Explanation Strategies for Trustworthy AI Diagnostic Systems: Examining Physicians' Explanatory Reasoning in Re-diagnosis Scenarios

Lamia Alam[1] and Shane T. Mueller[2]

Michigan Technological University, Houghton, MI 49931, USA
lalam@mtu.edu[1] shanem@mtu.edu[2]

Abstract

AI systems are increasingly being deployed to provide the first point of contact for patients. These systems are typically focused on question-answering, and suffer from many of the same deficiencies in explanation that have plagued medical diagnostic systems since the 1970s (Shortliffe, Buchanan, and Feigenbaum 1979). They provide information that patients or physicians may not need or would prefer to get in other ways. To provide better guidance about explanations in these systems, we report on an interview study in which we identified explanations that physicians used in the context of a re-diagnosis or a change in diagnosis. Five broad categories of explanation 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. We also present these in a diagnosis meta-timeline that identifies points at which we 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.

Introduction

Artificial Intelligence (AI) has the potential to revolutionize healthcare, and one potential area is initial diagnosis and first contact with patients. Researchers have known since the 1970s that transparency and explainability are necessary recurs to trustworthy systems, and this has been one of the largest impediments to the success of these systems (Clancey 1983) and it still remains one of the main challenges for these systems (Shaban-Nejad, Michalowski, and Buckeridge 2021).Without sufficient explanations, it is difficult for a physician or patient to understand how AI makes its decision, and thus whether to trust it. Research has suggested that many failures of AI systems in the medical domain stem from the lack of consideration of human issues in the design of the system (Patel et al. 2009). Nevertheless, Explainable AI (XAI) in healthcare is a nascent field (Lauritsen et al. 2019; Panigutti, Perotti, and Pedreschi 2020) that is not yet well-informed by physician-patient communication. In order to develop appropriate AI diagnostic explanations, it is important to understand the strategies physicians use to explain their diagnoses to their patients. Understanding physicians' explanatory reasoning may help AI developers create systems that make the patient-AI communication better, help patients comprehend the diagnosis process, and help physicians place trust in these systems as an aid for initial patient contact. To support this, we report on a study with physicians in which we identified explanation strategies during diagnosis. Based on these interviews, we will summarize themes of their explanation strategies to improve existing and future AI medical diagnostic systems and provide some design recommendations for patient-facing AI diagnostic systems.

Method

A more detailed description of the methods of this study are found in Alam (2020). We interviewed seven physicians with a variety of specialties and experience, with a focus on identifying incidents in which they made and changed diagnoses. We used an adapted Applied Cognitive Task Analysis (ACTA) technique (Crandall et al. 2006) to conduct incident-based interviews. What we can obtain from such qualitative analysis can be much different, and it is possible to analyze the data with relatively smaller samples. All methods were approved by the MTU institutional review board. Participants gave oral consent before the interview and agreed to have their interview audio recorded. Interviews were conducted either via phone/internet video or in-

 This study was conducted as part of a Master's thesis for LA at Michigan Technological University.

