay users, but also a tool to make all the information and visualizations interactive and attractive. In order to clarify the whole process and highlight the most important information of the project, the website is structured into three part orderly: introduction page, main page and about page.

- **Introduction page:** To briefly introduce the background of the project and the status quo of diagnosis applications in order to attract lay users for the first time. For the reason that some users do not have comprehensive knowledge about diagnostic applications and have doubt of the outputs, there is a section of user reviews (Figure.16) which is classified into “wrong result” and “extreme result” by the key words. The wrong result means those reviews that mention the wrong or inaccurate diagnoses, while the extreme result means those reviews complaining about the more severe outputs such as cancer and syndrome. This section explains to users why this project is important for them and why they should know some basic knowledge of this kind of application. After the user reviews section, the visualization of data collection is also displayed in the introduction page, which continues from reviews and provides the basis for the following visualization of the experiments in the main page.



WRONG RESULT | EXTREME RESULT

User
★☆☆☆☆
— WebMD

It’s always ***showing worng result***… my reports are saying something different and this app is saying something different…

User
★★☆☆☆
— Ada

All this did was worry. It told me that I had some disease when actually I've got a strep throat which is still bad but it didn't tell me that it told me I had hierophomi!

User
★☆☆☆☆
— Symptomate

Rubbish it told me i have tennis elbow when i had in fact had a stroke, don't use these medical apps people there ***WRONG***

User
★★★☆☆
— Healthily

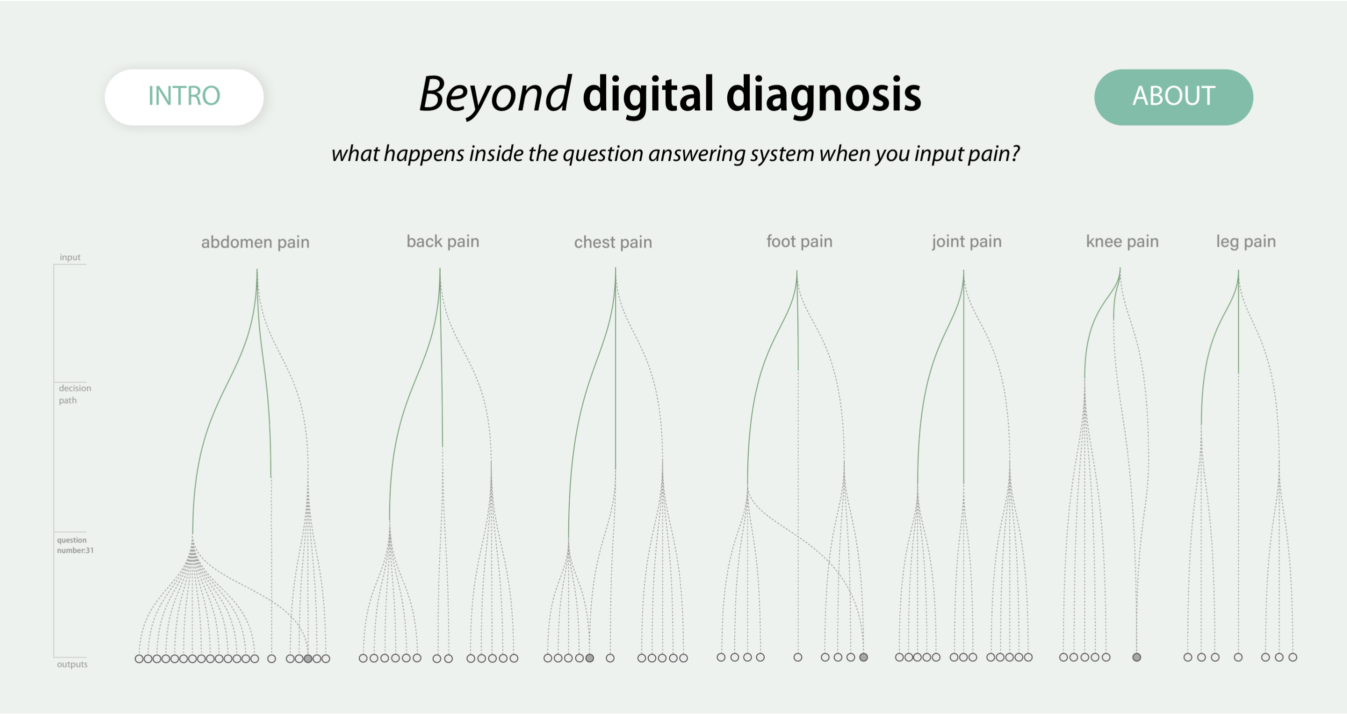
Good for others. ***Not right for me at all***. I came looking to see if my knee pain may be cause by something else, but since I mentioned I had a cough it said I had the flu, disregarding anything about my knee pain.

