Designing for AI Explainability in Clinical Context

Daniel Gruen, PhD[1], Shruthi Chari, MS[1], Morgan A. Foreman, BS[2], Oshani Seneviratne, PhD[1], Rachel L. Richesson, PhD[3], Amar K. Das, MD, PhD[2], Deborah L. McGuiness, PhD[1]

[1]Rensselaer Polytechnic Institute, Troy, NY; [2]IBM Research, Cambridge, MA;
[3]University of Michigan, Ann Arbor, MI

Abstract

The growing use of artificial intelligence in medical settings has led to increased interest in AI Explainability (XAI). While research on XAI has largely focused on the goal of increasing users' appropriate trust and application of insights from AI systems, we see intrinsic value in explanations themselves (and the role they play in furthering clinician's understanding of a patient, disease, or system). Our research studies explanations as a core component of bi-directional communication between the user and AI technology. As such, explanations must be understood and evaluated in context, reflecting the specific questions and information needs that arise in actual use. In this paper, we present a framework and approach for identifying XAI needs during the development of human-centered AI. We illustrate this approach through a user study and design prototype, which situated endocrinologists in a clinical setting involving guideline-based diabetes treatment. Our results show the variety of explanation types needed in clinical settings, the usefulness of our approach for identifying these needs early while a system is still being designed, and the importance of keeping humans in the loop during both the development and use of AI systems.

Introduction

The rapid growth of Artificial Intelligence (AI) in healthcare is built on the promise that AI can improve patient care and clinical practice (Matheny et al. 2019). The uptake of AI in healthcare, however, largely depends on usability, safety, workflow, and governance (Reddy et al. 2020; Shortliffe 2019). In particular, transparency and explainability have been identified as two necessary characteristics of AI systems in healthcare (Biran and Cotton 2017). Transparency refers to an understanding of the system's operation as a whole, including factors such as how it operates, how it was trained and on what data, how it has been tested, what knowledge it understands, how robust it is, where it has been shown to work well and where not (Adadi and Berrada 2018). Explainability refers to the ability of a system to provide information on how a *specific* result was obtained, including justifications of how the result makes sense and fits in with other knowledge (Chari et al. 2020a; Hoffman, Klein and Mueller 2018). Traditionally, explanations are provided by the system to help users evaluate and apply results, understand when the AI technology should be trusted, and in which situations they may be less accurate. Explanations also help ensure fairness, helping users make sure that only ethically justifiable considerations influence results and recommendations (Biran and Cotton 2017).

Recent approaches to AI system development use a *Human in the Loop (HITL)* framework that allows the user to change, correct, or update the system, with the system able to respond with new results (Holzinger 2016; Zanzotto 2019). In an ideal HITL system, AI technology works closely with human collaborators to construct a shared model of a situation and to jointly consider positive and negative solutions for a task, each drawing from their own abilities and knowledge. This perspective draws from the Distributed Cognition view, in which cognition is seen to take place not within the head of any one individual, but rather through the exchange and transformation of representations across multiple actors and artifacts (Hollan, Hutchins and Kirsh 2000).

To be effective partners in distributed cognition, each agent (human and otherwise) must be able to share information each possesses and proposed solutions to the problem at hand, as well as their rationales for solutions, considerations and concerns. Systems must therefore be constructed so as to empower users to question results and suggest competing hypotheses to be explored and evaluated together.

Within this framework, explanations have deeper value beyond the role they play in helping users determine which results should be trusted and applied; they contribute to the richness of the interaction between the various actors in the overall cognitive system. The design of explanations within medical AI systems must therefore be guided by understanding the value they will provide when addressing real clinical problems in context, taking into account what the users already know as well as the collaborative interactions they will enable.

Lim, Wang and others have echoed these ideas in their user-centered framework for designing explanations for the

context of use and the user's cognitive needs (Lim et al. 2019; Wang et al. 2019). Other work proposed a general question/answer-based approach to assist designers in determining what explanations a given AI system should be able to provide (Liao, Gruen and Miller 2020).

