frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 14 September 2023
DOI 10.3389/fcomp.2023.1235226

OPEN ACCESS

EDITED BY
Philippe Palanque,
Université Toulouse III Paul Sabatier, France

REVIEWED BY
Adel Khelifi,
Abu Dhabi University, United Arab Emirates
Ricardo Valentim,
Federal University of Rio Grande do Norte, Brazil

*CORRESPONDENCE
Christos Andreas Makridis
✉ christos.makridis@va.gov

RECEIVED 05 June 2023
ACCEPTED 21 August 2023
PUBLISHED 14 September 2023

CITATION
Makridis CA, Boese A, Fricks R, Workman D, Klote M, Mueller J, Hildebrandt IJ, Kim M and Alterovitz G (2023) Informing the ethical review of human subjects research utilizing artificial intelligence. *Front. Comput. Sci.* 5:1235226. doi: 10.3389/fcomp.2023.1235226

Informing the ethical review of human subjects research utilizing artificial intelligence

Christos Andreas Makridis[1]*, Anthony Boese[1], Rafael Fricks[1], Don Workman[2], Molly Klote[2], Joshua Mueller[1], Isabel J. Hildebrandt[3], Michael Kim[1] and Gil Alterovitz[1]

[1]United States Department of Veterans Affairs, National Artificial Intelligence Institute, Washington, DC, United States, [2]United States Department of Veterans Affairs, Veterans Health Administration, Washington, DC, United States, [3]United States Department of Veterans Affairs, VA Long Beach Healthcare System, Veterans Health Administration, Long Beach, CA, United States

Introduction: The rapid expansion of artificial intelligence (AI) has produced many opportunities, but also new risks that must be actively managed, particularly in the health care sector with clinical practice to avoid unintended health, economic, and social consequences.

Methods: Given that much of the research and development (R&D) involving human subjects is reviewed and rigorously monitored by institutional review boards (IRBs), we argue that supplemental questions added to the IRB process is an efficient risk mitigation technique available for immediate use. To facilitate this, we introduce AI supplemental questions that provide a feasible, low-disruption mechanism for IRBs to elicit information necessary to inform the review of AI proposals. These questions will also be relevant to review of research using AI that is exempt from the requirement of IRB review. We pilot the questions within the Department of Veterans Affairs–the nation's largest integrated healthcare system–and demonstrate its efficacy in risk mitigation through providing vital information in a way accessible to non-AI subject matter experts responsible for reviewing IRB proposals. We provide these questions for other organizations to adapt to fit their needs and are further developing these questions into an AI IRB module with an extended application, review checklist, informed consent, and other informational materials.

Results: We find that the supplemental AI IRB module further streamlines and expedites the review of IRB projects. We also find that the module has a positive effect on reviewers' attitudes and ease of assessing the potential alignment and risks associated with proposed projects.

Discussion: As projects increasingly contain an AI component, streamlining their review and assessment is important to avoid posing too large of a burden on IRBs in their review of submissions. In addition, establishing a minimum standard that submissions must adhere to will help ensure that all projects are at least aware of potential risks unique to AI and dialogue with their local IRBs over them. Further work is needed to apply these concepts to other non-IRB pathways, like quality improvement projects.

KEYWORDS

artificial intelligence, ethics, human subjects, institutional review board, research and development, trustworthy AI

1. Introduction

Research and development (R&D) efforts are a major driver of economic growth and social flourishing since they lead to the production and deployment of new ideas (Aghion and Howitt, 1992; Jones, 2002). The rapidly expanding use of artificial intelligence (AI) (AI Index Report, 2021) and the potential risks to society have grown exponentially, requiring the embedding of trustworthy AI principles that have been recently pioneered into the design and consideration of AI investments. Nowhere is this more important than in the R&D process, especially in research with human subjects.