User
★☆☆☆☆
— Ada

Sometimes it gives you the ***wrong information*** and makes you feel like that you got that diesease. 1 star

Figure.16 The section of user reviews in introduction page

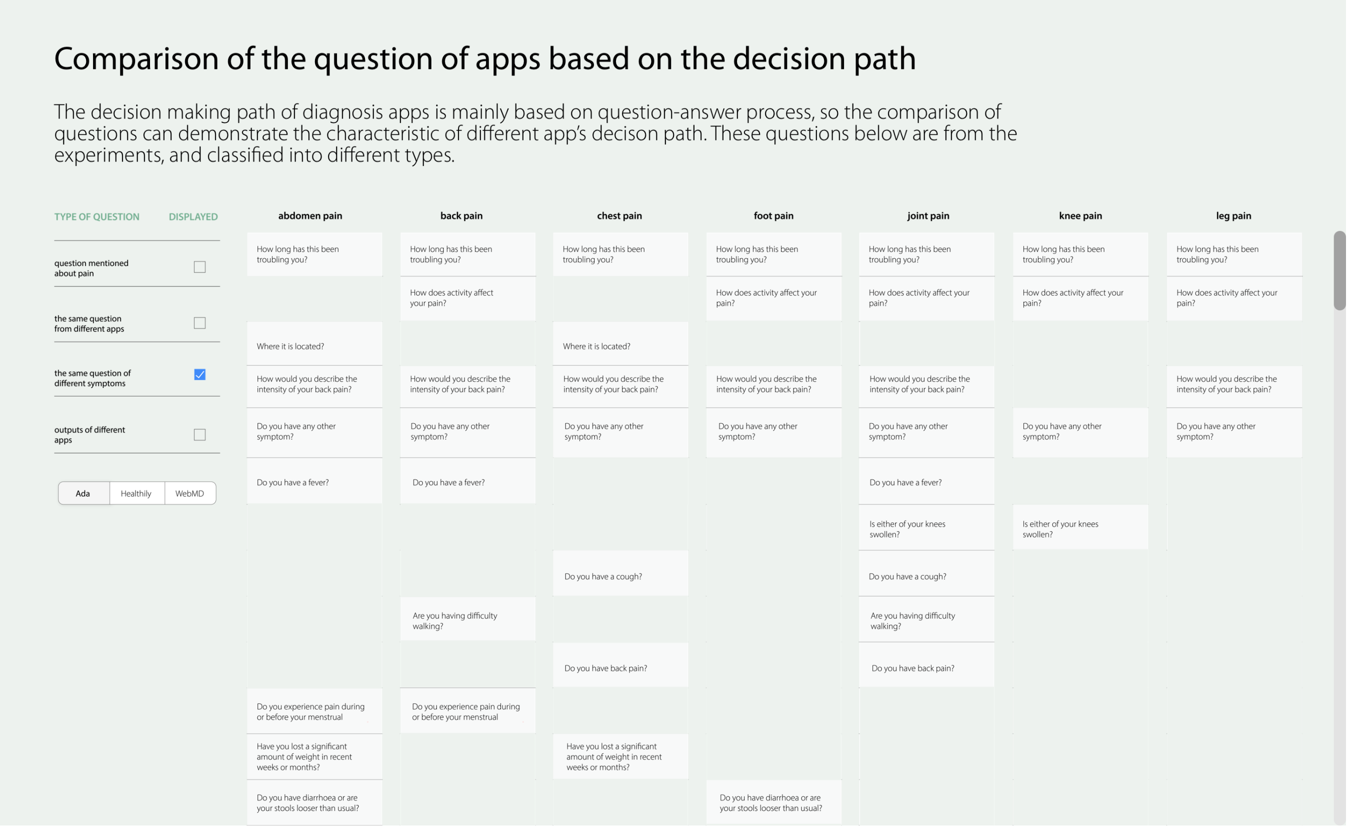
- **Main page:** The main page includes the visualization of all the experiments, which is the most essential part of this project. In the beginning of the main page, there is a general visualization(Figure.17) containing all the decision making paths and outputs of seven inputs, but it only shows the rough length of them that doesn't provide any detailed information. The aim of the general visualization is to let the user have an overview of the different diagnostic paths and realize how diverse these outputs are even with the same input. If they are interested in one specific input of them, they can click and see the detailed visualization with highlighting the questions and answers. Between the general visualization and detailed visualization of each input is the legend part, which tells the user how to read those visualizations effectively. The detailed visualization is interactive and if the user wants to know more about the questions, he/she can hover the mouse over that and the original question sentence will show up. However, it can only display one original question in one time and the main visual element is still the visualization, so at the end of the main page there is a section(Figure.18) of the comparison of all questions only by texts, which can let the user compare all types of question overall if they want to know the details of those questions beyond visualization.