In this paper, we show how a user-centered design approach that situates users in actual contexts of use is critical to uncovering the types of explanations a HITL AI system will need to be able to provide and learn, illustrated through an example involving clinical decisions around diabetes treatment.

Methods

To assess the various needs for explainability in a specific situation, we referenced a previously developed taxonomy of explanation types, drawn from literature in computer science, social sciences and philosophy (Chari et al. 2020a). This is summarized in Table 1.

Type	Description
Case based	Provides solutions based on actual prior cases that support the system's conclusions, and may involve analogical reasoning, relying on similarities between features of the case and of the current situation.
Contextual	Refers to information about items other than the explicit inputs and output, such as information about the user, situation, and broader environment that affected the computation.
Contrastive	Answers the question "Why this output instead of that output," making a contrast between the given output and the facts that led to it and an alternate output of interest and facts that would have led to it.
Counterfactual	Indicates what solutions would have been obtained with different inputs.
Everyday	Uses accounts that appeal to users and their general commonsense knowledge
Scientific	References the results of rigorous scientific methods, observations, and measurements (evidence) or underlying mechanisms of action (mechanistic).
Simulation based	Uses an imitation of a system or process and the results that emerge from similar inputs.
Statistical	Relates to the likelihood of the outcome based on data about the occurrence of events under specified (e.g., experimental) conditions.
Trace based	Provides information on the underlying sequence of steps used by the system to arrive at a specific result.

Table 1: A taxonomy of explanation types.

As an example case, we explored the potential for computational support and the resulting need for explainability in a future AI system aimed at supporting clinical decisions around relatively new secondary treatments for type 2 diabetes. Diabetes is a common condition familiar to clinicians, yet new treatment options and guidelines can present challenges with which AI potentially could help. Our approach consisted of three phases: (1) an interview with a panel of expert endocrinologists to understand the general role of guidelines in their clinical practice and any challenges they experience in using guidelines, (2) the development of a prototype design for a system that could address issues they raised and provide rationales for its recommendations, and (3) a subsequent walkthrough and review of the prototype to evaluate explanations generated by the system and surface situations in which users would want additional explanations or could provide rationales to the system.

Phase 1: Expert Panel Session

Three experienced endocrinologists were interviewed together using a semi-structured interview format. We inquired about their use and impressions of guidelines, and any concerns they had about applying guidelines to specific patients, such as concerns over differences between a patient and the cohorts in the studies on which the guidelines were based. We probed specifically about decisions and concerns related to newer diabetes treatments mentioned in the guidelines, what factors might lead them to question their use for a specific patient, and how they would determine how to proceed. We asked if they had the information they needed to make these determinations, and what other information could be useful. We also asked about ways technology could assist their decision making, and what they would need to know before trusting a new tool. The session was conducted using remote screen-sharing and recorded. A thematic analysis was conducted on the results of the expert panel session.

Phase 2: Prototype Design

We created a rough mockup showing the start of a possible AI system, based on what we learned during the panel session. This consisted of various screens (Fig. 1) including a profile screen with basic personal information, a summary screen of the patient's overall medical information, a timeline screen with notes, vitals, test results and medications specifically relevant to her diabetes, and an insights screen with guideline-based treatment options and the factors that were considered to arrive at those conclusions. On the insights screen, treatment options could be clicked on to reveal a pop-up screen showing the guideline-based decision path followed for that class of treatment and a list of specific medications with the option to request more information about each. The prototype was meant to serve as a foil to prompt feedback on the overall usefulness of features, the

importance of explanations of different types, and what additional information would be useful to have. We also wished to probe places where they would want to tell or teach the system something, such as to highlight additional factors about the patient it should consider or issues it was not considering.

We populated our prototype with data for a fictional type 2 diabetes patient, based on one used as a pharmacological training example (https://slideplayer.com/slide/12380430/). We added information and adjusted details to make the patient and edge case with some complexity and for whom questions might arise on how specific guidelines would apply.

