



**Michigan  
Technological  
University**

Michigan Technological University  
**Digital Commons @ Michigan Tech**

---

Dissertations, Master's Theses and Master's Reports

---

2020

## INVESTIGATING THE IMPACT OF EXPLANATION ON REPAIRING TRUST IN AI DIAGNOSTIC SYSTEMS FOR RE-DIAGNOSIS

Lamia Alam

*Michigan Technological University, [lalam@mtu.edu](mailto:lalam@mtu.edu)*

Copyright 2020 Lamia Alam

---

### Recommended Citation

Alam, Lamia, "INVESTIGATING THE IMPACT OF EXPLANATION ON REPAIRING TRUST IN AI DIAGNOSTIC SYSTEMS FOR RE-DIAGNOSIS", Open Access Master's Thesis, Michigan Technological University, 2020.  
<https://doi.org/10.37099/mtu.dc.etr/1029>

Follow this and additional works at: <https://digitalcommons.mtu.edu/etr>



Part of the [Health Communication Commons](#), [Health Information Technology Commons](#), [Health Psychology Commons](#), and the [Human Factors Psychology Commons](#)

INVESTIGATING THE IMPACT OF EXPLANATION ON REPAIRING TRUST IN  
AI DIAGNOSTIC SYSTEMS FOR RE-DIAGNOSIS

By  
Lamia Alam

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2020

© 2020 Lamia Alam

This thesis has been approved in partial fulfillment of the requirements for the Degree of  
MASTER OF SCIENCE in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Thesis Advisor: *Shane T. Mueller*

Committee Member: *Erich J. Petushek*

Committee Member: *Robert R. Hoffman*

Department Chair: *Kelly S. Steelman*

# Table of Contents

List of figures .....	vii
List of tables.....	viii
Acknowledgements .....	ix
Abstract .....	xi
1 Introduction.....	1
1.1 Problem Statement.....	1
1.2 Summary.....	3
2 Review of Literature .....	5
2.1 Medical Diagnosis Decision Making in Diagnosticians.....	5
2.1.1 Medical Decision Biases.....	6
2.1.2 Problem Solving Principles in Diagnostic Reasoning .....	8
2.2 Patient-centered Communication.....	11
2.3 AI Diagnostic Systems.....	13
2.4 Self-diagnosis Systems .....	14
2.4.1 Symptom Checkers .....	15
2.4.2 Medical Diagnosis Conversational Agents/Chatbots .....	17
2.5 Trust in AI-based Diagnosis .....	18
2.6 Explainable Artificial Intelligence.....	22
2.6.1 Explanation in Healthcare AI .....	22
2.7 Summary and Conclusion of the Literature Review.....	24



3	Study 1 .....	26
	3.1 Method .....	26
	3.1.1 Participants.....	26
	3.1.2 Procedure .....	27
	3.2 Qualitative Analysis Method .....	28
	3.2.1 Initial Coding .....	28
	3.2.2 Card Sorting.....	30
	3.2.3 Hierarchical Clustering.....	30
	3.3 Results and Discussion .....	31
	3.3.1 Prepare patient for later possibilities .....	32
	3.3.2 Tailoring Information to the Audience .....	33
	3.3.3 Using Case Information.....	34
	3.3.4 Emotional Connection and Rapport.....	35
	3.3.5 Test and Logical Construct to Support Diagnosis .....	36
	3.4 Generic Timeline for Explanation .....	37
	3.5 Summary.....	39
4	Study 2 .....	40
	4.1 Method .....	40
	4.1.1 Participants.....	40
	4.1.2 Procedure .....	40
	4.2 Results.....	44

4.3	Summary.....	48
5	Study 3 .....	49
5.1	Method.....	49
5.1.1	Participants.....	49
5.1.2	Procedure .....	50
5.2	Results.....	54
5.3	Summary.....	59
6	General Discussion .....	60
6.1	Summary.....	60
6.2	Guidance for Designing Explanation for Medical AI.....	61
6.2.1	Tailoring explanation to suit different patients.....	61
6.2.2	Tailoring explanation for different points of diagnosis ..	63
6.2.3	Testing effectiveness of explanation at crisis points .....	66
6.2.4	Integrating rationales with visualizations or examples...	68
6.3	Transparency and the Right to Explanation.....	70
6.4	Conclusion .....	72

7	Reference List .....	74
A	Study 3 Scenario: Control.....	94
B	Study 3 Scenario: Rationales .....	99
C	Study 3 Scenario: Rationales + Visuals .....	104
D	Study 3 Scenario: Rationales + Examples .....	109

## List of figures

Figure 1: Hierarchical clustering for physician explanation strategies.....	32
Figure 2: Generic timeline for explanation in re-diagnosis scenario .....	38
Figure 3: Week 5 Probability Chart Explanation from MediBot.....	42
Figure 4: Results for explanation satisfaction scales .....	46
Figure 5: Results from the statement ratings .....	47
Figure 6: Sample visual explanation.....	51
Figure 7: Sample example-based explanation .....	52
Figure 8: Rating for explanation satisfaction scales .....	55
Figure 9: Mean rating for Overall Satisfaction. ....	59
Figure 10: Different timepoints of explanation for AI.....	65
Figure 11: Visual explanation with a graphical representation of relative outcomes.....	69
Figure 12: Example-based explanation.....	70

## List of tables

Table 1: Major components of medical diagnosis .....	2
Table 2: Demographics of physician participants.....	27
Table 3: Explanation statements from Physicians' Interviews.....	29
Table 4: Decisions by MediBot for six weeks .....	43
Table 5: Post-hoc Analysis for statement ratings .....	48
Table 6 Diagnosis Scenario .....	53
Table 7: Results from Type- III factorial ANOVA for explanation satisfaction scales .....	56
Table 8: Significant differences between conditions at each Set according to the Tukey test, any pairing not mentioned was not significantly different for that Set. ....	57

## Acknowledgements

I would like to express my deepest appreciation to my mentor and advisor, Dr. Shane T. Mueller. He has always been very supportive and patient with me. He inspired me to take challenges and guided me through this journey. Thank you for taking me under your wings, thank you for teaching me everything about graduate research. You are the best supervisor anyone could have asked for. I am very happy that I will be continuing my PhD under your supervision.

I would like to thank my committee members, Dr. Erich Petushek and Dr. Robert Hoffman. Thank you for providing your valuable feedbacks. My sincere gratitude to Dr. Elizabeth Veinott for supporting my work and providing thoughtful comments on my research all the time.

I thank my lab mates Anne, Katy, Pomm for always boosting my confidence. Your feedbacks helped me build up my work more thoroughly. A special thanks to Anne and Katy for helping me out with my social anxieties. You people have kept me sane through everything. I forgot to mention another lab mate of mine, that is Tauseef. Thank you, husband, for doing all the chores when I am busy with work. Thank you for bringing my lunch when I forget it at home. Love you for bearing with all my tantrums.

Finally, I would like to thank my family. I believe I am one step towards fulfilling my dad's wish. He is not with us anymore, but I am sure he is watching over me and he is very happy about the completion of my master's degree. Annoying my mom with well-crafted pranks (most of the credits for these pranks goes to Tauseef) every day is kind of my stress reliever. I thank my Mamoni (my aunt) and Wahid uncle for being my guardian

in this foreign land. Because of you, I never had to worry about anything around here. Special thanks to my beloved cousin, Goda. Love you for always having my back and sharing my enthusiasm about food and family gossips. I am very happy to have very supporting in-laws. Last but not the least, I thank God Almighty for helping me pass the ordeal.

## **Abstract**

AI systems are increasingly being fielded to support diagnoses and healthcare advice for patients. One promise of AI application is that they might serve as the first point of contact for patients, replacing routine tasks, and allowing health care professionals to focus on more challenging and critical aspects of healthcare. For AI systems to succeed, they must be designed based on a good understanding of how physicians explain diagnoses to patients, and how prospective patients understand and trust the systems providing the diagnosis, as well as the explanations they expect. In this thesis, I examine this problem across three studies. In the first study, I interviewed physicians to explore their explanation strategies in re-diagnosis scenarios. I identified five broad categories of explanation strategies and I developed a generic diagnostic timeline of explanations from the interviews. For the second study, I tested an AI diagnosis scenario and found that explanation helps improve patient satisfaction measures for re-diagnosis. Finally, in a third study I implemented different forms of explanation in a similar diagnosis scenario and found that visual and example-based explanation integrated with rationales had a significantly better impact on patient satisfaction and trust than no explanations, or with text-based rationales alone. Based on these studies and the review of the literature, I provide some design recommendations for the explanations offered for AI systems in the healthcare domain.



# 1 Introduction

## 1.1 Problem Statement

Artificial intelligence (AI) is starting to be deployed in the healthcare industry to a significant level and the market value of AI in this industry is predicted to reach \$6.6 billion by 2021<sup>1</sup>. As the first point of contact for patients, AI systems might produce diagnoses and predictions about patient's health as well as can perform routine tasks and provide non-emergency medical advice. Based on the gathered data and information fed to their algorithms, they might generate a diagnosis and explain it to patients through text or voice-based interfaces very fast and accurately. Google's DeepMind trained a neural network to accurately detect over 50 types of eye diseases by analyzing 3D scans and then recommended the treatment for patients<sup>2</sup>. Many promising healthcare chatbots have been launched in the past few years that may efficiently play the role of personal health assistant<sup>3</sup>.

There are two major components of medical diagnosis: 1) diagnosis prediction and 2) communication, and these two have been studied in the relevant research areas for AI as well as physicians as shown in Table 1. Diagnosis not only depends upon the prediction accuracy, but it also involves communication with patients. Physicians make their predictions considering clinical uncertainties and trade-offs regarding the possible

---

<sup>1</sup> <https://www.forbes.com/sites/forbestechcouncil/2020/01/15/how-ai-is-revolutionizing-health-care/#b31fd2c403c3>

<sup>2</sup> <https://www.theverge.com/2018/8/13/17670156/deepmind-ai-eye-disease-doctor-moorfields>

<sup>3</sup> <https://medicalfuturist.com/top-12-health-chatbots>

outcomes of their decisions. Communication has been an integral part of diagnosis since a “patient-centered” approach is crucial for high-quality care by physicians. Clear communication helps to elicit patient perspectives, allows physicians to understand them within their psychosocial or cultural contexts, and reach an understanding of their problems so that physicians may use patients’ knowledge and experience to guide the interaction and the treatment.

Table 1: Major components of medical diagnosis

	<b>Diagnosis Prediction</b>	<b>Communication</b>
Physicians’ Diagnosis	Differential Diagnosis	Patient-centered Communication
AI Diagnosis	Prediction Algorithm	XAI

Though diagnosis prediction and communication are two major components of medical decision-making, they are often investigated separately in the context of diagnostic AI systems. AI researchers are building algorithms that can produce very accurate predictions for diagnosis but most of the time they do not emphasize on the communication through which the AI needs to engage with the patients. Explainable AI (XAI) can link these two aspects of medical diagnosis. It does not involve improving the performance of AI systems, rather it focuses on communicating the diagnosis predictions to the patients effectively. In many cases, AI gives a diagnosis that turns out to be wrong and it needs to re-diagnose the patient. But this does not mean that the diagnostic AI system is making a mistake, probably it is making the best decision at that point with the information it has. The AI system and its performance accuracy could be improved to some extent, but it still

might have to make those assumptions because the decision might be optimal even when it is wrong. For these situations, AI needs to explain why it is making its decisions using effective communication strategies. Like the patient-centered communication approach taken by physicians, AI needs to communicate its decisions responding to patients' wants, needs, and preferences to develop a shared understanding of the problems and the goals of treatment. To address this problem, I have come up with ways to understand how to build a better XAI system that links both prediction and communication for medical diagnosis.

## **1.2 Summary**

Next, in Chapter 2, I will present the literature review where I examine how human and AI approach the diagnostic process, from both prediction and communication perspectives. Some of the human diagnosis literature is about decision biases, decision errors, and communication styles. I have looked at the AI systems of diagnosis, and almost all of them are focused on the prediction accuracy of diagnosis, not the communication aspects. I have also examined the XAI literature. In the medical diagnosis area, XAI has been doing things such as setting up chatbots for AI systems and verbalizing rule sets. It has mostly focused on ways of communicating about how deep knowledge networks work. In the broader XAI area, the focus has been on linking algorithms for decision and for communicating information. That shows there are gaps between the two components of diagnostics. The strategies and advice for patient-centered communication are neither going into the AI systems nor into the XAI research. There have been some developments within the XAI area recently, but they are very nascent (Lauritsen et al., 2019; Panigutti et al., 2020; Tjoa and Guan, 2019) and not informed by the study of medical experts or of

how physicians actually generate diagnoses. Most of the AI developers are focused on building the algorithms but they do not know how to make the communication better and help people understand the diagnosis predictions.

I address these issues in this thesis. My research aimed to investigate how medical diagnosis works for physicians and for AI, and the question of how to design XAI systems that will link prediction and communication with patients together. In Chapter 3, I present the results of an interview study with seven physicians to explore their explanation strategies for communicating the diagnoses and treatments to their patients. I also developed a generic diagnosis timeline of explanations from the interviews. Next in Chapter 4, I present the results of an initial experiment using a diagnosis scenario that showed that explanations are effective for patients' satisfaction, trust, and perception of accuracy. In Chapter 5, I present another experiment where I implemented different forms of explanation in a similar diagnosis scenario. Text-based rationales (the "why" of the diagnosis) integrated with either visual explanation or example-based explanation had a significantly positive impact on patient satisfaction and trust compared to no explanations, or to rationales alone. Finally, in Chapter 6, I discuss the results of the experiments and make some design recommendations for the explanations offered by XAI systems in the healthcare domain.

## **2 Review of Literature**

In this chapter, I review relevant literature on diagnosis predictions and communication by physicians and by AI diagnostic systems. This involves three main areas of investigation: decision making and communication by physicians, AI systems for medical diagnosis, and XAI systems for medical diagnosis. To understand how AI systems may explain diagnoses, we first need to understand the typical approaches physicians take or are taught to take. I will first provide a basic overview of research on medical diagnosis decision making.

### **2.1 Medical Diagnosis Decision Making in Diagnosticians**

Clinical reasoning refers to a set of cognitive processes applied for medical problem evaluation and management (Barrows and Tamblyn, 1980). It is based on logic, probability, and value theory (Ledley and Lusted, 1959). The concept of probability emerges in medical diagnosis since the diagnostic process often ends up giving a “most likely” diagnosis. Clinical decisions are made under conditions of uncertainty (Weinstein and Fineberg, 1980). This uncertainty arises from gaps and errors in clinical data, the ambiguity of clinical data, variations in interpretations, uncertainty about the relationship between different types clinical information, and uncertainties about the patient's present condition. Sometimes there are several cognitive biases associated with the decision making, which can lead to inaccurate assessment of probabilities (Round, 2001). Besides the biases, there are several influential approaches to understand the overall problem-solving process of diagnosis reasoning that I will discuss in a later section.

### **2.1.1 Medical Decision Biases**

Decision-making processes vary and are often confounded by various assumptions and biases that lead judgments and decisions to violate commonly accepted normative principles. Several biases play a vital role in medical diagnosis such as representativeness, availability, and anchoring biases (Kahneman and Tversky, 1972). Representative bias overestimates or underestimates the similarity between medical conditions and gives undue weight to a relatively smaller or bigger sample. It sometimes manifests a greater effect on judgments than knowledge of the probabilities. It is important to be aware of base rates of occurrence of a particular condition to avoid this bias (Klein, 2005). A study found that due to representativeness heuristics triage decisions in emergency rooms (ER) to identify severely injured patients are affected and modifying physicians' heuristics might reduce predictable under-triage and improve outcomes after trauma (Kulkarni et al., 2019).

Availability bias refers to putting too much weight to easily available information or recently encountered events. Normally it leads to the correct diagnosis since the conditions come in mind easily project more likely to be true. But it is misleading when less likely conditions occur. In a study where physicians are asked to judge the probability of bacteremia in patients, the assumed probability was significantly higher for the physicians who recently had encountered with patients suffering from bacteremia (Poses and Anthony, 1991). To avoid this bias, physicians should always consider all the diverse factors that may influence the diagnosis. Another one is confirmatory bias, which refers to the tendency of looking for and remembering all the pieces of information that might fit with the expected condition. Contradictory information is often ignored or overlooked due

to this. If the physicians are always prepared to think that there is an alternative hypothesis and the current hypothesis may potentially be dismissed at any time, that is the way this bias can be avoided. Anchoring bias occurs when the initial probability is set to be considered too extreme to be adjusted later with subsequent information (Hogarth, 1980).

A major source of error in a dynamic environment such as medical diagnosis is the failure to revise a situation assessment when new evidence, or comes in whether that new evidence is confirming or disconfirming (De Keyser and Woods, 1990). Inappropriate diagnosis might persist even in the face of the new cues. This is called “Fixation Error”, which often happens because of failure to consider alternatives. Initial diagnostic assessment tends to be accurate as it seems consistent with the partial information available at that time. But if physicians do not revise their assessment in response to new evidence, the erroneous diagnosis is preserved, and a wrong treatment is engaged

Illusory correlation and overconfidence about judgments can also cause errors in medical decision making. Physicians sometimes perceive two events as causally related, while it might only be a coincidence or even non-existent (J. G. Klein, 2005). Physicians overestimate their skill in many cases even if they know most of the medical decisions involve some level of uncertainty. Positive information is also emphasized more than negative information since physicians actively search for data to support their current hypothesis, not to rule out them (Elstein et al., 1978). It is suggested that this positive bias occurs due to the intention of reducing costly errors (Friedrich, 1993).

All these decision biases mentioned above are an important part of diagnosis predictions. Many of these biases are demonstrated in contrived settings so that it could be

proved that they existed. But, in many cases they are reasonable for decision-making in the real world. These biases may influence the predictions of physicians about diagnosis, but most of the time the predictions are rational based on the evidences they have at that moment.

### **2.1.2 Problem Solving Principles in Diagnostic Reasoning**

Medical problems can be characterized as ill-structured, in the sense that the initial states, the definite goal state, and the necessary constraints are unknown at the beginning of the problem-solving process (Simon, 1973). There are heuristics that physicians use for diagnosis predictions, and a lot of research has documented approaches, rules of thumbs, strategies including hypothesis generation, pattern recognition, differential diagnosis, Occam's razor, Hickam's dictum, evidence-based medicine, Bayes' theorem for avoiding diagnostic errors. The typical best practice is to follow a "Differential Diagnosis" approach. William Osler is credited with introducing the "discipline of differential diagnosis" (Maude, 2014). According do this approach, physicians should consider different possible causes of a particular symptom for systematically solving a clinical problem.

But initially, problem-solving for diagnostic reasoning was viewed as a process of testing hypotheses. Generating few hypotheses early in the diagnostic process and using them to guide the collection of data used to lead the solution to diagnostic problems (Barrows et al., 1982; Kassirer and Gorry, 1978). However, research shows that expert physicians develop hypotheses much faster than novice physicians and the quality of their hypotheses are also higher than the novices(Elstein and Schwarz, 2002).



Pattern recognition in medical conditions, specific instances, general prototypes have also been noted in the problem-solving strategies involved with medical decision-making. Physicians categorize a new case by considering its resemblance to the memories of instances previously seen. This model is supported by the fact that clinical diagnosis is strongly affected by context (Brooks et al., 1991). The prototype model emphasizes that physicians construct a mental model for abstract sets of semantic relations between clinical features and diagnostic categories (Lemieux and Bordage, 1992). Pattern recognition in medical diagnosis is also related with Recognition-primed decision (RPD) model (Klein, 1999) that emphasizes on the expertise of physicians often have the intuitive ability to realize the characteristics of the case do not fit any familiar category and pose a novel challenge.

“Occam's razor” is a problem-solving principle which suggests that “Entities should not be multiplied without necessity”. It means physicians should try to fit all of the patient’s symptoms and clinical findings into one diagnosis (Miller, 1998). But this might result in garden path errors in diagnostic reasoning. A physician’s diagnostic knowledge may be grouped into several sets in memory with competing alternatives (Johnson et al., 1981). Following this method might make him overlook other cues when a medical condition is suggested by one specific cue by activating only that specific part of the available knowledge (Johnson et al., 1988). Fixation error also may occur due to this problem-solving principle.

