The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard Bioethics Collaborative

The chat is out of the bag: The future of AI in clinical research

Tuesday, October 10, 2023 | 1:00 – 3:30 PM EDT | Virtual Meeting

Meeting Summary

Introduction

Artificial intelligence (AI) and machine learning (ML) are tools that can be deployed for a variety of different purposes across different domains. Proposed uses of traditional AI in clinical research include helping to optimize recruitment and retention by performing eligibility analyses and matching individuals to trials, assisting in study design by predicting participant outcomes based on biomarkers and other factors and helping with drug selection, playing a role in monitoring through the use of AI-powered wearables and mobile applications, and by enabling more robust data analysis and assisting with data attribution in the case of missed research visits or missing data generally. Uses for generative large language models (LLMs), such as ChatGPT, are likely to center around informed consent and the potential for LLMs to assist with consent form development, simplify consent language, and provide support for participants via AI-powered chatbots during informed consent and throughout the study. The October 10, 2023, meeting of the Bioethics Collaborative aimed to clarify the current and potential uses of AI in clinical research, identify salient ethical challenges and issues, and provoke deliberation on how best to approach ethical issues with the use of AI in clinical research.

The meeting began with a brief presentation introducing different types of AI and ethical concerns related to using AI in clinical research, including justice and transparency. One distinction relevant to many discussions of AI and ML is the difference between traditional or predictive AI and generative AI. Both traditional and generative AI involve using large data sets to train algorithms. Traditional AI uses pattern recognition to produce insights and predictions, while generative AI creates new content based upon the data on which it was trained.[1] Predictive and generative AI have different strengths and capabilities and, therefore, lend

[1] *See* Marr B. The Difference Between Generative AI And Traditional AI: An Easy Explanation For Anyone. *Forbes*. https://www.forbes.com/sites/bernardmarr/2023/07/24/the-difference-between-generative-ai-and-traditional-ai-an-easy-explanation-for-anyone/. Published July 24, 2023.

themselves to different use cases. Generative AI is a relatively new and rapidly developing technology; it remains to be seen how it can best be used in clinical research.

Whether traditional or generative, AI is a tool that can be leveraged for a variety of different purposes across a variety of different domains. When discussing the ethics of its use, there are several questions we should bear in mind. What are we using it for, and what is our goal? Whose interests is it serving? Is it the best tool for the job? The hype surrounding AI – especially generative AI – could lead us to treat it differently when the actual ethical concerns and attendant mitigation levers may significantly overlap with those of more familiar technologies.

Ethical Considerations

One of the topics raised repeatedly throughout the meeting was the potentially inverse relationship between complexity and explainability. In general, more complex models tend to perform more accurately – something that is particularly desirable in healthcare settings where decisions informed by AI have the potential to significantly impact a person's well-being. The more complex the algorithm, however, the less explainable it tends to be. Explainability is important because it fosters trust and gives us additional and potentially helpful context for evaluating outputs. The issue is further complicated by the lack of consensus over how to define and understand "explainability." One participant suggested that explainability can roughly be split into three groups, depending on what the object of explanation is taken to be: data, models, and outcomes. For data explainability, what are the data sources upon which this algorithm was trained? For model explainability, how does the model work? For outcome explainability, what factors impact the model's output? Each form of explainability is desirable, but it may not be possible to fully satisfy demands for multiple forms of explainability simultaneously. Consequently, when a study team desires explainability for the use of AI in a trial – or when a regulatory body demands it – they must decide what form of explainability they want. There has been discussion on this topic in the clinical context, but there is no consensus among healthcare providers. The same disagreements are likely to arise among sponsors, investigators, and other research stakeholders.

The concept of fairness may be even more fraught with ethical conundrums than explainability, owing to its contested nature. While stakeholders continue to disagree on what true "fairness" looks like, it remains difficult to set benchmarks for achieving fairness in the AI context. For example, a study sponsor using an ML algorithm to identify potential study participants may want to ensure that all eligible individuals local to study sites are equally likely to be identified, on the value-based assumption that all people should have equal access to clinical research.

Alternatively, the sponsor may want to ensure that the study group's demographics reflect the demographics of the disease population, which may require actively recruiting more people from one demographic over another, on the value-based assumption that investigational therapies should be evaluated in the types of individuals most likely to use them. Both approaches embed defensible assumptions about fairness, but it is mathematically impossible for one algorithm to satisfy both definitions.[2] Sponsors and investigators need to carefully consider what type of "fairness" is most important for their study and work closely with ML engineers to ensure that the algorithms are designed with that definition in mind. It is not sufficient to simply state the desire for a "fair" ML model.

