PSNet
PATIENT SAFETY NETWORK

Artificial Intelligence and Patient Safety: Promise and Challenges

March 27, 2024
Tighe P, Mossburg S, Gale B. Artificial Intelligence and Patient Safety: Promise and Challenges. PSNet [internet]. 2024.
https://psnet.ahrq.gov/perspective/artificial-intelligence-and-patient-safety-promise-and-challenges

Introduction

Ensuring patient safety in modern healthcare is a complex task, with numerous interrelated factors contributing to numerous potential harms. These factors, including disorganized data, overburdened clinicians, and complex clinical cases, create challenges that require sophisticated solutions. The integration of artificial intelligence (AI) into health information technology (IT) systems offers the promise that some of the challenges can be reduced or overcome. AI can analyze a vast amount of data from various sources, optimize workflows, and offer evidence-based recommendations to clinicians. While certain specialties have already found success by implementing AI, and the research continues to progress, widespread AI adoption in daily clinical practice is still on the horizon. Not only is there a lack of peer-reviewed prospective evidence of its effects on patient care and the clinician experience, integrating AI poses a number of ethical and technical challenges. Nonetheless, the potential of AI to enhance patient safety and improve healthcare is great.

Definitions

“Artificial intelligence” or “AI” refers to the simulation of human intelligence in machines, enabling them to perform tasks that typically require human cognitive abilities. Machine learning (ML) is a subset of AI that focuses on creating algorithms that allow computers to learn from data. ML differs from traditional rules-based programming in that it does not rely on explicit programming of all possible scenarios, which is not feasible in the complex world of healthcare. Other more recent subsets of AI are deep learning (DL) and Large Language models (LLM). DL uses very large neural networks, principally in the processing of imaging and sequence data, and LLM generates new text, imaging, and video content based upon “prompts” from users. In recent years, ML has been the primary way that AI has been applied in healthcare, although this is rapidly changing with the advent of DL, LLMs, and other AI methods. For the rest of this

article, we will use the term “AI” to refer to AI, ML, DL, and LLM.

Current Applications for Patient Safety

Most uses of AI to increase patient safety are still in the research phase. In the area of medical imaging, however, AI-powered algorithms have demonstrated a remarkable ability to read and analyze medical images, potentially increasing diagnostic accuracy and efficiency and reducing diagnostic errors. A survey in 2020 by the American Academy of Radiology found that 30% of radiologists were currently using AI in their clinical practice and another 20% were planning to purchase it in the next one to five years.[1] The successes in this specialty can be attributed to the progress and development of AI image analysis technology more generally. Other fields outside healthcare have been training algorithms to recognize faces or landmarks in images, and that same technology can be adapted fairly easily to identify cancerous masses and other clinical conditions.

The applications of this technology are growing rapidly. One example is diagnosing diabetic retinopathy. Clinicians typically spend considerable time manually reviewing ophthalmological images, but an AI algorithm trained on vast datasets outperformed human ophthalmologists in detecting diabetic retinopathy.[2] Another example is the early detection of lung cancer in X-ray and CT scan images, where an AI algorithm significantly reduced false positives and false negatives compared to evaluations by six radiologists.[3] Despite these promising results, there is still a need for more peer-reviewed prospective evidence linking AI radiology products to improved patient outcomes. A recent systematic review of 100 commercially available products found that only 18% had validated their results in a clinical setting.[4]

Regulatory bodies and healthcare payment systems have started to recognize the potential of AI in medical imaging, as indicated by the U.S. Food and Drug Administration approval of 51 out of the 100 products reviewed in the aforementioned systematic analysis. Additionally, some healthcare payers, including CMS, have started covering specific AI-assisted diagnostic services, acknowledging the value and potential cost-saving benefits that these technologies bring to medical image analysis.[5]

Potential Contributions to Patient Safety

The potential for enhancing patient safety across various specialties and settings is substantial. A scoping review in 2021 explored the impact of AI on eight main patient safety domains, suggesting that AI's influence would be most pronounced in domains where existing strategies have proven insufficient and where integration and analysis of new, unstructured data is crucial for accurate predictions. Such domains include adverse drug events, clinical decompensation, and diagnostic errors.