Beyond **digital diagnosis**

what happens inside the question answering system when you input pain?

Figure.17 The general visualization of all diagnostic paths



Comparison of the question of apps based on the decision path

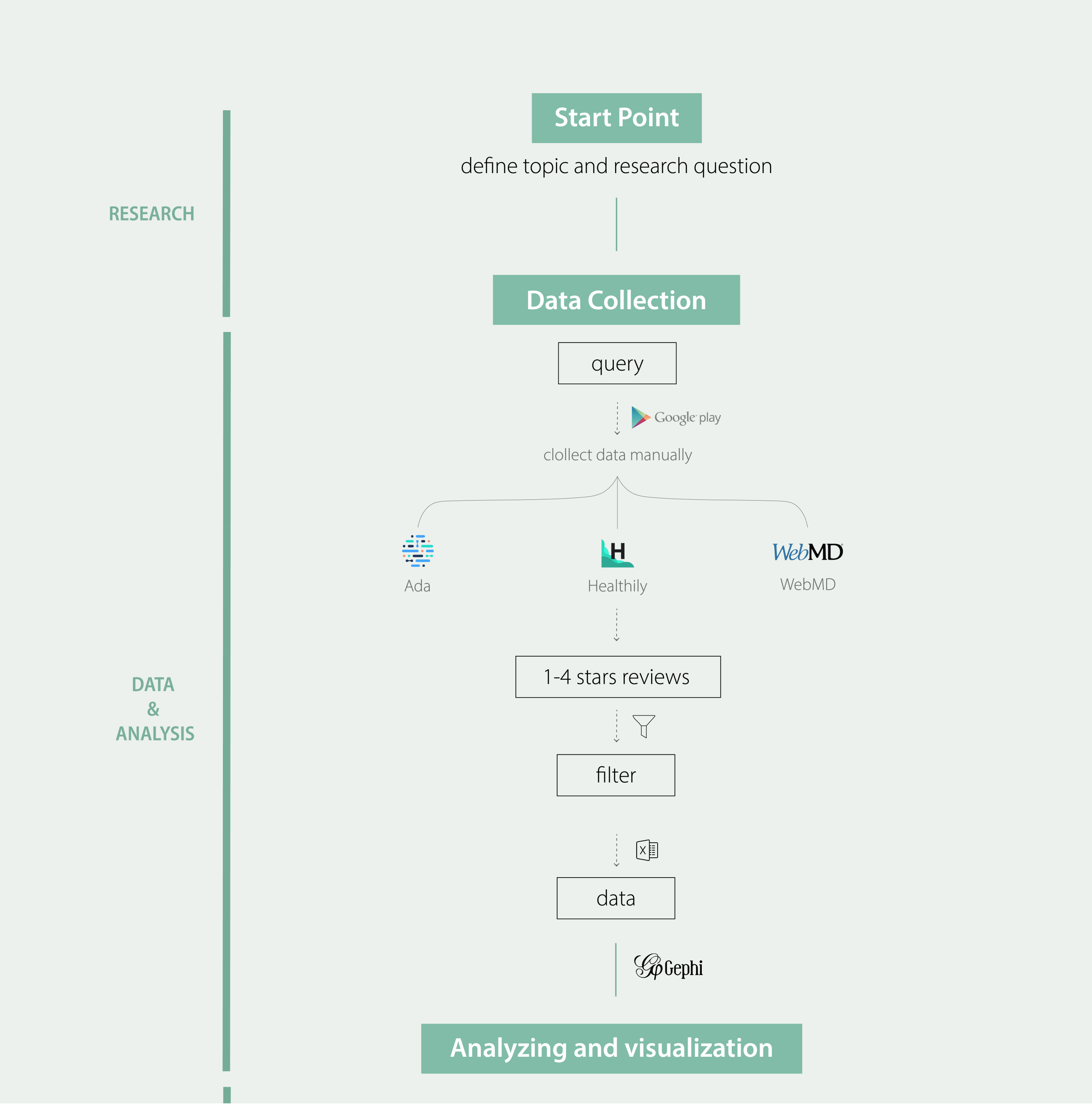
The decision making path of diagnosis apps is mainly based on question-answer process, so the comparison of questions can demonstrate the characteristic of different app's decison path. These questions below are from the experiments, and classified into different types.