Another important concept in this context is transparency. Transparency about how and when AI models are used in clinical research has the potential to foster trust even in the absence of explainability. Many decisions are made when designing an AI/ML model. These decisions, such as the definition of fairness and the tradeoff between accuracy and explainability, have important ramifications and should be made intentionally. One form of transparency is to provide clear rationale behind each such decision. This could help to alleviate distrust in the algorithm itself or in the study and study team using the algorithm. However, important questions remain about what should be disclosed to whom, and when. More discussion and work are needed on when the use of AI in research should be disclosed to ethics review bodies, to research participants, or to any other stakeholders.

Like any technology, AI has inherent limitations. First, the algorithm only knows what we teach it. ML algorithms can be extremely good at making predictions based on training data but may not perform well on data that differs from its training set. One participant pointed out that this idea of generalizability may actually be less of a concern with generative AI than it is with traditional AI, as generative AI algorithms are trained on large amounts of generalizable data (so long as those data remain representative). Second, AI is not concerned with the truth; it is concerned with statistical frequency. A meeting participant pointed out that this is a particular concern with generative AI for the same reason that generalizability is less of a concern. An algorithm trained on all information openly available on the internet – when a large percentage of that information is inaccurate – is likely to internalize those inaccuracies and make subsequent decisions/predictions based on them. Such inaccuracies may be difficult to identify

---

[2] *For further explanation and a concrete example, see* Verma S, Rubin J. Fairness Definitions Explained. In: Proceedings of the International Workshop on Software Fairness. FairWare '18. Association for Computing Machinery; 2018:1-7. doi:10.1145/3194770.3194776

since large language models are extremely skilled at presenting information in a human-like manner and can do so with the veneer of authority. Third, ML algorithms can compound existing biases. Because available data sets are often not representative, algorithms trained on those data will likely not perform well on new data that meaningfully differ from the training set.

The risks of exacerbating systemic inequities associated with using AI models should be considered. Algorithms are influenced by factors such as who is building them, who is sponsoring them, potentially competing interests of sponsoring organizations, and other hidden factors.[3] There need to be ongoing discussions about what risks and levels of risk are tolerable, as well as how to mitigate those risks. Several participants commented that people tend to accept very little risk or room for error regarding AI/ML algorithms. Minimizing risk and error is reasonable; however, we also appreciate that humans make mistakes and errors. This raises questions about the correct baseline against which to assess the error rates in AI/ML models and whether they should be held to a higher standard than human decision-making. One important set of questions, then, concerns how frequently humans make mistakes, what kinds of mistakes humans make, the significance and impact of those mistakes on study outcomes and individuals alike, and, most importantly in the current context, how to compare the results of these to mistakes of AI models. Whether human errors are qualitatively different from AI/ML algorithm errors remains unknown, as do the impacts of those errors. Likewise, whether to hold AI algorithms to an equivalent or lower error margin than humans requires both further empirical data and discussion.

Regulatory Considerations

Clinical trials, including those involving AI models, are regulated to ensure the safety and well-being of research participants. Meeting participants discussed the need for and challenges with oversight and regulation of such trials at both the institutional and federal levels. One of the concerns is that regulatory guidelines are typically created in response to technological advancements. The time interval between implementation and regulation leaves a gap for unchecked use of these technologies, which may lead to harm. There is reasonable concern that the potential for harm to occur during that gap is greater for AI than it has been with past

[3] *See* Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. doi:10.1126/science.aax2342

technologies because of how rapidly these powerful tools can be deployed, how rapidly they evolve, and the potential they will do so in a non-transparent fashion. It is extremely difficult, however, to regulate a technology when its applications, risks, and limitations are not yet fully understood – as is the case with AI in clinical research. Lawmakers are working to establish robust regulations,[4] but interim guidance in the meantime would be useful.

Several meeting participants noted that some AI algorithms being used in clinical research should be subject to oversight but appear to fall outside the purview of regulatory bodies, such as the Food and Drug Administration's Center for Drug Evaluation and Research (CDER). There was a discussion of how institutional review boards (IRBs) may play a role in oversight, and whether the scope of their purview includes AI/ML under the current regulations when most of the data sources are not "readily identifiable." A proposed "Ethics and Society Review (ESR)"[5] was discussed as a potential mechanism for addressing some of the gaps in current oversight bodies. The ESR process differs from IRB review in its consideration of the potential long-term impacts of different types of research and research interventions on society. The IRB regulations and remit have historically been focused on the protection of individual research participants rather than society at large. Societal implications may be a particularly important consideration for AI algorithms due to their potential to impact decision-making compounded by the risk of unrecognized systemic bias being introduced into that decision framework.

