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

Lamia Alam, Shane T. Mueller
Michigan Technological University
Contact: lalam@mtu.edu

Michigan Tech
Presented at the AAAI 2021 Workshop: Trustworthy AI for Healthcare
February 8- 9, 2021

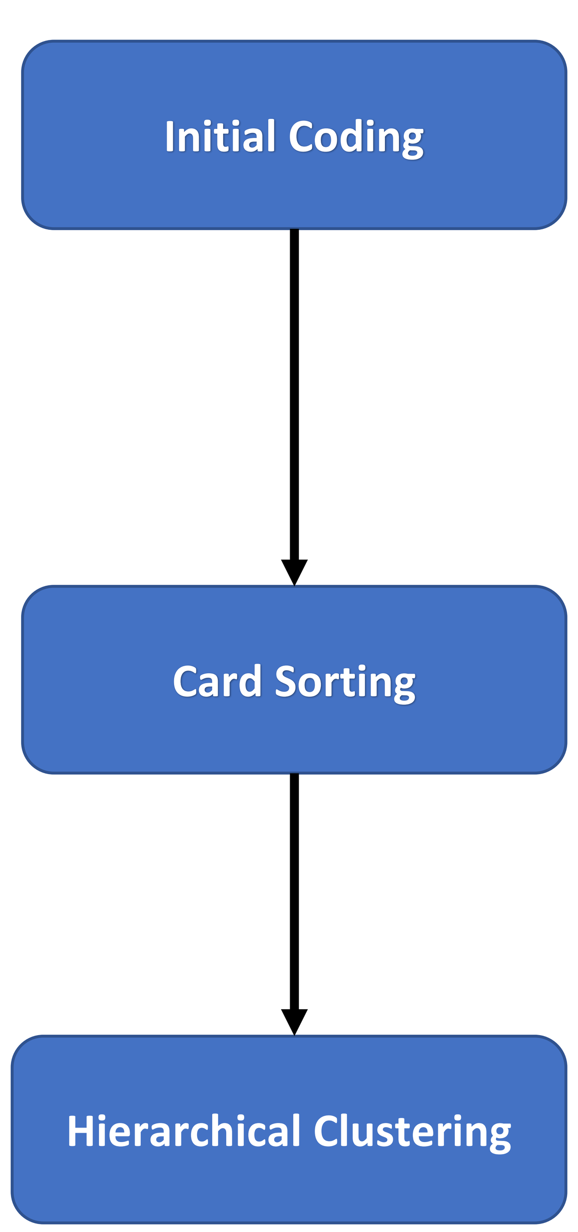
Abstract

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.

Method

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. Interviews were conducted either via phone/internet video or in-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.

Qualitative Analysis



Isolated the explanations from the transcripts and 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 κ= .9 and .88. Given the high agreement, a single rater coded the remaining interviews. 52 cases mapped into 24 categories of highly similar statements.

Five teams of students 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.