There has been an expansion in the use of AI in research that involves human subjects, particularly in health care (Yu et al., 2018; Matheny et al., 2019). Recent achievements include the early detection of sepsis, potentially reducing mortality by 20% (Goh et al., 2021; Henry et al., 2022) and rapid diagnosis of COVID-19 (Mei et al., 2020). AI models can already provide physician-level diagnosis of colorectal cancer, mesothelioma, lung cancer, and intra-cranial hemorrhaging (Courtiol et al., 2019; Huang et al., 2019; Wismüller and Stockmaster, 2020; Zhou et al., 2020). These examples demonstrate the current and future potential of AI tools as models, data sets, and computational capabilities improve.

For all this promise, however, AI offers new ethical challenges, particularly with building externally valid and reliable predictive models and deploying them with requisite safeguards and feedback loops. A major concern, for example, is that many models are trained on data samples that are not fully representative of the population, producing results that lead to a disparate impact (Obermeyer et al., 2019). While achieving external validity has always been a challenge, the stakes are higher with AI because predictive models are often automated and applied at great scale—meaning that inaccuracies or bias can be propagated rapidly and affect large groups of people. In health care, this can have disastrous consequences for patients, especially patients from underrepresented minorities and groups that already experience high rates of poor health outcomes (Buruk et al., 2020).

Growing recognition of these new risks have led to a proliferation of AI ethics frameworks, most notably, Executive Order (EO) 13960: Promoting the Use of Trustworthy Artificial Intelligence in the U.S. Federal Government, which establishes AI use and transparency requirements across federal agencies in the United States, but leaves the development of detailed compliance standards to other federal bodies, including agencies themselves. Resultingly, federal agencies have developed their own implementations of the EO.[1] We build on these various federal frameworks around the governance of trustworthy AI by creating a system by which to enact these principles during the institutional review board (IRB) review process (O'Shaughnessy, 2023).

Much of biomedical and social science requires review by an IRB (or another equivalent review authority) to ensure a minimal set of ethical standards are adhered to in research, for instance, the requirement of legally effective informed consent to participate. However, in the context of AI research, new complexities emerge for IRBs to manage. For example, a major challenge is the external validity of predictive models and preventing harm from unintended bias in those models. Bias can emerge through the introduction or exclusion of certain variables or sampling frame. Training models on samples that are not fully representative of the population can produce inaccurate, unreliable, and harmful results (Obermeyer et al., 2019). AI not trained on a properly representative sample may make recommendations that are better than the status quo, on average, but worse for certain sub-groups that may already be marginalized populations, amplifying existing inequities against already vulnerable patients (Buruk et al., 2020). While many IRBs are *infrastructurally* equipped to review and vet AI research, they often lack the requisite *subject matter expertise* in AI.

The primary contributions of this paper are to: (1) critically examine how the surge in AI and its vast possibilities creates a new opportunity to augment existing IRB processes and literature to meet this new reality; (2) give guidance regarding a modular series of questions that elicit information and manage against potential unintended consequences of AI that can be added into an existing IRB process; and (3) conduct a pilot of these questions in the Department of Veterans Affairs (VA)—the largest integrated healthcare system in the U.S.—and report on the outcomes as an illustrative example of how it can be used in other contexts. Importantly, we review the history and function of IRBs to provide some context for the reasons why IRB members need to be informed about AI in research under review by IRBs and to motivate our introduction of an AI IRB supplement.[2]

Our research contributes to an emerging literature on AI ethics that builds on general frameworks by taking principles into practice (e.g., O'Shaughnessy, 2023). Rather than laying out a prescriptive set of rules—which are important and necessary—our paper introduces the idea of an AI IRB module coupled with a real-world pilot that provides a streamlined set of questions that are meant to elicit relevant information about a project and facilitate a dialogue with the researchers so that both sides arrive at a reasonable solution—even if that means declining a study proposal.

Our research also fills a significant gap in risk mitigation and oversight of AI systems (Crossnohere et al., 2022), and the

Abbreviations: AI, Artificial Intelligence; R&D, Research and Development; IRB, Institutional Review Board; CFR, Code of Federal Regulations; EO, Executive Order; CIOMS, Council for the International Organizations of Medical Sciences; AAHRPP, Association for the Accreditation of Human Research Protection Programs; ORO, Office of Research Oversight; PII, Personally Identifiable Information; PHI, Protected Health Information; ISSO, Information System Security Officer; CROs, Clinical Research Organizations; VA, Veterans Affairs; NAII, National Artificial Intelligence Institute; VAMC, Veterans Affairs Medical Center.

