IRB Review of AI in Medical Research: What Study Teams Need to Know

AI-specific considerations for human subjects protection

Chuan Huang, PhD
Associate Professor of Radiology and Imaging Sciences
IRB Reviewer, 2016–Present

Additional affiliations: Biomedical Informatics; Biomedical Engineering; Winship Cancer Institute; Center for Systems Imaging Core

August 13, 2026

Prepared with assistance from OpenAI PRISM; all content originated with the presenter.

Disclaimer

- This presentation does *not* represent an official Emory IRB policy or decision.
- Institutional approaches to IRB review of AI studies are evolving.
- The content reflects the perspective of an AI expert reviewer and is intended to support discussion.

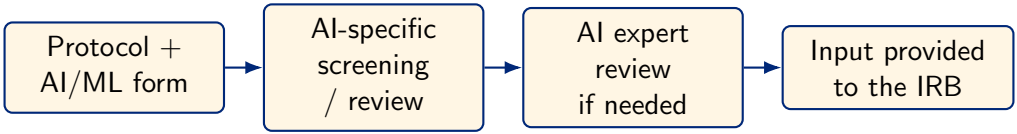
Why This Matters (for PIs / Study Teams)

With advanced AI, the IRB sometimes needs to look beyond how the data are handled. The **model and its outputs** may also affect privacy, disclosure, or participant-level decisions.

Common AI study examples:

- **Example A: Low-risk retrospective MRI baseline**
  A fixed U-Net model segments brain tumors on retrospective MRI scans.
- **Example B: Participant-facing generative AI**
  A chatbot interacts directly with participants and gives unscripted health-related advice about a cancer diagnosis.
- **Example C: PHI, LLM training, and model retention**
  An LLM is fine-tuned on local clinical notes to predict Alzheimer's risk.
- **Example D: Participant-level multimodal risk prediction**
  EHR and imaging data are combined to generate individual-level risk scores for cancer progression that inform clinical decision-making.

How AI Review Works Today



- The AI/ML form separates conventional machine learning from advanced AI that warrants **AI-specific screening**.
- **Human AI expert review** is added when the study presents material AI-specific concerns that require specialized judgment.
- The AI reviewer provides focused input; the IRB makes the determination.

What Traditional IRB Review Already Covers

IRB already reviews:

- Use of PHI and identifiers
- Data sharing and DUAs
- Consent and HIPAA waivers
- Clinical interventions and participant risk

Review principle: If AI does not add a new privacy, disclosure, safety, or participant-level issue, a separate layer of AI review is not needed.
Example A: A fixed U-Net segments brain tumors on retrospective MRI scans and derives imaging biomarkers, with no participant-level use.

When Does an AI Risk Become an IRB Risk?

An AI issue becomes IRB-relevant when it affects human subjects and could change an IRB-required protection or determination.

Common AI Issues

- Errors or hallucination
- Bias or poor generalizability
- Drift or adaptive behavior
- Privacy leakage or memorization
- Security or system failure

→

1. Does it affect participants?

- Privacy or disclosure
- Safety
- Recruitment, eligibility, or monitoring
- Diagnosis, treatment, or care

→

2. Could it change what the IRB requires or determines?

- Protocol safeguards or use restrictions
- Approved data-handling environment
- Human oversight, consent, or disclosure
- Approval or risk-benefit determination

AI-specific IRB review focuses on AI-related risks that can change participant protections or the IRB determination.

Example A: *Model errors are mainly scientific when retrospective outputs do not affect participants.*

Example B: *Hallucinated advice about a cancer diagnosis may create a participant safety issue.*

What AI-Specific Features May Matter to the IRB?

1. Model changes during use

Adaptive models continue learning or updating during the study instead of remaining fixed.

Example B: The chatbot model updates from participant interactions.

2. Information persists

Training may retain or reveal identifiable participant information.

Example C: Fine-tuning an LLM on clinical notes to predict Alzheimer's risk.