TYPE OF QUESTION	DISPLAYED
question mentioned about pain	☐
the same question from different apps	☐
the same question of different symptoms	☑
outputs of different apps	☐

abdomen pain	back pain	chest pain	foot pain	joint pain	knee pain	leg pain
How long has this been troubling you?	How long has this been troubling you?	How long has this been troubling you?	How long has this been troubling you?	How long has this been troubling you?	How long has this been troubling you?	How long has this been troubling you?
	How does activity affect your pain?		How does activity affect your pain?	How does activity affect your pain?	How does activity affect your pain?	How does activity affect your pain?
Where it is located?		Where it is located?				
How would you describe the intensity of your back pain?	How would you describe the intensity of your back pain?	How would you describe the intensity of your back pain?	How would you describe the intensity of your back pain?	How would you describe the intensity of your back pain?		How would you describe the intensity of your back pain?
Do you have any other symptom?	Do you have any other symptom?	Do you have any other symptom?	Do you have any other symptom?	Do you have any other symptom?	Do you have any other symptom?	Do you have any other symptom?
Do you have a fever?	Do you have a fever?			Do you have a fever?		
				Is either of your knees swollen?	Is either of your knees swollen?	
		Do you have a cough?		Do you have a cough?		
	Are you having difficulty walking?			Are you having difficulty walking?		
		Do you have back pain?		Do you have back pain?		
Do you experience pain during or before your menstrual	Do you experience pain during or before your menstrual					
Have you lost a significant amount of weight in recent weeks or months?		Have you lost a significant amount of weight in recent weeks or months?				
Do you have diarrhoea or are your stools looser than usual?			Do you have diarrhoea or are your stools looser than usual?			

Figure.18 The comparison of all types of question by texts

- **About page:** About page is the explanation and supplement of the project which is mostly explained by texts, containing the sections of “about this project”, “data and method” and “protocol”. “About this project” explains the background and intends of this project, and “data and method” demonstrates how the data is collected and analyzed, as well as the method of carrying out the experiments. The “protocol” section uses a visualization(Figure.19) to briefly illustrate the whole protocol of this project, from the start point and data collection to final visualizations.



