Trustworthy Artificial Intelligence in Healthcare

Umair Ali Khan & Ari Alamäki

The potential of AI for healthcare

Healthcare is one of the sectors where AI is in the highest demand. The emergence of new diseases, changing patterns of existing diseases, rising prevalence of chronic diseases, increasing workload of healthcare professionals, and growing amount of electronic healthcare data has necessitated the adoption of AI-enabled technologies in healthcare. The goal is to maximize the efficiency of healthcare centers, create better healthcare outcomes, produce a more accurate diagnosis and treatment plans, improve population health management, and free up time for healthcare professionals to focus on more important tasks. Especially, the advent of the Covid-19 pandemic has necessitated forecasting disease spreading, fostering preventative medicine, and new drug discovery in which AI can play a vital role.

AI has already shown its potential to be a game-changer in healthcare (Habli et al. 2020). AI can use patient data (patient's medical history, diagnoses, medications, treatment plans, vital signs recordings, radiology images, clinical notes, and lab reports) to discover the hidden patterns and intricate relationships among the data points to provide better diagnosis and treatment plans. Similarly, AI can analyze medical data to unveil the patterns for earlier diagnosis of a disease, identifying its biomarkers, revealing groups that are at particular risk for that disease, and finding treatments.

A few examples of AI-enabled healthcare solutions include predicting the likelihood to develop a disease, predicting heart disease which is one of the leading causes of death all over the world, predicting future illness from genomic data, diagnosing various critical diseases such as cancer by analyzing the electronic health records at earlier stages, robot-assisted surgeries, and clinical decision support systems, to name a few (Srinivasu et al. 2022).

Despite all these advancements, are AI solutions meeting acceptability in healthcare with the same enthusiasm with which they are being developed? Is the healthcare industry witnessing a widespread adoption of AI-enabled technologies? Despite the numerous value propositions of health technology providers and promises of research projects and pilots, there are also several failures and challenges at a practical level, such as the story of IBM Watson in healthcare.

IBM invested billions in Watson Health

IBM developed a successful AI-based supercomputer Watson, that defeated a human opponent in a quiz in 2011. It was able to read millions of books in a very short time compared to humans. Technologically, Watson was a question-answer machine that was able to manage a good structure of natural language based data. IBM invested billions of dollars in developing the next-generation AI platform for professionals in various sectors in the last decade. One major market for Watson was healthcare where AI is supposed to provide significant benefits since the industry is very data-intensive. After several pilots, real-life use cases, and numerous development iterations with physicians, medical researchers, and technologists, Watson still struggled in creating trustworthy recommendations for physicians.

Lohr (2021), in his New York Times article, opines that the underlying issues hampering Watson's performance include complex unstructured data, highly ambitious goals, technological inflexibility, and limitations. In addition, the approaches suitable to Watson's technological strengths are hard to come by. For instance, he states that genetic data used in real-life scenarios is often unstructured, complex, and sparse. Therefore, Watson lost the physician's trust in providing clinical recommendations due to its low decision-making accuracy which does not make it suitable for a safety-critical application. That said, Watson was still able to point out some better treatments missed out by physicians. Contrary to expectations, IBM Watson fell short of redeeming its value propositions with its original "AI physician" approach in real scenarios.

What are the barriers to the widespread adoption of AI in healthcare?

Despite AI's immense potential to solve complex problems and provide accurate predictions, there still exists a gap between application and research. AI solutions have still not acquired wide-scale acceptability among common citizens as well as healthcare providers (Goldfarb et al. 2022). In the case of the general public, the reason can be attributed to the lack of awareness and the inability to recognize AI's potential. However, a lack of interest from domain experts entails serious attention to this issue.

Several ethical and security issues need to be taken into account in developing and implementing AI applications in healthcare. For example, AI technologies must respect human autonomy and meet the requirements of trust, reliability, accountability, accuracy, explainability, and transparency (see e.g. Alamäki et al. 2019; Bærøe et al 2020; Markus et al. 2021). In addition, algorithm and data biases need to be mitigated (e.g. Kerasidou 2021), and privacy concerns managed (e.g., Lafky & Horan 2011; Wilkowska & Ziefle 2012). World Health Organization launched a guidance "Ethics and governance of artificial

intelligence for health" in 2021 that practically discusses those key issues. Below, we will discuss those topics in more detail.

1. Lack of trust in AI algorithms

Trust is strongly associated with functional understanding. Unless we do not get full insights into the functional mechanism of something, we cannot develop a high level of trust in it. AI works in mysterious ways in the sense that AI models are not explainable, and their decision paths cannot be tracked. In other words, one cannot know exactly how an AI model reaches a particular decision.

This black-box nature and lack of functional transparency are the main reasons preventing the wide-scale adoption of AI-enabled solutions in healthcare. Let alone the healthcare experts, lack of explanation is unacceptable even from the patient's point of view who has the right to know the basis of a treatment plan or the decision of risky surgery. Due to this limitation, it is also difficult to assess the quality of each decision and to disentangle AI's influence on decisions (Choi et al. 2022).

2. Reliability

The reliability of AI-enabled healthcare solutions is sometimes questionable. AI models generally perform well on the data for which they are trained, but they are prone to performance degradation under large data variations. When it comes to decision-making under diverse or unseen situations, a healthcare expert may opt to go for rigorous investigations and referrals.