Specific points of concern in the regulatory sphere include the challenge of regulating iterative AI/ML models that evolve over time. At what point does the model become sufficiently different from its baseline counterpart that it needs to be reviewed as a new product? Dual use and intentional misuse are also concerns. For example, an algorithm designed for therapeutic drug discovery could be modified relatively easily to create bioweapons.[6] There was some disagreement among participants about the feasibility of preventing intentional misuse. Bad

---

[4] *See* United States Food and Drug Administration. Using Artificial Intelligence & Machine Learning in the Development of Drug and Biological Products: Discussion Paper and Request for Feedback. 2023. https://www.fda.gov/media/167973/download?attachment

[5] *See* Bernstein MS, Levi M, Magnus D, Rajala BA, Satz D, Waeiss Q. Ethics and society review: Ethics reflection as a precondition to research funding. *Proceedings of the National Academy of Sciences*. 2021;118(52):e2117261118. doi:10.1073/pnas.2117261118

[6] *See* Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*. 2022;(4):189-191.

actors will always exist, but design features (e.g., digital watermarking) may help prevent successful misuse. Additionally, there have been successful efforts in regulating dual use in other spaces, such as gain-of-function research.[7] The feasibility of those guardrails in the AI/ML space, however, rested beyond the scope of the discussion.

Conclusion and Next Steps

The October session of the MRCT Center Bioethics Collaborative explored the ethical issues of using AI in clinical research. The meeting identified several decision-making fulcra with ethical implications that present at the design stage of an AI/ML model. For example, study designers and ML engineers must determine what form of "explainability" matters most to their study prior to the design and deployment of an AI/ML model because it may be impossible to successfully satisfy the criteria for multiple forms of explainability (i.e., data, models, and/or outcomes) with a single model. Similarly, the concept of "fairness" as a requirement of AI/ML use must also be considered at the design stage. Rational, human decision-makers must determine in advance what fairness "looks like" to the AI model – e.g., whether extending study access to all potential patients in a geographic locale is preferable to deliberately limiting study access to demographics most likely to be affected by the medical condition of interest. Such decision-makers should be prepared to describe which fairness criteria were selected and why. All of these human-mediated decisions occur prior to the deployment of the AI/ML model and therefore also raise concerns over transparency – which decisions and decision-making criteria should be disclosed, when, and to whom? – though it was generally accepted that transparency in these decisions will be an important component in building public confidence.

The use of AI in clinical research presents an important opportunity to address systemic inequities. However, given that AI/ML algorithms are trained on available datasets, which are often outdated and biased, careful and deliberate selection of training datasets will be critical to the promotion of equity in clinical research. Participants discussed the need to define an acceptable standard of error for AI models in clinical research. Doing this, however, will likely require a better understanding of the range and magnitude of human-mediated errors for comparison.

---

[7] *See* US Department of Human Health Services Science, Safety, Security. Dual Use Research of Concern: Gain-of-Function Research. Published June 3, 2021. Accessed November 9, 2023. https://www.phe.gov/s3/dualuse/Pages/GainOfFunction.aspx

As with many new technologies, AI is being used before the regulatory landscape can adapt to it. Consequently, there is real concern over a presumptive dark area between where existing regulatory guardrails end and the speed with which AI/ML models accumulate new capabilities. Some non-regulatory solutions were suggested – e.g., the ESR process – but there was a general desire for relevant interim guidance from regulatory bodies such as the FDA and EMA, even as those bodies work to produce more robust final guidance. Attendees also expressed concern over the potential for misuse of AI technologies by bad actors, especially in light of the extent to which AI/ML algorithms make decisions within a proverbial "black box" free from human oversight.

Until major regulatory bodies release comprehensive final guidance specific to the use of AI in clinical research, stakeholder organizations should remain deliberate in their use of AI and should be prepared to meet any potential transparency requirements that may emerge in the future. Diligent documentation of training datasets used, fairness and explainability criteria, and intended scope of use in the early design stages of AI/ML model development will be key factors to achieving ethical deployment of such exciting technology in clinical research.