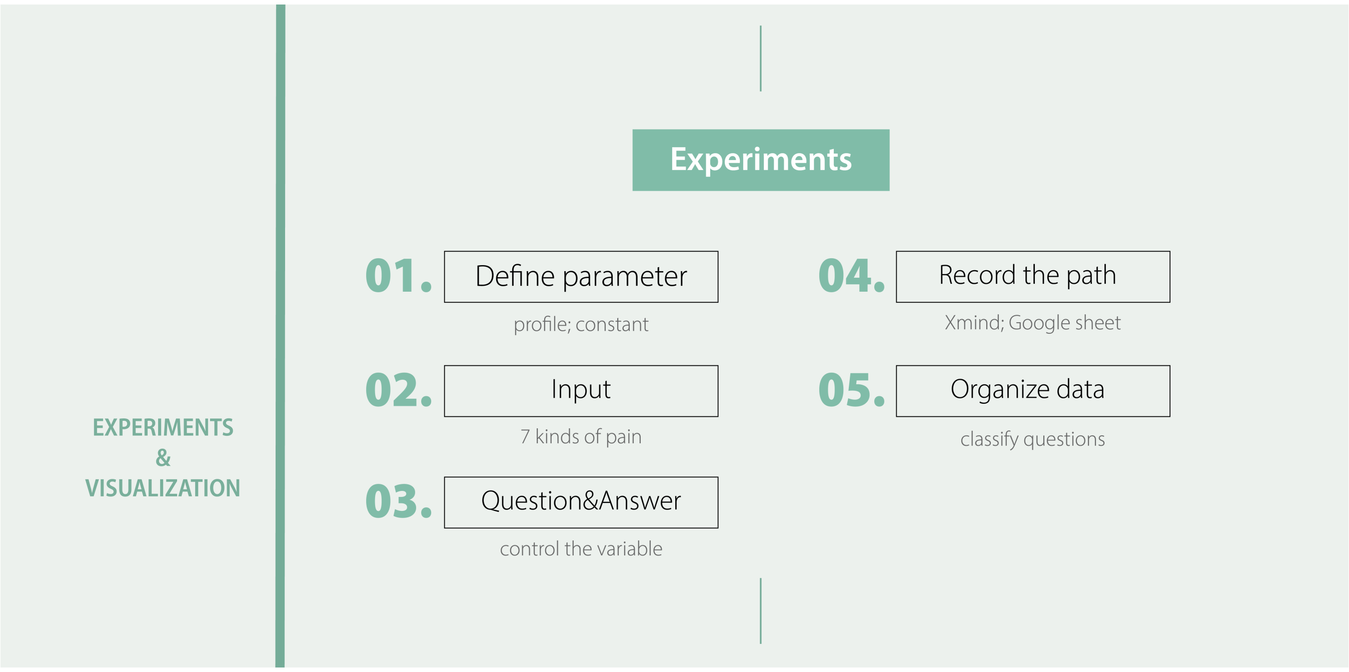


Figure.19 The whole protocol of the project

4. Conclusion and discussion

This project provides a perspective and a method of how visualization targeted at lay users can be applied for the algorithm of diagnosis applications or similar machine learning models, with a focus on visualizing the surface of question-answering systems rather than the inner algorithm routine. Beginning with the negative user reviews of diagnosis applications on Google Play Store, the project is carried out through data collection, data analysis, experiments and visualization result. All these steps are logical and linked with each other, in which the roughly browse of negative user reviews leads to the procedure of collecting more data and statistical analysis, and the data analysis provides the cause of doing experiments, and then the process and result of experiments contribute to the final visualization. During the process, to guarantee the outputs are not influenced by other irrelevant factors, all the 21 experiments are following the same regulation and method, and then the final visualizations also demonstrate these experiments in an understandable way. From the results of the experiments and visualization, we can summarize that the different decision making paths of diagnostic applications result in totally different outputs, and the diagnostic path is influenced by the key symptoms mentioned by questions and the types of questions. Another summary is that the diagnostic path is not the only determinant of diagnoses, like WebMD does not have any question-answering path but it also gives outputs.