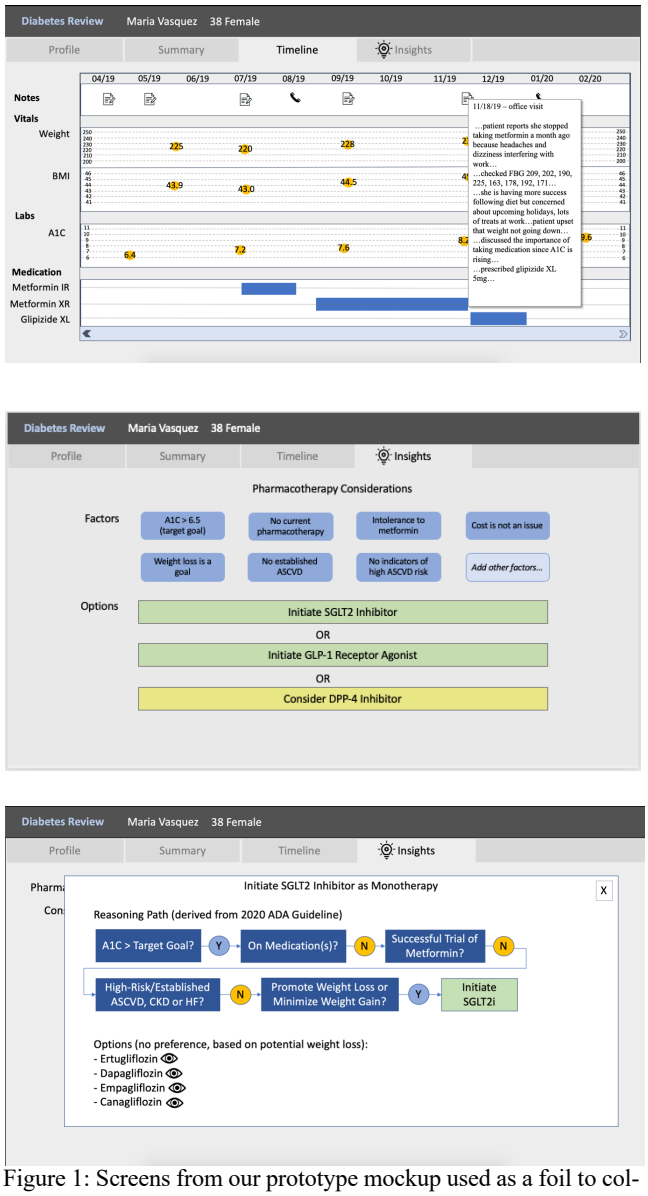

Figure 1: Screens from our prototype mockup used as a foil to collect requirements from the expert subjects.

Phase 3: Guided Walkthrough

We presented the prototype to two endocrinologists from our initial panel in individual sessions. We asked them to imagine that they were reviewing a summarized case history and insights from an AI system, with the goal of making a treatment recommendation for a patient that had been referred to them. We asked them to speak aloud and describe what they were thinking, as if instructing a medical student. We went through each screens, navigating and clicking on items as requested by the clinicians. We also asked directly about each screen. Questions we covered included: Is all the information on it useful to have? Is there anything missing? Is it displayed in a way that is useful or should it be shown differently? Are there other things they would want to know about how the information was obtained? We asked if there was anything they would want to tell the system so it could be more useful for this and future cases. We probed about the set of factors the system had identified on the Insights screen and if there were others they would want the system to include. We ended by asking what they would ultimately recommend for the patient.

Results

The three expert endocrinologists on our panel practiced at clinical sites affiliated with the Duke University School of Medicine. While the endocrinologists all focused on treating more complex patients with diabetes, they represented a diversity of practice regarding the use of guidelines to prescribe emerging diabetes treatments to their patients. One clinician was directly involved in the development of guidelines from literature, while another was skeptical about the extent to which they could be applied to the indigent population he treated.