1 There are also international frameworks, including the Organization for Economic Co-operation and Development (OECD) AI Principles, recommendations adopted by its member nations, including the U. S., and several non-member signatories to broaden the coalition of trustworthy AI principles adoption (OECD, 2021).

2 There are also other kinds of ethical oversight mechanisms (beyond the scope of this paper) that are being developed in contexts and for applications that are not related to human subjects' research that would require review by an IRB.

absence of streamlined best practices in health care informatics (van de Sande et al., 2022). From the patient's perspective, such best practices are especially important for non-AI subject matter experts to adhere to (Anderson and Anderson, 2019) consistently. Our tool, however, is not a substitute for more formal and specific guidelines, particularly among associations in sub-disciplines that are applicable to clinical trials and other interventions involving the use of AI, as described by Rivera et al. (2020).

2. Background on medical research ethics and institutional review boards

Much research in health care and biomedical informatics involves the engagement of human subjects, particularly through randomized control trials. However, since interacting with and administering treatments on human subjects creates risks and challenges, U.S. institutions, including universities, have created IRBs in accordance with federal guidelines to decide whether affiliated researchers can pursue their proposed projects and, if so, how they should go about doing it to ensure that the research is conducted safely and ethically, preserving the rights and interests of participants (Khin-Maung-Gyi, 2009). The existing structure and processes of IRBs offer a mechanism to address the issues of external validity and bias prevention. In the wake of the Tuskegee Syphilis Experiment in the mid-twentieth century, and the resulting congressional hearings, the National Research Act of 1974 set forth research requirements, including standardized requirements for the protection of human participants in medical research. These are codified in Title 45, Part 46 of the Code of Federal Regulations (45 CFR 46) Protection of Human Subjects. One of these requirements was the establishment of IRBs, creating a structured mechanism for review of the safety and ethicality of proposed studies, protecting the rights and welfare of research participants. The Act led to the Belmont Report, which highlighted three fundamental features for research design:

1) Respect for persons—subjects have autonomy and subjects with diminished autonomy are entitled to protection, and participation is voluntary. This principle is implemented by informed consent of the subject.
2) Beneficence—subjects are treated in an ethical manner, researchers do no harm, benefits of the research are maximized, and risks are minimized. This is implemented through a comprehensive and transparent review of the risk and benefits of the proposed research.
3) Justice—consideration of who receives the benefits of the research and who bears the burden. This impacts subject selection on the societal (i.e., social, racial, sexual, and cultural) and individual level (National Institutes of Health, 1979).

In 1991, 45 CFR 46 Subpart A was revised and the resulting section is referred to as the Common Rule, which was subsequently revised again in 2018. Research on AI in medical settings must be grounded in these requirements and principles. We put this into practice and provide a clear path for individual researchers and institutions to ensure high-quality, safe, and ethical AI research practices.

Ethical review boards are present around the world, both on the national level and in international organizations, and depending on their location, they operate in accordance with standards as the Declaration of Helsinki (World Medical Associations, 2008), the Council for International Organizations of Medical Sciences (CIOMS, 2016), and the World Health Organization. These boards also operate based on local procedures developed to meet regulatory and legal requirements, and those regulatory frameworks are predicated on sets of ethical principles (e.g., Belmont).

IRBs were instituted by the National Research Act and tasked with implementing these principles into practice. While IRBs are inherently decentralized, since they exist primarily within institutions and consist of different decision-makers subject to different state and local policies, all organizations performing federally funded research in the U.S. must meet the requirements described in 45 CFR 46, and if applicable, the requirements in 21 CFR parts 50 and 56 (Federal Drug Administration). Further, there are independent organizations, such as the Association for the Accreditation of Human Research Protection Programs (AAHRPP), which formalize and promote standardization and best practices across many non-governmental institutions. Similarly, the Department of Veterans Affairs Office of Research Oversight (ORO) enforces and promotes comparable quality control processes at the VA.[3]

Whether these processes are better centralized or decentralized is a subject of ongoing debate. For example, some researchers believe that a centralized oversight approach is preferable (Emanuel et al., 2004; Glickman et al., 2009), but innovations in distributed ledger technology may provide new ways to promote best practices in the research community through decentralization (Cong and He, 2019). We do not take a stand on the precise platform, but rather introduce a broader process that can be incorporated into all research compliance processes, whether centralized or decentralized.

