Purpose, Process, Performance: Designing for Appropriate Trust of AI in Healthcare

Position Paper

Natalie C. Benda, PhD

Weill Cornell Medicine, ncb4001@med.cornell.edu

Carrie Reale, MSN, RN-BC

Vanderbilt University Medical Center, Carrie.Reale@vumc.org

Jessica S. Ancker, PhD, MPH

Vanderbilt University Medical Center, Jessica.S.Ancker@vumc.org

Jessica Ribeiro, PhD

Florida State University, jribeiro@fsu.edu

Colin G. Walsh, MD, MA

Vanderbilt University Medical Center, Colin.Walsh@vumc.org

Laurie Lovett Novak, PhD, MHSA

Vanderbilt University Medical Center, Laurie.L.Novak@vumc.org

Fostering trust has been a challenge since the early days of artificial intelligence. Yet, some still conceptualize trust as a binary construct where the goal is to simply increase human trust in artificial intelligence. This view can result in overreliance in artificial intelligence, which can be harmful to patients and health professionals. Here, we describe the importance of determining and conveying the purpose, process, and performance of the AI to users so that they may trust AI appropriately, as opposed to absolutely. We support our position with results from two pre-implementation qualitative studies undertaken to determine stakeholder needs for artificial intelligence systems that 1) proactively identify high-need, high-cost patients 2) predict suicide attempts in an active-duty military clinic setting.

CCS CONCEPTS • artificial intelligence • trust • healthcare

ACM Reference Format:

Benda, N.C., Reale, C., Ancker, J.S., Ribiero, J., Walsh, C.G., Novak, L.L. 2021. Purpose, Process, Performance: Designing for Appropriate Trust of AI in Healthcare.

1 INTRODUCTION

The healthcare industry has long touted the potential of artificial intelligence (AI). AI has allowed for the development of highly accurate tools with application in various clinical settings.[4] Compared to the number of AI systems developed and validated, however, few are in use in clinical settings, and demonstrating the ability of AI to meaningfully improve health outcomes has been slow.[1; 5]

One issue in healthcare lies in the conceptualization that trust in AI is binary construct where the goal is to simply increase human trust in AI.[2] In a seminal review from the human factors literature, Lee and See describe trust in AI instead as a complex, dynamic entity, where the goal is to foster *appropriate* trust based on the purpose of the AI, its process for making recommendations, and its performance. Inappropriate trust in AI in a healthcare setting is problematic as it may lead to overreliance. Absolute reliance or overreliance in AI can results in patient harm if health professionals enact actions when the AI is inaccurate. In addition, overreliance can lead to "de-skilling" such that the human-in-the loop loses skills supported by the AI, in other words, degrading the clinical judgement skills of health professionals.[9] In this position paper, we describe the importance of determining and conveying the purpose, process, and performance of AI applications in healthcare. We support this position with empirical results from two pre-implementation qualitative studies involving AI systems.

2 METHODS – PRE-IMPLEMENTATION QUALITATIVE STUDIES

We present results from two pre-implementation qualitative studies designed to understand stakeholder needs for AI systems, specifically predictive algorithms. The first study focused on an algorithm designed to proactively identify high-need, high-cost patients, so that they could receive proper care to prevent unnecessary use of emergency or hospital care (preventable care study). The second study designed an algorithm for predicting suicide attempts to be implemented in an active-duty military clinic setting (military study). In these studies, we recruited interviewees from three stakeholder groups, including operational personnel (e.g., chief medical officers, chief strategic officers, practice managers), informatics personnel, and end users (e.g., primary care providers, nurse care managers, social workers, behavioral health specialists, medical assistants [corpsmen], nurses). See Benda et al. for the full methodological description and main results of the preventable care study and Reale et al. for the military study.[3; 7] Here, we present additional data from these studies to support our position for the importance of designing AI for appropriate trust in a healthcare setting.

3 PURPOSE, PROCESS, AND PERFORMANCE

Multiple authors have identified the core characteristics for trust in AI as purpose, process, and performance.[6; 9] Purpose explains why the automation was developed; process describes how the automation operates; performance refers to what the AI does, and how well the system supports people in achieving their goals.