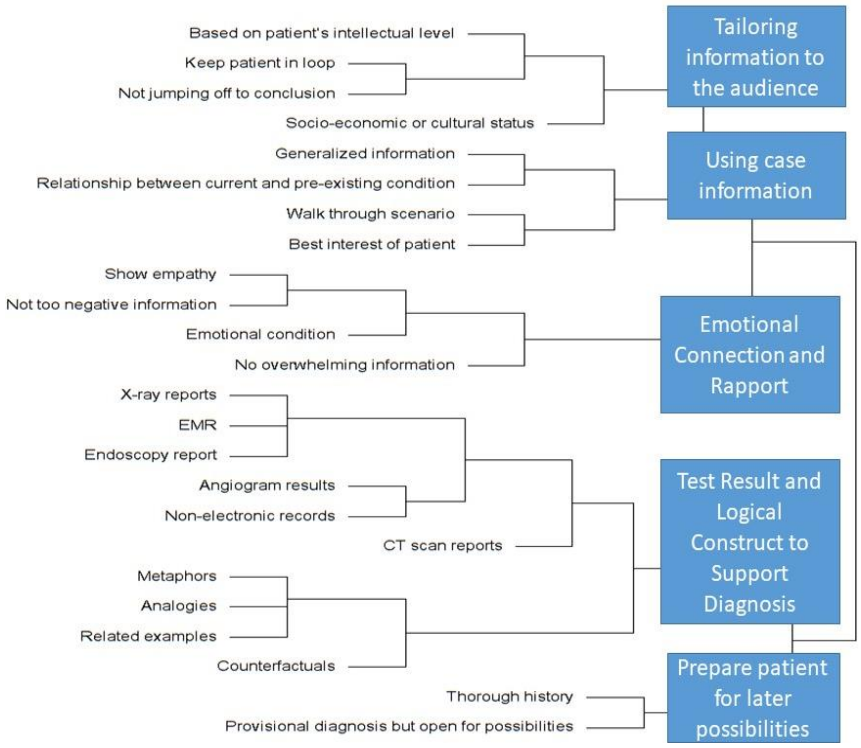

Figure 1: Hierarchical clustering for physician explanation strategies

person and lasted for 45-70 minutes. After initial background questions, we focused on 1-2 cases per physician that involved a 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.

Initial Coding

To analyze the interviews, we first isolated the explanations from the transcripts. We coded a statement as an explanation if it referred to some communication intended to help the patient understand a diagnosis. In a subset of two interviews, two independent raters identified each coded statement as either an explanation or non-explanation and achieved inter-rater reliability of $\kappa = .9$ and .88. Given the high agreement, a single rater coded the remaining interviews. We obtained 52 cases of explanation and mapped them into 24 categories of highly similar statements.

Card Sorting

Next, we conducted a card sort of the 24 categories. Five teams of students who were enrolled in graduate study at Michigan Technological University sorted the cards into 4-6 categories based on judged similarity. Each coding team derived their categories by consensus.

Hierarchical Clustering

We next used a hierarchical clustering approach, using as a dissimilarity measure the number of times any pair appeared in different themes across sorting teams. We then applied the *agnes* agglomerative clustering function in the *cluster* library (Maechler et al. 2013) of the R statistical computing language to compute a clustering hierarchy.

Results and Discussion

Five rough hierarchical themes emerged from the clustering analysis (see Figure 1), along with the 24 base codes. The similarity of a pair of themes is represented by the height of the branch that contains both themes. Next, we will briefly discuss each theme in turn.

Theme: Preparing Patients for Later Possibilities

Physicians often provided an initial 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 changes in diagnoses. Some AI systems do show probability distributions across different possibilities, which supports this same function. However, they are much less likely to do this in order

to create a narrative that will be followed up on later in the diagnosis.

Theme: Tailoring Information to the Audience

Physicians also reported that they often tailored their explanation to the individual, based on either socio-economic or cultural status, the intellectual level of their patients, their current emotional state, and other concerns, all of which were dependent on the patients and their ability to understand the information. XAI researchers have advocated for user models (Kass and Finin 1988), but tailoring can be done in simple ways as well, such as having the user select the complexity of the explanation they want.

Theme: Using Case Information

Physicians often generated their diagnoses over time using emerging information about the case. They then walked patients through the case scenario to help the patients understand the diagnoses. This diagnosis mode is similar to the explanation scripts initially explored by developers of expert diagnostic systems (see Clancey 1988), which would create text-based descriptions of the logical steps by which a diagnosis was reached. This mode of explanation is less common today in the XAI community, although Hendricks et al. (2016) demonstrated a way of using deep language models to generate similar explanations.

Theme: Using Test Results and Logical Constructs to Support Diagnosis

Physicians reported that they frequently used test results and medical records of the patients to support the diagnoses, which formed the basis for justifying and explaining diagnoses. The interviewees mentioned using an X-ray, CT scan, endoscopy, angiogram reports as visual aids, as well as medical records and test results (e.g. blood tests) to explain their diagnoses. Physicians also reported several higher-level strategies, including logical arguments, examples, analogies, metaphors, and counterfactuals to help patients understand a diagnosis. These correspond roughly to many forms of explanation being explored in the XAI research community. For example, the explanation given by a physician pointing out a critical area of an X-ray image is similar to the LIME algorithm highlights critical features in an image (see Ribeiro, Singh, and Guestrin 2016).

Theme: Build Emotional Connection and Rapport

Physicians often considered the emotional aspects of communication with the patient and their families. These were not always about providing explanations or information but involved empathetic strategies to ensure their patients knew the physicians listened and cared.