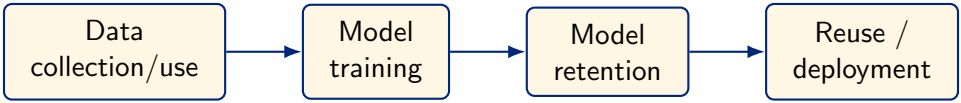
3. Outputs drive actions

Outputs may affect eligibility, monitoring, diagnosis, treatment, or care.

Example D: A cancer-progression risk score used for eligibility or monitoring.

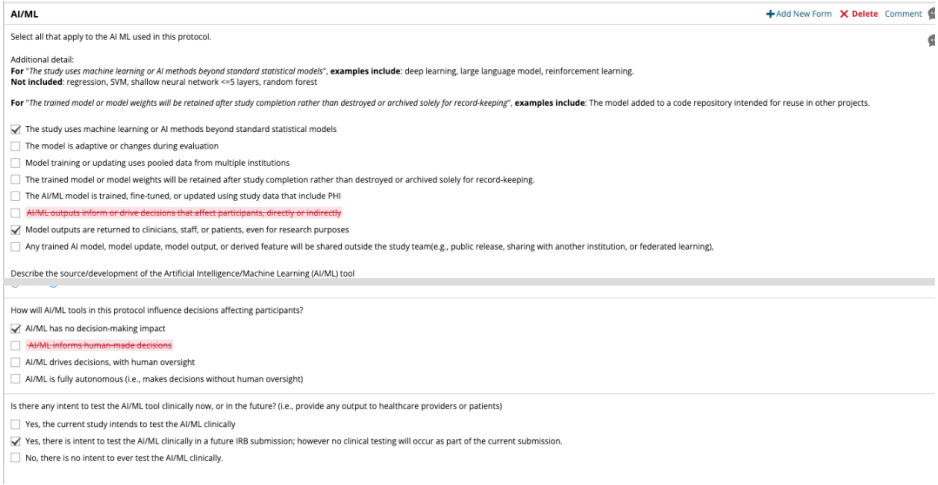
These matter only when they create a new participant risk beyond traditional IRB review.

Where Risk May Arise Across the AI Lifecycle



- **Data collection/use:** Does identifiable or re-identifiable information enter the AI workflow? **[Example A]**: retrospective MRI scans.
- **Model training:** Could the trained model reveal information about an individual participant? **[Example C]**: fine-tuning on clinical notes.
- **Model retention:** Does retention create a realistic disclosure or unauthorized-use concern? **[Example C]**: retained fine-tuned weights.
- **Reuse/deployment:** Could outputs affect participants, participant-level decisions, or clinical care? **[Example B/D]**: chatbot advice or cancer-progression risk scores.

How These Are Captured in the IRB AI/ML Form



AI/ML

+ Add New Form ✕ Delete Comment

Select all that apply to the AI ML used in this protocol.

Additional detail:
For "*The study uses machine learning or AI methods beyond standard statistical models*", **examples include**: deep learning, large language model, reinforcement learning.
Not included: regression, SVM, shallow neural network <=5 layers, random forest

For "*The trained model or model weights will be retained after study completion rather than destroyed or archived solely for record-keeping*", **examples include**: The model added to a code repository intended for reuse in other projects.

- [x] The study uses machine learning or AI methods beyond standard statistical models
- [ ] The model is adaptive or changes during evaluation
- [ ] Model training or updating uses pooled data from multiple institutions
- [ ] The trained model or model weights will be retained after study completion rather than destroyed or archived solely for record-keeping.
- [ ] The AI/ML model is trained, fine-tuned, or updated using study data that include PHI
- [ ] ~~AI/ML outputs inform or drive decisions that affect participants, directly or indirectly~~
- [x] Model outputs are returned to clinicians, staff, or patients, even for research purposes
- [ ] Any trained AI model, model update, model output, or derived feature will be shared outside the study team(e.g., public release, sharing with another institution, or federated learning).

Describe the source/development of the Artificial Intelligence/Machine Learning (AI/ML) tool

How will AI/ML tools in this protocol influence decisions affecting participants?