One of the most common potential applications of AI for patient safety is risk prediction—for example, predicting the likelihood that a patient will decompensate, have an adverse reaction to a medication, or develop a pressure ulcer and then alerting clinicians if the risk is high enough. AI-powered data models excel in this task, as they can process real-time data from various sources within the electronic health records (EHRs) and biometrics and dynamically adjust their predictions based on new data about patients. Integrating more biometric and sensor data into these models is expected to significantly enhance AI's risk-

predicting capabilities, as these data sources are often underutilized or too complex for human interpretation.[6] Another novel data source being explored is video taken in clinical environments.[7] In this application, cameras or movement sensors gather data on what is happening in the healthcare setting and can alert staff when a patient falls or a critical checklist step is skipped, for example.

In addition to risk prediction, AI can improve patient safety in other areas of clinical decision support (CDS), which provides clinicians with relevant information at the point of care so that they can make better informed decisions. For example, during patient examinations and EHR documentation, the AI system can suggest diagnoses or evidence-based treatment options or caution against potential treatment-related complications, thus reducing diagnostic errors and adverse drug events. While some non-AI clinical decision support systems already exist, integrating AI can enhance their capabilities and their impact on patient safety. LLMs in particular could improve clinical decision support because they excel at analyzing text and knowledge bases.

A final example of an application that may have a large impact on many clinicians' day-to-day life is AI auto-charting. Instead of needing to complete the EHR as they perform a procedure, with the computer between them and the patient, the AI system can be listening along and completing the chart for them.[8] This will not only reduce the documentation burden on clinicians, but it could also improve patient safety by streamlining and standardizing data collection and reducing documentation errors. Given the success of AI dictation and assisted documentation in other fields, its adaptation to healthcare holds promise, especially with the continued improvement of large language models and ongoing exploration of how to implement these technologies in live clinical practice.

Risks to Patient Safety

While the potential for improving patient safety is high, AI also comes with risks that must be carefully considered before and during implementation. First and foremost, as with all models, it is important to ensure that the AI model goals are in alignment with specific patient safety goals, such as identifying patient decompensation. If the AI model is not precisely aligned with this patient safety goal, it may either miss critical signs of decompensation or generate false alarms. AI prediction models must also be incorporated into broader process engineering programs to ensure that predictions can lead to reasonable, safe, and beneficial actions to improve upon the originally predicted outcome. Poor system design and inadequate workflow considerations during implementation can also lead to alert fatigue and mistrust of the system among patients and healthcare providers, undermining the intended benefits of AI.

Other important considerations include data quality, biases, privacy, and security. AI models are only as good as the data they are trained on, and if the training data are biased or underrepresent certain groups, the results of that model will not be equitable.[9] A systematic review by the AHRQ Evidence-based Practice Center found that algorithms can exacerbate racial and ethnic disparities, but also have the potential to reduce them.[10] Researchers and developers are attempting to mitigate the effects of bias in several different ways, including regular analysis of model metrics to detect bias, editing input variables, and by exploring the use of synthetic data, which involves creating artificial data that mimic real patient data but without the inherent biases.[10], [11], [12] Data-sharing privacy is another ethical consideration.[13] Healthcare

data are highly sensitive, as they contain personal and private information about patients, and sharing such data for AI model training and research purposes must be done with utmost caution and adherence to strict privacy and security measures.

Implementation Best Practices

Capitalizing on the potential of AI and overcoming the attendant risks requires responsible design and implementation of AI systems. Healthcare organizations would benefit from forming a multidisciplinary team consisting of data scientists, clinicians, ethicists, regulatory specialists, and IT professionals to collectively design, validate, implement, and monitor AI-powered solutions tailored to specific patient safety needs.

Building trust in AI technologies is an essential task of this multidisciplinary group,[14] and a vital component of the trust-building process is rigorous validation and ongoing monitoring of AI systems. This will involve (1) validating the models using data from the organization's patients to demonstrate applicability, accuracy, and the potential for clinical benefit for that population and setting, and (2) establishing robust quality assurance processes to continuously evaluate AI model performance (including biases) and ensure adherence to privacy and patient safety standards. Clinician engagement is essential throughout this process, including choosing clinical priorities, emphasizing usability and understandability in design, and providing clinicians with training on data science fundamentals and model functionality.[15]

Regulatory bodies should also play a role in setting guidelines and requirements for AI implementation in patient safety, which the National Artificial Intelligence Initiative[16] and the WHO[17] have already started.

