The Accuracy of Artificial Intelligence (AI) Chatbots in Telemedicine

Robert K. Swick

Spring Valley High School, 120 Sparkleberry Lane, Columbia, SC 29229

Partially due to the pandemic, artificial Intelligence (AI) chatbots in telemedicine are a recent advancement in the medical field and are pushing the medical community forward toward automated healthcare. The purpose of this research was to discover whether AI chatbots are effective in giving a patient an idea of what medical condition he/she might have before consulting a medical professional. The hypothesis was that of the four AI chatbots that were used--Symptomate, Ada, Isabel Symptom Checker, and K Health--K Health was the one that would be the most accurate. This prediction was made based on the user interface and the accessibility of the app. Chatbots were tested by first developing a set of medical symptoms after consulting with a medical professional. The predetermined medical symptoms were input into the AI chatbots which then gave a diagnosis. Out of thirty trials, the accuracy of the chatbots were as follows. Symptomate averaged 66% correct diagnoses, Isabel Symptom Checker averaged 86% correct diagnosis, K health and ADA scored with a mean of 80% accuracy. Since the F_u was less than F the null hypothesis failed to be rejected.

Introduction

Artificial Intelligence (AI) chatbots in telemedicine are a recent application in the medical field that are pushing the medical community forward in these hard times to automate healthcare (Amato et al., 2020; Hao, 2020; Miner et al., 2020). Covid-19 is an illness caused by a new coronavirus that can spread from person to person and has infected people throughout the world. Restrictions due to the virus have caused many problems with how people have been able to carry out simple everyday tasks, such as going to the doctor. Hao reported an increase in the use of AI chatbots that provides evidence of Covid complicating the medical field. Hao wrote that once the pandemic hit, the state governor ordered a 50% reduction of all government staff, forcing a cut of most call center employees (Hao, 2020). Hao explains that since many patients currently need to be quarantined or cannot go to the doctor's office because of health risks, the patients have found a new way to get medical diagnoses and treatment: AI chatbots. Amato et al. (2020) researched the advances of technologies in the AI chatbot field and found one of the most interesting challenges for modern eHealth is the application of intelligent recommendation systems. Additionally, other research has found that when chatbots are effectively designed and deployed, they are capable of sharing up-to-date information quickly, which may aid future pandemic preparedness (Miner et al., 2020).

In the field of chatbots in healthcare, many recent contributions have been made. Palanica (2019) states that many potential benefits for the uses of chatbots within the context of health care have been theorized, such as improved patient education and treatment compliance. There are several different brands of AI programs that run on these chatbots. Chatbots can be separated into two categories: those that are free and those that cost money. The way that chatbots work is after patients input their symptoms and answer a couple of questions, the chatbot then informs them whether or not to seek medical attention or contact a medical professional. One example of a chatbot is Ada, a phone-based app that is free to the public. Ada, created by CEO Mike Murchison, provides detailed questions to single out illnesses that it does not recognize. Symptomate is a desktop-based AI chatbot that works in a similar way as Ada. The difference between them is they use different patient records as a basis for their diagnoses. Symptomate was created by Infermedica, and works in the same exact way Ada works. A little girl, Isabel, who was struggling with a case of chickenpox, inspired a team of charity workers to create the chatbot, Isabel symptom checker, to help those in need. The Isabel Symptom Checker is also computer-based and is the easiest to use out of these four chatbots. Inputting a patient's symptoms and getting a diagnosis using Isabel Symptom Checker take less time; however, because of the convenience, this bot is less accurate with the answers that it provides a patient due to less specific information. K Health, a phone-based Apple app, was created in 2016 and run by Allon Bloch to provide people with a free healthcare solution as an alternative to waiting in lines at medical facilities. Available from the Apple app store K Health initially asks for symptoms and then asks specific questions based on the symptoms that were provided. After both of these steps, the app provides a diagnosis based on people who have answered similarly and urges users to seek help from a medical professional.