The statistical analysis of experiments identifies how the decision making process of diagnosis applications work and what factors can affect its outputs, and the visualizations display all the paths and outputs in an understandable way by highlighting the surface of the inner algorithm. However, the aim of this project is not to tell users which diagnosis application is more accurate or suggest people that do not trust these obscure Artificial Intelligence. This project only demonstrates the performance of different diagnostic paths and discloses their inner question-answering systems objectively in an approachable way. The approachable method of translating sophisticated machine learning models enables lay users to have a general idea of how diagnostic applications take high-stake decisions in their daily life and why the diagnoses are inaccurate and diverse sometimes. Therefore, they can be aware that behind those unfaithful outputs are different AI chatbot systems and databases, and they will develop their own justification and critical thinking of this kind of application and algorithm finally.

Nonetheless, there are still some limitations and more advances could be developed in the further work. For example, the experiments only contain three diagnosis applications based on the downloads on Google Play Store, but there are more than 1,000 different diagnosis applications on different platforms. Therefore, in order to get more extensive and weighted data and results in the further work, more diagnostic applications can be taken into account with the same method. Also, during the process of user reviews collection, the five-star reviews are excluded from the database because of its positivity and enormous amount. However, those positive user reviews can also provide few information of unfaithful or inaccurate diagnoses although they are really negligible, so all the reviews can be collected and analyzed in the future if time permits. As I mentioned before, one important thing always needs to be clarified is that, during the experiments all the profile background and parameters are not unmodifiable, which means they could be changed according to different profiles or users based on different needs. Hence, this step could be more rigorous such as restricting the numbers and determining the types of parameters according to the features of different symptoms.

On the whole, this project develops the approach of interpreting algorithms for lay users from a communication designer perspective, and the method can also be used for other applications with the AI chatbot system, such as online chatbots in the field of e-commerce and finance. In addition, more data collection and experiments need to be carried out in the further work with the deeper research in this field.

Acknowledgement

I would like to extend my sincere gratitude to my supervisors, Michele Mauri and Beatrice Gobbo, for their instructive advice and useful suggestion on my thesis. I am deeply grateful of their help in the completion of this thesis. I am also deeply indebted to all the other tutor and teachers in sharing useful materials for their direct and indirect help to me.

Special thanks should go to my families who always encourage and support me, as well as friends who have put considerable time and effort into their comments on the draft.

References

[1] Jutel, Annemarie & Lupton, Deborah. (2015). Digitizing diagnosis: a review of mobile applications in the diagnostic process. Diagnosis. 2. 10.1515/dx-2014-0068.

[2] Palanica, Adam; Flaschner, Peter; Thommandram, Anirudh; Li, Michael; Fossat, Yan (January 3, 2019). "Physicians' Perceptions of Chatbots in Health Care: Cross-Sectional Web-Based Survey". Journal of Medical Internet Research. 21(4): e12887. doi:10.2196/12887. PM C 6473203. PMID 30950796.

[3] Swapnil. Dambe. Chatbots for healthcare. Retrieved from https://www.engati.com/chatbots-for-healthcare

[4] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker. Interacting meaningfully with machine learning systems: Three experiments. International Journal of Human- Computer Studies, 67(8):639–662, 2009.

[5] A. W. Harley. An interactive node-link visualization of convolutional neural networks. In Int. Symp. on Visual Computing, pp. 867–877. Springer, 2015.

[6] J. Zhu, A. Liapis, S. Risi, R. Bidarra and G. M. Youngblood, "Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation," 2018 IEEE Conference on Computational Intelligence and Games (CIG), Maastricht, 2018, pp. 1-8, doi: 10.1109/CIG.2018.8490433.

[7] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018).

[8] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51, 5 (2019), 93

[9] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436.

[10] Jutel, A, (2011). Putting a Name to It: Diagnosis in Contemporary Society. Baltimore: Johns Hopkins University Press.

[11] Improving Diagnosis in Health Care. National Academies of Sci- ences, Engineering and Medicine. 2015. Available at: http://iom. nationalacademies.org/Reports/2015/Improving-Diagnosis-in- Healthcare.aspx. Accessed: 14 Jun 2016.

[12] Mandl KD, Bourgeois FT. The evolution of patient diagno- sis: from art to digital data-driven science. J Am Med Assoc 2017;318:1859–60.

[13] Millenson ML, Baldwin JL, Zipperer L, et al. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. Diagnosis 2018;5:95–105.