However, an AI model will only make a prediction based on its accumulated experience with past data. In this case, one wrong prediction can risk a precious human life. Although a decision-making mistake can also be made by a very experienced doctor, he has the leverage of trust that the AI models lack.

3. Unfair Behavior

AI models are known to develop a bias for certain population groups, gender, race, ethnicity, or age due to training data imbalance. Bias leads to unfair and wrong decisions which can jeopardize human lives or cause untoward situations. For example, an AI algorithm for detecting anxiety from facial expression analysis may perform better for Europeans in comparison to Asians just because the training data might have contained a larger number of European facial features than Asians.

4. Privacy and Security Issues

AI algorithms access a staggering amount of patient-related data and manipulate them in several ways which entail concrete security mechanisms and regulatory

assurances to maintain privacy and patient agency (Murdoch et al. 2021). Data breaches, cyber attacks, or unauthorized access to patient-related data can have a detrimental impact on one's social life.

5. Accountability

In the traditional healthcare system, the responsibility for a wrong decision lies with the healthcare professional. But when an AI system makes a wrong decision that jeopardizes human life, who should be held accountable? The healthcare expert, the designer of the AI system, the vendor?

Habli et. al. (2020) state that AI-enabled healthcare systems are mainly targeted to assist physicians in decision-making, and not steering the wheel themselves. It is still the physician's responsibility to either consider the AI system's recommendation for final decision-making or completely discard it. However, the physician is faced with the dilemma of not gaining insight into the AI system's decision-making mechanism and ascertaining how the system has generated a particular recommendation. This leaves a question of how far a clinician is responsible for a patient's harm.

Towards explainable and trustworthy AI

Undoubtedly, AI has immense potential to transform the healthcare sector. However, widespread trust and acceptance of AI-enabled healthcare solutions among stakeholders need to be promoted. It can be done through trustworthy AI – an ethical framework for AI formulated by the European High-Level Expert Group on AI. Trustworthy AI addresses all the aforementioned issues and concerns by envisaging a transparent design in which the outcomes of AI models are explainable and verifiable. Similarly, it also enlists the key requirements for ensuring privacy, data protection, accountability, technical accuracy and robustness, and ethical compliance.

Explainable AI models can achieve wide-scale acceptability in the healthcare sector because of their functional transparency. With the help of explainable AI models, a healthcare professional can validate the AI's outcomes and get insights into the model's decision mechanism. For example, a doctor can ascertain the quality of an AI-generated treatment plan. He can also trace how the AI system has generated this specific treatment plan.

A trustworthy AI healthcare system will not be aimed at replacing healthcare professionals, but at empowering them in decision-making through a human-AI collaborative framework in which AI and healthcare experts work together on common goals. The trust and acceptance of the AI systems developed by the trustworthy AI will also help eradicate other misconceptions and apprehensions,

for instance, the fear of AI taking over humanity and leading to large-scale unemployment and subjugation.

Trustworthy AI in healthcare will ensure human agency and oversight by enabling healthcare experts to make informed autonomous decisions regarding AI systems and ensuring that an AI system does not undermine human autonomy or causes other adverse effects.

Europe is already excelling in regulating AI and has taken the first step in the form of policy-making for trustworthy AI. AI researchers should explore novel ways to introduce trustworthy AI in healthcare to make human lives better.

References

AI, HLEG. 2019. High-level expert group on artificial intelligence. Ethics guidelines for trustworthy AI: 6/2019.

Alamäki A., Aunimo L., Ketamo H. & Parvinen L. 2019. Interactive machine learning: Managing information richness in highly anonymized conversation data. In Camarinha-Matos LM, Afsarmanesh H and Antonelli D (eds) Collaborative Networks and Digital Transformation. In proceedings of the 20th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2019, pp.173-183.

Bærøe, K., Miyata-Sturm, A., & Henden, E. 2020. How to achieve trustworthy artificial intelligence for health. Bulletin of the World Health Organization, 98(4), 257-262.

Choi, Sukwoong, et al. 2022. How Does AI Improve Human Decision-Making? Evidence from the AI-Powered Go Program. Evidence from the AI-Powered Go Program (April 2022). USC Marshall School of Business Research Paper Sponsored by iORB. Forthcoming.

Goldfarb, A., and F. Teodoridis. 2022. Why is AI adoption in health care lagging. Brookings Institute.

Habli, I., Lawton, T., & Porter, Z. 2020. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization, 98(4), 251.

Lafky D.B. & Horan TA. 2011. Personal health records: Consumer attitudes toward privacy and security of their personal health information. Health Informatics Journal, 17(1), 63-71.

Lohr, S. 2021. What Ever Happened to IBM's Watson? New York Times 16.7.2021.

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, 113, 103655.

Murdoch, B. 2021. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 22, 122.

Srinivasu, Parvathaneni Naga, et al. 2022. From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies. Mobile Information Systems.

Wilkowska W & Ziefle M. 2012. Privacy and data security in E-health: Requirements from the user's perspective. Health Informatics Journal; 18(3), 191-201.

World Health Organization. 2021. Ethics and governance of artificial intelligence for health: WHO guidance.