Since physicians do not always adhere to the principle of Occam's razor, there is a counterargument referred to “Hickam’s dictum”; a concept elaborated by an apocryphal

physician John Hickam<sup>4</sup>- "A man can have as many diseases as he damn well pleases." This explains that a patient may have multiple symptoms not just because of one medical condition, it may happen due to different conditions too. But following this method may cost a lot of effort, time and money for tests and examinations which might even be irrelevant (Jao, 2011)

Some of the research suggest Bayes' theorem as a theoretical approach to clinical reasoning as it provides a normative approach to the sequential processing of information implemented on medical decision making for updating hypotheses when new information is received (Round, 2001). The pretest probability is based on the general likelihood of the condition or physician's experience-based impression of the probability. Evidence-based medicine (EBM) seeks to establish a set of best practices for physicians by identifying the treatment of interest and researching the effectiveness of the treatment (Gray and Chambers, 1997). But there are also cognitive challenges involved in using EBM for diagnosis (Klein et al., 2016). Physicians trained in methods of EBM are more likely to use Bayes' theorem for diagnosis than untrained ones (Shaughnessy, 2007). However, it has been acknowledged that most of the physicians do not apply Bayesian strategies in daily practice, informal methods of opinions are still more applied in the complex real-time environments (Hammond et al., 1967; Wolf et al., 1985).

There has been a great deal of research into understanding diagnostic reasoning approaches. However, models of human diagnostic reasoning have generally been based on laboratory studies. Although, these studies have been influential in developing

---

<sup>4</sup> <https://radiopaedia.org/articles/hickams-dictum>

theoretical models of reasoning, they have not been adequately tested in ecologically valid contexts (Patel et al., 1994). In real world, physicians mostly rely on differential diagnosis method when it comes to diagnosis prediction. In contrast to the variety of research strategies for helping physicians make better diagnosis, the approach of AI-based diagnosis has been more focused on the actual data. It does not need to come up with strategies to prevent AI from having base-rate neglect or other biases. AI diagnostic systems operate based on their algorithms and make straight-forward decisions from the data they have.

## **2.2 Patient-centered Communication**

Communication is crucial in all steps of the healthcare process and it is especially important for both physicians and patients during diagnosis. Though medical educators and researchers have stressed the importance of communicating with patients and their families for a long time (Frank et al., 1996; Lansky, 1998; Lipkin et al., 1995; M. A. Stewart, 1995), The term “patient-centered communication” has emerged in more recent writing from the Institute of Medicine in 2001 (Medicine, 2001) defining patient-centered communication as *“a partnership among practitioners, patients, and their families ensures that decisions respect patients’ wants, needs, and preferences, and that patients have the education and support they need to make decisions and participate in their own care, as well as participate in quality improvement efforts”*. Although the definitions of patient-centered communication may vary (Epstein et al., 2005; Mead and Bower, 2000), the core concepts of patient-centered communication include “(1) eliciting and understanding patient perspectives (e.g., concerns, ideas, expectations, needs, feelings, and functioning), (2) understanding the patient within his or her unique psychosocial and cultural contexts, and

(3) reaching a shared understanding of patient problems and the treatments that are concordant with patient values” (Epstein and Street, 2007). Physicians’ explanations to the patients is a crucial part of the communication (Riccardi and Kurtz, 1983) and physicians who exhibit patient-centered communication behaviors gain a higher level of trust among patients (Fiscella et al., 2004).

Studies demonstrate that patient-centered communication is associated with improved healthcare outcomes, particularly in patients with chronic diseases (Naughton, 2018) and patients who feel understood by their physicians may be less anxious, have greater confidence in their physician’s abilities (Greenfield et al., 1985; Ong et al., 1995; Safran et al., 1998; Stewart et al., 2013). Several studies show that patient satisfaction is strongly associated with the communication behaviors that occur during the physician-patient interaction (Bertakis, 1977; Bredart et al., 2005; Buller and Buller, 1987; Korsch et al., 1968; Tallman et al., 2007; Wanzer et al., 2004). Apart from verbal communication, nonverbal communication behaviors such as eye contact, listening attentively also plays an important role in increasing patient satisfaction (Roter et al., 2006).

Patient-centered communication requires physicians to have the communication skills to elicit patients' preferences, to recognize and respond to their needs and their emotional concerns, to understand their personality, and to develop a shared understanding of their problem. As much as any technical skill, communication is a sophisticated procedure and it is required for the improvement of patient satisfaction.

## 2.3 AI Diagnostic Systems

AI-based medical decision support system assists medical decision making in multiple ways by dealing with patient data, providing aid to interpret diseases or to make decisions about clinical findings. They can predict patients' health and medical conditions based on their data and they can also make diagnoses based on the information they have.

Ever since the possibility of using machines for diagnosis appeared (Ledley and Lusted, 1959), prototypes of such systems started showing accuracy in many cases (Knill-Jones et al., 1973; Kruger et al., 1974; Warner et al., 1964). Medical decision support systems operate generally in two ways: 1) determine correct diagnosis for a patient, 2) make decisions about treatment, tests, and therapies for a patient (Shortliffe, 1987). Some systems provide aid for only one of these two, some others assist physicians in both ways. Decision support systems from the 1970s provide a strong foundation of work on such kind of expert systems (De Dombal et al., 1972; Pryor et al., 1975; Shortliffe, 1974).

A wide variety of techniques have been used in the design and implementation of medical decision support systems. Previously, machine-driven predictions used to depend on algorithms designed to extract specific features provided by expert medical professionals. Now, deep learning algorithms of AI allow for machines to receive data and self-develop complex functions to provide predictions about medical conditions (Fogel and Kvedar, 2018). Many researchers have been applying artificial intelligence for medical decision making. One of the advantages of using AI-based systems is the reliability and accuracy of them can be readily evaluated and gradually improved (London, 2019). Advanced forms of machine learning developed in the late 1990s which made

computational devices more accurate and robust. Thus, the clinical application of AI has been most rapid in image-intensive fields such as radiology, radiotherapy, pathology, dermatology, ophthalmology and image-guided surgery (Codella et al., 2018; Dominic et al., 2019; Esteva et al., 2017; Jha and Topol, 2016; Kaddioui et al., 2020; Kundu et al., 2017; Naylor, 2018; Piccini et al., 2020; Ting et al., 2019). Most of these AI systems have high accuracy in prediction and can make very fast diagnoses that make them capable of making the whole healthcare system faster than before. They also may ease pressure on physicians' workload so that they can focus on more challenging and critical aspects of healthcare.

Overall, AI systems are increasingly being deployed for diverse fields of medical diagnosis and they can perform with expert-level accuracy for predicting the diagnoses of patients. They have the potential to eventually speed up the diagnosis procedure and deliver cost-effective care at a large scale.

## **2.4 Self-diagnosis Systems**

Many self-diagnosis AI systems exist and are being used or deployed for use by various companies. Patients can directly use them for non-emergency medical advice or related healthcare queries. Almost all of these systems are dialogue-based and are focused on the prediction accuracy of diagnosis base on the information they gather from the users. There are two basic categories of self-diagnosis AI systems: 1) Symptom checker and 2) Diagnosis chatbots

### **2.4.1 Symptom Checkers**

Internet is increasingly being used for self-diagnosis as patients research their health concerns. Self-diagnosis usually starts with search engines like Google, Bing, or Yahoo (Fox and Duggan, 2013). But the search engines often provide unsupported confusing information or sometimes do not ask patients with serious symptoms to seek emergency care. Symptom checkers refer to more sophisticated programs based on computer algorithms that follow a systematic procedure to provide a potential diagnosis by asking a series of questions to understand the patient's condition. They require the users to input details about symptoms themselves and follow an algorithm to reach a decision based on the symptoms. The algorithms vary and may use branch logic, Bayesian inference, or other methods. Some symptom checkers only provide diagnostic advice, some offer triage advice and some of them can even identify the urgent condition and ask the users to seek emergency care (Saczynski et al., 2008). Symptom checkers may help patients save time, decrease anxiety, early recognition of a condition, and let them take control of their health. But there are some key concerns regarding the symptom checkers though. If they misdiagnose a patient and do not advise emergency care for critical patients, following their advice may prove fatal for those patients. There is an opposite scenario too. Asking every patient to seek care may cause the patients with minor conditions to visit hospitals, which will result in increasing the inappropriate burden on healthcare professionals. Most

of the symptom checkers can be accessed by websites such as WebMD<sup>5</sup>, Family Doctor<sup>6</sup>, Mayo Clinic<sup>7</sup>, Symcat<sup>8</sup>, Symptify<sup>9</sup>, and some of them also available as smartphone applications. These self-diagnosis applications provide a list of diagnoses, usually rank-ordered by most likelihood. Some of the apps have features to follow up on patients and ask them about their condition after certain intervals. A study using 45 standardized patient vignettes to evaluate the accuracy of 23 symptom checkers used in different regions of the world exhibited that correct diagnosis was listed first in 34% of evaluations (Semigran et al., 2015). The correct diagnosis was listed first in 51% of evaluations in the first three diagnoses and it was 58% within the first 20 differential diagnoses. The correct diagnosis was listed first for one of the most popular symptom checker WebMD 36% of the time and it increased to 62% within the first 20 diagnoses. Another study showed that the clinical diagnosis was matched with the first diagnosis on the lists of WebMD only 16% of the time for ENT symptoms (Farmer et al., 2011). Another study found mixed results. While 4 out of 21 patients with inflammatory arthritis were given the first diagnosis of inflammatory arthritis conditions such as psoriatic arthritis or rheumatoid arthritis, these diagnoses were listed in the first 5 diagnoses for 15 patients out of 21 (Powley et al., 2016). In a study to determine if symptom checkers give sufficient information for users to seek

---

<sup>5</sup> <http://symptoms.webmd.com>

<sup>6</sup> <http://familydoctor.org/familydoctor/en/health-tools/search-by-symptom.html>

<sup>7</sup> <https://www.mayoclinic.org/symptom-checker/select-symptom/itt-20009075>

<sup>8</sup> <http://www.symcat.com>

<sup>9</sup> <https://symptify.com/>



urgent care, it was found that 33% of them contained no critical symptom indicator (North et al., 2012).

Evidence suggests that the predictions of symptom checkers are inconsistent and not very accurate always and they do not focus on the communication aspects of diagnosis at all.

#### **2.4.2 Medical Diagnosis Conversational Agents/Chatbots**

Conversational agents or chatbots interact with patients through natural language and they can make diagnosis predictions based on patient information in the chat interface by asking relevant questions about symptoms and patient history.

The Healthcare sector is being challenged by increasing demand for healthcare services and an inadequate number of health professionals. Conversational agents can help to overcome this challenge by minimizing the need for human professionals to deal with the patients. In 1966 Joseph Weinbaum created the program ELIZA at the Massachusetts Institute of Technology (MIT) which was the first-ever chatbot, which was built to simulate a psychiatrist (Khan and Das, 2017). In the present day, conversational agents like Siri (Apple), Alexa (Amazon), Google assistant are ubiquitous. Alexa has skill sets to ask symptom questions about medical conditions from different sources including WebMD, Mayo Clinic. Recently the UK's National Health Service (NHS) has announced a partnership<sup>10</sup> with Alexa to offer health advice from the NHS website<sup>11</sup>. There are also

---

<sup>10</sup> <https://www.theverge.com/2019/7/10/20688654/amazon-alexa-health-advice-uk-nhs>

<sup>11</sup> <https://www.nhs.uk/>

specialized healthcare applications- Ada<sup>12</sup>, Your.MD<sup>13</sup>, Babylon<sup>14</sup>, which is only dedicated to healthcare services. Babylon has become the front-runner among those since it has been integrated with the UK's NHS. In a study, it was found that Babylon can identify the condition modeled by a clinical vignette with accuracy comparable to human doctors and gives safer advice than that of human physicians (Razzaki et al., 2018).

It was found in a study that conversational agents are inconsistent in medical diagnosis when they are asked simple questions about mental health, interpersonal violence, and physical health (Miner et al., 2016). They recognized and responded to some health concerns appropriately, but they responded incompletely or ineffectively to others.

Healthcare chatbots make natural language communication with the users or patients for predicting diagnoses based on their history and symptoms. AI researchers are building algorithms that make accurate diagnosis predictions, but they do not focus on patient-centered communication that plays an important role in improving patient satisfaction.

## **2.5 Trust in AI-based Diagnosis**

Users of AI diagnostic systems are basically of two types: 1) Healthcare professionals and 2) Patients or potential patients. The utilization of these systems depends

---

<sup>12</sup> <https://ada.com/>

<sup>13</sup> <https://www.your.md>

<sup>14</sup> <https://www.babylonhealth.com/>

on the trust and satisfaction of both types of users. Patient satisfaction and trust in these systems depend not only on the prediction accuracy but also on the communication aspects of these systems. How these systems communicate the diagnosis and relevant information to the patients is crucial for patient satisfaction and trust.

Understanding how users construct trust in medical decision support systems provides an insight into how these systems would be used, misused, or abused by physicians, healthcare providers, or patients (Parasuraman and Riley, 1997). Trust-building is a dynamic process, involving movement from initial trust-building to continuous trust (Wang and Siau, 2018). Despite the innovations in the field of AI-based medical decision systems, many healthcare professionals have been skeptical about the effectiveness of these systems (Friedman and Gustafson, 1977; Glantz, 1978). System developers concentrate on creating AI systems that can reach good and accurate decisions but that is only one part of the formula for system success. There are also logistical, mechanical, and psychological aspects of system implementation (Musen et al., 2014). Applications that are used to assist medical decision making are more acceptable to physicians than the ones used for automation of activities traditionally performed by physicians (Teach and Shortliffe, 1981). It indicates a distinction between assistance and replacement regarding the system use. Though it is suggested by researchers that AI-enabled healthcare systems would allow physicians to focus on their human abilities-building relationships, developing strategies to empathize and serving as a trusted advisor (Fogel and Kvedar, 2018).

Apart from physicians, it is very important to know how much patients would like to rely on these systems. According to the technology acceptance model (TAM), the

behavioral intention of using these systems will be influenced by perceived usefulness and perceived ease-of-use of patients and potential healthcare consumers (Davis, 1989; Davis et al., 1989). If users do not find it useful or easy to use, they might not have a positive attitude or intention to accept AI systems for diagnosis purposes. Promberger and Baron (2006) found that people are more likely to follow the recommendation of a physician than the recommendation of a computer. Participants of this study were given no information about the performance of either the physician or the computer. They assumed that the computer's performance was inferior to the physician's performance. The trust model of patients for AI in the healthcare domain depends on the accuracy of diagnosis, verification of that diagnosis, and doctors' involvement with the systems (Yanco et al., 2016). In a recent study, it was found that patients trusted human providers than AI providers and preferred having a human provider perform the service even if that meant there would be a greater risk of an inaccurate diagnosis or a surgical complication (Longoni et al., 2019). This study also suggested that if patients feel that the AI systems are providing them personalized care, it helps curb the resistance to AI-based diagnosis. When AI diagnostic systems were described as capable of tailoring care to each patient's unique characteristics participants were as likely to follow the recommendation of AI as that of a human provider. A survey of UK HealthWatch in 2018 was that two-thirds of people "would rather be treated by a human doctor who is more likely to make a mistake but offers compassion than by a robot doctor that rarely makes a mistake but lacks compassion"<sup>15</sup>.

In a study with three conversational assistants (Alexa, Siri, and Google Assistant) for user-initiated medical queries, it was found that there is some chance that in a use case they may act on the medical information provided by the conversation assistants (Bickmore et al., 2018). In another study, it was found that trust in conversational assistants not only depends on the performance of the application but also some other factors like privacy risks and effort expectations (Laumer et al., 2019). Nadarzynski et al. (2019) found in their study that there is a lack of familiarity and understanding of healthcare chatbots among people and people feel hesitant to use AI in healthcare due to their accuracy and security concerns. However, most participants were willing to use chatbots for minor health concerns and they were perceived as a convenient medical helpline that could facilitate the seeking of health information online. The lack of empathy in AI systems may also compromise patient or user engagement with them (Morris et al., 2018). Rather people have shown more willingness to use AI systems for general health information over specialist advice.

Though patient trust in AI-based systems in the healthcare domain has not been explored much until now, it is evident from the past research that patients may trust an AI-based system for assisting physicians or healthcare queries. There is still a greater need to improve patient trust for diagnosis or prediction of patient health. There is a lack of communication between AI diagnostic systems and the patients since these systems do not focus on a patient-centered communication approach. It is a barrier to improving patient satisfaction and trust in these systems as patient-centered communication improves patient satisfaction when they are treated by physicians.

## **2.6 Explainable Artificial Intelligence**

A key impediment to the use of AI-based systems is that they often lack transparency. The black-box nature of these systems allows powerful predictions, but it cannot be directly explained. Explainable AI (XAI) has the potential to make these back-box systems more transparent and improve user trust and satisfaction for these systems.

XAI aims to create a suite of techniques that produce more explainable models whilst maintaining high-performance levels (Adadi and Berrada, 2018). The importance of explanation in AI has been emphasized in literature over the past decades (Bellotti and Edwards, 2001; Dwork et al., 2012; Hoffman, Klein, et al., 2018; T. Miller, 2019; Otte, 2013; Wachter et al., 2017). Justification is one of the most effective ways of explanation that brings about changes in user attitudes toward the system and improves acceptance of AI-based system advice (Ye and Johnson, 1995). Enhancing the explanatory contents of AI systems can lead to easier use of such systems and ensure proper utilization of these systems for improving decision-making and problem-solving performance (Nakatsu, 2004). Several researchers have proposed methods on how to make explanations, or taxonomies of explanation, or descriptions of properties of explanation (Byrne, 1991; Felten, 2017; A. Kass and Leake, 1987; Kulesza et al., 2015; Swartout and Moore, 1993; Wallis and Shortliffe, 1981).

### **2.6.1 Explanation in Healthcare AI**

While the value of AI-based medical diagnosis systems is recognized mostly, there are impediments in the acceptance of such systems due to their “black box” nature. XAI has the potential to link the two important aspects of medical diagnosis: prediction and

communication. Good explanations from AI systems can help communicate the diagnosis predictions and relevant information with the right level of uncertainty leading to improved patient satisfaction and trust.

Underlying principles of the diagnostic systems may be understandable, but they lack an explicit declarative knowledge representation. In the context of AI systems, “understanding” refers to “functional understanding” of the system, not any low-level algorithmic understanding (Lipton, 2016). Even if a system is accurate in diagnosis, the neural network in the system may be biased stemming from the over or underrepresentation of classes of individuals. A model that is trained to predict the probability of death from pneumonia ranked asthma patients as having a lower probability than the general population (Caruana et al., 2015). It is because asthma patients receive aggressive medical care admitted to the intensive care unit (ICU). Without it, they would have a higher probability of death from pneumonia which would invalidate the model. Less accurate but transparent models with explanation might help to reflect patients’ medical needs in such cases. Often the best-performing methods are the least transparent, and the ones providing a clear explanation are less accurate (Bologna and Hayashi, 2017).

As explanations would help to facilitate transparency as well as trust in AI-based medical decision systems (Holzinger et al., 2017), systems are being developed for the medical domain with more interpretability (Che et al., 2016; Nemati et al., 2018; Zhang et al., 2017). These systems provide visual and semantic explanations for physicians or other healthcare professionals.

Overall, XAI is mostly verbalizing rule sets for AI diagnostics and focusing on ways of communicating deep knowledge networks. It has been linking algorithms for decision and communicating information, but it is not linking the prediction and communication aspects. XAI research for medical diagnosis is not informed by how physicians take a patient-centered approach for communicating the diagnoses.

## **2.7 Summary and Conclusion of the Literature Review**

While AI diagnostic systems may have sophisticated algorithms for predicting accurate diagnosis, they are not focusing on the communication aspects at all. There are gaps between the two important components of diagnostics: prediction and communication as the prediction accuracy of these systems are being investigated completely separately from communication. The strategies and advice for patient-centered communication are neither going into the AI systems nor in the XAI research. Most of the AI developers are building the algorithms but they do not know how to make the communication better and help people understand the diagnostic predictions. The key to the success of AI-based medical diagnostic systems is not only their accuracy and efficiency but also patient satisfaction, trust, and their perception of the accuracy of these systems. The current healthcare infrastructure is extremely burdened with a huge number of patients. Proper utilization of AI in diagnosis could reduce this burden immensely. But if patients do not trust these systems, they will not be willing to use them. There has been detailed research on patient-centered communication approach by physicians and how they communicate and explain their diagnosis to their patients, but none of them have gone into AI diagnostic systems. These systems have decision-making algorithms and not very effective



communication strategies (e.g. chatbots) but they are not focusing on effective communication strategies based on how physicians do it. In my thesis research, I take a first step in addressing this problem based on how physicians communicate their decisions to the patients. I investigate how medical diagnosis works for both physicians and AI and how to design XAI systems that will connect these two for both diagnosis predictions and communication with patients.

In the next chapter, I report an interview study with physicians to explore how they communicate and explain diagnostic decisions to their patients and how the patients react to the communication. As I will show, the results support the creation of a generic timeline of explanation by physicians at different diagnosis points.

