MULTI-REGIONAL CLINICAL TRIALS
THE MRCT CENTER OF
BRIGHAM AND WOMEN'S HOSPITAL
and HARVARD

wcg™

Framework for Review of Clinical Research Involving Artificial Intelligence

JUNE 2025

Acknowledgments

Our thanks to the following individuals who contributed their time and expertise to develop this resource, and without whom this would not have been possible.

MRCT Center & WCG Artificial Intelligence Task Force

LEADERSHIP

TREVOR BAKER, Program Manager, MRCT Center

BARBARA E. BIERER, Faculty Director, MRCT Center

DONNA L. SNYDER, Executive Physician, WCG

TASK FORCE

NICK BOTT, Head, Digital Ethics & Compliance, Global R&D, Takeda

DOROTHEE CAMINITI, Director, Bioethics, Markkula Center for Applied Ethics, Santa Clara University

CANSU CANCA, Director of Responsible AI Practice at the Institute of Experiential AI, Northeastern University

DAVID CLIFFORD, Head of Data Science and Applied Machine Learning, Biogen

CHERYL DANTON, IRB Managing Director, University of Chicago

CAROLINE DAVIS, Senior Program Manager, Ethics, Microsoft Research

TAMIKO ETO, Director, HRPP and IRB Research Operations, Mayo Clinic

KELLY FITZGERALD, IRB Executive Chair and Vice President, IBC Affairs, WCG

IASON GABRIEL, Senior Staff Research Scientist, Google DeepMind

MARY L. GRAY, Senior Principal Researcher, Microsoft Research

MISSY HEIDELBERG, Director Bioethics, Takeda

MARK LIFSON, Director, AI/ML Engineering, Mayo Clinic

ALEX JOHN LONDON, K&L Gates Professor of Ethics and Computational Technologies, Carnegie Mellon University

BRENNA LOUFEK, Director of AI, Regulatory & Quality, Mayo Clinic

CURRIEN MACDONALD, IRB Medical Chair Director, WCG

CAMILLE NEBEKER, Director of the Research Center for Optimal Ethics, University of California, San Diego

KEVIN NELLIS, Executive Director, Human Research Protection and Quality Assurance, The State University of New York, Downstate Health Sciences University

ANTONIA PATERSON, Science Manager, Responsible Development and Innovation, DeepMind

IRINA RAICU, Director, Internet Ethics Program, Markkula Center for Applied Ethics, Santa Clara University

STEPHEN ROSENFELD, Executive Director, North Start Review Board

REBECCA ROUSSELLE, Assistant Vice President for the Human Research Protection Program, Emory University

SUSAN RUBIN, The Ethics Practice

JOEL JIEHAO SEAH, PhD Candidate, Centre for Biomedical Ethics, National University of Singapore

SHARON SHRIVER, Senior Analyst, Policy Analysis & Legislative Support, American Cancer Society Cancer Action Network

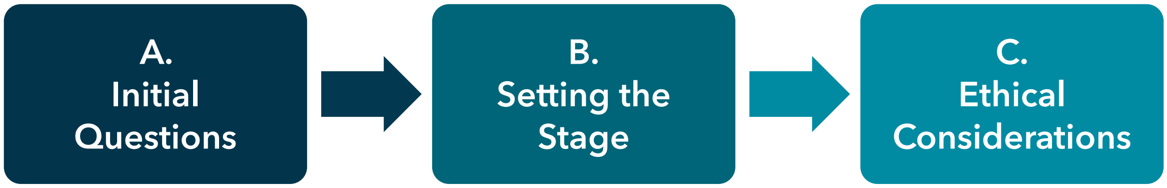
BENJY SILVERMAN, Senior IRB Chair, Mass General Brigham

MEGAN SINGLETON, Associate Dean, Human Research Protections, Johns Hopkins University

Abbreviations

AI	Artificial Intelligence
BAA	Business Associate Agreements
CDS	Clinical Decision Support
DA	Data Agreements
DUA	Data Use Agreements
FDA	Food and Drug Administration
FD&C	The Food, Drug, and Cosmetic Act
HHS	Department of Health and Human Services
HIPAA	The Health Insurance Portability and Accountability Act
IDE	Investigational Device Exemption
IND	Investigational New Drug
IRB	Institutional Review Board
LLM	Large Language Model
MTA	Material Transfer Agreements
NIST	National Institute of Standards and Technology
PI	Principal Investigator
PHI	Protected Health Information
REC	Research Ethics Committee
SaMD	Software as a Medical Device
SiMD	Software in a Medical Device
US	United States of America

IRB Approach and Considerations



Background: In recent years, the use of artificial intelligence (AI) in research involving human participants[1] has increased substantially. AI is transformative in its ability to perform repetitive tasks, analyze large amounts of data, and potentially augment and enhance clinical research decision-making. As with all research involving human participants, research using AI should be grounded in ethical principles and subject to regulations to protect those persons volunteering in research.

Challenge: The use of AI in research[2] involving human participants presents new ethical and regulatory challenges, including those related to privacy and data confidentiality, transparency, amplification of bias, and implications for maintaining human autonomy and oversight. Existing guidelines from the U.S. Food and Drug Administration (FDA),[3,4] the U.S. Department of Health and Human Services (HHS),[5] and the National Institute of Standards and Technology (NIST)[6] provide an overarching structure to protect the human participant's privacy and mitigate bias. While these guidelines provide the ethical foundations for the use of AI, additional, more specific tools are needed to enhance the protocol review process by entities charged with oversight to ensure consistency and thoroughness across institutions and that reflect accepted standards of ethical research. Those entities include institutional review boards (IRBs), also termed research ethics committees (RECs). This document aims to provide IRBs and other reviewing entities with practical, actionable steps to identify, assess, and mitigate potential risks to participants by the use of AI in research.