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

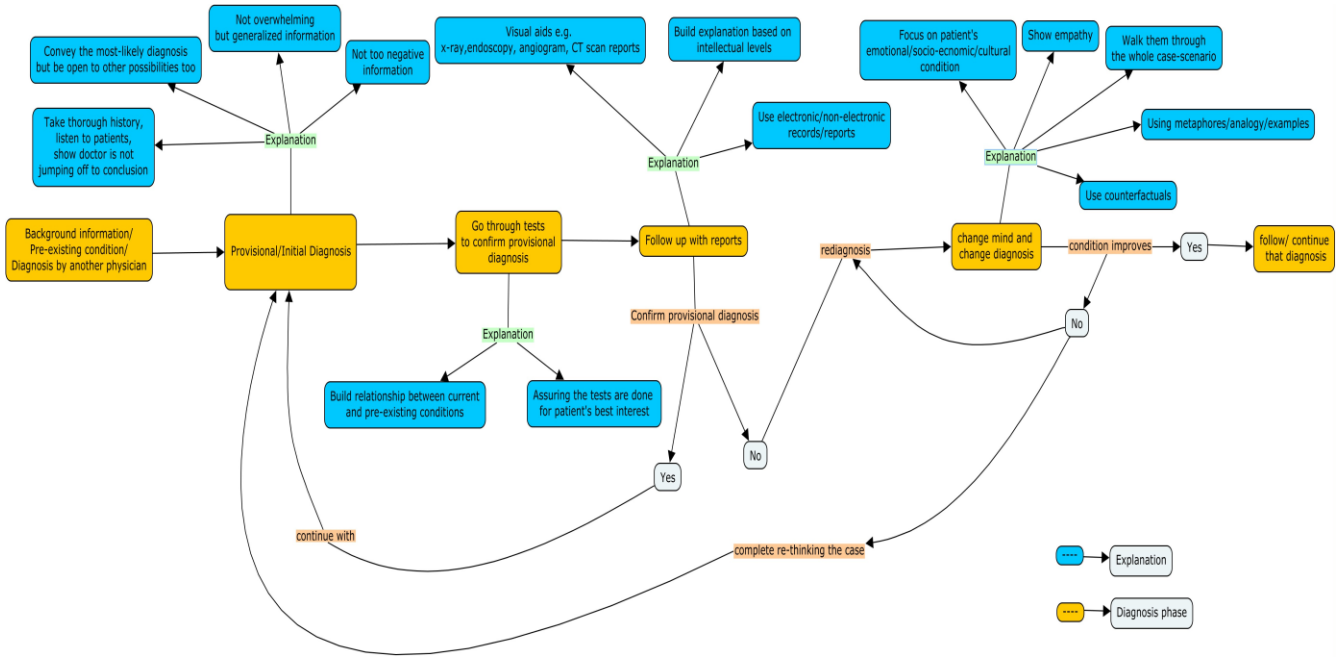

Figure 2: Meta-timeline for explanation in re-diagnosis scenarios

later explanations. Also, several physicians suggested at the beginning of the consultation, they did not want the patients to think about negative possibilities too much.

This may be an area where AI will never match the empathetic abilities of human diagnosticians. Nevertheless, researchers have investigated ways in which we treat computers as social actors (Lee and Nass 2010), which suggests that it may be possible to build social and emotional rapport between a human and a machine.

Diagnostic Meta-timeline of Explanation

The themes we presented show that there are many kinds of explanation used by physicians in the diagnostic process. Another outcome of our interviews is that these different explanations occur at different times. To help identify the typical points at which explanations emerge during diagnosis, we developed a generic diagnosis meta-timeline (see Figure 2) of explanation based on our interviews. This is a basic framework that encapsulates many of the commonalities we saw across diagnoses during the interviews. Although an AI researcher may be able to use this timeline as a basic flowchart for designing automated diagnostic systems, we see it more as a way of characterizing the explanations we observed at different times in diagnostic processes. During the initial phase of diagnosis, physicians conveyed the most-likely diagnosis to their patients, mentioned other possibilities, attempted to give general information, gained emotional rapport, and avoided 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.

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 might continue until the conditions of the patient improved, or the physician decided to reassess the symptoms from ground zero.

Considerations for Building Trustworthy Explanatory Diagnostic System

The first generation of AI medical diagnostic systems based on the 1980s expert systems framework failed. Many observers at that time rightly pointed to a lack of explainability as one of their main weaknesses, which led to the birth of the Explainable AI movement. Yet explanations in those systems were relatively simple to identify, as they came directly from human-generated rules. Today's diagnostic systems are becoming more difficult to understand, making explanations even more necessary. But the current XAI approaches remain algorithm-focused, without accounting for or modeling the explanation patterns of human physicians. Thus, the present study helps identify some of the goals and methods of explanation among human diagnosticians.