### **3 Study 1**

To develop appropriate diagnostic explanations for AI systems, it is important to consider the patient-centered communication issues and the strategies physicians use to explain their diagnosis to their patients. To address this, I report an interview study with physicians where I identified explanation strategies during diagnosis. Based on these interviews, I also developed a generic diagnosis timeline that identifies points at which I observed explanatory reasoning strategies. Altogether, this study suggests explanation strategies, approaches, and methods that might be used by medical diagnostic AI systems to improve user trust and satisfaction with these systems.

#### **3.1 Method**

##### **3.1.1 Participants**

I conducted semi-structured interviews with 7 physicians. One of my interviewees was a practicing physician in Cameron, two were from Bangladesh and four were from the United States. The interviewees belong to different specialties since I did not choose any specific specialty physician to interview. Their experiences in diagnosis varied from two years to thirty-five years. The demographics of the physicians are described in Table 2. Participants were recruited through personal contacts. They gave oral consent before starting the interviews and agreed about the interviews being audio-recorded. It was mentioned that all answers and opinions were treated anonymously and strictly confidentially and that the recording of the interviews would be used only for study purposes.

Table 2: Demographics of physician participants

<b>Country</b>	<b>Specialty</b>	<b>Experience</b>
Cameroon	General Physician	1 year
Bangladesh	Cardiologist	2 years
Bangladesh	General Physician	3 years
United States	Pediatrician	19 years
United States	Family Medicine	35 years
United States	Sports Medicine	12 years
United States	Family Medicine	3 years

### 3.1.2 Procedure

I conducted the interviews based on Applied Cognitive Task Analysis (ACTA) techniques (Militello and Hutton, 1998) either via phone/internet video or in-person and lasted for 45-70 minutes. I followed a simple incident-based approach (Crandall et al., 2006) for these interviews. After initial background questions, we focused on 1-2 cases per physician that involved re-diagnosis and had them discuss how they communicated this to the patients. The goal of these interviews was to understand the methods physicians used to communicate with patients to explain their decisions, changes in diagnosis, and their reasoning strategies.

Physicians were asked if they went through the re-diagnosis scenario while diagnosing their patients and how they communicated this to the patients. They were then

asked to describe one specific case where they did re-diagnose their patients and how they explained it to the patients. The goal of the interviews was to understand how physicians explain their decisions and probabilities of different conditions to the patients. The interviewees were also asked how the patients responded to the explanations and how the physicians would deal with the situation where the patients were not happy with the explanations. The transcription of the interviews was carried out immediately afterward to ensure that the explanation patterns were recorded.

### **3.2 Qualitative Analysis Method**

The qualitative analysis of the interviews was completed in three steps of coding:

1) Initial coding to identify explanation statements, 2) Card sorting by five groups of students, and 3) Hierarchical clustering analysis.

#### **3.2.1 Initial Coding**

To help develop a more detailed understanding of explanation and formalize what were the criteria of explanation, we identified 4 main purposes of these explanation statements. These included:

- 1) What helped the patient/patient's family to understand the diagnosis and re-diagnosis
- 2) What the physicians did to gain patients' trust
- 3) What factors physicians considered before talking to the patient party
- 4) What the ways were to satisfy patient with the diagnosis/treatment protocol/procedure

It helped us more clearly define our criteria for identifying explanations. Once these criteria for coding explanation were developed, two independent raters examined each statement of the transcripts and coded two of the interviews identifying statements that were explanations and achieved inter-rater reliability of  $\kappa = .9$  and  $.88^{16}$ . Given the high agreement, a single rater coded the remaining interviews. I obtained 52 cases of explanation and mapped them into 24 categories of highly similar statements. The explanation elements are shown in Table 3.

Table 3: Explanation statements from Physicians' Interviews

No.	Explanation Elements
E1	Based on the patient's intellectual level
E2	Show empathy
E3	X-ray reports
E4	Metaphors
E5	Emotional condition
E6	Generalized information
E7	Walkthrough scenario
E8	Thorough history
E9	Relationship between current and pre-existing condition
E10	Counterfactuals
E11	Best interest of patient
E12	Analogies

---

<sup>16</sup> For this, we did not assess the reliability of the criteria for what constituted explanations.

E13	EMR
E14	Socio-economic or cultural status
E15	Provisional diagnosis but open for possibilities
E16	Keep patient in loop
E17	Not giving overwhelming information
E18	Not jumping off to conclusion
E19	Angiogram results
E20	CT scan reports
E21	Endoscopy report
E22	Initially not too negative information
E23	Non-electronic records
E24	Related examples

Only seven of the twenty-four categories had more than two examples. Two of the seven categories had more than four examples.

### 3.2.2 Card Sorting

To identify common themes among the explanations, I conducted a card-sorting exercise with students enrolled in a graduate course at Michigan Technological University. Students in five groups sorted the cards into 4-6 self-identified groupings.

### 3.2.3 Hierarchical Clustering

To identify a common clustering of themes, I used a hierarchical clustering approach on the card sorting data. For each pair of explanations, I counted how many times they appeared in the same theme across groups, using this as a measure of similarity. I then

applied the *agnes* function in the *cluster* library (Maechler et al., 2013) of the R statistical computing language to compute a clustering hierarchy.

### **3.3 Results and Discussion**

Five hierarchical themes emerged from the clustering analysis. The clustering analysis is shown in Figure 1. I identified names for these main themes but include all 24 base codes in Figure 1. The main themes include: 1) Prepare the patient for later possibilities; 2) Tailor information to the audience; 3) Using case information to make a logical argument, 4) Using visualizations and testing results to support the diagnosis; and 5) Communication to build emotional connection and rapport.

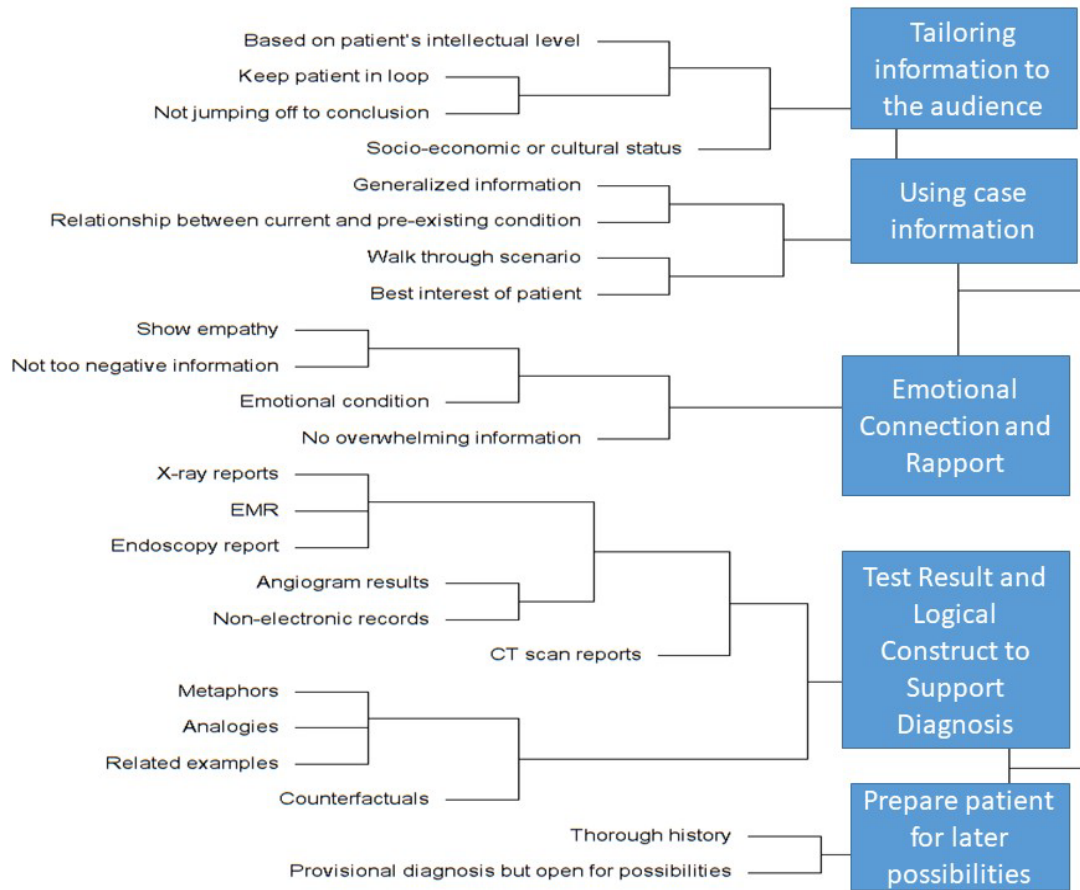


Figure 1: Hierarchical clustering for physician explanation strategies

### 3.3.1 Prepare patient for later possibilities

Initially, many physicians provided a provisional diagnosis based on the symptoms and history. This not only included the most-likely condition but also often included other possibilities. Thus, this kind of explanation prepares the patient to accept and understand possible future re-diagnoses. The physicians explain to the patients that there might be other possibilities, but at that moment they are thinking the provisional diagnosis is the most-likely condition:



*“I told the patient that there is a possibility that there could be other things going on or overlapping with the primary diagnosis and so in a few weeks if you don’t feel you are making progress, I want to see you back, re-examine you, take your history again. I kind of prepare the patient ahead of time that we might need to re-visit the diagnosis before actually doing it.”*

The statements reveal that physicians communicated their provisional diagnosis with their patients by not only giving information about the provisional diagnosis but also conveying that there were possibilities of other conditions that might need to be explored in the future.

### **3.3.2 Tailoring Information to the Audience**

Physicians reported that they often tailored their explanation to the individual, considering things as socio-economic or cultural status, the intellectual level of their patients, and their current emotional state. Thus, explanations were contextual and highly dependent on the patient and their ability to understand the information.

Interviewees reported customizing their explanations based on intellectual levels of the patients:

*“A lot of times based on the intellectual levels of patients I change my explanation. For average persons, the big thing is they need to find you relatable. They do not want to talk too above their level.”*

The socio-economic or cultural status of the patients was also an important basis of what explanations physicians give to the patients and their families:

*“You need to always understand where they are coming from. You definitely change explanation depending upon their socio-economic status.”*

### **3.3.3 Using Case Information**

Physicians often build their explanations based on the available information about the cases. They walk patients through the case scenario and make them understand how they are making diagnosis decisions. They often try to generalize the information at the beginning and go into greater detail later when they think it is reasonable:

*“Usually I keep it general. Unless during the examination I feel like there is a relatively high chance there is another thing going on.”*

Interviewees also stated that when they have to re-diagnose their patients, they explain this to the patient and their family by walking them through the entire scenario:

*“I told the mother your child came in with these symptoms. Now you look at him and think he is getting better. But you need to know he is not better and look at the reports.”*

Interviewees also mentioned explaining the medical conditions and the re-diagnosis based on the relationship between pre-existing and current conditions of the patients, since pre-existing conditions can either mask or be the true cause of the current condition. In both cases, often the provisional diagnosis did not improve the patient's condition, so physicians needed to clarify the reasoning behind re-diagnosis by explaining the relationship between pre-existing and current conditions:

*“I told the daughter of the patient her mother might not have a severe cardiac attack. Whatever she had gone through because of the dialysis that her body was not supporting.”*

### **3.3.4 Emotional Connection and Rapport**

Physicians consider the emotional aspects of communication to the patient and their families. These were not always about providing explanations or information but involved empathic strategies to ensure their patient knew the physician listened and cared:

*“In a situation like that the person has to really believe that you believe them, you care for them. You have to honestly feel that way. They can sense that. I try to have empathy for the person.”*

Physicians suggested that patients might initially be anxious and not in a condition to understand the reasoning and explanation, and their explanations at this point differ from later explanations. How they explain things initially is usually different from how they do it later when the patient is stable:

*“It was a calmer environment than the first night and now we could have a calmer conversation. And I can explain to her what was going on my mind.”*

Also, several physicians suggested at the beginning of the consultation, they do not want the patients to think about negative possibilities too much:

*“I don’t want to put on their mind something very negative. I say if it does not work, we will rule out other problems but I am not going to give up on you.”*

### 3.3.5 Test Results and Logical Construct to Support Diagnosis

The interviewees stated that they often use various tools to support the facts or the data of the diagnosis. Tools physicians use can be divided into two categories: 1) Reports and Medical Records, and 2) Logical Constructs. The interviewees mentioned about using X-ray, CT scan, endoscopy, angiogram reports as visual aids for the patients and their families:

“The patient’s family was brought to the monitoring room. We showed the video and recordings of the angiogram and explained to them what the patient’s problem was.”

The interviewees also mentioned using the electronic and non-electronic medical records to explain patients’ condition to them and their families:

*“I always pull out the EMR and show them what I am thinking and what are the results.”*

Physicians often use logical constructs as tools for explaining their decisions and the diagnosis to the patients. They use related examples, analogies, or metaphors so that the patients or their families can understand the situation better:

*“I said that the cells in your body are like the police. The police in his body increased, you have more policemen, more defense in his body but we could not really tell what they were defending against. But this test tells us exactly what those cells were defending against.”*

Interviewees also stated that they use counterfactuals when they have to deal with re-diagnosis and communicate that with the patients:

*“I would tell my patient why I overlooked some symptoms and why it went wrong. I would be telling him why I went for IBS, not for Celiac disease.”*

### **3.4 Generic Timeline for Explanation**

To help identify the typical points at which explanations emerge during diagnosis, I developed a generic timeline of explanation in re-diagnosis scenarios. This is presented in Figure 2. based on the interviews. The timeline shows where the explanations fit in each phase of diagnosis and communicating with the patients. The explanations in the figure are color coded according to the five themes emerged from the interviews.

During the initial phase of diagnosis, physicians tended convey the most-likely diagnosis to their patients, mention other possibilities, attempt to give general information, gain emotional rapport, and avoid discussion of negative possibilities.

When the physicians ordered tests to confirm a provisional diagnosis, they often assured patients that it would be in patients’ best interest. During the follow-up phase, physicians typically used these testing results to explain the condition to the patient.

During the follow-up phase with the reports, physicians often use x-ray, endoscopy, angiogram, CT scan reports as visual aid to explain the condition to the patients. Using electronic or non-electronic medical records are also another tool used by physicians in this phase. They also think about patients’ intellectual level while they are building these explanations for them.

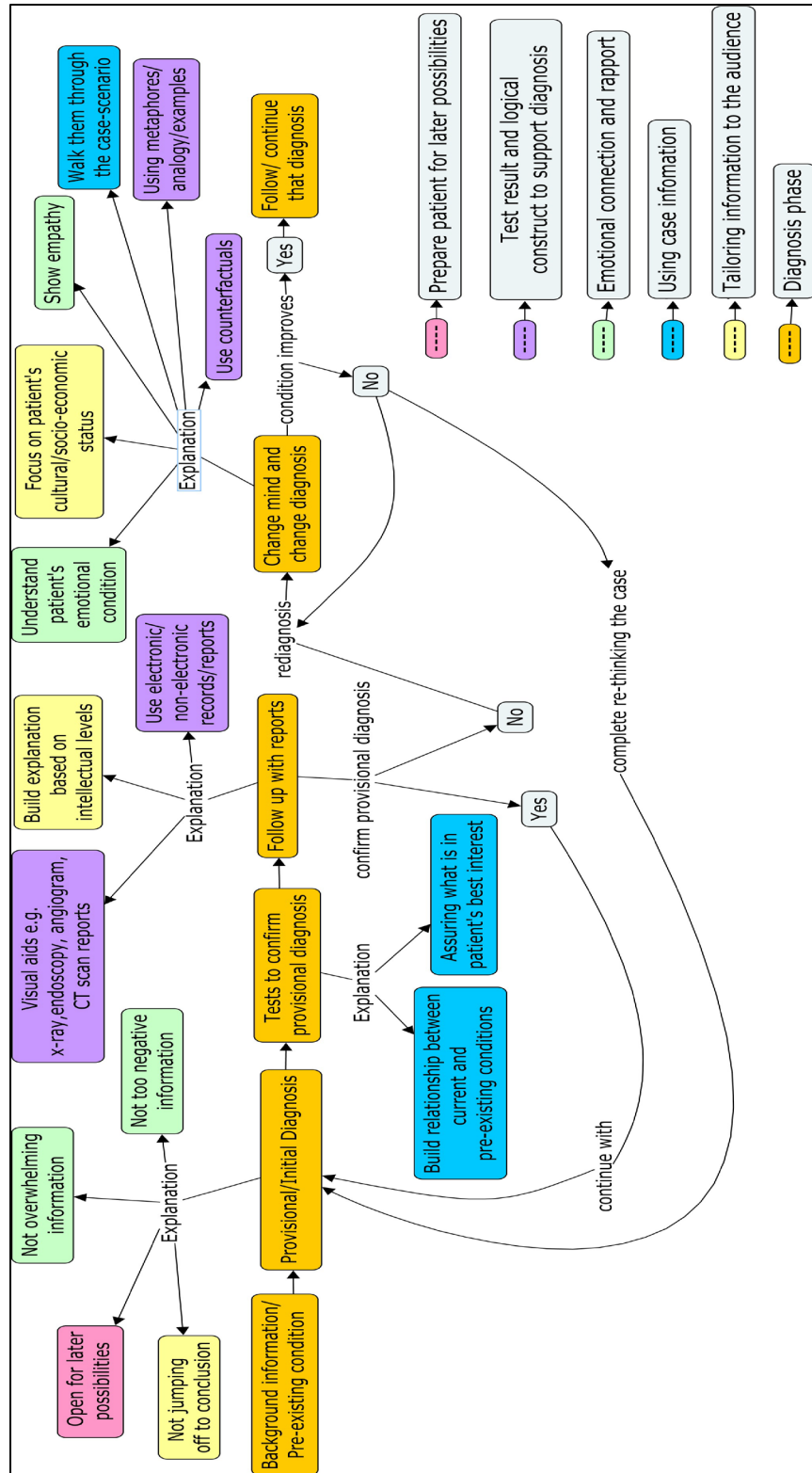


Figure 2: Generic timeline for explanation in re-diagnosis scenario

When the diagnosis did not work and the physicians needed to change the diagnosis, they often focused on patients' emotional, cultural, or socio-economic status since often the reaction of the patients depends on these factors. They reported trying to be compassionate and empathetic to their patients and use counterfactuals to make the patients understand what would have happened had they taken another course of action. This phase may continue until the conditions of the patient improved, or the physician decided to reassess the symptoms from ground zero.

### **3.5 Summary**

The explanation strategies and methods I identified in this interview study show that building good explanations requires the physicians to know the emotional condition and level of understanding of the patients. Expert physicians often apply these approaches, but current AI approaches ignore these communication aspects. They may need to focus on these for improving patient satisfaction and trust. The study I present in the next chapter will evaluate the effectiveness of explanations for AI diagnostic systems in a diagnosis scenario.

## **4 Study 2**

XAI has the potential to link diagnosis predictions and communication for AI diagnosis systems through appropriate explanations. I hypothesize that explanations will induce greater satisfaction, trust, understanding, and perceptions of accuracy. To investigate this, I will test a diagnosis scenario in this chapter. In this scenario, a simulated AI system gives a most likely but incorrect diagnosis, but later it changes the diagnosis to the correct disease. In comparison to a no-explanation condition, justifications that visualized disease likelihoods improve overall satisfaction and trust, both before and after re-diagnosis. In contrast, pre-test global explanations using example diagnoses do not show the same benefits. Results suggest that explanations can be effective at improving patient understanding of diagnoses, but not all explanations are equal.

### **4.1 Method**

#### **4.1.1 Participants**

Eighty undergraduate students at Michigan Technological University took part in the study in exchange for partial course credit.

#### **4.1.2 Procedure**

I created a diagnosis scenario in which a simulated AI system gives a most likely but incorrect diagnosis, but later changes the diagnosis to the correct disease. I created a simulated AI system called MediBot.ai. The scenario was based on gastrointestinal disorders. The participants played the role of patients in the scenario, instructed to say they



were suffering from abdominal pain, cramps, diarrhea, fatigue, and joint pain. MediBot concluded that the patient was suffering from Irritable Bowel Syndrome (IBS). It advised patients to follow a specific diet chart and come back for follow up next week.

In the scenario, the patient felt better after one week, but the condition started getting worse after that. When the patient did not feel good even after three consecutive weeks, MediBot realized that the patient might not be suffering from the “most likely” medical condition IBS that it had diagnosed. MediBot changed its diagnosis and looked out for the less likely conditions. It asked the patient to go through a few pathological tests, and it turned out that the patient was suffering from Celiac disease which occurs due to gluten allergy. Participants had to communicate with MediBot through six simulated weeks but the study took around 20 minutes to complete. To maintain certain intervals between the simulated weeks, they were given crosswords to solve during the intervals. After they solved one crossword, they were asked to start following up with MediBot and play their role as patients again.