Conclusion

The integration of AI into healthcare holds great potential for improvements in patient safety, as demonstrated by the advancements in medical imaging analysis. In the near future, clinicians could be assisted with many of their daily tasks, including risk prediction and documentation. However, there are many risks to overcome before that point as well as concerns about data quality, bias, privacy, and interpretability of models. Rigorous validation, responsible implementation, and continuous monitoring by multidisciplinary teams are necessary to address these risks and realize the full potential of AI.

Patrick Tighe, MD, M
Anesthesiologist
Executive Director, Quality and Patient Safety Initiative
University of Florida Health
Gainesville, FL

Bryan M. Gale, MA
Researcher
American Institutes for Research (AIR)
Columbia, MD

Sarah E. Mossburg, RN, PhD
Senior Researcher
AIR
Arlington, VA

References

1.Allen B, Agarwal S, Coombs L, Wald C, Dreyer K. 2020 ACR Data Science Institute Artificial Intelligence Survey. *J Am Coll Radiol.* 2021;18(8):1153-1159. https://doi.org/10.1016/j.jacr.2021.04.002

2. Lim JI, Regillo CD, Sadda SR, et al. Artificial intelligence detection of diabetic retinopathy: subgroup comparison of the EyeArt system with ophthalmologists' dilated examinations. *Ophthalmol Sci.* 2022;3(1):100228. https://doi.org/10.1016/j.xops.2022.100228

3. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography [published correction appears in *Nat Med*. August 2019;25(8):1319]. *Nat Med.* 2019;25(6):954-961. https://doi.org/10.1038/s41591-019-0447-x

4. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol.* 2021;31(6):3797-3804. https://doi.org/10.1007/s00330-021-07892-z

5. Chen MM, Golding LP, Nicola GN. Who will pay for AI?. *Radiol Artif Intell.* 2021;3(3):e210030. https://doi.org/10.1148/ryai.2021210030

6. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med.* 2021;4(1):54. https://doi.org/10.1038/s41746-021-00423-6

7. Yeung S, Downing NL, Fei-Fei L, Milstein A. Bedside computer vision: moving artificial intelligence from driver assistance to patient safety. *N Engl J Med.* 2018;378(14):1271-1273. https://doi.org/10.1056/NEJMp1716891

8. Rajkomar A, Kannan A, Chen K, et al. Automatically charting symptoms from patient-physician conversations using machine learning. *JAMA Intern Med.* 2019;179(6):836-838. https://doi.org/10.1001/jamainternmed.2018.8558

9. Agarwal R, Bjarnadottir M, Rhue L, et al. Addressing algorithmic bias and the perpetuation of health inequities: an AI bias aware framework. *Health Policy Technol.* 2023; 12(1). https://doi.org/10.1016/j.hlpt.2022.100702

10. Tipton K, Leas BF, Flores E, et al. Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare. Comparative Effectiveness Review No. 268. (Prepared by the ECRI-Penn Medicine Evidence-based Practice Center under Contract No. 75Q80120D00002.) AHRQ Publication No. 24-EHC004. Rockville, MD: Agency for Healthcare Research and Quality; December 2023. DOI: https://doi.org/10.23970/AHRQEPCCER268

11. Rojas JC, Fahrenbach J, Makhni S, et al. Framework for Integrating Equity Into Machine Learning Models: A Case Study. *Chest*. 2022;161(6):1621-1627. doi:10.1016/j.chest.2022.02.001

12. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng.* 2021;5(6):493-497. https://doi.org/10.1038/s41551-021-00751-8

13. World Economic Forum. *Why Artificial Intelligence Design Must Prioritize Data Privacy.* Cologny, Switzerland: World Economic Forum; 2022. https://www.weforum.org/agenda/2022/03/designing-artificial-intelligence-for-privacy/

14. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res.* 2020;22(6):e15154. https://doi.org/10.2196/15154

15. Rojas JC, Teran M, Umscheid CA. Clinician Trust in Artificial Intelligence: What is Known and How Trust Can Be Facilitated. *Crit Care Clin*. 2023;39(4):769-782. doi:10.1016/j.ccc.2023.02.004

16. National Artificial Intelligence Initiative. Strategy Documents [database online]. Accessed August 22,2023. https://www.ai.gov/strategy-documents/

17. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health.* Geneva, Switzerland: World Health Organization; 2021. ISBN: 9789240029200.