[1] Here, the term "participant(s)" is used to describe a living individual about whom an investigator is conducting research through interaction or intervention, or obtains, uses, studies, analyzes, or generates identifiable information or biospecimens. Whenever this document refers specifically to regulatory language, the term "subject" appears in place of "participant." See also [45 CFR § 46.102 (e)]

[2] In the course of research, the AI or AI system is deployed: (i) as a study intervention; or (ii) to have a direct influence on the intervention, so as to bring about a change - temporary or permanent - in the research participant (e.g., in their condition, actions, or outcomes).

[3] Good Machine Learning Practice for Medical Device Development: Guiding Principles | FDA.gov

[4] Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products | FDA.gov

[5] IRB Considerations on the Use of Artificial Intelligence in Human | HHS.gov

[6] Towards a Standard for Identifying and Managing Bias in Artificial Intelligence | NIST

Purpose of the Framework: This framework supports IRBs in evaluating whether research involving AI meets regulatory requirements and aligns with longstanding foundations of ethical research, including but not limited to principles from the Belmont Report (Respect for Persons, Beneficence, and Justice). By offering a structured decision process, this tool is intended to:

1. Prompt critical questions about the role of AI in a research protocol
2. Guide IRBs through key decisions to determine appropriate oversight
3. Identify and address potential risks and benefits associated with the use of AI in research
4. Promote consistency and clarity in IRB reviews, communications, and outcomes

IRBs should recognize that complete information about AI systems may not always be available, particularly when proprietary algorithms or external data sources are used. The IRB should seek consultation if there are complex issues or questions with which they are unfamiliar. This framework complements existing regulatory tools and aims to ensure that research involving AI protects human participants and their data while advancing scientific innovation. The framework described here is relevant and can be adapted to other entities responsible for the conduct and oversight of clinical research; for simplicity, we have limited the perspective to that of the IRB. We have also restricted the discussion to U.S. regulations and jurisdiction; local regulations, guidelines, and processes should always be considered.

This framework is organized into sections. The first is to consider whether a protocol that deploys AI is subject to regulatory oversight in the U.S., and it presents both a set of initial questions and a prototype decision tree that will help in that determination. The second section sets the stage to understand the type of AI use and the potential risks of each. The third section presents ethical implications and points to consider in the exercise of thoughtful and complete IRB review.

Download the Toolkit

https://mrctcenter.org/wp-content/uploads/2025/06/2025_AI-Toolkit_06-23-2025.pdf

Toolkit Components

A. Initial Questions for Determining Required IRB Oversight

The foundational questions are designed to help IRB reviewers determine whether a proposed research protocol involving AI requires IRB oversight under current U.S. regulations. By addressing these high-level questions, IRBs can assess whether the use of AI aligns with ethical and regulatory standards that protect human participants.

Instructions for Use: This decision tree is intended to provide a step-by-step guide for IRB reviewers to assess protocols involving the use of AI in the research or as the "subject" of research itself (e.g., the research question centers around the development, utility, efficacy, and/or safety and risk of the AI algorithm). Each question includes context for evaluation and recommended next steps. Note that many of the defined terms are sourced from the Common Rule; FDA considerations, if they apply, appear in later sections of this framework.

- **Question:** Presents the key question to address at each review step.
- **Context:** Defines the standards or definitions guiding a decision.
- **Next Steps:** Outlines the reviewer's next actions based on the answers.

Question	Context and comments	Next Steps
1. Is the activity considered "research" under US federal definitions?[7]	Research is a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge [45 CFR § 46.102 (l)] The answer to this question may not be obvious. In general, activities intended to improve local processes (e.g., local QI/QA activities) are not typically considered research, but if the scope is to apply the lessons from such activities more broadly, then they may constitute research.	**Yes:** Proceed to question 2. - Consider reviewing questions in the Discovery stage for AI technology in early development. **No:** IRB review is not generally required.
2. Does the research involve human participants?	Human Subjects refer to living individuals about whom an investigator obtains **data** or biospecimens through intervention or interaction, or obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens. [45 CFR § 46.102 (e)]	**Yes:** Proceed to question 3. - Consider additional questions for AI systems in the Translation or Deployment stage. **No:** IRB review is not generally required.
3. What is the intended use of the AI technology in the research study?[8]	Types of AI deployments include: **Administration of Research** (e.g., data analysis support, recruitment, transcribing interviews) **AI as the Intervention** (e.g., clinical decision-making or therapeutic intervention, AI-enabled medical devices)	**If for Administration of Research:** Refer to Part D of the framework **If AI is the Intervention** Proceed to question 4.

[7] 45 CFR 46 | HHS.org. see https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46

[8] Intended use is the purpose or purposes for which an AI health technology supplier specifies that they intend the technology to be used. It is usually specified by the manufacturer, person, or organization legally responsible (Alderman et al. 2025).