Participants were divided into three groups. The control group received no explanation of why MediBot was making any decision in any week. Another group received a pre-diagnosis global explanation about how MediBot does diagnosis in general. It included two examples: 1) A success case of first diagnosis 2) A failure case for the first diagnosis, but eventually successful second diagnosis. Another group received local explanations about each decision and prediction of MediBot. Local explanations explained why the MediBot made a particular decision for a particular case. For this group, MediBot showed a probability chart of the likely conditions of the patient in each week that is shown

in Figure 3. This a visual explanation for the experimental group. Although we call this visualization, we were basically representing probability or likelihood of different outcomes but doing it visually. It also had descriptive text explanation about why it was making a particular decision to help patients see the relationship between the symptoms and its decision. For the first week, IBS seemed to be the “most likely” condition for the patient; Celiac, Crohn’s, bowel infection, and arthritis were less likely conditions. As the patient’s condition did not improve over time, the probability of IBS went down and the probability of Celiac increased as the patient suffered from joint pain, as shown in Figure 3.

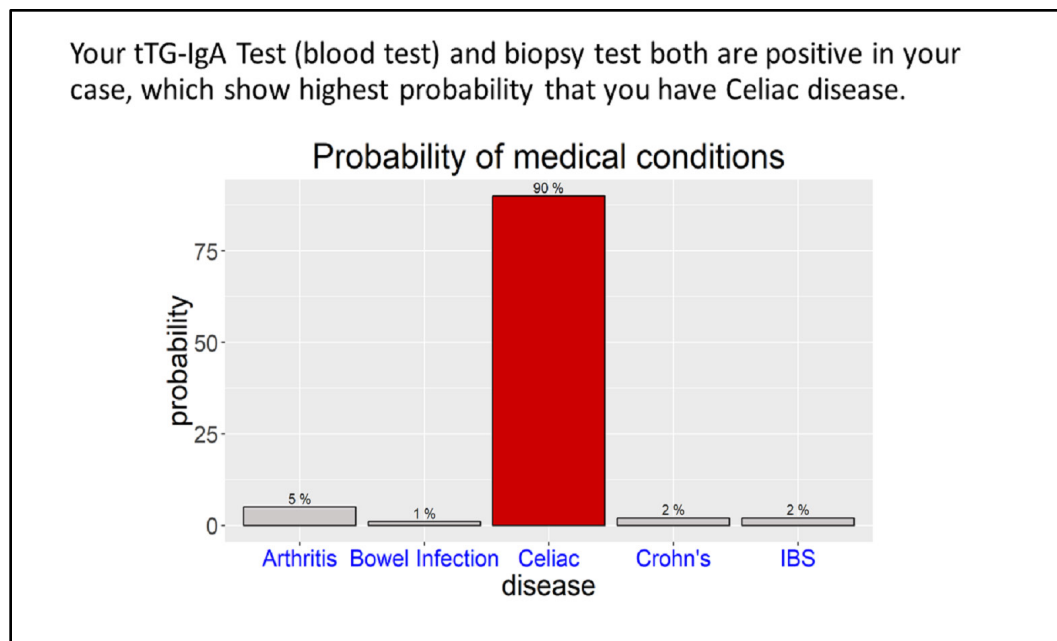


Figure 3: Week 5 Probability Chart Explanation from MediBot

Table 4 lists the entire scenario for a patient across six weeks of diagnosis. After each simulated week, participants were asked about their satisfaction, trust, perception of

accuracy, sufficiency, usefulness, and completeness for the explanations received from MediBot. These are some of the key attributes of explanations identified in the literature and are referred to as “Explanation Satisfaction Scale” attributes (Hoffman, Mueller, et al., 2018). Participants also rated their agreement about the following 4 statements on a 5-point Likert scale after completing their role as patients and finishing the study. The statements were:

- i) I do not understand what MediBot is doing
- ii) I think MediBot is behaving erratically
- iii) I understand MediBot is following a systematic elimination method
- iv) I understand why MediBot changed its mind between week 4 and week 5

Participants answered a question each simulated week as part of the knowledge test. It was asked to understand if they were paying attention to the scenario of the experiment. The questions were based on the medical condition and treatment advice given in that specific week; one question was asked in each week at the end of the treatment advice.

Table 4: Decisions by MediBot for six weeks

Week No.	Decision of MediBot
1	The patient consults with MediBot for diarrhea, abdominal pain, cramps, fatigue, and joint pain. MediBot suspects it is IBS, it advises to follow a strict diet chart and come back next week.

2	Patient feels better after following the diet chart, MediBot advises to continue the diet chart and come back next week.
3	Patient again starts feeling worse suffering from diarrhea, bloating, fatigue, and joint pain. MediBot thinks that the IBS diet chart takes some time to adjust with the body and joint pain might occur due to arthritis. It advises to continue the diet chart and come back next week.
4	Patient's condition is not improving. MediBot changes its mind and the recurrence of the symptoms lead it to consider that patient might be suffering from some other condition. It asks patient to go through some tests and come back with the report next week.
5	Patient comes back with the report and it is found that patient is suffering from Celiac disease, which occurs due to gluten allergy. Celiac has similar symptoms like IBS but shows some additional symptoms like joint pain and fatigue. MediBot asks the patient to consult a nutritionist to follow a gluten-free diet and come back for a follow up next week.
6	Patient feels great due to following a gluten-free diet.

## 4.2 Results

Both the control and the global explanation groups expressed less satisfaction, trust, perception of accuracy, sufficiency, usefulness, and completeness than the local explanation group, as shown in Figure 4. It was more evident in the week-3 when the patient condition was getting worse but MediBot was still considering the possibility of IBS more than other conditions. But all three groups showed almost the same perception of how accurate the system was at that time. The control group and global explanation groups did not differ from. All the satisfaction scores dropped at the end of the diagnosis in the last week, which means week 6. A two-sample Welch t-test for comparing local explanations with the control group and local with a global explanation group confirmed that satisfaction, sufficiency, completeness, and usefulness are significantly better for the

local explanation than the control group at week 3 ( $p < .05$ ). Completeness judgments were also significantly higher for local explanation than the control group at week 4 and 5 ( $p < 0.05$ ). According to the t-test, the local explanation also led to greater satisfaction, sufficiency, completeness, and usefulness than a global explanation at week 3 ( $p < .05$ ). Though the mean scores for trust and accuracy were lower for the control and global explanation group at week 3, they were not statistically significantly different ( $p > .05$ ). However, perceived accuracy was significantly higher for the local explanation in Week 1,2,4 and 5 than the control group. It was significantly better for the local explanation in Week 1 and 4 than the global explanation group. In Weeks 1,4 and 5, trust was also rated higher for the local explanation group than the control group.

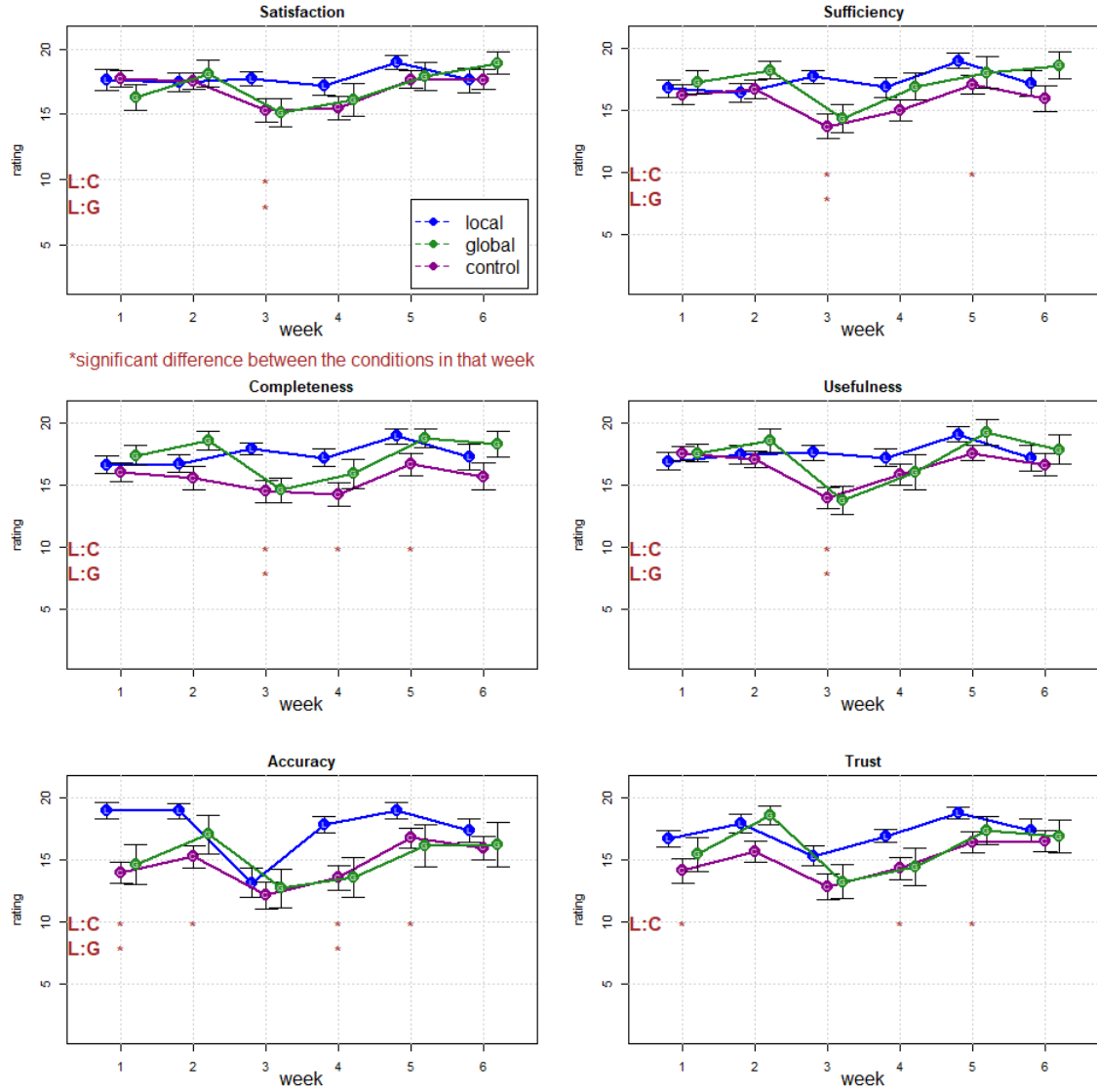


Figure 4: Results for explanation satisfaction scales

The results of the ratings of participants for the four statements at the end of the study are shown in Figure 5. I examined the results with one-way ANOVA. All the statements were significantly different ( $p < 0.05$ ) for the three groups except “I think MediBot is behaving erratically” ( $F(2,77) = 2.3$ ,  $p = 0.11$ ).

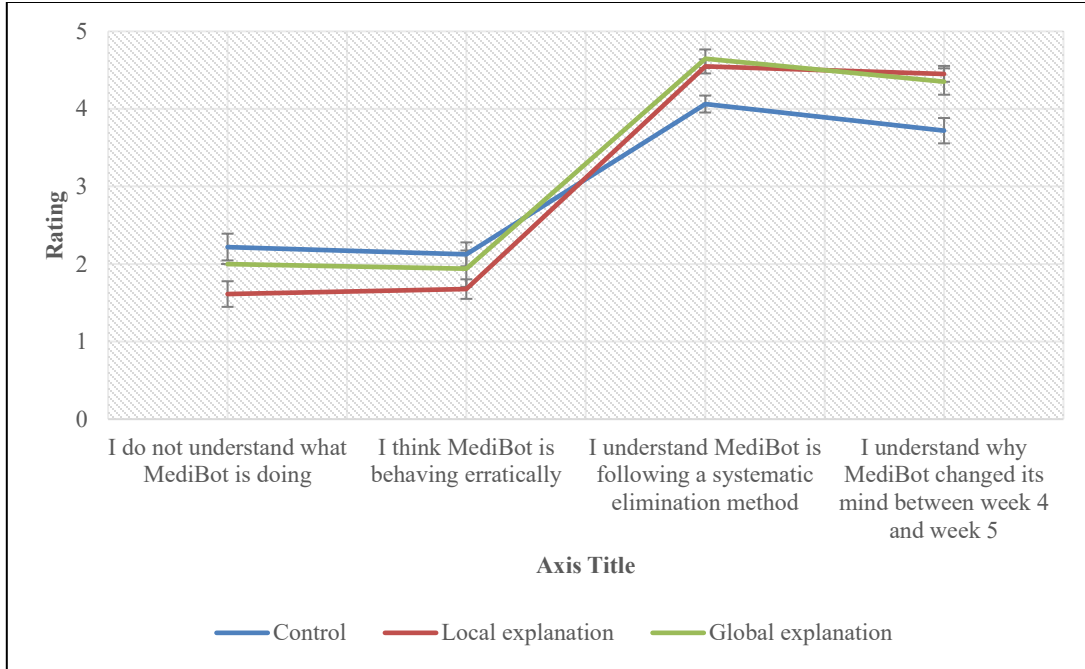


Figure 5: Results from the statement ratings

Post-hoc Tukey test on the three groups analyzed pairwise differences for each statement and the results are presented in Table 5. The global explanation was rated as significantly better than the local explanation for the statements: 1) I do not understand what MediBot is doing and 2) I understand MediBot is following a systematic elimination method. Thus, even if the pre-diagnosis global explanation was not helping much for improving satisfaction, it provided an overall understanding of the general method of diagnosis by the AI system. But it did not help much in the critical situation where the AI realized what was wrong and changed its diagnosis. The global explanation was better than the control group in this situation but local explanations were the most helpful at this point.

Table 5: Post-hoc Analysis for statement ratings

	Control- Local Ex	Control-Global Ex	Local-Global
I do not understand what MediBot is doing	$p < 0.05$	$p < 0.05$	$p < 0.05$
I think MediBot is behaving erratically	$p = 0.09$	$p = 0.74$	$p = 0.55$
I understand MediBot is following a systematic elimination method	$p < 0.05$	$p < 0.05$	$p < 0.05$
I understand why MediBot changed its mind between week 4 and week 5	$p < 0.05$	$p < 0.05$	$p = 0.9$

### 4.3 Summary

The study shows that local justifications for each diagnosis expressed as probability charts provide better understanding and satisfaction about the diagnosis for a particular case or a patient. Pre-diagnosis global explanation did not help to raise satisfaction measures, but it may help patients understand the general method for diagnosis as much as local justifications. In the next chapter, I present a third study using a similar diagnosis scenario to investigate whether different forms of explanation in an AI diagnostic system affect patient satisfaction, trust, and perception of accuracy differently.



## **5 Study 3**

Explanations in AI diagnostic systems may come in different forms such as text-based rationales, visualizations, examples, or contrasts. The goal of this third study was to investigate whether different forms of explanation in an AI diagnostic system affect patient satisfaction, trust, and perception of accuracy. I implemented three forms of explanation: rationales, visuals + rationales, and examples + rationales in a diagnosis scenario similar to the one in Study 2. In this scenario, a simulated AI system gave a most likely but incorrect diagnosis, but later it changed the diagnosis to the correct disease. I hypothesize that text-based rationales integrated with visuals or examples will induce greater satisfaction, trust, perception of accuracy, completeness, sufficiency, and usefulness of the explanations than only text-based rationales and the control group. Results show that the overall satisfaction for visual + rationales, example +rationales groups are not different from each other but they both had a better effect than rationales only and the control group. In most cases, rationales were no better than the control group.

### **5.1 Method**

#### **5.1.1 Participants**

One hundred and thirteen undergraduate students at Michigan Technological University took part in the study in exchange for partial course credit.

### 5.1.2 Procedure

The study was conducted online, and it took 15-20 minutes to complete. Participants gave their consent online before taking part in the study. They played the role of a patient suffering from a gastrointestinal disorder interacting with the simulated AI system “MediBot.AI”. I modified a few things in the diagnosis scenario used in Study 2. After interviewing physicians in Study 1, I found that IBS is considered a diagnosis of exclusion, and patients are often diagnosed with IBS by the process of ruling out some other medical condition. The scenario for Study 3 started with a patient suffering from abdominal pain, cramps, bloating, diarrhea, fatigue, and joint pain who has recently been exposed to a natural water source and had no family history of gastrointestinal diseases. MediBot predicted that the patient was suffering from Giardia and asked for tests for confirmation. But the test came negative, and then MediBot predicted that it might be IBS and asked to follow the IBS diet chart. The patient’s condition was inconsistent for a few weeks even after following the diet chart, then eventually MediBot figured out that the patient was actually suffering from Celiac disease and confirmed it from tests.

Participants were divided into four groups and received different forms of explanation during the interaction with the AI for diagnosis:

- 1) One group received only rationales as explanation
- 2) One group received visual + rationales explanation
- 3) One group received example-based + rationales explanation
- 4) Control group received no explanation.

Rationales are the narrative justifications of how MediBot made decisions. Visual explanations include figures of likelihood of each suspected disease based on features MediBot used to make decisions as shown in Figure 6. These visualizations were akin to the LIME algorithm(Ribeiro et al., 2016). The features were of patient's symptoms and medical history. The rationales group all the justifications included in the visual explanation, only the figures were removed from the explanations.

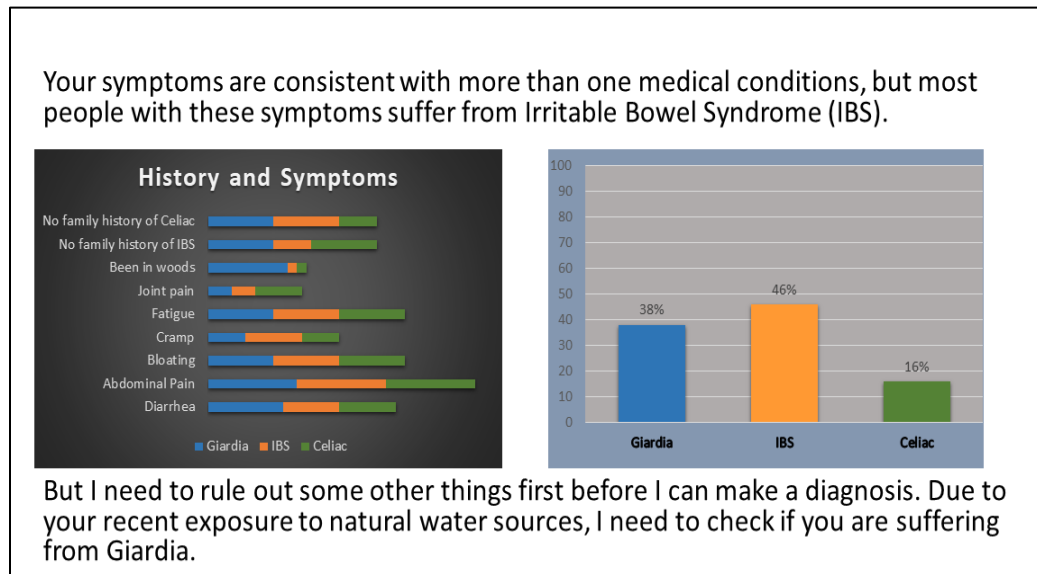


Figure 6: Sample visual explanation

Example-based explanation included examples of similar cases diagnosed by MediBot in the past, as illustrated in Figure 7 . It had one exception, a contrast explanation in week-5 to explain why it did not consider Celiac disease the most-likely condition at the beginning of the consultation.

Your symptoms are consistent with more than one medical conditions, but most people with these symptoms suffer from Irritable Bowel Syndrome (IBS). But I need to rule out some other things first before I can make a diagnosis. Due to your recent exposure to natural water sources, I need to check if you are suffering from Giardia.