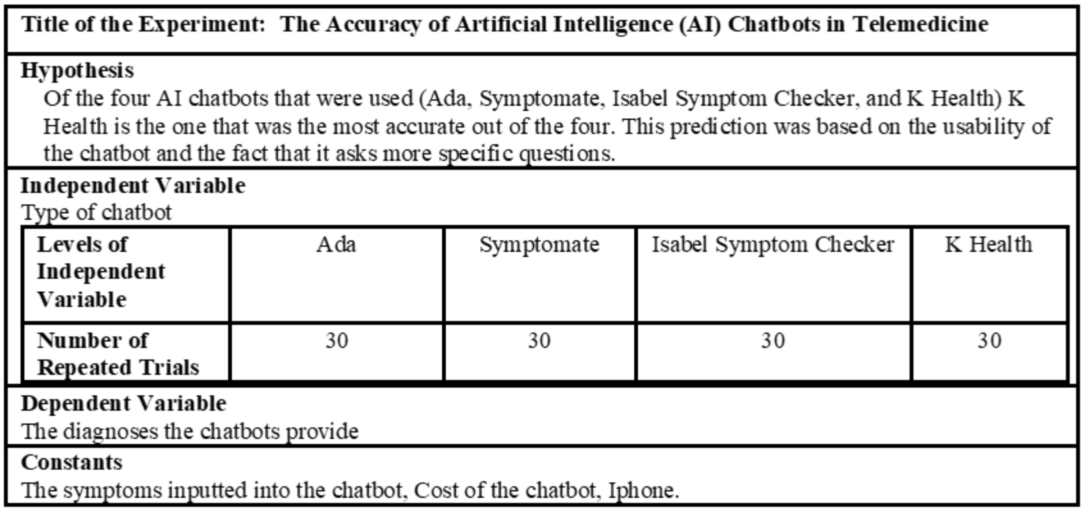
These four chatbots can be used by people in most situations because they are all free and can either be found on the web or in the Apple app store. In the past, medical AIs were primarily used in hospitals as triage systems to tell nurses which patients may need to be seen first, but as AI has developed so has the use of this technology. Some new applications of this technology include using image technology to predict whether a user has a condition that needs to be treated and how it might progress (Heaven, 2020).

The purpose of this research was to find out whether AI chatbots are effective in giving a patient an idea of what medical condition he/she might have before consulting a medical professional, as well as to find out which AI chatbot was most accurate with a certain set of symptoms. It was hypothesized that when the symptoms of real medical conditions were put into the AI chatbots Ada, Symptomate, Isabel Symptom Checker, and K Health, the AI chatbot K Health would recognize that the patient needs to contact a doctor and would diagnose them more accurately than the other four chatbots. This prediction was based on the usability of the chatbot and the fact that K Health asks more specific questions. To test the accuracy of these chatbots the symptoms of real medical problems were input into the four different AI chatbots, and the resulting diagnoses were written down and compared to the actual diagnoses a medical professional would give.

Methods

First, the chatbot to be tested was installed on a mobile phone or desktop. Then medical symptoms were provided by MayoClinic.org, a website that was recommended by a medical professional. Thirty different sets of symptoms were then inputted into Ada. An example of the symptoms that were used in this experiment are chest pain that radiates into the neck and left arm, as well as shortness of breath and diaphoresis, which should result in a diagnosis of myocardial infarction/or a heart attack. The other 29 medical symptoms used in this experiment, are listed in figure A1 in the Appendix. The diagnosis the first chatbot gave was then recorded in a data table. The same procedure was repeated for each remaining chatbot. Due to privacy constraints, medical symptoms were not from real patients; they came from medical conditions that were suggested by a medical professional or mayoclinic.org. (M.Miller, personal communication, December 15, 2020). The Data collected in the experiment was then coded as a correct response as a 1 and an incorrect response as a 0. The encoded data was then put through a one-way ANOVA test.

Figure 1. Experimental Design Diagram (EDD)



Title of the Experiment: The Accuracy of Artificial Intelligence (AI) Chatbots in Telemedicine

Hypothesis

Of the four AI chatbots that were used (Ada, Symptomate, Isabel Symptom Checker, and K Health) K Health is the one that was the most accurate out of the four. This prediction was based on the usability of the chatbot and the fact that it asks more specific questions.

Independent Variable

Type of chatbot

Levels of Independent Variable	Ada	Symptomate	Isabel Symptom Checker	K Health
Number of Repeated Trials	30	30	30	30