4. What is known about the AI algorithm? (i.e., origins, and "marketed" or intended use)	Consider details in the protocol on whether the AI is a pre-existing tool (e.g., available commercially, open source, developed locally), and whether the current use is consistent with the study's intended use, or developed specifically for this research intervention.	**Sufficient Details in Protocol:** Proceed to question 5. **Insufficient Details in Protocol:** Request additional information about the AI system's developmental stage, intended use, and validation.[9] Refer to Part B for more information on AI developmental stages.
5. Has a risk analysis of the AI technology been conducted? Is there adequate evidence of risk considerations within the protocol?[10]	Risks could include impacts on clinical decision-making, amplification of bias, data confidentiality, identifiability, and privacy that could affect human participants.	**Minimal Risk:** Document risks and proceed to question 6. **Risks Identified:** Determine if they can be minimized or require further review. **More than minimal risk:** Full Board IRB review is required. Consider supplemental questions in Part C here.

[9] See Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations | FDA.gov for more information on AI Algorithm model cards.

[10] The FDA's Draft Risk-Based Credibility Assessment Framework outlines a process for determining risk across the research lifecycle.

6. Does the research qualify for exemption under the Common Rule?	Exempt categories may include benign behavioral interventions, educational practice studies, or secondary research of identifiable or linkable data.[11] [45 CFR § 46.104]	**Yes:** Document the exemption. In cases where limited review[12] is conducted, refer to Part C, particularly information on Informed Consent. Otherwise, conclude the review. **No:** Proceed to question 7.
7. Does the research involve more than 'minimal risk' to human participants?	Minimal risk means the probability and magnitude of harm or discomfort anticipated in the research are not greater than those ordinarily encountered in daily life or routine exams. [45 CFR § 46.102(j)]	**Yes:** Full Board IRB review is required. • Consider supplemental questions in Part C. **No:** Consider eligibility for expedited review. [45 CFR § 46.110]

[11] Note that benign behavioral interventions "are brief in duration, harmless, painless, not physically invasive, not likely to have a significant adverse lasting impact on the subjects, and the investigator has no reason to think the subjects will find the interventions offensive or embarrassing." [45 CFR § 46.104 (d)(3)(ii)]

[12] See conditions for limited IRB review at § 46.104(d)(2)(iii), (d)(3)(i)(C), or (d)(7) or (8).

B. Setting the Stage: Review Guide for the Stage of AI Development

This guide supports IRB reviewers in assessing research protocols involving AI by aligning review considerations with the stage of development of the AI system. The prompts below correspond to three phases (Discovery, Translation, and Deployment),[13] and they are intended to help reviewers understand how each stage presents different implications regarding ethical considerations, data requirements, and regulatory oversight.

In addition to posing questions for each stage of AI development for clinical research, this section also includes potential questions regarding the data sources, collection methods, and the identifiability of data.

[13] Eto T, Lifson M, Vidal D. Pre-print: A novel, streamlined approach to the IRB review of artificial intelligence human subjects research (AI HSR). Whitepaper. September 2024. https://purl.stanford.edu/zj025zw1714

Discovery

The **Discovery** stage marks the conceptualization and early development of AI algorithms in research. It involves gathering and early analysis of training data to explore potential use cases. Refer to more specific considerations for Data Sources and Identifiability for additional considerations.

Questions for the Discovery Stage	Considerations for the Discovery Stage
Where/What are the sources of data for this research? What considerations have been made regarding data identifiability or the linkability of individual participants' data?[14]	• Review the provenance of the data, including how and under what terms (e.g., secondary use) the data were originally collected, ensuring compliance with relevant regulations (e.g., 45 CFR 46).[15] • Confirm informed consent includes provisions for current and future uses of data or has received a waiver of consent under the regulations. • For use of Protected Health Information (PHI), confirm that applicable HIPAA Authorizations or waivers/alterations are adequate.[16]
Does the research involve secondary use of data or integration of external datasets?	• Evaluate governance structures to ensure appropriate agreements [e.g., business associate agreements (BAA) and data use agreements (DUA), data agreements (DA), material transfer agreements (MTA)]. • Assess privacy risks to human participants, including reidentification, from combining datasets and secondary use of data.

[14] Here, the concept of linkable data reflects the growing potential of re-identification of individuals, even from datasets in which individual identifiers are removed, in an era of advanced integration of multiple datasets. For additional information, see the Identifiability of Data section in this framework.

[15] The HIPAA Privacy Rule. U.S. Department of Health and Human Services, Office for Civil Rights | HHS.gov

[16] HIPAA for Professionals | HHS.gov

Translation

The **Translation** stage involves advancing AI systems in research from 'conceptual development' to 'validation,' emphasizing performance testing and identifying risks. This stage is pivotal in establishing the accuracy and reliability of AI systems before they are deployed in clinical settings.

Questions for the Translation Phase	Considerations for the Translation Phase
Is the AI algorithm's intended use and purpose clearly defined?	• Confirm that the AI's appropriate use protocol includes comprehensive details of the AI system's role, objectives, and how it interacts with study participants or research staff. • Ensure alignment with trial goals and regulatory frameworks.
Is the downstream intended use of the AI system proposed to diagnose, alleviate, mitigate, treat, cure, or prevent a disease, disorder, or injury in humans, or is it for exploratory purposes?	• Refer to the anticipated intended use in the protocol and consider whether the study would require an Investigational New Drug (IND) or Investigational Device Exemption (IDE).[17] • Ensure research objectives are clearly outlined and assess whether exploratory research or preparation for research activities pose risks to human participants.
How will performance metrics (e.g., accuracy, false positive/negative rates) be evaluated? How will risks of harm be evaluated for demonstration that they have been mitigated?	• Confirm that metrics are representative of the sample population. • Ensure plans are included to address discrepancies and potential bias amplification. • Performance metrics should also be considered in the Discovery stage to establish whether continued development should occur.