[14] Enrico C. Paper Review: the Babylon Chatbot [Internet]. The Guide to Health Informatics 3rd Edition, 2018. Available: https://coiera.com/2018/06/29/paper-review-the-babylon-chatbot/ [Accessed 29 May 2019].

[15] Balkanyi L, Cornet R. The interplay of knowledge representation with various fields of artificial intelligence in medicine. Yearb Med Inform 2019;28:027–34.

[16] Aboueid S, Liu RH, Desta BN, et al. The use of artificially intelligent Self-Diagnosing digital platforms by the general public: Scoping review. JMIR Med Inform 2019;7:e13445.

[17] Magrabi F, Ammenwerth E, McNair JB, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform 2019;28:128–34.

[18] Semigran HL, Linder JA, Gidengil C, et al. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480.

[19] Correll, Michael. (2019). Ethical Dimensions of Visualization Research. CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1-13. 10.1145/3290605.3300418.

[20] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Sat- terthwaite, Oded Nov, and Enrico Bertini. 2017. Showing People Behind Data: Does Anthropomorphizing Visualizations Elicit More Empathy for Human Rights Data?. In Proceedings of the 2017 CHI Con- ference on Human Factors in Computing Systems. ACM, 5462–5474.

[21] Narayanan, Menaka & Chen, Emily & He, Jeffrey & Kim, Been & Gershman, Sam & Doshi velez, Finale. (2018). How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation.

[22] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, “Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2018.

[23] Sevastjanova, Rita & Beck, Fabian & Ell, Basil & Turkay, Cagatay & Henkin, Rafael & Butt, Miriam & Keim, Daniel & El-Assady, Mennatallah. (2018). Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models.

[24] J. Krause, A. Perer, and E. Bertini. Using visual analytics to interpret predictive machine learning models. arXiv preprint arXiv:1606.05685, 2016.

[25] U.Ehsan, B.Harrison, L.Chan, and M.O.Riedl. Rationalization: Aneural machine translation approach to generating natural language explanations. CoRR, abs/1702.07826, 2017.

[26] Heinrichs B, Eickhoff SB. Your evidence? Machine learning algorithms for medical diagnosis and prediction. Human Brain Mapping. 2020 Apr;41(6):1435-1444. DOI: 10.1002/hbm.24886.

[27] Kononenko, Igor. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. Artificial intelligence in medicine. 23. 89-109. 10.1016/S0933-3657(01)00077-X.

[28] Krause, Josua & Perer, Adam & Bertini, Enrico. (2016). Using Visual Analytics to Interpret Predictive Machine Learning Models.

[29] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, and others. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion (2019).

[30] Mittelstadt, Brent & Allo, Patrick & Taddeo, Mariarosaria & Wachter, Sandra & Floridi, Luciano. (2016). The Ethics of Algorithms: Mapping the Debate. Big Data & Society.

[31] Aboueid, Stephanie & Liu, Rebecca & Desta, Binyam & Chaurasia, Ashok & Ebrahim, Shanil. (2019). The use of artificially-intelligent self-diagnosing digital platforms by the general public: A scoping review. 10.2196/preprints.13445.

[32] El-Assady, Mennatallah & Jentner, Wolfgang & Kehlbeck, Rebecca & Schlegel, Udo & Sevastjanova, Rita & Sperrle, Fabian & Spinner, Thilo & Keim, Daniel. (2019). Towards XAI: Structuring the Processes of Explanations.

[33] Nahum D. Gershon and Ward Page. 2001. What Storytelling can do for Information Visualization. Commun. ACM, 44, 8, 31–37.

[34] Zezhong Wang, Shunming Wang, Matteo Farinella, Dave Murray-Rust, Nathalie Henry Riche, and Benjamin Bach. 2019. Comparing Effectiveness and Engagement of Data Comics and Infographics. In Proc. of ACM CHI.

[35] Edward Segel and Jeffrey Heer. 2010. Narrative Visualization: Telling Stories with Data. IEEE
Trans. Vis. Comput. Graph., 16, 6, 1139–1148.