Dependent Variable

The diagnoses the chatbots provide

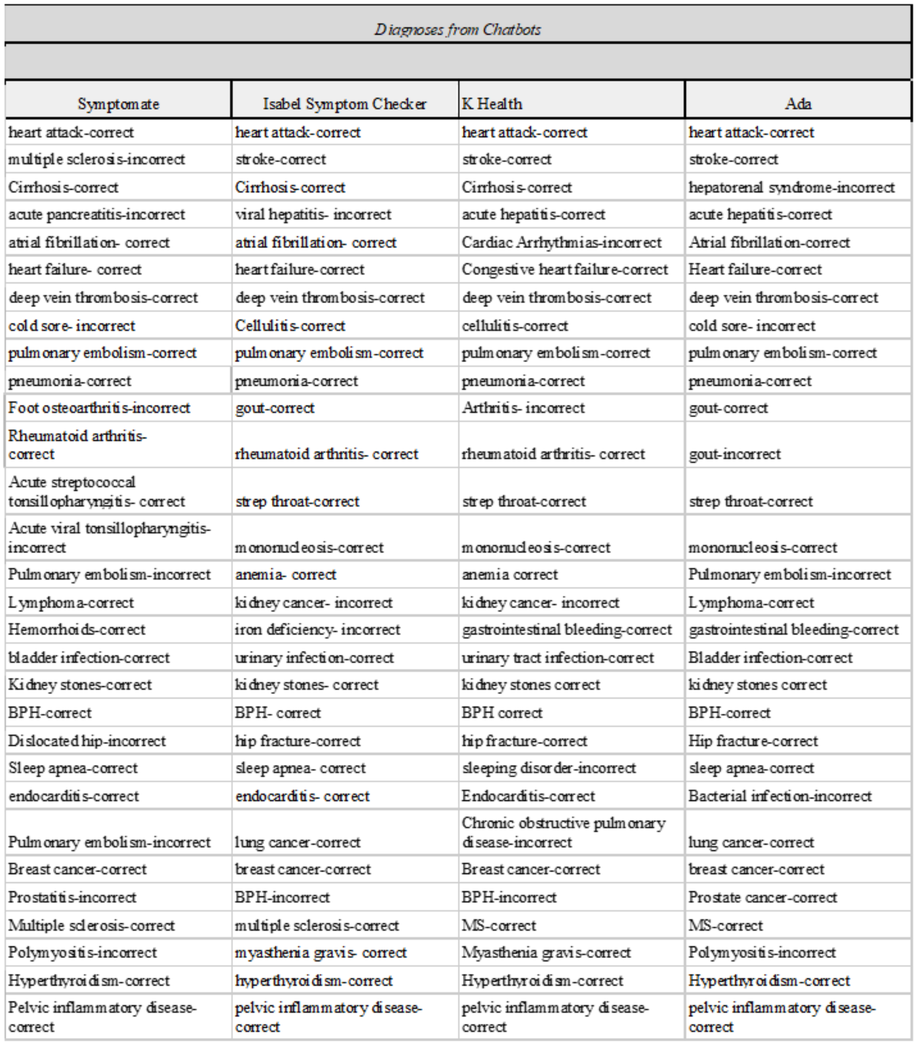
Constants

The symptoms inputted into the chatbot, Cost of the chatbot, Iphone.

Results

The diagnoses for a collection of the same symptoms inputted into the four different AI Chatbots are shown in Table 1. Each chatbot provided diagnoses in a unique way. For example, the Ada Chatbot gave three possible medical diagnoses while the Isabel symptom checker gave 15 for the same set of symptoms. Table 2 shows the accuracy of the chatbots after all the symptoms were input. Isabel Symptom Checker was the most accurate chatbot but, due to the vagueness of the diagnoses given, the most reliable chatbot was Ada. Figure 2 depicts the ranking of the chatbots accuracy in percentage, and Isabel Symptom Checker can clearly be determined as the most accurate. Table 3 depicts the one-way ANOVA analysis of the raw data (F=1.24, P=0.298, α=0.05). The null hypothesis failed to be rejected (0.298>0.05).