[17] How to Determine if Your Product is a Medical Device | FDA.gov

Deployment

Deployment refers to the use of a tested and validated AI system within a research context, with the immediate concern of negatively influencing clinical decision-making and/or making harmful changes to the participant's diagnosis, treatment, disease prevention, or well-being. The concerns include risks to participant safety and/or privacy, and/or amplification of bias, particularly when the AI system may be considered Software as a Medical Device (SaMD).[18,19] This stage requires heightened scrutiny and ongoing human oversight to ensure the protection of human participants.

AI algorithms in the deployment stage of interventional clinical trials can be used as companion diagnostic devices, providing information essential for the safe and effective use of a corresponding drug or biological product. They can also assist in the selection of participants, treatment decisions, monitoring, and detection of adverse events or efficacy signals. The range of uses presents a challenge for regulatory and oversight authorities, including IRBs, regarding risk assessment and regulatory device determinations. The utilization of AI algorithms as companion diagnostics in clinical trials can significantly contribute to the advancement of precision medicine.

Refer to the "Algorithm Stability" section for additional considerations.

[18] Software as a Medical Device (SaMD) | FDA.gov

[19] Software as a Medical Device (SAMD): Clinical Evaluation Guidance for Industry and Food and Drug Administration Staff | FDA.gov

Questions for the Deployment Phase	Considerations for the Deployment Phase
Is there human oversight during the trial? How are AI outputs monitored in a safe and timely manner? Will the investigators and relevant healthcare professionals involved review the AI's outputs, particularly if the outputs will influence the human participant's clinical or research care (e.g., diagnosing, alleviating, treating, or preventing a disease)? Should human participants be provided with clinically relevant results, and/or should follow-up with a healthcare provider be recommended?	• Ensure oversight roles and responsibilities are clearly outlined, including protocols for investigators and healthcare professionals to review AI outputs promptly before they influence clinical or research decision-making, or introduce other important changes to participant's care.
Does the AI algorithm or software that incorporates the AI meet the requirements for regulatory review as a device or Software as a Medical Device (SaMD)? Does the intended use of the AI system meet the criteria of software functions that the FDA does not regulate as medical devices per section 201(h) of the Food, Drug, and Cosmetic (FD&C Act)?[20,21]	• Certain devices, such as those used for clinical decision support (CDS), general wellness, or medical device storage, may not be considered devices.[22] • Software in a Medical Device (SiMD) and Mobile Medical Applications are subject to different levels of oversight, which depend on the device classification and risk.[23] • Consider whether the software meets the criteria for Non-Device CDS software functions or other software excluded from IDE regulations.

[20] Changes to Existing Medical Software Policies Resulting from Section 3060 of the 21st Century Cures Act | FDA.gov

[21] Policy for Device Software Functions and Mobile Medical Applications | FDA.gov

[22] Examples of Software Functions That Are NOT Medical Devices | FDA.gov

[23] Policy for Device Software Functions and Mobile Medical Applications | FDA.gov

Questions for the Deployment Phase	Considerations for the Deployment Phase
Does the intended AI use meet the four (4) criteria for a Non-Device Clinical Decision Support (CDS) software?[24] Does the intended AI meet the criteria for Medical Device Data Systems, Medical Image Storage Devices, and Medical Image Communications Devices that is either not a device or where enforcement discretion might be applicable?[25]	• Consider Good Machine Learning Practices, Technology Assessments, and requirements of IDE approvals. • Consider whether an IND or IDE will be needed for the trial. • Consider whether a separate regulatory device determination (IDE/Abbreviated IDE) is needed for an AI algorithm companion diagnostic.[26]
What measures exist in the protocol to ensure transparency of any changes to clinical or research care that are introduced by the AI application?	• Detail plans for transparency in decision-making processes, including the understanding of the AI outputs for clinicians and participants.
Is the AI "locked" or adaptive (involving continuous "learning" as more data is included in the training)? Are there limits to allowed adaptations? What metrics are used to assess the AI system's performance over time?	• Require procedures for monitoring and validating updates to adaptive algorithms. • Establish prospective timelines in collaboration with the sponsor or PI that would prompt re-review by an IRB or return to an acceptable performant model if model performance changes, preventing performance "drift."

[24] Clinical Decision Support Software – Guidance for Industry and Food and Drug Administration Staff | FDA.gov
The four criteria are:
(1) Not intended to acquire, process, or analyze a medical image or a signal from an in vitro diagnostic device or a pattern or signal from a signal acquisition system;
(2) Intended for the purpose of displaying, analyzing, or printing medical information about a patient or other medical information;
(3) Intended for the purpose of supporting or providing recommendations to an HCP about prevention, diagnosis, or treatment of a disease or condition;
(4) Intended for the purpose of enabling an HCP to independently review the basis for recommendations that such software presents so that it is not the intent that the HCP rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient.