3. Overview of ethical and operational requirements

The expansion of AI has introduced new challenges and opportunities that require a refinement in the process in which R&D is conducted, specifically to ensure the ethical application of AI on human subjects. The potential costs to subjects range from the inadvertent release of personally identifiable information (PII) or personal health information (PHI), to the loss of life due to an inappropriate recommendation or course of treatment. Researchers frequently develop models with AI-driven recommendations to improve the delivery of health care using patient information.[4]

3 https://www.va.gov/ORO/IRB.asp

4 See, for example, Atkins et al. (2022) for a comprehensive review in the area of medical informatics and the use of data from the Department of Veterans Affairs. Makridis et al. (2021b) also provides additional background on predictive models for clinical use and their relative efficacy with an application over COVID-19.

However, who is ultimately liable for these recommendations and what process exists for evaluating their efficacy? If there is not a transparent process to guarantee that the outcomes are generating value for the patient, then even if "good recommendations"—that is, as defined by what will positively affect patient wellbeing—are followed, they may still render distrust among patients. Furthermore, the absence of transparency may also increase the likelihood of recommendations that are not suited for patients by creating distance between how they feel and how AI models are built. For example, Agarwal et al. (2023) report the results of an experiment with radiologists where they randomized the provision of supplementary information from AI models, finding that humans—when supplemented by AI—did not perform any better than their counterparts.

Even properly designed and implemented clinical trials with AI tools may still fail to achieve desired aims, including inferior performance relative to a human, particularly when the model is trained on a limited sample or incomplete data. For example, a clinical support tool might be designed for a specific population, but changes within the population served by a medical center could attract a different set of patients for which the tool is not calibrated. Here, the onus is on providers to decide if the tool remains appropriate for their population. Preliminary work on liability, when basing treatment on AI recommendations, points to a precedent in case law. A physician is exposed to liability when their actions deviate from the standard of care and result in patient harm, regardless of the algorithm's accuracy in recommending non-standard care (Price et al., 2019).

Without transparency in how the tool was developed and evaluated, a physician may discard its recommendations entirely—good or bad—rather than face liability. In fact, distrust may be the norm even when good recommendations are perceived to come from unexplainable AI (Gaube et al., 2021). Transparency into the development and evaluation process may help physicians understand where and when to apply a specific tool, how the recommendations are generated, and move a specific algorithm closer to the standard of practice as other laboratory-developed diagnostics have done in the past, thus improving patient outcomes in well-determined settings. While these goals for AI applications are exciting and promising, they need to be disciplined with processes that simultaneously mitigate risk and liability, particularly with the scale of AI.

The considerations have led to an explosion of research activity in the area of "trustworthy AI," where the term "trustworthy" generally refers to systems where their design and implementation satisfies the highest possible standards of protection for those affected by their use.[5] Several of such frameworks have been created by U.S. federal agencies and international organizations, enumerating principles that AI systems must adhere to for their use to be considered responsible. Most notably, EO 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government states, "The ongoing adoption and acceptance of AI will depend significantly on public trust. Agencies must therefore design, develop, acquire, and use AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable laws" (White House, 2020). Table 1 provides a summary.

The application of these principles will look different for every organization—and even sovereign nations—since contexts will vary. Our focus, however, is on AI research and within the lens of the Department of Veterans Affairs—as we apply the principles of trustworthy AI within the health care context (e.g., see Makridis et al., 2021a). Both the development and application of such principles are important: merely outlining principles without mapping them to action conceals substantial political and normative differences (Mittelstadt, 2019).