The endocrinologists invoked specialized knowledge in diabetes management, and thus often deal with complex cases not managed well by primary care physicians. As one said, *"We see people who are not responding as expected."*

Role for AI

The endocrinologists on our panel saw specific value for AI technologies for general practitioners without their specialized knowledge and when dealing with new medications and/or new guidelines. For example, saying it is *"hard to shift thinking with new medication because of side effects and unknown side effects. Easier to stay with what you know,"* and a challenge to know *"what is it offering that we don't already have, and if it is offering something new, then I look at risk benefit ratios."*

There was some disagreement on the extent to which guidelines should be adhered to in practice, with one saying that they: *"try to stay with guideline unless there is any reason not to,"* while another physician *"feels serious*

limitations with CPGs because so many people don't fit guidelines." Assistance with determining when guidelines would and would not apply was seen as valuable. Endocrinologists also saw value in the system's proposed ability to identify within notes snippets of information relevant to the diabetes treatment decision.

Rationales and Explanation Types

We transcribed conversations during the walkthrough as endocrinologists reviewed the prototype and worked to understand the patient and determine and justify a treatment plan**.** From the transcripts, we identified rationales, namely instances in which explanations were given to support an assertion, recommendation or decision. These included those that the endocrinologists mentioned in questioning insights provided by the system or when discussing information and explanations they would have liked the system to provide. The clinicians provided their recommendations on the patient case and discussed the reasons behind them. These supported, contradicted, or added to the recommendations and explanations provided in the prototype.

In all, we identified 43 instances of rationales, and categorized them in terms of the scheme shown above. For example, we identified 5 examples of **contrastive** explanations (*"[choose a GLP-1 class] because a DPP4 isn't going to be enough for her"*); and 4 examples of **counterfactual** ones *(what if the patient actually were not metformin intolerant"* or *what if the patient were to become pregnant.).* We saw 4 examples of **mechanistic** explanations (*This class of drugs, "in a case such as this …will make her more hungry and lead to further weight gain.")* **Case-based** explanations, in which a specific prior case is referenced, were only seen as valuable in very rare, atypical situations.

In addition to the previously identified categories, we noted 16 instances in which physicians referred to general treatment principles and lessons learned from experiential knowledge. We classified such rationales as **"clinical pearls"** (Lorin et al. 2008), the term used to reference a well-known practice in medicine of crystalizing bits of information, rules or heuristics to be taught explicitly and shared among practitioners. For example, one clinical pearl involved fears of an increased risk of fungal infection in an overweight patient with hyperglycemia, learned from years of experience with similar patients and medications.

Need for Human Input

We saw multiple situations in which the clinicians would want to interact with the AI system to ask questions, obtain explanations or explore alternative scenarios. In addition, there were many instances in which the clinicians would want to provide information to the system, including identifying factors they noticed in the patient record that the system should have included, or correcting input assumptions they felt were inaccurate. These included questioning whether the patient were truly intolerant to a particular drug based on a relatively brief prior experience with it, or even whether the patient's rapid changes were consistent with their current diagnosis of Type 2 Diabetes. Clinical pearls represented a clear example of situations in which the clinicians would want to be able to teach the system explicit lessons from their experience, and in turn have such lessons presented to them when relevant, much as they would do when teaching medical students or sharing information and learning from colleagues.

Design Iteration

Lessons from our situated review were used to drive changes to the design, to support the sharing of rationales and requests for explanations we uncovered. For example, we created a mechanism for users to add factors for consideration and toggle them on or off to explore counterfactuals, while ensuring they weren't confused with the true patient information. We also added a similar mechanism to ask for contrastive explanations for alternate outputs, as well as adding support for clinical pearls and other explanation types.

We are currently reviewing our updated designs with practitioners, and are implementing a working prototype connected to an experimental clinical reasoning system capable of providing explanations of various kinds.