Table 1. Diagnoses Provided by Each Chatbot



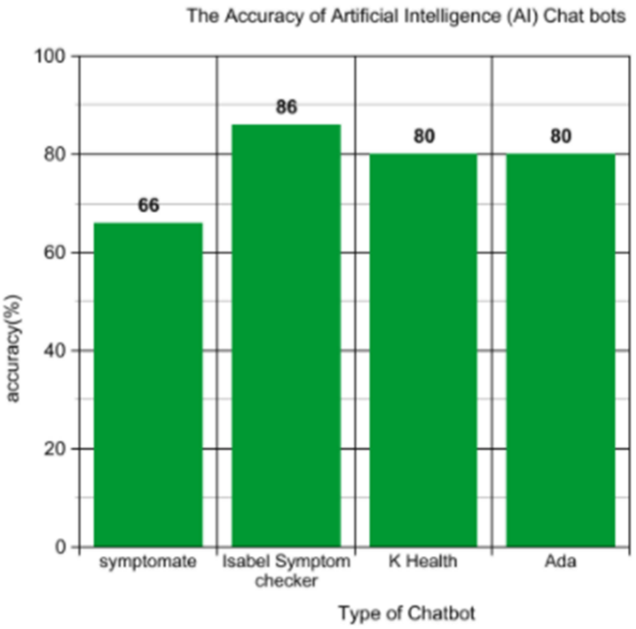
Diagnoses from Chatbots			
Symptomate	Isabel Symptom Checker	K Health	Ada
heart attack-correct	heart attack-correct	heart attack-correct	heart attack-correct
multiple sclerosis-incorrect	stroke-correct	stroke-correct	stroke-correct
Cirrhosis-correct	Cirrhosis-correct	Cirrhosis-correct	hepatorenal syndrome-incorrect
acute pancreatitis-incorrect	viral hepatitis- incorrect	acute hepatitis-correct	acute hepatitis-correct
atrial fibrillation- correct	atrial fibrillation- correct	Cardiac Arrhythmias-incorrect	Atrial fibrillation-correct
heart failure- correct	heart failure-correct	Congestive heart failure-correct	Heart failure-correct
deep vein thrombosis-correct	deep vein thrombosis-correct	deep vein thrombosis-correct	deep vein thrombosis-correct
cold sore- incorrect	Cellulitis-correct	cellulitis-correct	cold sore- incorrect
pulmonary embolism-correct	pulmonary embolism-correct	pulmonary embolism-correct	pulmonary embolism-correct
pneumonia-correct	pneumonia-correct	pneumonia-correct	pneumonia-correct
Foot osteoarthritis-incorrect	gout-correct	Arthritis- incorrect	gout-correct
Rheumatoid arthritis-correct	rheumatoid arthritis- correct	rheumatoid arthritis- correct	gout-incorrect
Acute streptococcal tonsillopharyngitis- correct	strep throat-correct	strep throat-correct	strep throat-correct
Acute viral tonsillopharyngitis-incorrect	mononucleosis-correct	mononucleosis-correct	mononucleosis-correct
Pulmonary embolism-incorrect	anemia- correct	anemia correct	Pulmonary embolism-incorrect
Lymphoma-correct	kidney cancer- incorrect	kidney cancer- incorrect	Lymphoma-correct
Hemorrhoids-correct	iron deficiency- incorrect	gastrointestinal bleeding-correct	gastrointestinal bleeding-correct
bladder infection-correct	urinary infection-correct	urinary tract infection-correct	Bladder infection-correct
Kidney stones-correct	kidney stones- correct	kidney stones correct	kidney stones correct
BPH-correct	BPH- correct	BPH correct	BPH-correct
Dislocated hip-incorrect	hip fracture-correct	hip fracture-correct	Hip fracture-correct
Sleep apnea-correct	sleep apnea- correct	sleeping disorder-incorrect	sleep apnea-correct
endocarditis-correct	endocarditis- correct	Endocarditis-correct	Bacterial infection-incorrect
Pulmonary embolism-incorrect	lung cancer-correct	Chronic obstructive pulmonary disease-incorrect	lung cancer-correct
Breast cancer-correct	breast cancer-correct	Breast cancer-correct	breast cancer-correct
Prostatitis-incorrect	BPH-incorrect	BPH-incorrect	Prostate cancer-correct
Multiple sclerosis-correct	multiple sclerosis-correct	MS-correct	MS-correct
Polymyositis-incorrect	myasthenia gravis- correct	Myasthenia gravis-correct	Polymyositis-incorrect
Hyperthyroidism-correct	hyperthyroidism-correct	Hyperthyroidism-correct	Hyperthyroidism-correct
Pelvic inflammatory disease-correct	pelvic inflammatory disease-correct	pelvic inflammatory disease-correct	pelvic inflammatory disease-correct

Table 2. Accuracy of All Chatbots

	Symptomate	Isabel Symptom Checker	K Health	Ada
Accuracy	20/30	26/30	24/30	24/30
Percentage	66%	86%	80%	80%

Table 2 displays the number of correct diagnoses for each chatbot out of 30 sets of input symptoms. The overall average accuracy of 23/30 or 77.5% correct medical diagnoses.

Figure 2. Accuracy of all chatbots



This is a graphical representation of how well each chatbot performed compared to the others. The data suggested that the Isabel Symptom Checker was the most accurate of the four chatbots.