4. Sample AI supplement for institutional review board applications

Based on the principles that we have presented and discussed so far, we now create a sample IRB module focused on adherence to principles of trustworthy AI. These questions will be uploaded to the National Artificial Intelligence Institute (NAII) at the Department of Veterans Affairs website for ongoing updates with additional templates for other researchers and stakeholders.[6]

4.1. Sample IRB module supplement

1. Are any artificial intelligence (AI) tools being developed in the project (e.g., data are being used to train and/or validate an AI tool)?
   ☐ Yes ☐ No

   If yes, complete the following:
   a. Describe how the data used to train/validate the AI tool is representative of the population that the algorithm is designed to impact. Be sure to describe the underlying population of interest and the sample that was used to train the model.
   b. Describe data wrangling tools and analytical controls that will be used to actively mitigate the potential effects of statistical bias and other misattributions of cause in or on the data.
   c. Describe plans to ensure that data used to train the AI tool will be used to perform effective AI decision-making, as well the metrics that will be used to evaluate effectiveness.

2. List the different sources of data used to train or validate any AI tool(s) and the way that the sample is constructed (e.g., with nationally-representative sample weights).
3. Is the AI algorithm intended to be used for commercial profit? ☐ Yes ☐ No

If yes, describe plans to reciprocate to the Veteran community if Veteran data are used in the development of any AI tools.

5 See, for example, Makridis et al. (2021a) for an early application of these principles to Veterans.

6 https://www.research.va.gov/naii/

TABLE 1 Summary of trustworthy principles of AI.

Principle	Explanation
Lawful and respectful of our nation's values	Agencies shall design, develop, acquire, and use AI in a manner that exhibits due respect for our nation's values and is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties.
Purposeful and performance-driven	Agencies shall seek opportunities for designing, developing, acquiring, and using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed.
Accurate, reliable, and effective	Agencies shall ensure that their application of AI is consistent with the use cases for which that AI was trained, and such use is accurate, reliable, and effective.
Safe, secure, and resilient	Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation.
Understandable	Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate.
Responsible and traceable	Agencies shall ensure that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI. Agencies shall ensure that AI is used in a manner consistent with these principles and the purposes for which each use of AI is intended. The design, development, acquisition, and use of AI, as well as relevant inputs and outputs of particular AI applications, should be well documented and traceable, as appropriate and to the extent practicable.
Regularly monitored	Agencies shall ensure that their AI applications are regularly tested against these principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or this order.
Transparent	Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders, including Congress and the public, to the extent practicable and in accordance with applicable laws and policies, including with respect to the protection of privacy and of sensitive law enforcement, national security, and other protected information.
Accountable	Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI, and shall monitor, audit, and document compliance with those safeguards. Agencies shall provide appropriate training to all agency personnel responsible for the design, development, acquisition, and use of AI.

4. Are there any big data repositories that primarily store Veteran data used to train the AI algorithm?
   ☐ Yes ☐ No

If yes, describe how any developed AI tools are planned to benefit Veterans.

5. Describe plans to ensure that the privacy and security of Veteran data are maintained during and after the research process, particularly through the application of AI and how that interacts with existing guidance from an information system security officer (ISSO).
6. Describe any decisions or recommendations that the AI tools would be making for human subjects, how these conclusions are reached, and any limitations to them.
7. Describe if research volunteers are specifically being informed of the use of the tool through the informed consent process.
8. Will study participants be impacted by decisions made by any AI tools?
   ☐ Yes ☐ No
   If yes, complete the following:

   a. Describe any risks associated with the application of AI to the project or human subjects, including the potential impact severity on study participants.
   b. Categorize the type of decision-making by AI:
   ☐ AI is used to help inform human decisions.
   ☐ AI drives decisions with human oversight.
   ☐ AI is fully autonomous (i.e., no oversight).

5. Piloting the AI IRB supplement in practice

The AI IRB supplemental questions were piloted at the VA across the NAII AI Network locations, including Washington, D.C.; Kansas City, MO; Tampa, FL; and Long Beach, CA. In this section, we now review a specific AI-related project that was submitted and evaluated with the supplemental IRB module, as well as discussing the overall experience with the module. We focus on a project that aimed to achieve early detection of lung cancer using non-invasive screening methods.