Let's look at a similar case C30117:

Case C30117	Your Case
Patient had diarrhea, abdominal pain, bloating, cramps, fatigue	You have diarrhea, abdominal pain, bloating, cramps, fatigue
Patient had been exposed to natural water sources	You have been in woods and exposed to natural water sources
Patient went through Giardia antigen test and Giardia lamblia was detected in the test.	Please go through Giardia antigen test and come back with the result

Figure 7: Sample example-based explanation

Participants interacted with MediBot for six simulated weeks and received an explanation about its prediction and diagnosis each week. After each simulated week, participants were asked to rate their satisfaction, trust, perception of accuracy, sufficiency, usefulness, and completeness for the explanations, each on a 7-point Likert scale:

- 1) I am satisfied with the explanation of my diagnosis (satisfaction)
- 2) The explanation of my diagnosis had sufficient detail (sufficiency)
- 3) The explanation for my diagnosis was complete (completeness)
- 4) The explanation for my diagnosis was useful (usefulness)
- 5) The system let me know how accurate the diagnosis was (accuracy)
- 6) The explanation let me judge whether I should trust the diagnosis or not (trust)

The summary of the diagnosis scenario is presented in Table 6:

Table 6 Diagnosis Scenario

Week No.	Predictions and diagnoses
1	Patient consulted with MediBot for diarrhea, abdominal pain, bloating, cramps, fatigue, and joint pain. MediBot suspected it's Giardia and asked to do test for confirmation
2	Giardia test was negative, MediBot changed the diagnosis, suspected IBS, asked to follow IBS diet chart and asked to come back after two weeks
3	Patient started feeling better after following the diet chart, MediBot advised to continue the diet chart and come back next week.
4	Patient's condition became worse again with diarrhea, abdominal pain, bloating, fatigue and joint pain. MediBot suspected it was not IBS, the recurrence of the symptoms led it to consider that patient might be suffering from some other condition. It asked patient to cut down gluten from diet chart and come back next week.
5	Gluten-free diet improved patient's condition. MediBot suspected patient was suffering from Celiac disease and asked to do tests for confirmation.
6	Tests confirm that patient had Celiac disease. MediBot advised the patient to consult a nutritionist to follow a gluten free diet.

The entire scenario for the Control Group is described in Appendix A in detail, the scenario for the Rationales Group is described in Appendix B, and the scenario for the Visuals +Rationales Group and Visuals + Examples Group are described in Appendix C and Appendix D, respectively.

## 5.2 Results

In order to simplify the results, I organized the ratings for all six weeks for all six attributes into three sets: Week 1 and 2 averaged into Set 1, Weeks 3 and 4 averaged into Set 2 and Weeks 5 and 6 averaged into Set 3. The mean rating for all six attributes (satisfaction, trust, perception of accuracy, sufficiency, usefulness, and completeness) over the Sets for the four Explanation Type Groups of explanations are shown in Figure 8. Overall, the control group was worse than any other group (rationales, rationales + visuals, rationales + examples) in week 1-2 (set 1) overall six dimensions. During this time period, the AI system was ruling out one condition and predicting another (IBS). But when the system started predicting wrong and was not helping in the improvement of patient condition at week 3-4 (set 2), only rationales did not help either. For all these attributes, both control and rationale only explanations are worse than rationales + visuals and rationales + examples. But at the end when the system comes to a resolution about Celiac disease; all the explanation groups converge over satisfaction, sufficiency and usefulness. But control group was still worse than other groups for completeness, trust and perception of accuracy at week 5-6 (set 3).

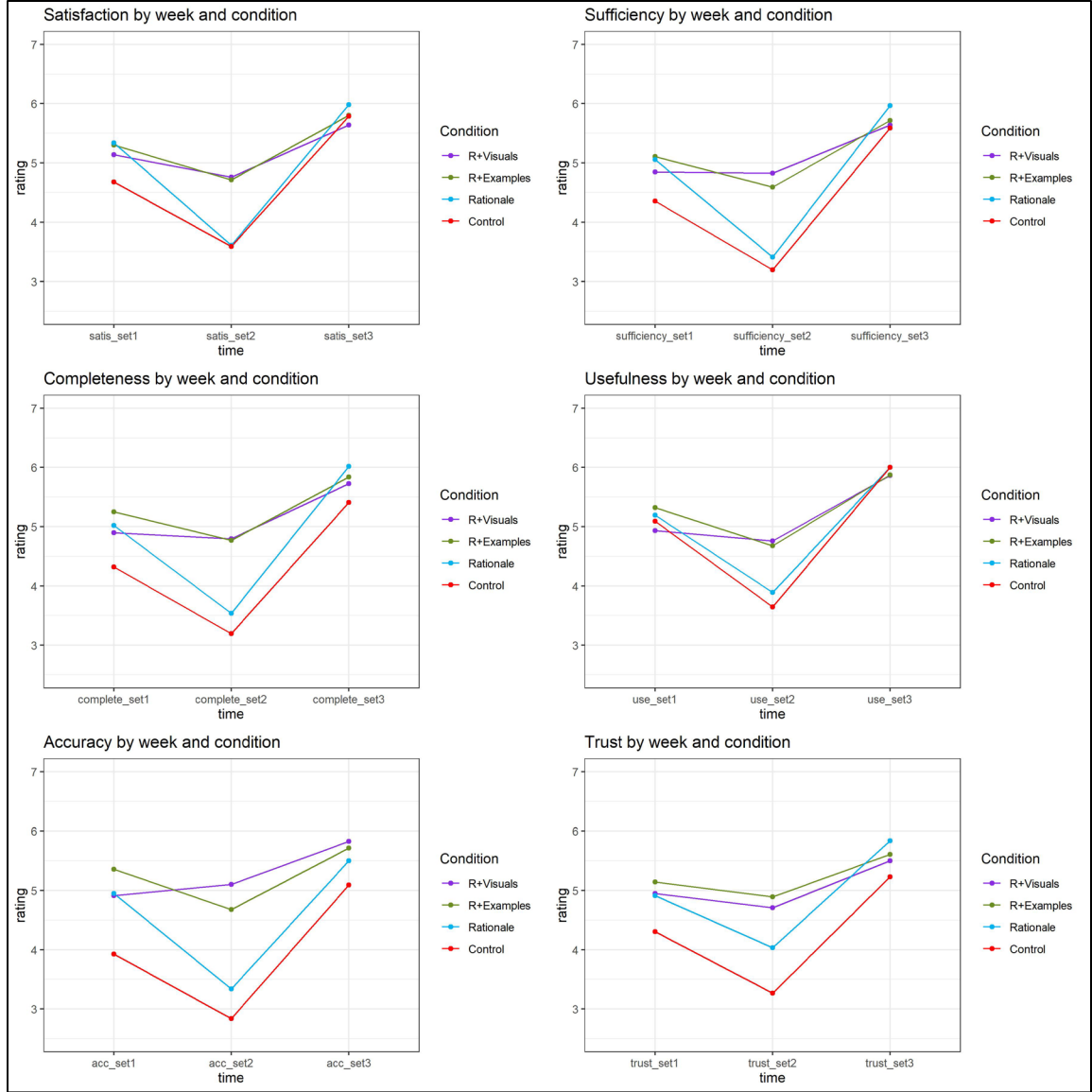


Figure 8: Rating for explanation satisfaction scales

I examined the rating for each dimension of explanation satisfaction scales with a Type-III factorial ANOVA using the R package *car* (J. Fox et al., 2012). The results are shown in Table 7. There was a statistically significant difference in average satisfaction, sufficiency, completeness, usefulness, trust, and perception of accuracy of the system yield by time sets (week 1-2, week 3-4, week 5-6) and there was a significant time by condition

(types of explanation) interaction in each case. There was an only a statistically significant effect of Explanation Condition for completeness and accuracy.

Table 7: Results from Type- III factorial ANOVA for explanation satisfaction scales

	<b>Time</b>	<b>Condition</b> (Types of explanation)	<b>Time: Condition</b>
Satisfaction	$F(2,327) = 24.62$ $p < 0.05$	$F(3,327) = 1.89$ $p = 0.13$	$F(6,327) = 3.52$ $p < 0.05$
Sufficiency	$F(2,327) = 26.57$ $p < 0.05$	$F(3,327) = 2.17$ $p = 0.09$	$F(6,327) = 4.25$ $p < 0.05$
Completeness	$F(2,327) = 21.72$ $p < 0.05$	$F(3,327) = 2.76$ $p < 0.05$	$F(6,327) = 3.15$ $p < 0.05$
Usefulness	$F(2,327) = 27.24$ $p < 0.05$	$F(3,327) = 0.53$ $p = 0.66$	$F(6,327) = 2.55$ $p < 0.05$
Accuracy	$F(2,327) = 21.7$ $p < 0.05$	$F(3,327) = 6.31$ $p < 0.05$	$F(6,327) = 3.8$ $p < 0.05$
Trust	$F(2,327) = 17.61$ $p < 0.05$	$F(3,327) = 2.41$ $p = 0.07$	$F(6,327) = 2.03$ $p = 0.06$

To understand the differences between the Explanation Conditions at each Set, I conducted Tukey post-hoc tests for each of the six scales using the R package *agricolae* (de Mendiburu and de Mendiburu, 2019). The results are shown in Table 8.



Table 8: Significant differences between conditions at each Set according to the Tukey test, any pairing not mentioned was not significantly different for that Set.

	Week 1-2	Week 3-4	Week 5-6
Satisfaction	None	Visuals and examples were better than control and rationales	None
Sufficiency	None	Visuals and examples were better than control and rationales	None
Completeness	None	Visuals and examples were better than control and rationales	None
Usefulness	None	Visuals were better than Control	None
Accuracy	Examples were better than control, no other groups differed	Visuals and examples were better than control and rationales	None
Trust	None	Visuals and examples were better than Control	None

For Set 1, there is no significant differences between any pair of explanation types over all six dimensions except accuracy. For Set 2, there are no significant differences between rationales + visuals and rationales + examples for satisfaction, sufficiency, completeness, trust, and accuracy. But they both are better than control and rationales for satisfaction, sufficiency, completeness, and accuracy. For Set 3, there are no differences between any of the explanation types. The explanations at the end, once the system determined the right diagnosis are almost irrelevant. Only during crisis weeks when the

system was getting it wrong, there were statistically significant differences between explanation conditions. Including visual explanations and examples to rationales generally improved the ratings of the explanations. It is almost the same pattern over all dimensions except usefulness. The exceptions are not strong enough to cover separately.

Because of the high similarity between the results of six scales, I ran one additional test for the grand average of all the six attributes, overall satisfaction. For overall satisfaction, visuals and example-based explanations were rated as better than rationales alone and the no Explanation condition in Set 2 as shown in Figure 9. According to Type-III factorial ANOVA, there was a statistically significant difference in overall satisfaction yield by time ( $F(2,327) = 28.75, p < 0.05$ ), condition ( $F(3,327) = 2.75, p < 0.05$ ) and there was also a statistically significant time by condition interaction ( $F(6,327) = 3.76, p < 0.05$ ). According to the Tukey post-hoc test, there are no significant differences between any pair of explanation types for Set 1. For Set 2, there are no significant differences between rationales + visuals and rationales + examples but they both are better than control and rationales. For Set 3, there are no differences between any of the explanation types. The results show that at the beginning when the system started differential diagnosis and was one the way of ruling out Giardia before going for IBS, the types of explanation did not have much effect on overall satisfaction. But when the system was not performing well and predicting the wrong diagnosis, the differences between explanation types became evident. And again, at the end when the system resolved all the issues, the overall satisfaction almost converged for all four conditions. It did not matter whether they got any explanation or what type of explanations they got.

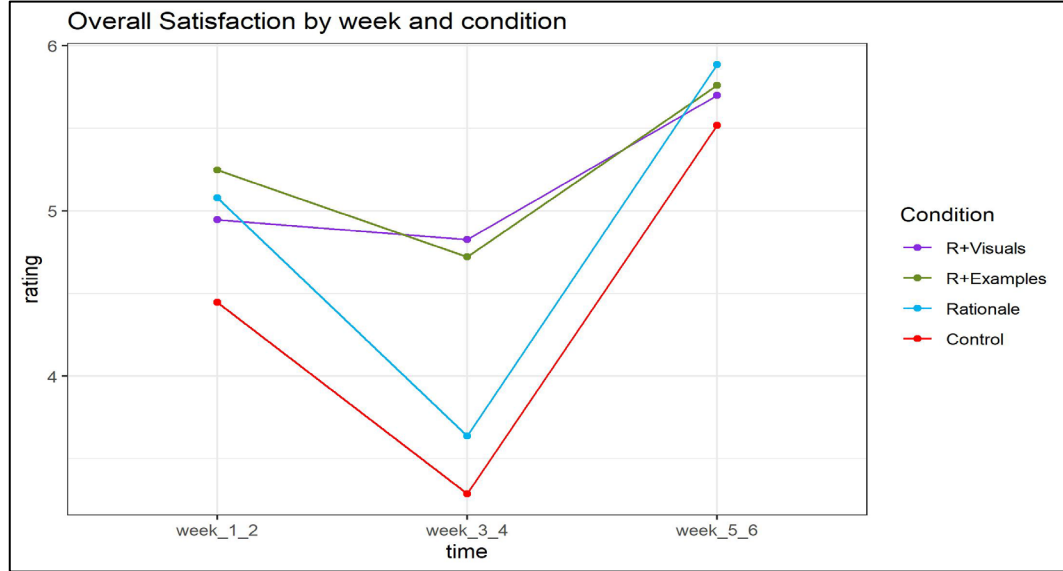


Figure 9: Mean rating for Overall Satisfaction.

## 6 Summary

The study shows that types of explanation matter only during crisis weeks. Explanations, especially richer explanations are needed when the system is predicting it wrong, adding visual and examples with rationales generally improve satisfaction at this point. Once the system came to a resolution or gets things right, the explanation almost did not matter anymore. In the next chapter, I provide some recommendations for designing better XAI systems.

## 7 General Discussion

### 7.1 Summary

I first conducted an interview study with physicians to learn how they communicate with their patients when they change their diagnosis. Five themes or broad categories of explanation strategies and purposes emerged: 1) explanations intended to prepare the patient for later possibilities; 2) ways to tailor information to the audience; 3) use of case information to make a logical argument, 4) use of test results and logical constructs to support the diagnosis; and 5) communication intended to build emotional connection and rapport. I presented these in a generic diagnosis timeline that identifies points at which explanatory reasoning strategies were observed. I ran a study based on a re-diagnosis scenario with a simulation of patient-AI interaction, with participants playing the role of the patients.. The goal was to understand whether the explanation had any effect on "patient" satisfaction at all in this situation. The result showed promise in that the control group (No Explanation Condition) gave a lower rating on all six dimensions- satisfaction, sufficiency, completeness, usefulness, accuracy, and trust. Local justifications for each diagnosis along with graphical information (probability of correct diagnosis) led to better understanding and satisfaction about the diagnosis.

I then ran a small sample of participants in the same scenario with pre-diagnosis global explanation only (no explanation during the diagnosis). Pre-diagnosis global explanation did not help to improve satisfaction measures, although it might help patients understand the general method for diagnosis as much as local justifications.

The final study involved a visualized and example-based explanation for this study. Participants were divided into 4 groups and received different forms of explanation during the interaction with the AI for diagnosis: control, rationales, visual + rationales, example + rationales. Results show that the overall satisfaction for visual + rationales, example + rationales groups are not different from each other but they both had a greater effect than rationales only and the control group. In most cases, rationales were no better than the control group.

## **7.2 Guidance for Designing Explanation for Medical AI**

Based on the results, I could derive some guidance for designing explanations for AI systems in the medical domain:

1. Tailoring explanation to suit different patients or healthcare consumers
2. Tailoring explanation for different points of diagnosis
3. Testing the effectiveness of explanatory contents at crisis points of diagnosis
4. Integrating rationales with either visualizations or examples for explanation

### **7.2.1 Tailoring explanation to suit different patients**

If an AI system is designed for patient diagnosis and patients or their families are the primary user of that system, there is a need to tailor the explanations of that AI system to suit different patients. The need for a user model in the AI system has been discussed within explainable AI systems in general (Brézillon, 1994; Cawsey, 1993; R. Kass and Finin, 1988; Weiner, 1989). The user model ensures that the explanations generated from AI systems will users or modified to match users' mental models. Some user models are

also about tracking the users over time(Kelly and Belkin, 2002). This has also been discussed in the healthcare domain but from a different perspective. Darlington (2011) advocated the need for explanation in the healthcare expert system considering the user requirements of different stakeholders of the healthcare domain such as- physicians, patients, administrators, medical researchers. Personalization of explanation in AI systems has been discussed in XAI literature to draw attention to the lack of human aspects consideration in AI systems (Miller, 2019). One explanation cannot satisfy every user and therefore, there is a need to personalize these explanations. And, AI can achieve this goal by having an interactive environment where it can receive information about different aspects of its users. Google's people + AI Guidebook has described the best practices for designing human centered AI products and acknowledging the importance of interaction and explainability<sup>17</sup>. There are also other AI systems that personalize explanation in interactive environment (Akula et al., 2019; Schneider and Handali, 2019; Sokol and Flach, 2020), but the necessity of tailoring explanation to the need of different patients has not been discussed in the past literature. In my interview study, physicians often stated that they develop their explanation considering the patient's emotional, cultural, or socioeconomic status, they also have to keep in mind the intellectual level of the patients. This indicates that the AI system for medical diagnostics also needs to tailor its explanation to suit different patients and their families. If it can recognize the emotional, cultural, or socio-economic status of the patients by interacting with its user, it may also recognize their need and adapt its explanations according to that.

---

<sup>17</sup> <https://pair.withgoogle.com/>

### **7.2.2 Tailoring explanation for different points of diagnosis**

The explanation provided by AI systems might differ at different timepoints. For any complex task, the explanation might differ based on the parts of the task. But user models for explanation are mostly developed considering the persona-based generation of explanation over time. Brézillon (1994) discussed the importance of context for explanations but did not mention anything specific about task timepoint-dependent explanation. User models within intelligent tutoring systems (ITS) often use knowledge space theory (Doignon and Falmagne, 1985) to understand where the learners stand in terms of their knowledge or learning. ITSs can conduct knowledge assessment based on what students know and can teach them after inferring what they are prepared to learn (Burton and Brown, 1979; Clancey, 1984; Sleeman and Brown, 1982). ITS and similar user model-based systems focus on the knowledge of its users and generates explanations based on it but do not address the issue of timepoint-based explanations directly. I have found from my interview study that explanation differs due to the timeline of the diagnosis within the explanation schema for physicians. Physicians talked about how they change their explanations based on the timeframes of diagnosis. Explanations at the initial point of diagnosis are often related to explaining differential diagnosis- giving one diagnosis but preparing patients for later possibilities, giving generalized information, or providing triage rationales only. More technical or logical explanations are presented later, when patients are stabilized, or their families are calmer than the initial state. Physicians make a very specific diagnostic argument at this point, an explanation incorporating all the information about the patient- history, symptom, diagnosis, health improvement. This can happen either

when a patient's condition is improving or is at a critical point where physicians need to convince the patient that they are doing what is in the patient's best interest. The AI system also needs to tailor its explanation for a different point of diagnosis to communicate with the patients and their families. Figure 10 is the modified version of Figure 2 that highlights the different points of explanation for AI diagnosis. At the initial point of diagnosis, AI may provide explanations to prepare patients for later possibilities. It may not give too much negative information and it should try to focus on the generalized information about the patient's condition. While presenting the results of diagnostic tests, the explanation may need to explain why the tests were given by describing the relationship between pre-existing and current conditions. When there are follow up diagnostic reports, an explanation should use the reports (and perhaps also the patient's medical records) as a visual aid to explain the situation. Logical or technical arguments may start from these points to be more specific about the patient's condition.