[25] Medical Device Data Systems, Medical Image Storage Devices, and Medical Image Communications Devices | FDA. gov

[26] Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product | FDA.gov

Algorithm "Stability"

Algorithm "stability" in this context refers to whether the AI system is fixed, "locked," or is designed to continuously adapt (i.e., change/improve) its outputs when it is provided with more training data either externally or from the algorithm's outputs, or when its algorithm is updated or enhanced. Data shifts refer to changes in the statistical nature or distribution of all or some of the data included in the dataset that, when substantial, can result in a mismatch between the data used to train the algorithm and the context in which it is intended to be used, affecting its performance and contributing to the amplification of bias.[27]

Questions for Algorithm "Stability"	Considerations Questions for Algorithm "Stability"
Will the AI algorithm involve continuous updating ("learning"), are there limits on adaptation, will updates be made periodically, or is it "locked" or fixed such that no changes will be introduced to the AI algorithm with time? What oversight mechanisms are in place to ensure it remains locked, or to monitor and audit updates if adaptive?	• If the model involves continual learning, adaptations, or changes over time, inquire about schedules for additional model training, criteria for re-review, and mechanisms in place to ensure stability over time. This may be identified via a model card,[28] a Device Performance Monitoring Plan, or other means if not FDA-regulated.[29]
If there are continuous data updates to the model, what mechanisms are in place to assess their impact on risks to participants?	• Request timelines for ongoing risk assessments before deployment.
Will the IRB be notified, and will the IRB review the changes to ensure that the research meets its objectives and that no additional risks to participants are introduced?	• Develop procedures for prompting the submission of a protocol amendment and IRB re-review based on prospectively identified risk levels.

[27] Alderman JE, Palmer J, Laws E, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. Lancet Digit Health. 2025

[28] Here, a model card is a structured report of relevant technical characteristics of an AI model and benchmark evaluation results relevant to the intended application domains. Model cards also provide information about the context in which models are intended to be used and details of how their performance was assessed. | FDA Digital Health and Artificial Intelligence Glossary

[29] Draft Guidance for Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations |FDA.gov

Identifiability of Data

Current regulations in the United States regarding the protection of human participants (2018 Common Rule) define identifiable private information as "*private information for which the identity of the subject is or may readily be ascertained by the investigator or associated with the information.*"[30] However, especially due to advances in generative AI [and particularly large language models (LLMs)], it is increasingly likely that participants might be identifiable in the context of broadly deployed AI. Therefore, in addition to addressing the regulatory differentiation of the terminology 'identifiable', 'de-identified', or 'anonymized' data, IRBs might consider using the terms 'linkable' or 'not linkable' to indicate whether the training data (e.g., health records) used in the development of AI systems can be linked back to an individual human participant based on the type of data collected.

Questions for Identifiability of Data	Considerations for Identifiability of Data
Are the data: 1) Identifiable; 2) De-identified (e.g., a code linking back to the human subject exists, but is unavailable to the investigator or algorithm); or 3) Anonymized (e.g., cannot be linked back to the human subject)?	• Ensure protections such as differential privacy and encryption are in place to mitigate re-identification risks.[31] • Ensure the research complies with applicable HIPAA privacy and security regulatory requirements.[32]
Are there any direct identifiers in the data (e.g., name, email address, audio or video recordings) or information that might be combined, such as age, gender, sexual orientation, place of employment, telemetry, etc.?[33] What measures are in place to mitigate the risk of re-identification, and how likely is it that someone with access and intent could re-link the data to an individual?	• Assess safeguards, including data governance processes such as data use agreements, secure access protocols, data minimization and obfuscation, and data anonymization techniques. • Consider whether the data can be used in a secure compute platform. • Ensure regular audits are in place to evaluate the risk of data linkage.

[30] The Common Rule (2018 Revision)
[31] Differential Privacy | Harvard University Privacy Tools Project
[32] HIPAA for Professionals | HHS.gov
[33] "Dataset Reflection Questions" | Microsoft Research

Questions for Identifiability of Data	Considerations for Identifiability of Data
If the data are identifiable or de-identified, were consent and HIPAA authorizations provided for the use of the data? Is a waiver of the requirements for informed consent and/or a HIPAA waiver/alteration appropriate?	• Confirm that informed consent and HIPAA authorizations adequately address data use, including secondary uses or future uses, and is conveyed in a meaningful way to participants. • Ensure participants understand the potential implications of secondary use (see last question of "Data Sources and Collection") and agree to any future uses involving their data.

Data Sources and Collection

The integrity of AI-driven research relies heavily on the quality, provenance, and management of the data used to train, validate, and deploy AI systems. The increasing use of secondary data sets, obtained from public and privately sourced data, coupled with the complexities of obtaining informed consent, necessitates a thorough review by IRBs to ensure compliance with ethical and regulatory standards.

Questions for Data Sources and Collection	Considerations for Data Sources and Collection
What is the source of the data, and is it publicly available or collected directly from human participants?	• Publicly available data may not require consent, but data collected directly from participants must comply with informed consent and HIPAA authorization requirements.[34]
If primary data is collected, will informed consent or a waiver of consent be obtained?	• Ensure informed consent clearly outlines privacy risks and whether data will be used solely for the current research project, future AI development, or shared with third parties. • Evaluate whether a waiver may be permissible.
Does the informed consent cover data use for the current research project, future algorithm development, and/or other stated purposes?	• Confirm informed consent explicitly includes provisions for intended uses of participant data that align with ethical standards (e.g., transparency).

[34] The conventional regulatory interpretation of "publicly available" data, such as census records, differs significantly from modern contexts (e.g., social media, online forums). Although such data may be technically public, ethical considerations remain when individuals do not reasonably anticipate that their disclosures would be used for research. While this falls outside the scope of current regulations, an ethically grounded approach asks whether individuals understood how their data might be accessed or reused, even if it was shared in a public setting.