Lung cancer is a leading cause of death among Veterans (Zullig et al., 2012). Detection using a non-invasive screening for early-stage disease changes in suspicious pulmonary nodules could lead to better treatment and outcomes in patients with lung cancer. The study would specifically compare performance of a proprietary *in-vitro* blood test, referred to as a "liquid biopsy," to detect circulating tumors cells with the standard of care transthoracic/transbronchial biopsy or fine needle aspiration outcome for the diagnosis of lung cancer. The possible role of AI in the study was raised because the name of the sponsor and project both included the term "AI".

The IRB used the AI supplemental questions as guidance for the study reviews. The initial protocol did not provide clear language regarding how the proprietary "liquid biopsy" was developed and whether it involved machine learning. Furthermore, the protocol did not outline the intended use of radiology imaging obtained as part of routine clinical care and whether machine learning would be applied to images to develop a new product. Due to the lack

of transparency in the study, the local IRB had to assume that AI was involved in the research and the IRB reviewers applied the AI supplemental questions and requested specific responses from the sponsor regarding the use of AI. However, after several rounds of communication between the sponsor, local investigators, and the IRB, it was discovered that the sponsor, in fact, did not have any product or technology using AI.

While the sponsor decided to withdraw the study from one site, they pursued it at another with a more detailed project intake process based on the AI IRB module. However, even at the new site, the project eventually closed. This case highlights the lack of transparency among some sponsors in the AI sphere. Many companies with designs on conducting AI research are either start-ups or are primarily technology firms, who often have little to no experience in the clinical research realm. This experience emphasizes the importance of having many of the new AI sponsors partner with established Clinical Research Organizations (CROs), who have a well-established foothold in the clinical research regulatory arena and with AI, or providing the investigators and sponsor with clear guidance on the required content for studies with AI components.

After the IRB committee shared these concerns with the prospective vendor, the vendor made the decision to withdraw the study. The decision to remove or decline proceeding with a study is not necessarily a negative outcome, as not every study can or should be approved. The AI IRB module still provided an important litmus test that equipped the IRB committee to ask relevant questions. As this case demonstrated, the AI IRB module is an effective tool in preventing potential risks of inadequate use of AI at the R&D level. Saving time and resources, as well as safeguarding the trust invested in us by the VA participants and protecting their private information. Establishing a standardized set of questions to facilitate the review process for studies will save time and minimize risks, especially as not every Veterans Affairs Medical Center (VAMC) IRB, or even IRB in non-VA institutions have in place such requisite background. Moreover, the existence of such questions also provides potential vendors with guidelines for more transparency into the review process and will increase the quality of initial submissions.

This is one example of the questions identifying a study of concern. In practice, the majority of AI studies where these questions, or some version of these questions are applied, have not had controverted review. The AI IRB supplemental questions received a highly positive reaction from IRB and R&D committee members. "Any time I review a new study, AI and various aspects of AI are always in the back of my head. When I do see a study with a potential AI component, I have the IRB module for unbiased, structured reference," according to one member in leadership. "Standardized modules will provide standardization; facilitate comparison across sites or peer review by other sites; structure for evaluation with educational elements to guide reviewers not so well-versed with AI; guidance to investigators on formulating the submission project and thus avoiding unnecessary back-and-forth and delays, and potentially discouraging people from doing research."

Furthermore, at the Kansas City VA Medical Center (KCVA), seven AI-related studies have gone through the IRB process and four of them were reviewed through the lens of the supplemental AI module. "Before the module, whenever we used to review a project, we didn't take a look from an AI perspective, but now we do… the module has helped us ask the right questions," said Dr. Vikas Singh, a neurologist and IRB reviewer at KCVA. Another advantage of the IRB module has been its ability to help increase transparency among projects through ex-ante disclosure and enumeration of the different AI components, such as how the data will be used and stored even after the completion of the study. To date, there has been no objection from the investigators on the use of these supplemental questions and they have helped streamline the process for review.

6. Conclusion

The expansion of AI has led to more use in clinical trials. AI applications offer significant prospects for improving the health and wellbeing of patients, particularly promoting preventative behavior that allows individuals to mitigate risk and avoid serious health challenges years in advance. However, there are also many new risks that AI poses even before it is ever deployed—that is, in the R&D process. Therefore, an IRB process aimed to address these new, distinct risks is urgently needed.