Table 3. ANOVA Summary Table

Source	Sum of Squares	d.f	Mean Square	F	P
Between	0.63	3	0.21	1.24	0.298
Within(error)	19.73	116	0.17		
Total	20.36	119			

This table depicts the process used to show that the type of chatbot was not significant (P=0.298 a=0.05). The null hypothesis was failed to be rejected (0.298>0.05).

Discussion

The purpose of this research was to determine which types of the (AI) Chatbots, Ada Symptomate, Isabel Symptom Checker or K Health, provided the most accurate diagnosis for given sets of symptoms. These results could allow patients to choose the most accurate chatbot and allow medical professionals to suggest a specific chatbot. After conducting this experiment, it was found that Isabel Symptom Checker was the most accurate out of all the chatbots. The hypothesis that of the four AI chatbots that were used--Symptomate, Ada, Isabel Symptom Checker, and K Health--K Health would be the one that was the most accurate out of the four. This hypothesis was not supported by the data collected in this study. Table 2 shows the accuracy for all chatbots in context to the raw data. When graphed (Figure 2) the chatbots are compared by their overall accuracy in percentages with Isabel Symptom Checker having the best score of 86%. A one-way analysis of variance showed that the type of chatbot was not significant, F=1.24, F_u=2.76, p=0.298. $H_o= u_1= u_2= u_3= u_4$, H_a=at least one mean is different. Since 0.298>0.05 the H_o has failed to be rejected.

Potential benefits of using chatbots include increased patient knowledge before visits and increased awareness (Palanica, 2019). Palancia (2019) conducted a similar experiment that found averages of AI chatbot effectiveness similar to the current study. Amato et al., (2020) conducted research using a similar method to the current study. They used a total of 16,733 patient records and tested their symptoms with the chatbots to see if the chatbots were correct in their diagnoses. The Chatbot scored with a mean of 74.65%. This further contributes to this paper's support in justifying the methods used in the current study. Furthermore, Abd-Alrazaq et al. (2020) stated that chatbots have been used as triage systems in immediate health care offices. They are used to qualify who needs medical attention the most and who goes where in the hospital. Heaven (2020) claims that chatbots as a whole are a growing industry and are indeed lifesaving technology. This research pulls together information and relevance from many sources. It uses this research to prove that chatbots are effective at diagnosing diseases and informing the public about what they might have.

Sources of error that could affect results include wrong symptoms regarding the diagnoses given. Another source of error could be using an uncredited source for finding symptoms. Some improvements to the procedure could be made, such as finding a less time-consuming way to test a chatbot without altering results, or using more symptoms to get a clearer result. In future studies more chatbots could be tested to show a greater variety in statistical evidence. More symptoms could also be used and as a result the relevance of the current research could be shown more clearly.

Acknowledgements

First of all, I would like to thank Dr. Miles Miller for helping me with finding symptoms and credible resources to use while experimenting. He helped me with finding out what is credible on the internet as being used as a background for specific medical symptoms. I would also like to thank Dr. Nick Patel for helping me refine my research process and know what to look for in AI chatbots as a whole. I would like to thank my teachers Dr. Wyatt and Lindsey Rega who helped me with my statistical analysis of my data and pushed me along the research process. I would also like to thank my parents for pushing me along the research pathway. And lastly, I would like to thank my brother, who is always there to wake me up in the morning and prepare me for the day ahead.

References

Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical metrics used to evaluate health care chatbots: Scoping review. *Journal of Medical Internet Research*, *22*(6), e18301. https://doi.org/10.2196/18301

Amato, F., Marrone, S., Moscato, V., Piantadosi, G., Picariello, A., & Sansone, C. (2017, November). Chatbots meet eHealth: Automatizing healthcare. *WAIAH@ AI* IA,* 40-49.

Hao, K. (2020, May 14). The pandemic is emptying call centers. AI chatbots are swooping in. *MIT Technology Review; MIT Technology Review*. https://www.technologyreview.com/2020/05/14/1001716/ai-chatbots-take-call-center-jobs-during-coronavirus-pandemic/Diseases and Conditions - Mayo Clinic . (2021). Retrieved 6 February 2021, from https://www.mayoclinic.org/diseases-conditions/index?letter=A

Heaven. (23 Sept., 2020). We're not ready for AI, says the winner of a new $1M AI prize." *MIT Technology Review*, *MIT Technology Review,* www.technologyreview.com/2020/09/23/1008757/interview-winner-million-dollar-ai-prize-cancer-healthcare-regulation/. Accessed 6 Jan. 2021.