The explanation strategies and methods we identified in this study reveal that building good explanations for diagnosis and re-diagnosis scenarios requires the clarification of the symptoms and medical conditions as well as understanding the emotional, cultural, intellectual, socio-economic status of the patients. Expert human physicians often apply these approaches.

This study suggests several tangible pieces of design advice for AI researchers hoping to create usable diagnostic systems that will be informative for patients or physicians:

- **Tailor Explanations to the patient.** One theme that emerged from this study is that physicians often tailored their explanations considering the need of different patients. The need for user models and personalization of explanation in the AI systems has been discussed (Kass and Finin 1988; Weiner 1989; Miller 2019), but this is a difficult problem with no clear and easy solution.

- **Tailor Explanations During Diagnosis.** AI systems should not consider an explanation one single event. It unfolds throughout a diagnosis and may take different forms at different time points. 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. Moreover, initial explanations might be simplified if the patients or the family members are under stress, with more complex information waiting until later, when patients are stabilized, or their families are calmer.

- **Consider multiple forms of explanation.** Most researchers focus on developing and validating a single kind of explanation system. Physicians are not restricted in this way, and use many kinds, from logical arguments to visualizations to examples, and analogies. An explanatory diagnostic system must be prepared to do this as well, offering a variety of things that serve as explanations. Along with small explanations along the way, physicians often create a completely logical argument that draws on all existing information to help the patient

understand why a diagnosis is being made, why other diagnoses are not, what treatment options exist, and what will be done if the current diagnosis fails.

Current AI approaches to explanation ignore many of the human needs for explanation and fail to address many of the explanation modes we identified in this study. Consequently, this research suggests the importance of several areas of explanation when developing AI medical diagnosis systems. Some of the things we discovered really are the core of XAI (Hoffman et al. 2018; Mueller et al. 2019)- use of examples, counterfactuals, visual aids. But these are the pieces of explanation, it is not the whole explanation that patients need. AI systems in healthcare need to put it together at the right time, tailoring it for different patients at different points of diagnosis to ensure proper utilization of these systems.

References

Alam, L. 2020. Investigating the Impact of Explanation on Repairing Trust in Ai Diagnostic Systems for Re-Diagnosis (Publication No. 28088930) [*Master's Thesis, Michigan Technological University*]. ProQuest Dissertations Publishing.

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. 1988. The knowledge engineer as student: Metacognitive bases for asking good questions. In *Learning issues for intelligent tutoring systems* (pp. 80–113). Springer. https://link.springer.com/chapter/10.1007/978-1-4684-6350-7_5

Crandall, B., Klein, G., Klein, G. A., and Hoffman, R. R. 2006. *Working minds: A practitioner's guide to cognitive task analysis*. Mit Press.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and 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

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *ArXiv Preprint ArXiv:1812.04608*.

Kass, R., and Finin, T. 1988. The Need for User Models in Generating Expert System Explanation. *Int. J. Expert Syst.*, *1*(4), 345–375.

Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., and Thiesson, B. 2019. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *ArXiv Preprint ArXiv:1912.01266*.

Lee, J.-E. R., and Nass, C. I. 2010. Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives* (pp. 1–15). IGI Global.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., and Studer, M. 2013. Package 'cluster.' *Dosegljivo Na*.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv Preprint ArXiv:1902.01876*.

Panigutti, C., Perotti, A., and 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.

Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., and Abu-Hanna, A. 2009. The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine*, *46*(1), 5–17. https://doi.org/10.1016/j.artmed.2008.07.017

Ribeiro, M. T., Singh, S., and 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

Shaban-Nejad, A., Michalowski, M., and Buckeridge, D. L. Explainability and Interpretability: Keys to Deep Medicine. *In Explainable AI in Healthcare and Medicine (pp. 1-10)*. Springer, Cham.

Shortliffe, E. H., Buchanan, B. G., and Feigenbaum, E. A. 1979. Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. *Proceedings of the IEEE*, *67*(9), 1207–1224.

Weiner, J. L. 1989. The effect of user models on the production of explanations. *Expert Knowledge and Explanation: The Knowledge-Language Interface*, 144–156.