Here, we pilot a new tool to address this need by introducing novel questions to the IRB review process. The next phase is to develop these questions into an AI IRB module with an extended application, review checklist, informed consent, and other informational materials. These questions are anchored to the federal principles that govern trustworthy AI within the Department of Veterans Affairs, but aim to be broadly applicable in other health care research environments.

These questions focus on eliciting information from the researchers about study design, data and statistical strategy, safeguards and risk-benefit analysis, while providing non-AI subject matter experts with a set of streamlined best practices. We subsequently pilot the IRB tool within the VA and find strong support for its efficacy, particularly among non-AI subject matter experts in IRB and R&D committees. Further, our questions provide guidance to sponsors of the expectations of the IRB. Although there is no method for completely eliminating risk, our approach provides a low-cost and accessible way to embed ethics and transparency into the design and development of AI.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CM led up the writing of the draft and contributed to the design of the IRB module. AB contributed to writing. RF, DW, MKl, and JM reviewed the draft. IH and MKi reviewed the draft and contributed to the design of the IRB module. GA reviewed the draft

and initiated the project. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the U.S. Department of Veterans Affairs National Artificial Intelligence Institute.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). "Combining human expertise with artificial intelligence: Experimental evidence from radiology," in *NBER Working Paper*. doi: 10.3386/w31422

Aghion, P., and Howitt, P. (1992). A model of growth through creative destruction. *Econometrica* 60, 323–251. doi: 10.2307/2951599

AI Index Report (2021). *Measuring Trends in Artificial Intelligence*. Stanford, CA: Stanford University Human-Centered AI Institute. https://aiindex.stanford.edu/report/

Anderson, M., and Anderson, S. L. (2019). How should ai be developed, validated, and implemented in patient care? *AMA J. Ethics*. 21, E125–130. doi: 10.1001/amajethics.2019.125

Atkins, D., Makridis, C. A., Alterovitz, G., Ramoni, R., and Clancy, C. (2022). Developing and implementing predictive models in a healthcare system: traditional and artificial intelligence approaches in the Veterans Health Administration. *Annu. Rev. Biomed. Data Sci.* 5, 393–413. doi: 10.1146/annurev-biodatasci-122220-110053

Buruk, B., Ekmekci, P. E., and Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med. Health Care Philos*. 23, 387–399. doi: 10.1007/s11019-020-09948-1

CIOMS (2016). *International Ethical Guideline for Health Related Research Involving Humans*. Geneva.

Cong, L. W., and He, Z. (2019). Blockchain disruption and smart contracts. *Rev. Financ. Stud.* 32, 1754–1797.

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* 25, 1519–1525. doi: 10.1038/s41591-019-0583-3

Crossnohere, N. L., Elsaid, M., Paskett, J., Bose-Brill, S., and Bridges, J. F. (2022). Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. *J. Med. Internet Res.* 24, 8. doi: 10.2196/36823

Emanuel, E. J., Wood, A., Fleischman, A., Bowen, A., Gets, K., Grady, C., et al. (2004). Oversight of human participants research: identifying problems to evaluate reform proposals. *Ann. Intern. Med*. 141, 282–291. doi: 10.7326/0003-4819-141-4-200408170-00008

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., et al. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digi. Med.* 4, 31. doi: 10.1038/s41746-021-00385-9

Glickman, S. W., McHutchison, J., Peterson, E. D., Cairns, C. B., Harrington, R., Califf, R. M., et al. (2009). Ethical and scientific implications of the globalization of clinical research. *N. Engl. J. Med.* 360, 816823. doi: 10.1056/nejmsb0803929

Goh, K. H., Wang, L., Yeow, A. Y. K., Poh, H., Li, K., Yeow, J. J. L., et al. (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat. Commun.* 12, 711. doi: 10.1038/s41467-021-20910-4

Henry, K. E., Adams, R., Parent, C., Soleimani, H., Sridharan, A., Johnson, L., et al. (2022). Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nat. Med.* 28, 1447–1454. doi: 10.1038/s41591-022-01895-z