Used 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

Figure 1: *Hierarchical clustering for physician explanation strategies*

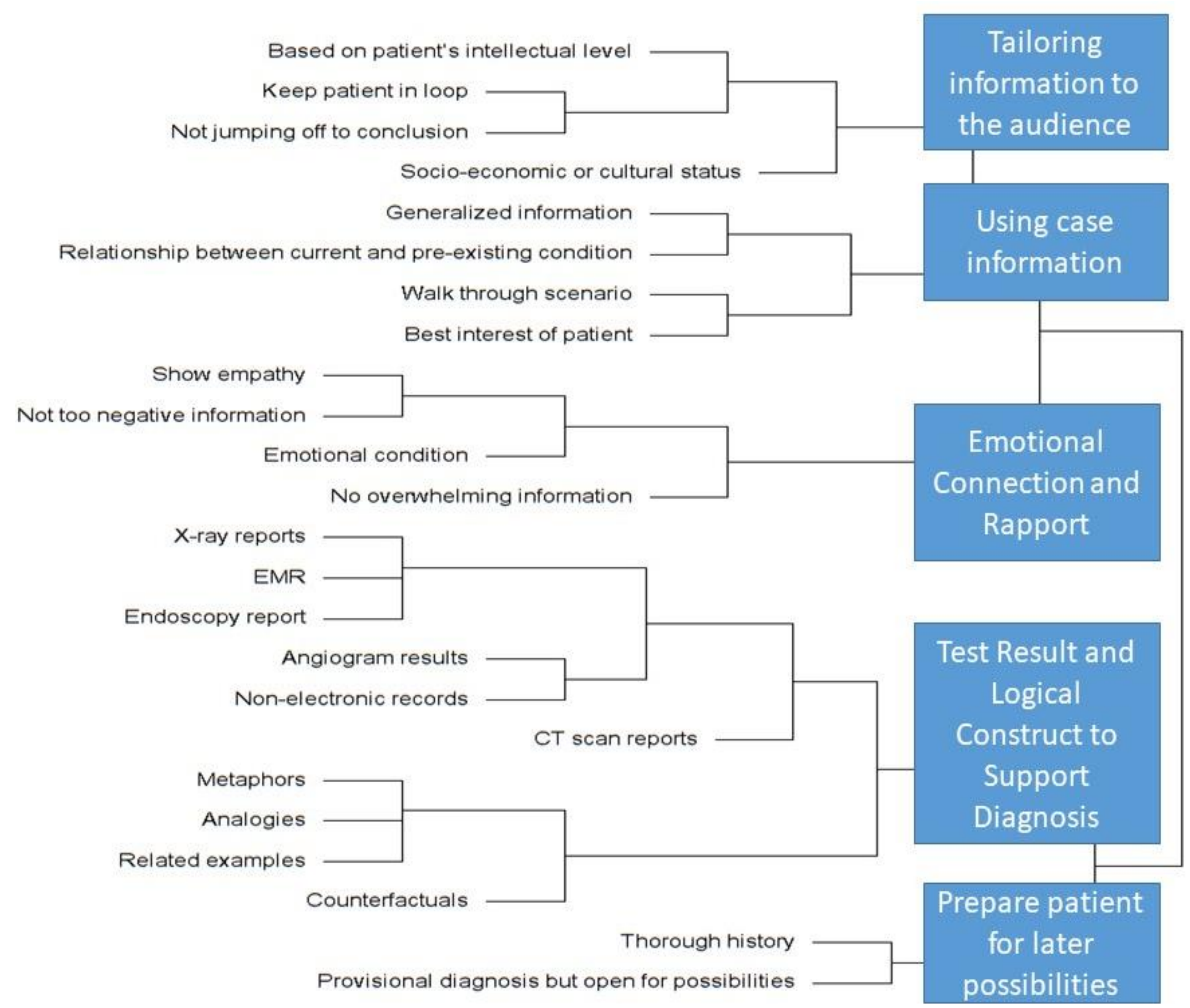
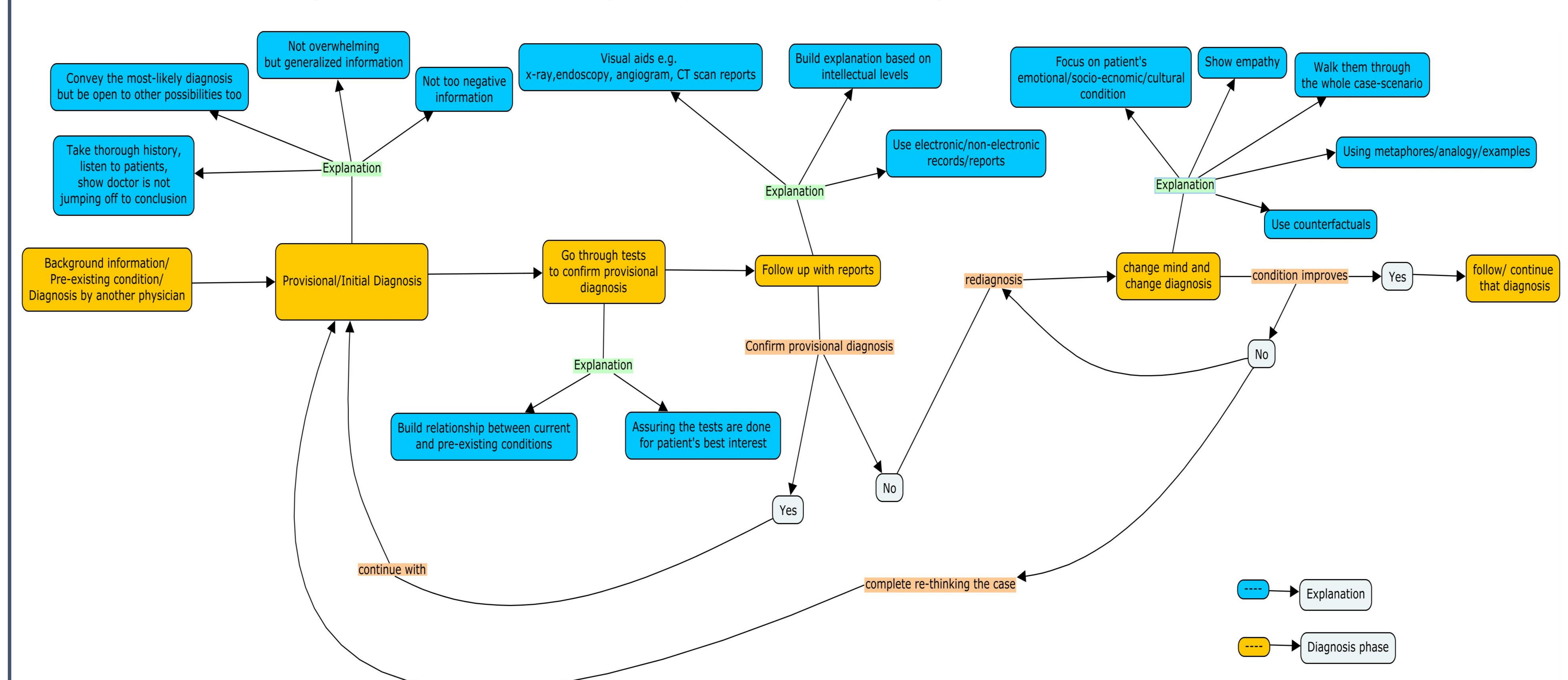


Figure 2: *Meta-timeline for explanation in re-diagnosis scenarios*



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.

Discussion

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.

Design Recommendation

1. Tailor Explanations to the patients

2. Tailor Explanations During Diagnosis

3. Consider Multiple Forms of Explanation

References

Crandall, B., Klein, G., Klein, G. A., and Hoffman, R. R. 2006. *Working minds: A practitioner's guide to cognitive task analysis*. Mit Press.
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., and Studer, M. 2013. Package 'cluster.' *Dosegljivo Na*.
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.