Discussion

In this paper, we present a methodological framework for identifying specific needs for explanations required to build an effective HITL AI system. We show the value of a three-phase process, including a panel discussion to identify needs for explanations, creation of a rudimentary incomplete prototype, and the use of the prototype as a probe to understand explanation needs in the context of specific usage situations. In particular, the use of an *edge case*, which included attributes that drew into question the applicability of the guidelines for that specific patient, revealed opportunities for human users to inform a guideline-based system about factors that could influence its reliability, and which should be taken into consideration in the future.

Our work extends prior research on AI explainability (Chari et al. 2020a; Liao, Gruen and Miller 2020; Wang et al. 2019) and demonstrates the pragmatic value of using an iterative human-centered design approach. Applying an HITL AI framework provides guidance on how to enrich a system's capability to generate clinically relevant explanations. This includes ongoing work (Chari et al. 2020b) to ensure capabilities exist to represent a range of explanation types as system outputs and inputs, both for their immediate value to a particular decision and for the larger educational role they play in enabling knowledge sharing among humans and AI systems.

Acknowledgements

This work is partially supported by IBM Research AI through the AI Horizons Network. We are indebted to the time, expertise and insights of Susan Spratt, MD., Jennifer B. Green, MD, and the late Mark N. Feinglos, MD, all of Duke University. This work is dedicated to the memory of Dr. Feinglos, who sadly passed away as results were being analyzed. We thank our colleague from IBM Research, Ching-Hua Chen, and from RPI, Rebecca Cowan, who assisted in the document preparation.

References

Adadi A, Berrada M, 2018. Peeking inside the Black-Box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access; 6:52138-52160.

Biran O, Cotton C, 2017. Explanation and justification in machine learning: a survey. IJCAI-17 Workshop on Explainable AI (XAI).

Chari S, Gruen DM, Seneviratne O, McGuinness DL. Directions for explainable knowledge-enabled systems, 2020. In: Tiddi I, Lecue F, Hitzler P (eds.), Knowledge Graphs for eXplainable AI Foundations, Applications and Challenges. Amsterdam: IOS Press; pp 245 – 261.

Chari S, Seneviratne O, Gruen DM, Foreman MA, Das AK, McGuinness DL. Explanation Ontology: A Model of Explanations for User-centered AI. Intl. Semantic Web Conf. 2020, pp 228 – 243.

Hoffman RR, Klein G, Mueller ST, 2018. Explaining explanation for "Explainable AI". Proc Human Factors and Ergonomics Society Annual Meeting., pp. 197-201.

Hollan J, Hutchins E, Kirsh D, 2000. Distributed cognition: toward a new foundation for human-computer interaction research. ACM Trans. Comput.-Hum. Interact.; 7:174-196.

Holzinger A., 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform.; 3:119-131.

Liao QV, Gruen DM, Miller S, 2020. Questioning the AI: toward design practices for explainable AI user experiences. ACM CHI Conference on Human Factors in Computing Systems (CHI), in press.

Lim, B.Y., Yang, Q., Abdul, A., & Wang, D, 2019. Why these explanations? selecting intelligibility types for explanation goals. ACM IUI Workshops.

Lorin MI, Palazzi DL, Turner TL, Ward MA, 2008. What is a clinical pearl and what is its role in medical education?. Medical Teacher.; 30:870-874.

Matheny M, Thadaney IS, Ahmed M, Whicher D, 2019. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington DC: National Academy of Medicine.

Reddy S, Allan S, Coghlan S, Cooper P, 2020. A governance model for the application of AI in health care. J Am Med Inform Assoc. Mar 1;27:491-497.

Shortliffe EH, 2019. Artificial Intelligence in medicine: weighing the accomplishments, hype, and promise. Yearbook Med Inform.; 28: 257-262.

Wang D, Yang Q, Abdul A, Lim BY, 2019. Designing theory-driven user explainable AI. ACM CHI Conference on Human Factors in Computing Systems (CHI 19). Paper 601, pp. 1-15.

Zanzotto FM, 2019. Human-in-the-Loop AI. J AI Res.; 64:243-252.