Huang, P., Lin, C. T., Li, Y., Tammemagi, M. C., Brock, M. V., Atkar-Khattra, S., et al. (2019). Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. *Lancet Digit. Health* 1, 7. doi: 10.1016/S2589-7500(19)30159-1

Jones, C. (2002). Sources of U.S. economic growth in a world of ideas. *Am. Econ. Rev.* 92, 220–239. doi: 10.1257/000282802760015685

Khin-Maung-Gyi, F. (2009). The history and role of institutional review boards: local and central IRBs, a single mission. *AMA J. Ethics*. Available online at: https://journalofethics.ama-assn.org/article/history-and-role-institutional-review-boards-local-and-central-irbs-single-mission/2009-04

Makridis, C. A., Hurley, S., Klote, M., and Alterovitz, G. (2021a). Ethical applications of artificial intelligence: evidence for health research on veterans. *J. Med. Inter. Res. Med. Informat.* 9, 6. doi: 10.2196/28921

Makridis, C. A., Strebel, T., Marconi, V., and Alterovitz, G. (2021b). Designing COVID-19 mortality predictions to advance clinical outcomes: evidence from the department of Veterans Affairs. *BMJ Health Care Inform*. 28, e100312. doi: 10.1136/bmjhci-2020-100312

Matheny, M., Thadaney Israni, S., Ahmed, M., and Whicher, D. (2019). *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Washington, DC: National Academy of Medicine.

Mei, X., Lee, H.-C., Diao, K.-Y., Huang, M., Lin, B., Liu, C., et al. (2020). Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* 26: 1224–1228. doi: 10.1038/s41591-020-0931-3

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI perspectives. *Nat. Mach. Intellig.* 1, 501–507. doi: 10.1038/s42256-019-0114-4

National Institutes of Health (1979). "The National Commission for the protection of human subjects of biomedical and behavioral research," in *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Available online at: http://ohsr.od.nih.gov/guidelines/belmont.html (accessed May 30, 2023).

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 25, 447–453. doi: 10.1126/science.aax2342

OECD (2021). "State of implementation of the OECD AI principles: insights from national AI policies," in *OECD Digital Economy Papers*. Paris.

O'Shaughnessy, M. R. (2023). Five policy uses of algorithmic explainability. *arXiv*.

Price, W., Nicholson, S. G., Dipl-Jur, U., and Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. *J. Am. Med. Assoc.* 322, 17651766. doi: 10.1001/jama.2019.15064

Rivera, S. C., Liu, X., Chan, A-W., Denniston, A. K., Calvert, M. J., and Group, S-A. W. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* 26, 1351–1363. doi: 10.1136/bmj.m3210

van de Sande, D., Van Genderen, M. E., Smit, J. M., Huiskens, J., Visser, J. J., Veen, R. E. R., et al. (2022). Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Inform.* 29, 100495. doi: 10.1136/bmjhci-2021-100495

White House (2020). "Executive order on promoting the use of trustworthy artificial intelligence in the Federal Government," in *Federal Register: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*. Washington, DC.

Wismüller, A., and Stockmaster, L. (2020). "A prospective randomized clinical trial for measuring radiology study reporting time on Artificial Intelligence-based detection of intracranial hemorrhage in emergent care head CT," in *Proc. SPIE 11317, Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging 113170M*. Washington, DC. Available online at: https://spie.org/about-spie/the-society

Yu, Kun-Hsing, B., Andrew, L., and Keohane, I. S. (2018). Artificial intelligence in healthcare. *Nat. Biomed. Eng.* 2, 719–731. doi: 10.1038/s41551-018-0305-z

Zhou, D., Tian, F., Tian, X., Sun, L., Huang, X., Zhao, F., et al. (2020). Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. *Nat Communic.* 11, 2961. doi: 10.1038/s41467-020-16777-6

Zullig, L., Jackson, G., Dorne, R., Provenzale, D., McNeil, R., Thomas, C., et al. (2012). Cancer incidence among patients of the U.S. Veterans Affairs Health Care System. *Military Med.* 177, 693–701. doi: 10.7205/MILMED-D-11-00434