Miller, M. (2020, december 12) Personal interview [Zoom Meeting] Miner, A. S., Laranjo, L., & Kocaballi, A. B. (2020). Chatbots in the fight against the COVID-19 pandemic. *Nature PartnerJournals Digital Medicine*, *3*(1).https://doi.org/10.1038/s41746-020-0280-0

Palanica, A., Flaschner, P., Thommandram, A., Li, M., & Fossat, Y. (2019). Physicians'perceptions of chatbots in health care: Cross-sectional web-based survey. *Journal of Medical Internet Research*, *21*(4), e12887. https://doi.org/10.2196/12887

Wang, W. & Siau, K., 2019. *Trust in Health Chatbots*. *ResearchGate* [online]https://www.researchgate.net. Available at: <https://www.researchgate.net/profile/Keng_Siau/publication/333296446_Trust_in_Health_Chatbots/links/5ce5dd35299bf14d95b1d15b/Trust-in-Health-Chatbots.pdf> [Accessed 6 January 2021].

Appendix

Figure A1. List of Medical Symptoms Used with Chatbots

Myocardial infarction- chest pain radiating into neck and L arm, shortness of breath, diaphoresis

Stroke- difficulty speaking, facial droop, limb weakness

Cirrhosis- fatigue, muscle wasting, jaundice, abdominal swelling (ascites), leg swelling, dilated veins (caput medusa, varices)

acute hepatitis- fever, nausea/vomiting, jaundice

atrial fibrillation- irregular heart rhythm, palpitations, shortness of breath, occasionally leg swelling

congestive heart failure- shortness of breath particularly with exertion, cough with clear sputum (sometimes pink), leg swelling, inability to lay flat secondary to shortness of breath

deep vein thrombosis- unilateral leg pain, swelling, erythema

cellulitis- red skin, often fever, tender to touch, sometimes blistering, red streaking

pulmonary embolism- cough, hemoptysis, pleuritic pain, chest pain, sometimes lightheadedness

pneumonia- cough, fever, purulent sputum

gout- severe joint pain, swelling/erythema, often just in big toe

rheumatoid arthritis- joint pain, swelling, deformity, often symmetric (both hands, etc)

strep throat- fever, sore throat, tonsillar exudates, no cough

mononucleosis (epstein-barr virus)- sore throat, fatigue, fever, malaise. Often with abnormal blood work and enlarged spleen (on exam)

anemia- pallor, fatigue, lightheadedness, shortness of breath

lymphoma- fever, night sweats, weight loss. Palpable adenopathy

gastrointestinal bleeding- dark black tarry stools, red blood per rectum, pallor, fatigue, lightheadedness, shortness of breath

urinary tract infection- burning with urination, increased frequency, sometimes fever, bladder spasm

nephrolithiasis (kidney stone)- severe flank pain, sometimes blood in urine, costovertebral angle tenderness. Pain can be colicky or in paroxysms

benign prostatic hypertrophy- male difficulty with urination, hesitancy, dribbling, incomplete emptying, straining to urinate

hip fracture- hip/groin pain after fall, internal rotation of foot

sleep apnea- morning headaches, frequent awakenings at night, snoring, frequent nodding off during the day, not feeling rested in the am

endocarditis- fever, cough, rash, splinter hemorrhages in fingernails, weakness, weight loss, nodular lesions on palms.

cancers: lungs, breast, stomach, prostate (all can have some differentiating symptoms)

-lungs: cough, weight loss, hemoptysis

-breast: palpable breast mass, weight loss

-prostate- bph symptoms as above, weight loss, bone pain

multiple sclerosis- vision changes (double, vision loss), numbness or weakness, electric shock sensations with neck movement, fatigue, slurred speech, dizziness

myasthenia gravis- weakness worsening with repetitive exertion, eye lid lag

hyperthyroidism- weight loss, diarrhea, tremor, insomnia, manic behavior

pelvic inflammatory disease- abdominal pain, fever, purulent vaginal discharge

pancreatitis- severe abdominal pain radiating to the back, worse with eating, sometimes diarrhea with fatty meals, dehydration

cholecystitis- fever, nausea, pain with eating, tenderness in R upper quadrant of abdomen

pericarditis- chest pain, pleuritic pain, pain worse with laying flat and feels better when sitting forward. Sometimes fever

Symptoms that were used in the current study were found on Mayoclinic.org and were double checked by a medical professional. (Mayo Clinic, 2021)