Questions for Data Sources and Collection	Considerations for Data Sources and Collection
If the protocol involves secondary use of data, was it originally collected for research purposes aligned with the study's intended use of an AI system?	• Confirm that consent initially obtained to collect data for the original dataset applies to the proposed research or constitutes a reasonable extension of the permitted use. • If not, the IRB should consider whether the secondary use of the data is appropriate.
Are there agreements in place (e.g., BAAs or DUAs) for the use of secondary or external data?	• Ensure that appropriate agreements are in place to authorize the use of secondary or external datasets, and that these agreements meet regulatory standards.
Is the data source reliable?	• Consider if reasonable assurance is provided that datasets (primary or secondary) are appropriately sized, reliable, and representative of the population being studied.[35]
Is it clear to the human participant that their research data (which may include personal data) will or may be retained for future algorithm development even if they were to subsequently withdraw from the research?	• Confirm that this information is included in the informed consent, with processes to ensure the potential human participant understands this resulting implication should they choose to participate.

[35] Draft Guidance for Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations | FDA.gov

C. Ethical Considerations

As a threshold consideration, the ethical foundations for human participant research involving AI should be grounded in the key ethical principles of 'Respect for Persons', 'Beneficence', and 'Justice' outlined in the Belmont Report;[36] the Regulation for the Protection of Human Participants (U.S. Common Rule),[37] the International Ethical Guidelines for Health-related Research Involving Humans;[38] and the Declaration of Helsinki,[39] which underpins the ethical rules for conducting biomedical and behavioral research with participants in the United States.

The validation and deployment of AI in research settings amplifies existing ethical complexities (related to AI utilization in human-centric work), introducing risks related to data privacy, bias amplification, transparency, and potentially diminishing human oversight in the clinical decision-making process. For instance, continuous learning algorithms can introduce algorithmic drift that changes the model's outputs over time, necessitating continuous monitoring and validation. Similarly, large datasets used in AI systems often raise privacy concerns, particularly when anonymization methods may no longer suffice to protect individuals in an era of advanced integration of multiple datasets.

To address these challenges, ethical research involving AI as the intervention requires deliberation of the following ethical considerations during IRB review: (1) Human Agency and Oversight; (2) Technical Robustness and Safety; (3) Privacy, Confidentiality, and Data Governance; (4) Transparency; (5) Representativeness and Fairness; (6) Informed Consent.[40, 41] These principles emphasize the need for explainable and auditable AI systems representative of the study population of interest. By first connecting these considerations specifically as they arise in and apply to AI with the foundational principles for human research identified in the Belmont Report, specific questions are then proposed to support the IRB's review of protocols where AI systems are embedded in the intervention.

[36] The Belmont Report | HHS.gov

[37] The Common Rule (Revised 2018) | HHS.gov

[38] International Ethical Guidelines for Health-related Research Involving Humans | CIOMS

[39] Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Participants | WMA

[40] Diaz-Rodriguez N, Del Ser J, et al., Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion. https://www.sciencedirect.com/science/article/pii/S1566253523002129

[41] Alderman JE, Palmer J, Laws E, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. Lancet Digit Health. 2025;7(1):e64-e88. doi:10.1016/S2589-7500(24)00224-3

Human Agency and Oversight

Human agency and oversight in research involving AI ensure that the autonomy of all parties, including patients, researchers, and clinicians, is respected, and human judgment remains central to decision-making processes, particularly when decisions impact diagnosis, treatment, disease prevention, or well-being. As AI systems are increasingly integrated into clinical workflows, it is critical to ensure that researchers maintain control over how AI outputs influence study and human participant (and patient) outcomes. Safeguards, for example, protocols for clinician oversight and criteria for intervening in AI outputs, must be in place to prevent over-reliance on AI, de-skilling, and deference to its outputs and recommendations. Risk assessments should be included in research protocols to identify potential harms to both participants and intended end users. Training for researchers and clinicians to understand the AI's limitations and the explainability/interpretability of its outputs should also be detailed.

1. Will AI be deployed autonomously or unsupervised, or will human decision-making and/or oversight be required as part of the process before any changes are made in the research, the AI outputs, or its recommendations?

2. Have the researchers considered how AI outputs could or would be integrated into clinical workflows and decision-making processes, and how it might influence a healthcare professional's decisions?

3. What safeguards are in place to maintain human judgment in critical decisions?

4. Have the researchers conducted a risk assessment to identify potential risks and/or harms to prospective study participants arising from AI's use? What are the expected benefits and potential risks of deploying AI in this specific clinical research context?

5. What training and education will be provided to study team members and clinical staff interacting with the AI algorithm and/or using its outputs?

Technical Robustness and Safety

AI systems used as the intervention in research must demonstrate technical robustness to ensure safety and reliability. This includes comprehensive validation testing (see translation) to identify potential harms before deployment and protocols in place for monitoring performance (e.g., accuracy) throughout the study. Continuous or adaptive learning algorithms present unique challenges, such as algorithm stability, necessitating more rigorous oversight mechanisms and predefined thresholds for identifying when the algorithm has consequently changed. If the risk to participants has changed, IRB re-review is indicated. The consideration of algorithm stability and shift is connected to Transparency in that communicating the limitations—and potential variability—of the AI system to clinicians, study participants, and regulators is vital to prevent unintended consequences or harm.

1. What are the AI's limitations, and how will these limitations be communicated to different parties such as clinicians, participants/patients or their legally authorized representatives, administrators, oversight bodies, and the public?

2. Has adequate validation testing and relevant evaluations been done before deployment to minimize potential downstream harms?

3. What processes will be in place for researchers to identify and respond to unexpected or adverse events related to output and/or performance?

4. If the AI is simultaneously trained and developed during the research project, what checks and balances will be in place to ensure that the AI outputs continue to be reliable, and without additional risk to participants? Will the research project include automated alerts for human review or layered decision approval systems?