3.1 Purpose

Prior AI creation, developers must first choose a purpose that addresses an important need within the healthcare environment, considers potential unintended consequences, and provides a comparative advantage over current practice. Systems that fulfill important needs better than existing approaches may be worthy of user trust.

In both the preventable care and military study, participants unsurprisingly found the aims of the respective AI systems, preventing unnecessary utilization and suicide prediction/prevention, to be valuable use cases. Leaders in the military study specifically felt urgency about the growing problem of suicide in active duty and retired military personnel. However, a seemingly straightforward clinical problem is layered with cultural implications in the military setting, particularly related to being "deployable," as described here by a primary care provider.

> "That's what I tell my patients when I talk to them, "You have diabetes…you've got to take it because that's an acceptable diagnosis. Mental health is not an acceptable diagnosis." And by us in this room, we know that it's life, it's just as important as menopause, and having a virus and having a flu. But in the community, other people – non-medical people – to put that stigma of a mental health label attached to them, and their job may depend on it, the stigmatism from command, their friends, their peers."

With AI and sophisticated computational models, many developers seem to be attempting to automate processes simply because there is a method for doing so (e.g. deep learning). To be useful in a healthcare setting, however, AI must provide a comparative advantage over current clinical practices. In the preventable care study, some participants expressed concern that AI was not needed to identify high-need, high-cost patients:

> "We all know a high utilizing patient when we see one. You can run the algorithms…but you can tell who's going to be a high utilizer and who's not sort of almost when you meet them."

In the military study, the AI algorithm held appeal because of its ostensibly more objective representation of risk as derived from electronic health record data. Suicide risk can be difficult to detect through screening and other provider assessments. One consequence of the stigma of behavioral health issues was that patients often were not honest about their thoughts. Staff characterized the typical responses of patients to depression screening questions (about the incidence and frequency of certain patterns of thinking) as "no, never, none." A provider described the frustration of not having documentable information:

> "The thing is, as soon as they hear speak, talk, write, look – they're out…many of them come in and they're like, "No, I don't [want] this in my record." But I can't do anything if it's not in the record."

In the military study, the participants described the painful tension between the algorithm "flagging" patients inappropriately (with attendant stigma or career-impacting consequences) and the current state, which is often not being aware that a patient is high risk. One provider described her two roles, military officer and clinician. She was responsible for taking care of patients and also for ensuring the physical and behavioral "fitness" of deployed personnel. Even with the new risk score, clinical judgement would be essential in documenting any confirmed risk.

Once AI has been created for an appropriate purpose, this purpose must be conveyed to end users. Conveying the purpose of AI helps prevent misuse, which occurs when a system is used inappropriately. Participants in the preventable care study, for example, explicitly described perils of overreliance on AI.

> "I can probably envision somebody you know taking it too much into account …rely on that solely without you know, precursory chart review or just to get some more qualitative details."

3.2 Process

Because healthcare professionals are highly knowledgeable about the component medical data and are themselves diagnosticians, it is important for trust that they understand process-related information in AI, including data inputs, analysis procedures, and outputs. The medical informatics community has recently advocated for improved *explainability*, such that the AI describes the process used to reach decisions, make recommendations, and take actions.[5] Participants in the preventable care study highlighted the criticality of transparency for process-based information in developing trust and utilizing the system.

> "I think just the more transparent it is, the better. Anytime people see a score, they always want to know how it's calculated, and I think transparency around that is critical."

The concept of explainable processes within AI has previously seemed at odds with AI processes becoming more sophisticated, and therefore, more challenging to explain. Our participants, however, indicated sufficient explainability may be achieved without having users understand concepts like machine learning, as one participant in the preventable care study discussed:

> "I think the best predictive models these days are black boxes. Machine learning is probably the best or one of these various techniques for predictive modeling...But, you can tell people what most important set of variables are, but not tell them how they're actually be used. Because it's too complicated."

Participants also noted that it was important to understand that AI's process because knowing some of the underlying predictors may affect interventions they select for the patient.

> "I think what I would do that day would change if I knew and had a nice synopsis of why they went to the emergency room and what was everybody's impression of that emergency room stay."

The AI's purpose may, however, further complicate the explainability of its process if the goal is to predict something inherently difficult to assess like mental health behavioral outcomes, as was the case in the military study. Participants had a hard time conceptualizing how a computer could even make this type of determination when current work processes rely heavily on face-to-face interactions with the patient.