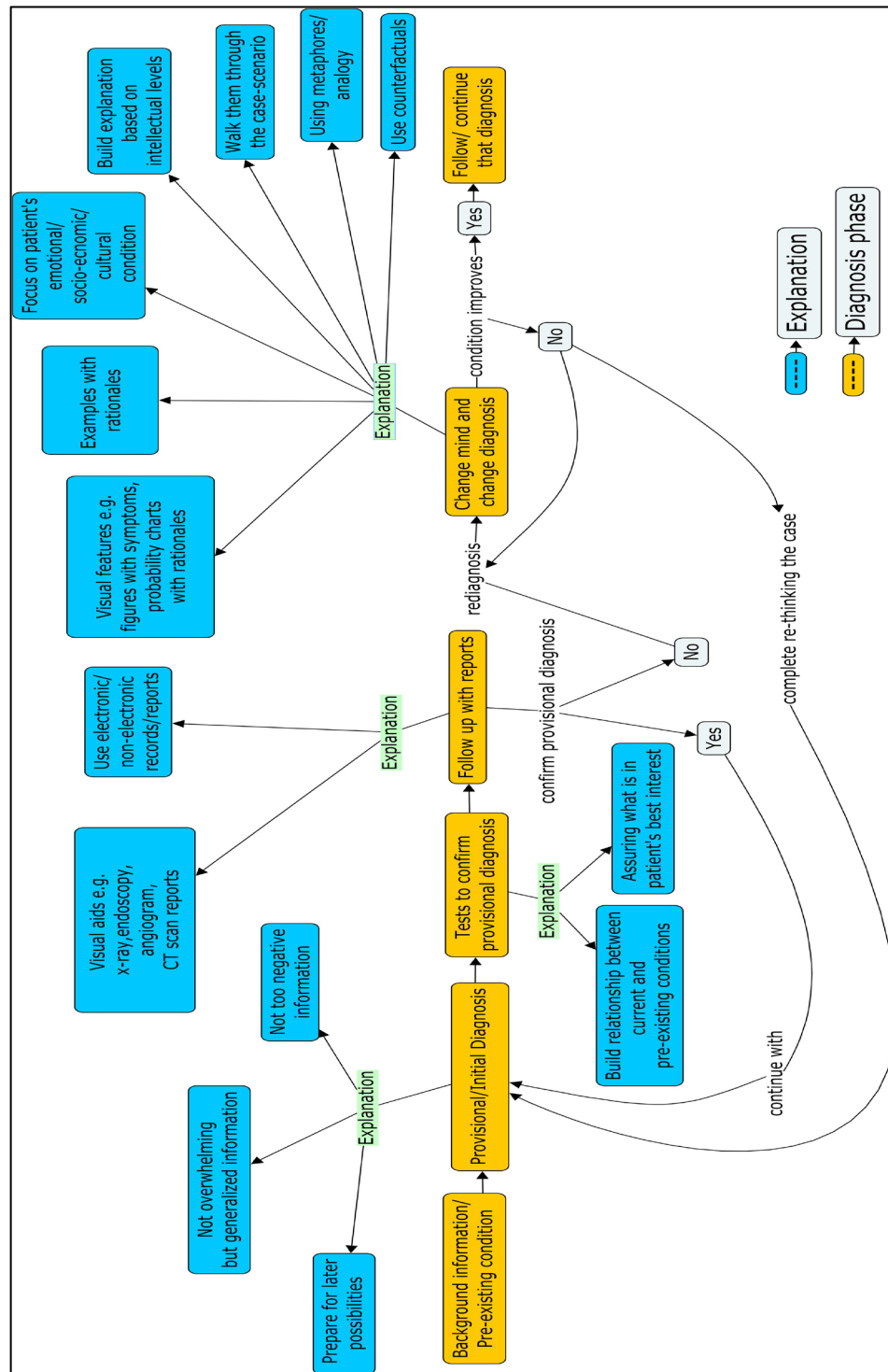


Figure 10: Different timepoints of explanation for AI

When the AI understands that it was diagnosing incorrectly and changes its diagnosis the crisis point starts. The AI would need to present a more logical argument at this point to convince the patient that it is going to do the right thing now, and why the previous diagnosis made sense at the time. Such an explanation may improve a patient's trust and satisfaction if it employs visualization, examples, metaphors, analogy, counterfactuals at this point.

As mentioned earlier, the AI also needs to focus on user models while providing all these technical or logical arguments. It may describe the entire scenario in brief to patients and walk them through it to explain what it did and why it did. The explanations at this point will be more detailed than explanations presented at previous stages and will need to be tailored to the need of different patients. The explanation at the time point when the patient is improving or at reaching the end of consultation will also be different from what it is at the initial stage of diagnosis.

### **7.2.3 Testing the effectiveness of explanatory contents at crisis points**

Explanatory contents of AI systems are found effective when the systems can explain their reasoning and users can explain back corrections for the systems (Kulesza et al., 2015). It enables them to get the most benefit from systems by building useful mental models. Users demand more explanation from intelligent systems in critical situations (Lim and Dey, 2009) though it is very challenging to satisfy them in these situations. Users look for “what else” explanation hoping that the system will do some more to handle the critical situation.

In Study 3, I found that the types of explanation mattered the most at the crisis points of diagnosis. Explanations led to greater satisfaction and trust when the prediction of the AI system had been incorrect. There was not much difference between the four conditions (control, rationales, rationales + visuals, rationales + examples) at the beginning of the scenario. Differences emerged when the patient data showed that that AI diagnosis might have been wrong, and the treatment was not helping improve the patient's condition. But when the AI system came to a resolution about an alternative diagnosis (Celiac disease), there was no big difference seen. This means that if someone wants to test the effectiveness of the explanatory contents of a diagnostic system, it should be tested at the crisis points of diagnosis rather than testing it at the end when the system comes to a resolution. It could be true for any other expert systems as well. Often user satisfaction and trust are measured at the end when user feedbacks are asked. This would not be effective unless user satisfaction is also measured at the crisis points when the system itself is not very effective. Overall, what type of explanation is beneficial should be determined by testing its effectiveness at the crisis points. According to the results of my study, that is how proper explanatory content is going to be found for XAI systems.

Effectiveness of explanation is evident at the crisis points of diagnosis in my study and this could be also true for expert systems and XAI systems in general. Explanations would lead to better satisfaction especially when a system makes an error. From a psychology perspective, errors lead to beneficial learning if they are followed by corrective feedback (Metcalfe, 2017). Errors refine learners' mental models about intelligent systems and improve the usability of such systems, since errors highlight boundary conditions and

problem areas (Mueller and Klein, 2011). The way the effectiveness of explanations is measured in much of the recent research (Kulesza et al., 2015), it might not capture the critical points. The crisis point is critical in healthcare, it is the point when AI may lose contact with patients.

System developers should concentrate on investing more effort into explanations in cases where the system goes wrong. This might be the most challenging thing for explanations to handle and might require core changes in XAI architecture. In the scenario of my study, if the system diagnoses the disease correctly, people tend to be satisfied whether they receive an explanation. But the satisfaction diminishes during the crisis or when the system is performing poorly. And those are not minimal consequences, because that is when the users stop using the system even if the system is doing the same thing a physician would do in any critical situation. If users are not getting a good explanation of why the system did what it did, they will stop using the system and go for physician resources for the same thing that AI diagnostic system did or would do. So, more focus should be invested in better explanations for times when things go wrong or AI systems may make errors.

#### **7.2.4 Integrating rationales with either visualizations or examples**

The contents of explanations can differ for justifications, rationales, and visualizations. Explanations like rationales from the 80s (Buchanan and Shortliffe, 1984; Clancey, 1983) did not help much alone in my experiment. The explanations were effective when these rationales were integrated with visual explanations or example-based explanations. Since I did not test visuals-only or examples-only conditions, it is unclear

whether these two separately will be better than rationales-only or not. But I have found examples of visual explanations integrated with rationales that were effective and improved satisfaction, trust, and perception of its accuracy. Figure 11 is one instance of a visual example combines with rationales, there could be other effective forms of visualization (Goyal et al., 2016; Rajani and Mooney, 2015; Ramanishka et al., 2017; Selvaraju et al., 2016). Visual explanations have also been used in XAI literature with relevant sentences (Hendricks et al., 2016), but those are not necessarily rationales or justifications. In my study, the visual explanation is comprised of two parts: the left one is a graphical representation of patient history, symptoms, and their relative outcomes for three different diseases. The right one is a probability chart formed from the outcomes of the left figure. This visualization was more helpful than only rationales at the point when the AI system was performing worst and the patient condition was not improving after following its advice.

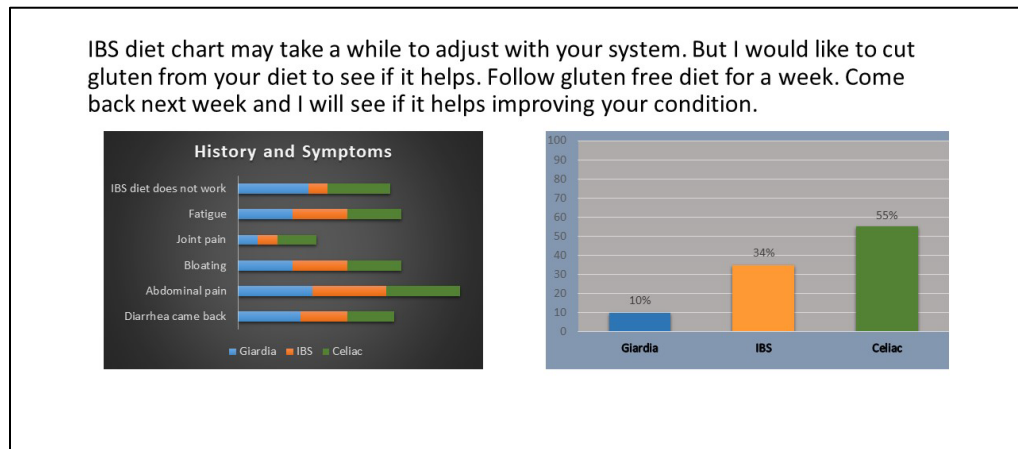


Figure 11: Visual explanation with a graphical representation of relative outcomes

Case-based reasoning or an example-based explanation has been utilized in some XAI systems (Doyle et al., 2003; Nugent et al., 2009), and it has also been used for medical

diagnostic systems (Cunningham et al., 2003). For integrating examples with rationales, I used similar example-based explanations where the system gives an example of a previous case similar to the current case (Figure 12). This helped in improving satisfaction, trust, and perception of its accuracy when integrated with rationales at crisis points of diagnosis.

IBS diet chart may take a while to adjust with your system. But I would like to cut gluten from your diet to see if it helps. Follow gluten free diet for a week. Come back next week and I will see if it helps improving your condition. Let's look at a similar case C10635:	
Case C10635	Your case
Patient had diarrhea, abdominal pain, joint pain, fatigue	You have diarrhea, abdominal pain, joint pain, fatigue
Patient was diagnosed with IBS	You have been diagnosed with IBS
Patient condition did not improve by following IBS diet chart, so I prescribe Gluten free diet	Your condition did not improve by following IBS diet chart for two weeks

Figure 12: Example-based explanation

I did not find any difference in ratings of satisfaction or trust between visual explanation and example-based explanation in my study. They had the same degree of effect on satisfaction dimensions and led to higher satisfaction than in the Control (No explanation) Condition and the Rationale Explanation Condition. And again, it is not clear how visual and example-based explanation will affect satisfaction when rationales are removed from the explanation.

### 7.3 Transparency and the Right to Explanation

One aspect of XAI systems within the context of healthcare is the transparency issues. The General Data Protection Regulation (GDPR) contains a specific right to

explanation under Article 22 for individuals within the European Union (EU)<sup>18</sup>. However, the consequences of these regulations leave many questions unanswered. Ordish and Hall (2019) posed some questions about the uncertainties of how the right to explain requirements of GDPR will apply to AI within the context of healthcare:

- 1) When is explanation required? Is explanation required before the data is processed and/or after processing?
- 2) What is to be explained? Must data controllers explain the model and how it functions as a whole and/or must they provide an explanation of individual decisions post-processing?
- 3) What kind of explanation is required? Might counterfactual explanations (that describe the nearest possible world where the result sought was obtained) suffice?

I think this research may clarify and provide some reasonable answers to these questions. Often it is not exactly how it has been framed within the XAI community. But, if we look at how physicians explain things and the types of explanation are useful in our patient scenario, we can think about each of these questions with respect to the conclusions I have just made. For example, the first question was “*when is explanation required?*” and whether it should be before the data is handled or after processing. What we found is slightly different, it is not necessarily before or after, explanations are most useful in the moment of the crisis. At the beginning or after the AI systems make a resolution, explanation is almost irrelevant in our patient scenario.

---

<sup>18</sup> <https://gdpr-info.eu/art-22-gdpr/>

The second question is about “*what is to be explained?*” and whether the explanation should function as a whole, or it should be provided for individual decisions like local justifications. We have looked at both and we found advantages for both but a little differently. The local explanation helped assessment of trust in the moment and the global explanation helped assessment of understanding of the system as a whole afterwards.

The last question was about “*What kind of explanation are required?*”. The requirements for types of explanation pose like simple dichotomies within the context of XAI but if we look at it within the context of human explanation, it is much richer, much more complex. In our interviews, physicians expressed that all kinds of explanation are important depending when, how, and what the goals of diagnosis and treatment are. A lot of them are useful as they have different purposes. But we found in the patient scenario that in the moments of crisis, people may prefer richer form of explanations than plain text-based rationales within the context of AI systems.

## **7.4 Conclusion**

XAI systems for medical diagnostics need to consider the human aspects of the system and its explanation. Personalizing the explanations and tailoring them to user needs and expectations will help ensure the proper utilization of these systems. Otherwise, in cases where AI is capable enough to do the same thing the physicians do, patients will not be willing to use it rather they will prefer to go to the physicians. Crisis points are crucial in healthcare diagnosis. If AI is making an accurate diagnosis and patients get positive results by using it, an explanation is not a matter of concern at that time. But if AI makes



any error in crisis points but cannot generate satisfactory explanations about it, people will not think much before losing trust on it and stop using it for diagnosis. In my study, I found that visual or example-based explanations integrated with rationales help at this point. Design recommendations provided would help build better XAI systems for medical diagnosis. The future direction of this work could be to investigate if visual or example explanations are effective without adding rationales to them or what are the best ways to incorporate these explanations within an AI system that is already developed.

## 8 Reference List

- Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Akula, A. R., Todorovic, S., Chai, J. Y., & Zhu, S.-C. (2019). Natural Language Interaction with Explainable AI Models. *ArXiv Preprint ArXiv:1903.05720*.
- Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. *Clinical and Investigative Medicine. Medecine Clinique et Experimentale*, 5(1), 49–55.
- Barrows, Howard S., & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education*. Springer Publishing Company.
- Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction*, 16(2–4), 193–212.
- Bertakis, K. D. (1977). The communication of information from physician to patient: A method for increasing patient retention and satisfaction. *J Fam Pract*, 5(2), 217–222.
- Bickmore, T. W., Trinh, H., Olafsson, S., O’Leary, T. K., Asadi, R., Rickles, N. M., & Cruz, R. (2018). Patient and consumer safety risks when using conversational assistants for medical information: An observational study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research*, 20(9), e11510.

- Bologna, G., & Hayashi, Y. (2017). Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning. *Journal of Artificial Intelligence and Soft Computing Research*, 7(4), 265–286.
- Bredart, A., Bouleuc, C., & Dolbeault, S. (2005). Doctor-patient communication and satisfaction with care in oncology. *Current Opinion in Oncology*, 17(4), 351–354.
- Brézillon, P. (1994). Context needs in cooperative building of explanations. *First European Conference on Cognitive Science in Industry*, 443–450.
- Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, 120(3), 278–287. <https://doi.org/10.1037/0096-3445.120.3.278>
- Buchanan, B., & Shortliffe, E. (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*.
- Buller, M. K., & Buller, D. B. (1987). Physicians' communication style and patient satisfaction. *Journal of Health and Social Behavior*, 375–388.
- Burton, R. R., & Brown, J. S. (1979). An investigation of computer coaching for informal learning activities. *International Journal of Man-Machine Studies*, 11(1), 5–24.
- Byrne, R. M. (1991). The Construction of Explanations. In *AI and Cognitive Science '90* (pp. 337–351). Springer.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.

- Cawsey, A. (1993). User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3(3), 221–247.
- Che, Z., Purushotham, S., Khemani, R., & Liu, Y. (2016). Interpretable deep models for ICU outcome prediction. *AMIA Annual Symposium Proceedings*, 2016, 371.
- Clancey, W. J. (1983). The epistemology of a rule-based expert system—A framework for explanation. *Artificial Intelligence*, 20(3), 215–251.
- Clancey, W. J. (1984). Methodology for building an intelligent tutoring system. *Methods and Tactics in Cognitive Science*, 51–84.
- Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2018). Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). *ArXiv:1710.05006 [Cs]*.  
<http://arxiv.org/abs/1710.05006>
- Crandall, B., Klein, G., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. Mit Press.
- Cunningham, P., Doyle, D., & Loughrey, J. (2003). An evaluation of the usefulness of case-based explanation. *International Conference on Case-Based Reasoning*, 122–130.
- Darlington, K. W. (2011). Designing for Explanation in Health Care Applications of Expert Systems. *SAGE Open*, 1(1), 2158244011408618.  
<https://doi.org/10.1177/2158244011408618>

- Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3), 319–340. JSTOR. <https://doi.org/10.2307/249008>
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982–1003.
- De Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., & Horrocks, J. C. (1972). Computer-aided diagnosis of acute abdominal pain. *Br Med J*, 2(5804), 9–13.
- De Keyser, V., & Woods, D. D. (1990). Fixation Errors: Failures to Revise Situation Assessment in Dynamic and Risky Systems. In A. G. Colombo & A. S. de Bustamante (Eds.), *Systems Reliability Assessment* (Vol. 6, pp. 231–251). Springer Netherlands. [https://doi.org/10.1007/978-94-009-0649-5\\_11](https://doi.org/10.1007/978-94-009-0649-5_11)
- de Mendiburu, F., & de Mendiburu, M. F. (2019). Package ‘agricolae.’ *R Package Version*, 1–2.
- Doignon, J., & Falmagne, J. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23(2), 175–196. [https://doi.org/10.1016/S0020-7373\(85\)80031-6](https://doi.org/10.1016/S0020-7373(85)80031-6)
- Dominic, D. P., Gopal, D. G., & Abbas, A. M. (2019). Chapter 11—Combining Predictive Analytics and Artificial Intelligence With Human Intelligence in IoT-Based Image-Guided Surgery. In V. E. Balas, L. H. Son, S. Jha, M. Khari, & R. Kumar (Eds.), *Internet of Things in Biomedical Engineering* (pp. 259–289). Academic Press. <https://doi.org/10.1016/B978-0-12-817356-5.00014-0>

- Doyle, D., Tsymbal, A., & Cunningham, P. (2003). *A review of explanation and explanation in case-based reasoning*. Trinity College Dublin, Department of Computer Science.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
- Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *Bmj*, 324(7339), 729–732.
- Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving an analysis of clinical reasoning*.
- Epstein, R. M., & Street, R. L. (2007). *Patient-centered communication in cancer care: Promoting healing and reducing suffering*. National Cancer Institute; Bethesda. MD: 2007. NIH Publication.
- Epstein, Ronald M., Franks, P., Fiscella, K., Shields, C. G., Meldrum, S. C., Kravitz, R. L., & Duberstein, P. R. (2005). Measuring patient-centered communication in patient–physician consultations: Theoretical and practical issues. *Social Science & Medicine*, 61(7), 1516–1528.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
- Farmer, S. E. J., Bernardotto, M., & Singh, V. (2011). How good is Internet self-diagnosis of ENT symptoms using Boots WebMD symptom checker? *Clinical*

*Otolaryngology*, 36(5), 517–518. <https://doi.org/10.1111/j.1749-4486.2011.02375.x>

Felten, E. (2017). *What does it mean to ask for an “explainable” algorithm?* Freeom to Tinker. <https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/>

Fiscella, K., Meldrum, S., Franks, P., Shields, C. G., Duberstein, P., McDaniel, S. H., & Epstein, R. M. (2004). Patient trust: Is it related to patient-centered behavior of primary care physicians? *Medical Care*, 1049–1055.

Fogel, A. L., & Kvedar, J. C. (2018). Artificial intelligence powers digital medicine. *NPJ Digital Medicine*, 1(1), 5.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., & Graves, S. (2012). Package ‘car.’ *Vienna: R Foundation for Statistical Computing*.

Fox, S., & Duggan, M. (2013). Health online 2013. *Health*, 2013, 1–55.

Frank, J. R., Jabbour, M., Tugwell, P., Boyd, D., Labrosse, J., & MacFadyen, J. (1996). Skills for the new millennium: Report of the societal needs working group, CanMEDS 2000 Project. *Ann R Coll Physicians Surg Can*, 29(4), 206–216.

Friedman, R. B., & Gustafson, D. H. (1977). *Computers in clinical medicine, a critical review*.

Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review*, 100(2), 298.

- Glantz, S. A. (1978). Special Feature Computers in Clinical Medicine: A Critique. *Computer*, 5, 68–77.
- Goyal, Y., Mohapatra, A., Parikh, D., & Batra, D. (2016). Towards Transparent AI Systems: Interpreting Visual Question Answering Models. *ArXiv Preprint ArXiv:1608.08974*. <https://arxiv.org/abs/1608.08974>
- Gray, J. A., & Chambers, L. W. (1997). Evidence-based healthcare: How to make health policy & management decisions. *Canadian Medical Association. Journal*, 157(11), 1598.
- Greenfield, S., Kaplan, S., & Ware Jr, J. E. (1985). Expanding patient involvement in care: Effects on patient outcomes. *Annals of Internal Medicine*, 102(4), 520–528.
- Hammond, K. R., Kelly, K. J., Schneider, R. J., & Vancini, M. (1967). Clinical inference in nursing: Revising judgments. *Nursing Research*, 16(1), 38–45.
- Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. *European Conference on Computer Vision*, 3–19. [http://link.springer.com/chapter/10.1007/978-3-319-46493-0\\_1](http://link.springer.com/chapter/10.1007/978-3-319-46493-0_1)
- Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining Explanation For “Explainable Ai.” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 197–201. <https://doi.org/10.1177/1541931218621047>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *ArXiv Preprint ArXiv:1812.04608*.
- Hogarth, R. M. (1980). Judgement, drug monitoring, and decision aids. *Monitoring for Drug Safety. Philadelphia: Lippincott*, 439–478.



- Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *ArXiv Preprint ArXiv:1712.09923*.
- Jao, C. (2011). *Efficient decision support systems-practice and challenges in biomedical related domain*.
- Jha, S., & Topol, E. J. (2016). Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *Jama*, 316(22), 2353–2354.
- Johnson, P. E., Duran, A. S., Hassebrock, F., Moller, J., Prietula, M., Feltovich, P. J., & Swanson, D. B. (1981). Expertise and Error in Diagnostic Reasoning\*. *Cognitive Science*, 5(3), 235–283. [https://doi.org/10.1207/s15516709cog0503\\_3](https://doi.org/10.1207/s15516709cog0503_3)
- Johnson, P. E., Moen, J. B., & Thompson, W. B. (1988). Garden path errors in diagnostic reasoning. In *Expert system applications* (pp. 395–427). Springer.
- Kaddioui, H., Duong, L., Joncas, J., Bellefleur, C., Nahle, I., Chémaly, O., Nault, M.-L., Parent, S., Grimard, G., & Labelle, H. (2020). Convolutional Neural Networks for Automatic Risser Stage Assessment. *Radiology: Artificial Intelligence*, 2(3), e180063. <https://doi.org/10.1148/ryai.2020180063>
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454.
- Kass, A., & Leake, D. (1987). *Types of Explanations*. YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE.
- Kass, R., & Finin, T. (1988). The Need for User Models in Generating Expert System Explanation. *Int. J. Expert Syst.*, 1(4), 345–375.

- Kassirer, J. P., & Gorry, G. A. (1978). Clinical problem solving: A behavioral analysis. *Annals of Internal Medicine*, 89(2), 245–255.
- Kelly, D., & Belkin, N. J. (2002). A user modeling system for personalized interaction and tailored retrieval in interactive IR. *Proceedings of the American Society for Information Science and Technology*, 39(1), 316–325.
- Khan, R., & Das, A. (2017). *Build better chatbots: A complete guide to getting started with chatbots*. Apress.
- Klein, D. E., Woods, D. D., Klein, G., & Perry, S. J. (2016). Can we trust best practices? Six cognitive challenges of evidence-based approaches. *Journal of Cognitive Engineering and Decision Making*, 10(3), 244–254.
- Klein, G. A. (1999). *Sources of power: How people make decisions*. MIT press.
- Klein, J. G. (2005). Five pitfalls in decisions about diagnosis and prescribing. *BMJ : British Medical Journal*, 330(7494), 781–783.
- Knill-Jones, R. P., Stern, R. B., Girmes, D. H., Maxwell, J. D., Thompson, R. P. H., & Williams, R. (1973). Use of sequential Bayesian model in diagnosis of jaundice by computer. *Br Med J*, 1(5852), 530–533.
- Korsch, B. M., Gozzi, E. K., & Francis, V. (1968). GAPS IN DOCTOR-PATIENT COMMUNICATION: I. Doctor-Patient Interaction and Patient Satisfaction. *Pediatrics*, 42(5), 855–871.
- Kruger, R. P., Thompson, W. B., & Turner, A. F. (1974). Computer diagnosis of pneumoconiosis. *IEEE Transactions on Systems, Man, and Cybernetics*, 1, 40–49.

- Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137. <https://doi.org/10.1145/2678025.2701399>
- Kulkarni, S. S., Dewitt, B., Fischhoff, B., Rosengart, M. R., Angus, D. C., Saul, M., Yealy, D. M., & Mohan, D. (2019). Defining the representativeness heuristic in trauma triage: A retrospective observational cohort study. *PLoS ONE*, 14(2). <https://doi.org/10.1371/journal.pone.0212201>
- Kundu, S., Kolouri, S., Erickson, K. I., Kramer, A. F., McAuley, E., & Rohde, G. K. (2017). Discovery and visualization of structural biomarkers from MRI using transport-based morphometry. *ArXiv:1705.04919 [Cs]*. <http://arxiv.org/abs/1705.04919>
- Lansky, D. (1998). Measuring what matters to the public. *HEALTH AFFAIRS-MILLWOOD VA THEN BETHESDA MA*, 17, 40–41.
- Laumer, S., Maier, C., & Gubler, F. T. (2019). *CHATBOT ACCEPTANCE IN HEALTHCARE: EXPLAINING USER ADOPTION OF CONVERSATIONAL AGENTS FOR DISEASE DIAGNOSIS*.
- Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., & Thiesson, B. (2019). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *ArXiv Preprint ArXiv:1912.01266*.

- Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis. *Science*, 130(3366), 9–21.
- Lemieux, M., & Bordage, G. (1992). Propositional versus structural semantic analyses of medical diagnostic thinking. *Cognitive Science*, 16(2), 185–204.
- Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. *Proceedings of the 11th International Conference on Ubiquitous Computing*, 195–204. <https://doi.org/10.1145/1620545.1620576>
- Lipkin, M., Putnam, S. M., Lazare, A., Carroll, J. G., & Frankel, R. M. (1995). *The medical interview: Clinical care, education, and research*. Springer.
- Lipton, Z. C. (2016). The mythos of model interpretability. *ArXiv Preprint ArXiv:1606.03490*.
- London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1), 15–21.
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4), 629–650. <https://doi.org/10.1093/jcr/ucz013>
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., & Studer, M. (2013). Package ‘cluster.’ *Dosegljivo Na*.
- Maude, J. (2014). Differential diagnosis: The key to reducing diagnosis error, measuring diagnosis and a mechanism to reduce healthcare costs. *Diagnosis*, 1(1), 107–109.
- Mead, N., & Bower, P. (2000). Patient-centredness: A conceptual framework and review of the empirical literature. *Social Science & Medicine*, 51(7), 1087–1110.

- Medicine, I. of. (2001). Committee on Quality of Health Care in America. Crossing the quality chasm: A new health system for the 21st century. *National Academies Press*.
- Metcalf, J. (2017). Learning from errors. *Annual Review of Psychology*, 68, 465–489.
- Militello, L. G., & Hutton, R. J. (1998). Applied cognitive task analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics*, 41(11), 1618–1641.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Miller, W. T. (1998). Letter from the editor: Occam versus Hickam. *Seminars in Roentgenology*, 3, 213.
- Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, 176(5), 619–625.
- Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, 20(6), e10148.
- Mueller, S. T., & Klein, G. A. (2011). Improving Users' Mental Models of Intelligent Software Tools. *IEEE Intelligent Systems*, 26(2), 77–83.  
<https://doi.org/10.1109/MIS.2011.32>

- Musen, M. A., Middleton, B., & Greenes, R. A. (2014). Clinical decision-support systems. In *Biomedical informatics* (pp. 643–674). Springer.
- Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health*, 5, 2055207619871808.
- Nakatsu, R. T. (2004). Explanatory power of intelligent systems: A research framework. *Proceedings of the IFIP International Conference on Decision Support Systems, Prato, Italy*, 568–577.
- Naughton, C. A. (2018). Patient-Centered Communication. *Pharmacy: Journal of Pharmacy Education and Practice*, 6(1).  
<https://doi.org/10.3390/pharmacy6010018>
- Naylor, C. D. (2018). On the prospects for a (deep) learning health care system. *Jama*, 320(11), 1099–1100.
- Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical Care Medicine*, 46(4), 547–553.
- North, F., Ward, W. J., Varkey, P., & Tulledge-Scheitel, S. M. (2012, April 5). *Should You Search the Internet for Information About Your Acute Symptom?* [Research-article]. <https://Home.Liebertpub.Com/Tmj>. <https://doi.org/10.1089/tmj.2011.0127>
- Nugent, C., Doyle, D., & Cunningham, P. (2009). Gaining insight through case-based explanation. *Journal of Intelligent Information Systems*, 32(3), 267–295.

- Ong, L. M., De Haes, J. C., Hoos, A. M., & Lammes, F. B. (1995). Doctor-patient communication: A review of the literature. *Social Science & Medicine*, 40(7), 903–918.
- Ordish, J., & Hall, A. (2019). *PHG Foundation is an exempt charity under the Charities Act 2011 and is regulated by HEFCE as a connected institution of the University of Cambridge. We are also a registered company No. 5823194, working to achieve better health through the responsible and evidence based application of biomedical science.*
- Otte, C. (2013). Safe and interpretable machine learning: A methodological review. In *Computational intelligence in intelligent data analysis* (pp. 111–122). Springer.
- Panigutti, C., Perotti, A., & Pedreschi, D. (2020). Doctor XAI: An ontology-based approach to black-box sequential data classification explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 629–639.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Patel, V., Arocha, J. F., & Kaufman, D. (1994). Diagnostic reasoning and medical expertise. *Psychology of Learning and Motivation-Advances in Research and Theory*, 31(C), 187–252.
- Piccini, D., Demesmaeker, R., Heerfordt, J., Yerly, J., Di Sopra, L., Masci, P. G., Schwitter, J., Van De Ville, D., Richiardi, J., Kober, T., & Stuber, M. (2020). Deep Learning to Automate Reference-Free Image Quality Assessment of Whole-Heart MR

- Images. *Radiology: Artificial Intelligence*, 2(3), e190123.  
<https://doi.org/10.1148/ryai.2020190123>
- Poses, R. M., & Anthony, M. (1991). Availability, wishful thinking, and physicians' diagnostic judgments for patients with suspected bacteremia. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 11(3), 159–168. <https://doi.org/10.1177/0272989X9101100303>
- Powley, L., McIlroy, G., Simons, G., & Raza, K. (2016). Are online symptoms checkers useful for patients with inflammatory arthritis? *BMC Musculoskeletal Disorders*, 17(1), 362. <https://doi.org/10.1186/s12891-016-1189-2>
- Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5), 455–468.
- Pryor, T. A., Morgan, J. D., Clark, S. J., Miller, W. A., & Warner, H. R. (1975). Help—A Computer System for Medical Decision Making. *Computer*, 8(1), 34–38.
- Rajani, N. F., & Mooney, R. J. (2015). Using explanations to improve ensembling of visual question answering systems. *Training*, 82, 248–349.
- Ramanishka, V., Das, A., Zhang, J., & Saenko, K. (2017). Top-down visual saliency guided by captions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7206–7215.
- Razzaki, S., Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D., Sangar, D., Taliercio, M., Butt, M., & Majeed, A. (2018). A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *ArXiv Preprint ArXiv:1806.10698*.



- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <http://dl.acm.org/citation.cfm?id=2939778>
- Riccardi, V. M., & Kurtz, S. M. (1983). *Communication and counseling in health care*. Charles C. Thomas Publisher.
- Roter, D. L., Frankel, R. M., Hall, J. A., & Sluyter, D. (2006). The expression of emotion through nonverbal behavior in medical visits. *Journal of General Internal Medicine*, 21(1), 28–34.
- Round, A. (2001). Introduction to clinical reasoning. *Journal of Evaluation in Clinical Practice*, 7(2), 109–117. <https://doi.org/10.1046/j.1365-2753.2001.00252.x>
- Saczynski, J. S., Yarzebski, J., Lessard, D., Spencer, F. A., Gurwitz, J. H., Gore, J. M., & Goldberg, R. J. (2008). Trends in Prehospital Delay in Patients With Acute Myocardial Infarction (from the Worcester Heart Attack Study). *The American Journal of Cardiology*, 102(12), 1589–1594. <https://doi.org/10.1016/j.amjcard.2008.07.056>
- Safran, D. G., Taira, D. A., Rogers, W. H., Kosinski, M., Ware, J. E., & Tarlov, A. R. (1998). Linking primary care performance to outcomes of care. *The Journal of Family Practice*, 47(3), 213–220.
- Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *ArXiv Preprint ArXiv:1901.00770*.


- Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *ArXiv Preprint ArXiv:1610.02391*.  
<https://arxiv.org/abs/1610.02391>
- Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ*, 351, h3480.  
<https://doi.org/10.1136/bmj.h3480>
- Shaughnessy, A. F. (2007). Clinical epidemiology: A basic science for clinical medicine. *Bmj*, 335(7623), 777–777.
- Shortliffe, Edward H. (1987). Computer programs to support clinical decision making. *Jama*, 258(1), 61–66.
- Shortliffe, Edward Hance. (1974). *MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection*. STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE.
- Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, 4(3–4), 181–201.
- Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. London: Academic Press.  
<https://hal.archives-ouvertes.fr/hal-00702997/>
- Sokol, K., & Flach, P. (2020). One Explanation Does Not Fit All. *KI - Künstliche Intelligenz*. <https://doi.org/10.1007/s13218-020-00637-y>
- Stewart, M. A. (1995). Effective physician-patient communication and health outcomes: A review. *CMAJ: Canadian Medical Association Journal*, 152(9), 1423.

- Stewart, M., Brown, J. B., Weston, W., McWhinney, I. R., McWilliam, C. L., & Freeman, T. (2013). *Patient-centered medicine: Transforming the clinical method*. CRC press.
- Swartout, W. R., & Moore, J. D. (1993). Explanation in second generation expert systems. *Second Generation Expert Systems*, 543, 585.
- Tallman, K., Janisse, T., Frankel, R. M., Sung, S. H., Krupat, E., & Hsu, J. T. (2007). Communication practices of physicians with high patient-satisfaction ratings. *The Permanente Journal*, 11(1), 19.
- Teach, R. L., & Shortliffe, E. H. (1981). An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, 14(6), 542–558.
- Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., & Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2), 167–175. <https://doi.org/10.1136/bjophthalmol-2018-313173>
- Tjoa, E., & Guan, C. (2019). A survey on explainable artificial intelligence (XAI): Towards medical XAI. *ArXiv Preprint ArXiv:1907.07374*.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). *Counterfactual explanations without opening the black box: Automated decisions and the GDPR*.
- Wallis, J. W., & Shortliffe, E. H. (1981). *Explanatory power for medical expert systems: Studies in the representation of causal relationships for clinical consultations*. DTIC Document.

- Wang, W., & Siau, K. L. (2018). *Living with Artificial Intelligence: Developing a Theory on Trust in Health Chatbots-Research in Progress*.
- Wanzer, M. B., Booth-Butterfield, M., & Gruber, K. (2004). Perceptions of health care providers' communication: Relationships between patient-centered communication and satisfaction. *Health Communication, 16*(3), 363–384.
- Warner, H. R., Toronto, A. F., & Veasy, L. G. (1964). Experience with Baye's theorem for computer diagnosis of congenital heart disease. *Annals of the New York Academy of Sciences, 115*(2), 558–567.
- Weiner, J. L. (1989). The effect of user models on the production of explanations. *Expert Knowledge and Explanation: The Knowledge-Language Interface*, 144–156.
- Weinstein, M. C., & Fineberg, H. V. (1980). *Clinical decision analysis*.
- Wolf, F. M., Gruppen, L. D., & Billi, J. E. (1985). Differential diagnosis and the competing-hypotheses heuristic: A practical approach to judgment under uncertainty and Bayesian probability. *JAMA, 253*(19), 2858–2862.
- Yanco, H. A., Desai, M., Drury, J. L., & Steinfeld, A. (2016). Methods for Developing Trust Models for Intelligent Systems. In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust Intelligence and Trust in Autonomous Systems* (pp. 219–254). Springer US. [https://doi.org/10.1007/978-1-4899-7668-0\\_11](https://doi.org/10.1007/978-1-4899-7668-0_11)
- Ye, L. R., & Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly, 15*7–172.

Zhang, Z., Xie, Y., Xing, F., McGough, M., & Yang, L. (2017). Mdnet: A semantically and visually interpretable medical image diagnosis network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6428–6436.

## A Study 3 Scenario: Control

<p>Hi I'm MediBot.ai, a medical diagnostic expert system. I would like to help you with your medical condition.</p> <p>Please click the screen to continue.</p>	<p>Which of the symptoms are bothering you most?</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/>Diarrhea</li><li><input checked="" type="checkbox"/>Bloating</li><li><input checked="" type="checkbox"/>Abdominal pain</li><li><input checked="" type="checkbox"/>Joint Pain</li><li><input type="checkbox"/>Itchy Rash</li><li><input type="checkbox"/>Head Cold</li><li><input checked="" type="checkbox"/>Cramps</li><li><input checked="" type="checkbox"/>Fatigue</li></ul>
<p>Do you have family history of any of the following conditions?</p> <ul style="list-style-type: none"><li><input type="checkbox"/>gastrointestinal malignancy</li><li><input type="checkbox"/>Irritable Bowel Syndrome (IBS)</li><li><input type="checkbox"/>Celiac disease</li><li><input type="checkbox"/>Crohn's</li><li><input checked="" type="checkbox"/>None</li></ul>	<p>Have you been exposed to natural water source recently?</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/>Yes</li><li><input type="checkbox"/>No</li></ul> <p>Have you been to woods for hiking or camping recently?</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/>Yes</li><li><input type="checkbox"/>No</li></ul>
<p>Please go through the following test and come back with result :</p> <p><input type="checkbox"/>Giardia antigen test</p> <div data-bbox="492 1268 727 1388"><p><b>DID YOU KNOW?</b></p><p>About 75 million American adults have high blood pressure—that's 1 in every 3 adults</p></div>	<p>Follow up-1</p> <p>How are you feeling?</p> <ul style="list-style-type: none"><li><input type="checkbox"/>Great</li><li><input type="checkbox"/>Moderately great</li><li><input type="checkbox"/>Okay</li><li><input checked="" type="checkbox"/>Bad</li><li><input type="checkbox"/>Very bad</li></ul>

### Follow up-1

What are the symptoms that are bothering you now?

- ✓Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

### Follow up-1

You have come back with you reports. Giardia lamblia is not detected in your antigen test. You are not suffering from Giardia.

### Follow up-1

Based on your current symptoms, I think you are suffering from Irritable Bowel Syndrome (IBS).

Please follow the advices below and come back two weeks later:

- Avoid chocolate, fried foods, alcohol, caffeine, carbonated drinks, the artificial sweetener, fructose
- Eat the soluble kind of fiber
- Drink plenty of water



Drinking 2 cups of cold water on a empty stomach can boost metabolism by 30%

### Follow up-2

Since you are back after two week, I would like to follow up your condition.

Are you following the IBS diet chart?

- ✓Yes
- No

### Follow up-2

How are you feeling after following the diet?

- ☐Great
- ☐Good
- ✓Moderate
- ☐Bad
- ☐Very bad

### Follow up-2

What are the symptoms that are bothering you now?

- Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

Continue the diet and come back for follow up next week



The most likely time for a post-stroke seizure to happen is within the first 24 hours

### Follow up-3

Welcome back!!!

Since you are back after one week, I would like to follow up your condition.

Are you still following the IBS diet chart?

- ☒ Yes
- ☐ No

How are you feeling after following the diet?

- ☐ Great
- ☐ Good
- ☐ Moderate
- ☐ Bad
- ☒ Very Bad

What are the symptoms that are bothering you now?

- ☒ Diarrhea
- ☒ Abdominal pain
- ☒ Joint Pain
- ☒ Bloating
- ☐ Cramps
- ☐ Cold
- ☒ Fatigue

I would like to cut gluten from your diet. Follow gluten free diet for a week. Come back next week and I will see if it helps improving your condition.



The prevalence of diagnosed diabetes in the U.S. increased by 382% from 1988 to 2014

### Follow up-4

Welcome back!!!

You are back after another week and I want to follow up your condition.

Did you follow gluten-free diet this week?

- ☒ Yes
- ☐ No



## Follow up-4

How are you feeling after following the diet?

- ☐ Great
- ☒ Good
- ☐ Moderate
- ☐ Bad
- ☐ Very Bad

## Follow up-4

Which of the symptoms are still bothering you

- ☐ Diarrhea
- ☒ Joint pain
- ☒ Abdominal pain
- ☐ Bloating
- ☐ Cramps
- ☐ Fatigue

## Follow up-4

I have started to think that you may have Celiac disease. However I need you to go through some pathological tests for confirming my diagnosis.



In the United States, someone has heart attack every 40 seconds

- Go for a tTG-IgA Test (blood test)
- Biopsy of the small intestine
- Continue gluten-free diet
- Come back with the report

## Follow up-5

You are back after another week and I will assess your medical reports.

Your reports of tTG-IgA Test (blood test) and biopsy of the small intestine are positive for Celiac disease.

Are you still following gluten-free diet?

- ☒ Yes
- ☐ No

How are you feeling after following the diet?

- ☒ Great
- ☐ Good
- ☐ Moderate
- ☐ Bad
- ☐ Very bad

What are the symptoms still bothering you?