Privacy, Confidentiality, and Data Governance

Research involving AI relies on vast amounts of data, raising complex issues around privacy and confidentiality. To protect the information of participants collected as part of a research study, IRBs must ensure that protocols outline robust data governance strategies, including encryption, anonymization, and differential privacy methods, as appropriate. As data linkage capabilities advance, IRBs should assess whether data is linkable and verify safeguards against re-identification risks. Clear data retention and deletion protocols should be established and communicated to the participants, particularly when participants withdraw from a study. Finally, IRBs should have sufficient information about the dataset, including concepts of traceability, linkability, and auditability, to ensure that the training data supports the intended use of the AI system. Data cards and model cards are increasingly used to document data use during the discovery and translation stages before deploying an AI system.[42]

1. What safeguards are in place to protect the privacy and confidentiality of individuals whose data is used in the AI algorithm?

2. How will the development of the AI be documented, including its training data and performance metrics?

3. How will data retention and deletion be handled, especially if participants withdraw from the study, noting that data incorporated into an AI algorithm cannot be removed?

4. Are the proposed AI activities in accordance with the institution's AI governance principles, guidance, or policy?

[42] Draft Guidance for Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations | FDA.gov

Transparency

Transparency in research involving AI is closely tied to the concept of explainability and interpretability. Clear documentation and communication of the AI system's capabilities, limitations, and decision pathways can mitigate risks, ensuring that reviewers, investigators, sponsors, and participants remain informed. Explainability and interpretability ensure that researchers, clinicians, and participants can understand how AI systems generate their outputs, fostering trust and accountability. Black-box AI models, which operate with limited visibility into decision-making processes, raise ethical concerns about their potential for bias, unexplainable errors, and unanticipated consequences that could cause harm.[43] IRBs should assess whether protocols include methods to support the explainability and interpretability of AI models to clarify the rationale behind decisions, particularly when AI outputs may influence the care of participants and future patients. Explainability and interpretability help clarify the rationale behind AI-supported decisions and support risk assessment, transparency, and informed consent.

1. How will the intended use of AI be communicated to participants during informed consent?

2. Is it appropriate not to inform the participants in some cases? For example, when AI is incorporated in software that is part of a device, and the algorithm is fixed and not modified by the research?

3. To what extent can the AI's decision-making process be explained, understood, and interpreted?

4. How will the potential biases or errors be addressed?

5. What information about the AI system will be made available to participants, researchers, and the public?

[43] While the latest LLM 'reasoners' may offer some form of transparency/explainability of its responses via its Chain-of-Thought (CoT) outputs, there are also concerns regarding how "faithful" and/or "complete" the model's CoT outputs are. See Measuring Faithfulness in Chain-of-Thought Reasoning | arXiv.org

Representativeness and Fairness

Ensuring fairness in research involving AI requires that datasets reflect the intended population for the use of the AI or device. IRBs should review protocols to ensure that underrepresented populations are not unfairly excluded, disproportionately affected, or disenfranchised by AI-driven decisions. Protocols should also include strategies to identify and mitigate biases in the training data, including using model and data cards.

1. Are or were the data used to develop the algorithm sufficient and representative of the population affected by the disease or condition, or of the general population? If not, why not? How is or was "representativeness" evaluated, and is it appropriate for what the algorithm will ultimately assess?

2. Are there historically underrepresented populations (e.g., as a result of data source limitations, systemic exclusion, or selection bias) who will be affected, and will there be a differential impact on specific vulnerable or underrepresented populations by using AI? Has this been identified and discussed in the protocol?

3. Will accessibility (e.g., mobile device and internet access) be required to interact with the AI? Will access or digital literacy be an issue?

    - What safeguards will be in place to address such issues?

    - Will the researchers provide appropriate access to hardware, software, training, and the internet for eligible participants to take part in the research?

    - What ongoing technical, financial, and cultural support is provided?

Informed Consent

To uphold respect for persons in human participant research involving AI, the informed consent process must adapt to address risks and benefits unique to AI systems. Plain language explanations should be provided to participants, detailing how the AI operates within the proposed research and how participants' data will be used within the study and beyond. Additionally, consent processes should be in place to assess the participants' comprehension and understanding of the explanations provided.

1. Are there unique risks and benefits associated with the use of AI that should be outlined in the informed consent document? Are those risks and benefits, in plain language, understandable to the potential participant?

2. Does the informed consent document adequately outline, in plain language, any potential privacy and confidentiality issues related to the sharing of data for the development, translation, or deployment of the AI algorithm?

3. Are there ample opportunities for participants to ask questions about the use of AI and the participants' data prior to enrolling in the study?

4. Is there a clear understanding on the participants' part that data collected during the study cannot be removed from the study database once the data is collected?

5. If the research would otherwise meet requirements for a waiver of the requirements for informed consent, does the AI used in this research project nevertheless warrant obtaining consent/authorization from human participants? What would a reasonable person want to know about the research?

6. Is the intended use compatible with what the human participants agreed to at the time that the data were collected?[33]

7. How well aligned is the use with the original reasons for collecting the data (e.g., EHR records)?[33] Does the informed consent request broad or open-ended data use permissions, and if so, are the permissions proportionate to the study's goals?