> "So I'm thinking…what are you going to use to track the system you're talking about? What are you going to use to like gauge how they are, you know? It's more of like a subjective thing than objective."

Previous work has indicated that users better trust systems that they understand.[8] The participant's aforementioned challenge in conceptualizing the military study algorithm underscores the importance of conveying process-based information so users may understand and begin to trust the AI appropriately.

3.3 Performance

Performance in AI is commonly measured using statistics, such as AIC, AUC, PPV, and NPV. It is unclear, however, if performance statistics are conveyed to users, or if these statistics are meaningful to them. Healthcare organizations in the preventable care study wanted to understand, "the amount of people that use the information to make a different intervention or change a decision."

In addition to understanding how AI affects decisions, healthcare organizations have a desire to understand AI's impact on health outcomes and potential unintended consequences of various approaches to using the information in individual and population level care. One participant in the preventable care study simply said, "In other words, show me that if I react to that score today, that I will actually prevent an ED visit within the month."

The organizations' desires to capture the unintended consequences reflect their concerns about the ethics of their own actions, which will be taken in an environment of uncertainty because the technology is so new.

Frontline personnel can sometimes best capture those tensions, as does a corpsman in the military study describing the potential implications for himself (i.e., given that he is also sometimes a patient in the system):

> "Because if you have a system that even though it's not like disqualifying them, but still going to flag them. If I show up to my appointment and I said two years ago I was feeling kind of down and I show up to my appointment two years from now. And it's like flag, this guy, suicide risk."

4 CONCLUSION AND FUTURE DIRECTIONS

Our work and evidence from other industries highlights the importance of designing AI systems so that they may be trusted appropriately based on the purpose, process, and performance of the AI. We argue that inappropriate trust, particularly overreliance, may pose patient safety issues and degrade the clinical judgement of health professionals. Purpose, process, performance must be conveyed to the end users to foster appropriate trust. A key discussion point for the workshop and area of future research will involve *how* to best convey information about the AI to end users.

ACKNOWLEDGMENTS

The preventable care study was supported by HSB-1604-35187 from the Patient Centered Outcomes Research Institute (PI; R Kaushal). The military study was funded through the Military Suicide Research Consortium (MSRC), an effort supported by the Office of the Assistant Secretary of Defense for Health Affairs (Award No. W81XWH-10-2-0181). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the MSRC, the Department of Veterans Affairs, or the Department of Defense.

REFERENCES

[1] AMARASINGHAM, R., PATZER, R.E., HUESCH, M., NGUYEN, N.Q., and XIE, B., 2014. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff (Millwood) 33*, 7 (Jul), 1148-1154. DOI= http://dx.doi.org/10.1377/hlthaff.2014.0352.

[2] ARMSTRONG, K., 2018. If you can't beat it, join it: uncertainty and trust in medicine American College of Physicians.

[3] BENDA, N., DAS, L., ABRAMSON, E., BLACKBURN, K., THOMAN, A., KAUSHAL, R., and ANCKER, J., 2020. "How did you get to this number?" Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study *Journal of the American medical Informatics Association 27*, 5, 709-716.

[4] BRIGANTI, G. and LE MOINE, O., 2020. Artificial intelligence in medicine: today and tomorrow. *Frontiers in medicine 7*, 27.

[5] GORDON, L., GRANTCHAROV, T., and RUDZICZ, F., 2019. Explainable artificial intelligence for safe intraoperative decision support. *JAMA surgery 154*, 11, 1064-1065.

[6] LEE, J.D. and SEE, K.A., 2004. Trust in automation: Designing for appropriate reliance. *Human Factors 46*, 50-80.

[7] REALE, C., NOVAK, L.L., ROBINSON, K., SIMPSON, C., RIBEIRO, J., FRANKLIN, J., RIPPERGER, M., and WALSH, C., 2020. User Centered Design of a Machine Learning Intervention for Suicide Risk Prediction in a Military Setting (In Press). *AMIA Annual Symposium Proceedings*.

[8] SHERIDAN, T.B., 1992. Telerobotics, automation, and human supervisory control.

[9] ZUBOFF, S., 1988. In the age of the smart machine.