- ☐ Diarrhea
- ☐ Joint pain
- ☐ Abdominal pain
- ☐ Bloating
- ☐ Cramps
- ☐ Fatigue
- ☒ None

You have Celiac disease , an autoimmune disorder. Visit a nutritionist who can help you plan a healthy gluten-free diet.



Reading can lower levels of cortisol, or other unhealthy stress hormones, by 67 percent.

## Advice

A strict, lifelong gluten-free diet is the only way to manage celiac disease. In addition to wheat, foods that contain gluten include:

- Barley
- Bulgur
- Durum
- Farina
- Graham flour
- Malt
- Rye
- Semolina
- Spelt (a form of wheat)
- Triticale

## B Study 3 Scenario: Rationales

<p>Hi I'm MediBot.ai, a medical diagnostic expert system. I would like to help you with your medical condition. Please click the screen to continue.</p>	<p>Which of the symptoms are bothering you most?</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Diarrhea</li><li><input checked="" type="checkbox"/> Bloating</li><li><input checked="" type="checkbox"/> Abdominal pain</li><li><input checked="" type="checkbox"/> Joint Pain</li><li><input type="checkbox"/> Itchy Rash</li><li><input type="checkbox"/> Head Cold</li><li><input checked="" type="checkbox"/> Cramps</li><li><input checked="" type="checkbox"/> Fatigue</li></ul>
<p>Do you have family history of any of the following conditions?</p> <ul style="list-style-type: none"><li><input type="checkbox"/> gastrointestinal malignancy</li><li><input type="checkbox"/> Irritable Bowel Syndrome (IBS)</li><li><input type="checkbox"/> Celiac disease</li><li><input type="checkbox"/> Crohn's</li><li><input checked="" type="checkbox"/> None</li></ul>	<p>Have you been exposed to natural water source recently?</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Yes</li><li><input type="checkbox"/> No</li></ul> <p>Have you been to woods for hiking or camping recently?</p> <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Yes</li><li><input type="checkbox"/> No</li></ul>
<p>Your symptoms are consistent with more than one medical conditions, but most people with these symptoms suffer from Irritable Bowel Syndrome (IBS). But I need to rule out some other things first before I can make a diagnosis. Due to your recent exposure to natural water sources, I need to check if you are suffering from Giardia.</p> <p>Please go through the following test and come back with result :</p> <ul style="list-style-type: none"><li><input type="checkbox"/> Giardia antigen test</li></ul>	<p>Follow up-1</p> <p>How are you feeling?</p> <ul style="list-style-type: none"><li><input type="checkbox"/> Great</li><li><input type="checkbox"/> Moderately great</li><li><input type="checkbox"/> Okay</li><li><input checked="" type="checkbox"/> Bad</li><li><input type="checkbox"/> Very bad</li></ul>

### Follow up-1

What are the symptoms that are bothering you now?

- ✓Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

### Follow up-1

You have come back with you reports. Giardia lamblia is not detected in your antigen test. You are not suffering from Giardia.

### Follow up-1

Since the possibility of Giardia has been ruled out, your symptoms and the duration of your condition indicates it is most likely that you are suffering from IBS. Please follow the advices below and come back two weeks later:

- Avoid chocolate, fried foods, alcohol, caffeine, carbonated drinks, the artificial sweetener, fructose
- Eat the soluble kind of fiber
- Drink plenty of water

### Follow up-2

Since you are back after two week, I would like to follow up your condition.

Are you following the IBS diet chart?

- ✓Yes
- No

### Follow up-2

How are you feeling after following the diet?

- ☐Great
- ☐Good
- ✓Moderate
- ☐Bad
- ☐Very bad

### Follow up-2

What are the symptoms that are bothering you now?

- Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

<p>IBS diet chart may take a while to adjust with your system. Continue the diet and come back for follow up next week</p>	<h3>Follow up-3</h3> <p style="text-align: center;">Welcome back!!!</p> <p>Since you are back after one week, I would like to follow up your condition.</p> <p>Are you still following the IBS diet chart?</p> <p><input checked="" type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p>
--	--

<p>How are you feeling after following the diet?</p> <p><input type="checkbox"/> Great</p> <p><input type="checkbox"/> Good</p> <p><input type="checkbox"/> Moderate</p> <p><input type="checkbox"/> Bad</p> <p><input checked="" type="checkbox"/> Very Bad</p>	<p>What are the symptoms that are bothering you now?</p> <p><input checked="" type="checkbox"/> Diarrhea</p> <p><input checked="" type="checkbox"/> Abdominal pain</p> <p><input checked="" type="checkbox"/> Joint Pain</p> <p><input checked="" type="checkbox"/> Bloating</p> <p><input type="checkbox"/> Cramps</p> <p><input type="checkbox"/> Cold</p> <p><input checked="" type="checkbox"/> Fatigue</p>
--	---

<p>IBS diet chart may take a while to adjust with your system. But I would like to cut gluten from your diet to see if it helps. Follow gluten free diet for a week. Come back next week and I will see if it helps improving your condition.</p>	<h3>Follow up-4</h3> <p style="text-align: center;">Welcome back!!!</p> <p>You are back after another week and I want to follow up your condition.</p> <p>Did you follow gluten-free diet this week?</p> <p><input checked="" type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p>
---	--

## Follow up-4

How are you feeling after following the diet?

- ☐ Great
- ☒ Good
- ☐ Moderate
- ☐ Bad
- ☐ Very Bad

## Follow up-4

Which of the symptoms are still bothering you

- ☐ Diarrhea
- ☒ Joint pain
- ☒ Abdominal pain
- ☐ Bloating
- ☐ Cramps
- ☐ Fatigue

## Follow up-4

Based on the positive effect of gluten-free diet, I have started to think that you may have Celiac disease. Since you do not have family history of Celiac, I did not consider it the most likely condition.

It's good that some of your symptoms have started improving after following gluten-free diet. It may take a few more days for all the symptoms to clear up. However I need you to go through some pathological tests for confirming my diagnosis.

- Go for a tTG-IgA Test (blood test)
- Biopsy of the small intestine
- Continue gluten-free diet
- Come back with the report

## Follow up-5

You are back after another week and I will assess your medical reports.

Your reports of tTG-IgA Test (blood test) and biopsy of the small intestine are positive for Celiac disease.

Are you still following gluten-free diet?

- ☒ Yes
- ☐ No

How are you feeling after following the diet?

- ☒ Great
- ☐ Good
- ☐ Moderate
- ☐ Bad
- ☐ Very bad

What are the symptoms still bothering you?

- ☐ Diarrhea
- ☐ Joint pain
- ☐ Abdominal pain
- ☐ Bloating
- ☐ Cramps
- ☐ Fatigue
- ☒ None

You have Celiac disease , an autoimmune disorder. People with celiac disease are unable to process proteins called gluten, which are found in foods containing wheat, rye, and barley, for example. It causes damage to small intestine. Visit a nutritionist who can help you plan a healthy gluten-free diet.

## Advice

A strict, lifelong gluten-free diet is the only way to manage celiac disease. In addition to wheat, foods that contain gluten include:

- Barley
- Bulgur
- Durum
- Farina
- Graham flour
- Malt
- Rye
- Semolina
- Spelt (a form of wheat)
- Triticale

## C Study 3 Scenario: Rationales + Visuals

Hi I'm MediBot.ai, a medical diagnostic expert system. I would like to help you with your medical condition.  
Please click the screen to continue.

Which of the symptoms are bothering you most?

- ☒ Diarrhea
- ☒ Bloating
- ☒ Abdominal pain
- ☒ Joint Pain
- ☐ Itchy Rash
- ☐ Head Cold
- ☒ Cramps
- ☒ Fatigue

Do you have family history of any of the following conditions?

- ☐ gastrointestinal malignancy
- ☐ Irritable Bowel Syndrome (IBS)
- ☐ Celiac disease
- ☐ Crohn's
- ☒ None

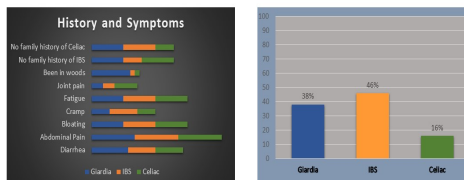
Have you been exposed to natural water source recently?

- ☒ Yes
- ☐ No

Have you been to woods for hiking or camping recently?

- ☒ Yes
- ☐ No

Your symptoms are consistent with more than one medical conditions, but most people with these symptoms suffer from Irritable Bowel Syndrome (IBS).



But I need to rule out some other things first before I can make a diagnosis. Due to your recent exposure to natural water sources, I need to check if you are suffering from Giardia.

Please go through the following test and come back with result :

- ☐ Giardia antigen test

Follow up-1

How are you feeling?

- ☐ Great
- ☐ Moderately great
- ☐ Okay
- ☒ Bad
- ☐ Very bad



### Follow up-1

What are the symptoms that are bothering you now?

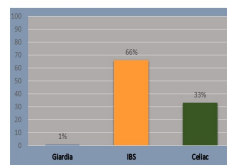
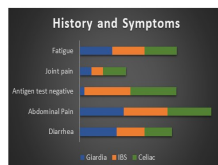
- ✓Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

### Follow up-1

You have come back with you reports. Giardia lamblia is not detected in your antigen test. You are not suffering from Giardia.

### Follow up-1

You have come back with you reports. Giardia lamblia is not detected in your antigen test. So I have to rule out the Giardia as your possible condition.



### Follow up-2

Since you are back after two week, I would like to follow up your condition.

Are you following the IBS diet chart?

- ✓Yes
- No

### Follow up-2

How are you feeling after following the diet?

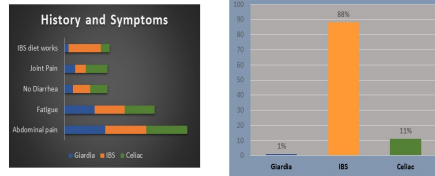
- ☐Great
- ☐Good
- ✓Moderate
- ☐Bad
- ☐Very bad

### Follow up-2

What are the symptoms that are bothering you now?

- Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

IBS diet chart may take a while to adjust with your system. Continue the diet and come back for follow up next week



### Follow up-3

Welcome back!!!

Since you are back after one week, I would like to follow up your condition.

Are you still following the IBS diet chart?

- ☒ Yes
- ☐ No

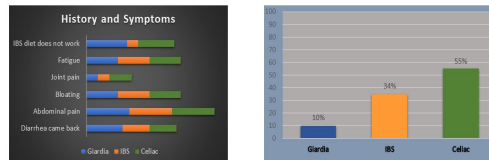
How are you feeling after following the diet?

- ☐ Great
- ☐ Good
- ☐ Moderate
- ☐ Bad
- ☒ Very Bad

What are the symptoms that are bothering you now?

- ☒ Diarrhea
- ☒ Abdominal pain
- ☒ Joint Pain
- ☒ Bloating
- ☐ Cramps
- ☐ Cold
- ☒ Fatigue

IBS diet chart may take a while to adjust with your system. But I would like to cut gluten from your diet to see if it helps. Follow gluten free diet for a week. Come back next week and I will see if it helps improving your condition.



### Follow up-4

Welcome back!!!

You are back after another week and I want to follow up your condition.

Did you follow gluten-free diet this week?

- ☒ Yes
- ☐ No

## Follow up-4

How are you feeling after following the diet?

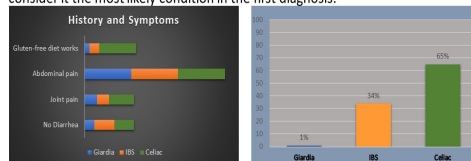
- ☐ Great
- ☒ Good
- ☐ Moderate
- ☐ Bad
- ☐ Very Bad

## Follow up-4

Which of the symptoms are still bothering you

- ☐ Diarrhea
- ☒ Joint pain
- ☒ Abdominal pain
- ☐ Bloating
- ☐ Cramps
- ☐ Fatigue

Based on the positive effect of gluten-free diet, I have started to think that you may have Celiac disease. Since you do not have family history of Celiac, I did not consider it the most likely condition in the first diagnosis.



It's good that some of your symptoms have started improving after following gluten-free diet. It may take a few more days for all the symptoms to clear up. However I need you to go through some pathological tests for confirming my diagnosis.

- Go for a tTG-IgA Test (blood test)
- Biopsy of the small intestine
- Continue gluten-free diet
- Come back with the report

## Follow up-5

You are back after another week and I will assess your medical reports.

Your reports of tTG-IgA Test (blood test) and biopsy of the small intestine are positive for Celiac disease.

Are you still following gluten-free diet?

- ☒ Yes
- ☐ No

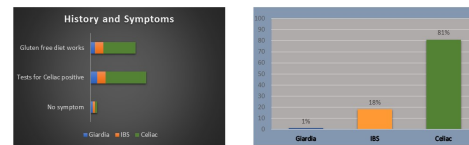
How are you feeling after following the diet?

- ☒ Great
- ☐ Good
- ☐ Moderate
- ☐ Bad
- ☐ Very bad

What are the symptoms still bothering you?

- Diarrhea
- Joint pain
- Abdominal pain
- Bloating
- Cramps
- Fatigue
- ✓ None

You have Celiac disease, an autoimmune disorder. People with celiac disease are unable to process proteins called gluten, which are found in foods containing wheat, rye, and barley, for example. It causes damage to small intestine. Visit a nutritionist who can help you plan a healthy gluten-free diet.



## Advice

A strict, lifelong gluten-free diet is the only way to manage celiac disease. In addition to wheat, foods that contain gluten include:

- Barley
- Bulgur
- Durum
- Farina
- Graham flour
- Malt
- Rye
- Semolina
- Spelt (a form of wheat)
- Triticale

## D Study 3 Scenario: Rationales + Examples

Hi I'm MediBot.ai, a medical diagnostic expert system. I would like to help you with your medical condition.  
Please click the screen to continue.

Which of the symptoms are bothering you most?

- ☒ Diarrhea
- ☒ Bloating
- ☒ Abdominal pain
- ☒ Joint Pain
- ☐ Itchy Rash
- ☐ Head Cold
- ☒ Cramps
- ☒ Fatigue

Do you have family history of any of the following conditions?

- ☐ gastrointestinal malignancy
- ☐ Irritable Bowel Syndrome (IBS)
- ☐ Celiac disease
- ☐ Crohn's
- ☒ None

Have you been exposed to natural water source recently?

- ☒ Yes
- ☐ No

Have you been to woods for hiking or camping recently?

- ☒ Yes
- ☐ No

Your symptoms are consistent with more than one medical conditions, but most people with these symptoms suffer from Irritable Bowel Syndrome (IBS). But I need to rule out some other things first before I can make a diagnosis. Due to your recent exposure to natural water sources, I need to check if you are suffering from Giardia.

Let's look at a similar case C30117:

Case C30117	Your Case
Patient had diarrhea, abdominal pain, bloating, cramps, fatigue	You have diarrhea, abdominal pain, bloating, cramps, fatigue
Patient had been exposed to natural water sources	You have been in woods and exposed to natural water sources
Patient went through Giardia antigen test and Giardia lamblia was detected in the test.	Please go through Giardia antigen test and come back with the result

### Follow up-1

How are you feeling?

- ☐ Great
- ☐ Moderately great
- ☐ Okay
- ☒ Bad
- ☐ Very bad

### Follow up-1

What are the symptoms that are bothering you now?

- ✓Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

### Follow up-1

You have come back with you reports. Giardia lamblia is not detected in your antigen test. You are not suffering from Giardia.

### Follow up-1

Since the possibility of Giardia has been ruled out, your symptoms and the duration of your condition indicates it is most likely that you are suffering from IBS.

Let's look at a similar case C30209:

Case C30209	Your Case
Patient had diarrhea, abdominal pain, bloating, cramps, fatigue	You have diarrhea, abdominal pain, bloating, cramps, fatigue
Patient had been exposed to natural water sources	You have been in woods and exposed to natural water sources recently
Giardia lamblia was not detected in the antigen test, so Giardia was ruled out.	Giardia lamblia is not detected in your antigen test, so Giardia is ruled out.
Patient was eventually diagnosed with IBS	Your symptoms indicates it is most likely IBS

### Follow up-2

Since you are back after two week, I would like to follow up your condition.

Are you following the IBS diet chart?

- ✓Yes
- No

### Follow up-2

How are you feeling after following the diet?

- ☐Great
- ☐Good
- ✓Moderate
- ☐Bad
- ☐Very bad

### Follow up-2

What are the symptoms that are bothering you now?

- Diarrhea
- ✓Abdominal pain
- ✓Joint Pain
- Bloating
- Cramps
- Cold
- ✓Fatigue

IBS diet chart may take a while to adjust with your system. Continue the diet and come back for follow up next week.

Let's look at a similar case C30209

Case C30209	Your case
Patient was diagnosed with IBS	You have been diagnosed with IBS
Patient had abdominal pain and fatigue even after one week of starting to follow IBS diet chart	You have abdominal pain and fatigue after one week of following IBS diet chart

### Follow up-3

Welcome back!!!

Since you are back after one week, I would like to follow up your condition.

Are you still following the IBS diet chart?

☒ Yes

☐ No

How are you feeling after following the diet?

☐ Great

☐ Good

☐ Moderate

☐ Bad

☒ Very Bad

What are the symptoms that are bothering you now?

☒ Diarrhea

☒ Abdominal pain

☒ Joint Pain

☒ Bloating

☐ Cramps

☐ Cold

☒ Fatigue

IBS diet chart may take a while to adjust with your system. But I would like to cut gluten from your diet to see if it helps. Follow gluten free diet for a week. Come back next week and I will see if it helps improving your condition. Let's look at a similar case C10635:

Case C10635	Your case
Patient had diarrhea, abdominal pain, joint pain, fatigue	You have diarrhea, abdominal pain, joint pain, fatigue
Patient was diagnosed with IBS	You have been diagnosed with IBS
Patient condition did not improve by following IBS diet chart, so I prescribe Gluten free diet	Your condition did not improve by following IBS diet chart for two weeks

### Follow up-4

Welcome back!!!

You are back after another week and I want to follow up your condition.

Did you follow gluten-free diet this week?

☒ Yes

☐ No

## Follow up-4

How are you feeling after following the diet?

- ☐ Great  
☒ Good  
☐ Moderate  
☐ Bad  
☐ Very Bad

## Follow up-4

Which of the symptoms are still bothering you

- ☐ Diarrhea  
☒ Joint pain  
☒ Abdominal pain  
☐ Bloating  
☐ Cramps  
☐ Fatigue

## Follow up-4

Based on the positive effect of gluten-free diet, I have started to think that you may have Celiac disease. Since you do not have family history of Celiac, I did not consider it the most likely condition. Let's look at case C107409 and compare your case with it:

Case 107409		Your case
Patient had diarrhea, abdominal pain, joint pain, fatigue	✓	You have diarrhea, abdominal pain, joint pain, fatigue
Patient reported about family history of Celiac disease in the first week of consultation	✗	You have no family history of Celiac disease
Patient was diagnosed with Celiac. With family history of Celiac, I choose Celiac in first diagnosis as the most likely one.	✗	You were diagnosed with IBS. Without family history of Celiac, I rule out Celiac in first diagnosis.

- Go for a tTG-IgA Test (blood test)
- Biopsy of the small intestine
- Continue gluten-free diet
- Come back with the report

## Follow up-5

You are back after another week and I will assess your medical reports.

Your reports of tTG-IgA Test (blood test) and biopsy of the small intestine are positive for Celiac disease.

Are you still following gluten-free diet?

- ☒ Yes  
☐ No

How are you feeling after following the diet?

- ☒ Great  
☐ Good  
☐ Moderate  
☐ Bad  
☐ Very bad



What are the symptoms still bothering you?

- Diarrhea
- Joint pain
- Abdominal pain
- Bloating
- Cramps
- Fatigue
- ✓None

You have Celiac disease , an autoimmune disorder. People with celiac disease are unable to process proteins called gluten, which are found in foods containing wheat, rye, and barley, for example. It causes damage to small intestine. Visit a nutritionist who can help you plan a healthy gluten-free diet. Let's look at a similar case C20974:

Case C20974	Your case
Patient had abdominal pain, diarrhea, joint pain, fatigue for few weeks	You had abdominal pain, diarrhea, joint pain, fatigue, bloating for few weeks
Patient was first diagnosed with IBS but IBS diet chart did not work	You were diagnosed with IBS first, IBS diet chart did not work for you
tTG-IgA Test (blood test) and biopsy of the small intestine were positive for the patient and diagnosed with Celiac eventually	Your tTG-IgA Test (blood test) and biopsy of the small intestine are positive for Celiac disease.

### Advice

A strict, lifelong gluten-free diet is the only way to manage celiac disease. In addition to wheat, foods that contain gluten include:

- Barley
- Bulgur
- Durum
- Farina
- Graham flour
- Malt
- Rye
- Semolina
- Spelt (a form of wheat)
- Triticale