Artificial Intelligence Deployed in the Administration of Research

Background

Beyond its use as the intervention in research or the "subject" of research itself, AI is increasingly being used as a tool to facilitate or augment the administration of research. Examples include, but are not limited to: AI-enhanced data analysis; human subject recruitment; use of LLMs to help develop protocols, subject facing materials (e.g., informed consent forms, or recruitment materials), and research instruments (e.g., questionnaires, data collection tools); transcription of interviews and generation of transcripts; LLM-generated responses to participant questions about the research; and other operational roles where AI is not the primary intervention. Although the use of AI in the administration of research could fall out of the IRB's purview on the ethical review of research protocols, these questions are intended to help Principal Investigators (PIs), researchers, and institutions maintain adequate human oversight and establish ethical safeguards as these use cases evolve.

Checklist

AI System Purpose and Role

- □ Does the protocol clearly describe the presence and role of AI in the administration of the research (e.g., recruitment, informed consent development, or data collection)?
    - o If not, what elements of the use of AI are suggested?
- □ How has the AI system been validated for its intended purpose for the administration of the research?

Bias and Fairness

- □ How has the protocol or process addressed potential biases in the AI-driven processes (e.g., recruitment, enrollment, or eligibility evaluation)?
- □ How has the recruitment and enrollment plan ensured equitable human participant selection? Were the burdens and benefits of the research distributed equitably?

Data Privacy and Security

- □ How were safeguards put in place to protect sensitive data collected by AI systems [e.g., Business Associate Agreements (BAAs) or Data Use Agreements (DUAs)]?
- □ How were data storage, sharing, or use protocols clearly explained in the protocol?

Transparency and Accountability

- □ Who is responsible and/or accountable for any concerns related to the use of AI?

Human Oversight

- □ Does the protocol describe a mechanism for human oversight of AI processes, ensuring clear and timely decisions, and that they can be reviewed or overridden if necessary?
- □ Are users adequately informed of the ethical implications of using AI in the administration of the research?

Download the Toolkit

https://mrctcenter.org/wp-content/uploads/2025/06/2025_AI-Toolkit_06-23-2025.pdf

Additional References

1. Alderman JE, Palmer J, Laws E, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *Lancet Digit Health*. 2025;7(1):e64-e88. doi:10.1016/S2589-7500(24)00224-3

2. Blau W, Cerf VG, Enriquez J, et al. Protecting scientific integrity in an age of generative AI. *Proc Natl Acad Sci U S A*. 2024;121(22):e2407886121. doi:10.1073/pnas.2407886121

3. Bouhouita-Guermech S, Gogognon P, Bélisle-Pipon JC. Specific challenges posed by artificial intelligence in research ethics. *Front Artif Intell*. 2023;6:1149082. Published 2023 Jul 6. doi:10.3389/frai.2023.1149082

4. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. Annu Rev Biomed Data Sci. 2021;4:123-144. doi:10.1146/annurev-biodatasci-092820-114757

5. Diaz-Rodriguez N, Del Ser J, et al., Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion. https://www.sciencedirect.com/science/article/pii/S1566253523002129

6. Doerr M, Meeder S. Big Health Data Research and Group Harm: The Scope of IRB Review. *Ethics Hum Res*. 2022;44(4):34-38. doi:10.1002/eahr.500130

7. Goldberg CB, Adams L, Blumenthal D, et al. To do no harm - and the most good - with AI in health care. Nat Med. 2024;30(3):623-627. doi:10.1038/s41591-024-02853-7

8. Greene KK, Theofanos MF, Watson C, Andrews A, and Barron E, "Avoiding Past Mistakes in Unethical Human Participants Research: Moving From Artificial Intelligence Principles to Practice," in *Computer*, vol. 57, no. 2, pp. 53-63, Feb. 2024, doi: 10.1109/MC.2023.3327653

9. Lavin, A., Gilligan-Lee, C.M., Visnjic, A. *et al.* Technology readiness levels for machine learning systems. *Nat Commun* 13, 6039 (2022). https://doi.org/10.1038/s41467-022-33128-9

10. McCradden MD, Anderson JA, Zlotnik Shaul R. Accountability in the Machine Learning Pipeline: The Critical Role of Research Ethics Oversight. *Am J Bioeth*. 2020;20(11):40-42. doi:10.1080/15265161.2020.1820111

11. Youssef A, Nichol AA, Martinez-Martin N, et al. Ethical Considerations in the Design and Conduct of Clinical Trials of Artificial Intelligence. *JAMA Netw Open*. 2024;7(9):e2432482. Published 2024 Sep 3. doi:10.1001/jamanetworkopen.2024.32482

Appendices

Appendix A: Discovery, Translation, Deployment

Development Stage	Research Purpose	Ethical Implications	Potential AI Impact
Discovery (e.g., Phase 1, preclinical studies)	Developing an AI algorithm	• Data collection or availability • Source of data, whether consent obtained, identifiability of data, privacy protections	• Clarity of purpose • Intended use and user • Consider whether the research is not research, not human research, or exempt. • Consider whether an institution is engaged.
Translation (e.g., Phase II)	Piloting, evaluating, and/or validating AI	• Data availability and use • Stability of the AI algorithm • Technical robustness • Intended end users of the AI system	• Clarity of purpose • Intended use and user • Stable or "learning" algorithms • Validation metrics • Consider the need for an IND/IDE
Deployment (e.g., Phase III, Pivotal Trial Testing)	• AI algorithm used in the conduct of the research • AI algorithm used as a companion diagnostic in the research	• Status of the device/SaMD • AI stability • Human oversight and accountability • Transparency	• Device or Software as a Medical Device (SaMD) • Clinical decision-making support • Support precision medicine • Consider the need for an IND/IDE