- [x] AI/ML has no decision-making impact
- [ ] ~~AI/ML informs human-made decisions~~
- [ ] AI/ML drives decisions, with human oversight
- [ ] AI/ML is fully autonomous (i.e., makes decisions without human oversight)

Is there any intent to test the AI/ML tool clinically now, or in the future? (i.e., provide any output to healthcare providers or patients)

- [ ] Yes, the current study intends to test the AI/ML clinically
- [x] Yes, there is intent to test the AI/ML clinically in a future IRB submission; however no clinical testing will occur as part of the current submission.
- [ ] No, there is no intent to ever test the AI/ML clinically.

How to Interpret the AI/ML Screening Items

Across all items, first establish what data enter the AI system, what may be retained or shared, and who will receive or use the outputs.

- **Advanced AI used:** Does this study belong in AI-specific review?
- **Adaptive during evaluation:** Does the model continue learning or updating rather than remain fixed? A changing model makes performance and risk harder to evaluate consistently.
- **Multi-institutional training:** What actually moves between institutions?
- **Model/weights retained:** Could keeping the model create a realistic privacy or reuse issue?
- **Training/updating with PHI:** Could identifiable information be retained or later disclosed through the trained artifact?
- **Outputs returned:** Who will see participant-level outputs, and what will they do with them?
- **Decisions affecting participants:** Could AI influence eligibility, safety, diagnosis, treatment, or care?
- **Sharing outside the team:** Could anything shared reveal participant information?

What Study Teams Should Include in the IRB Submission

- **Method:** Name the actual advanced AI method. Do not rely on labels such as "algorithm" or "machine learning."
- **Data status:** State what information enters the AI system and whether it remains identifiable or linkable.
- **AI-assisted de-identification:** If AI removes identifiers, state how the output will be checked before further use or sharing.
- **External processing:** If identifiable information leaves Emory, name the approved environment or service.
- **Model change/retention/sharing:** Describe these when they could affect privacy, disclosure, participant-level decisions, or clinical care.
- **Outputs:** State who receives them and whether they affect participants or clinical care.
- **Future use:** Separate current protocol activities from future work requiring a new submission or amendment.

Example A — Clear Submission Wording

"Retrospective MRI scans will be processed on an Emory server using a fixed pretrained U-Net to segment brain tumors and derive imaging biomarkers. The model will not be updated; outputs will be used only for retrospective research analysis and will not affect participants or clinical care."

Multi-Site Model Development: What Study Teams Should Describe

For multi-site AI studies, clearly describe what will move between institutions.

This may include:

- Identifiable or coded data
- Model updates or weights
- Derived features
- Participant-level outputs
- A reusable trained model

What to include in the IRB submission

- **What is shared:** Specify what data or model-related information will cross institutional boundaries.
- **Participant identifiability:** State whether any shared information could identify or reveal information about an individual participant.
- **Retention and reuse:** Explain whether shared models, weights, or outputs will be retained or reused after the current study.
- **Agreements:** Identify any applicable data use or other institutional agreements.

Illustrative Contrast — Example D

- **Lower concern:** Sites train locally and share only securely aggregated updates; participant records and individual risk scores stay local.
- **Still requires care:** Federated learning moves only model updates or weights, but these may still reveal or memorize participant information.

Multi-site training is not automatically higher risk. What matters is what moves between institutions and whether it changes participant privacy or confidentiality risk.

Where the AI Runs: Local vs. External Services

Classify the environment by where processing occurs and who controls it—not by how users access the AI.

Either may be accessed through a webpage, API, CLI, desktop app, agent, plugin, or integrated tool.

Locally Hosted / Self-Hosted AI

- Runs on hardware operated by the institution or study team
- Model execution occurs within the locally controlled environment
- Study data are not sent to a third-party AI service

Example C: The Alzheimer's-risk task runs on a local GPU server using the open-weight DeepSeek-R1 model.

External / Cloud AI Service

- Model execution occurs on infrastructure operated by a third party or cloud provider
- Data or prompts are transmitted to that external environment
- The service may still be institutionally approved for PHI

Example C: The same task uses Emory-approved Microsoft Copilot in Work mode.

"Local" describes where the model is hosted and who operates the infrastructure. It does not mean "approved." An external cloud service may also be institutionally approved.

External AI Services: Start With the Data

For an external AI service, the first question is what data are being sent to it.

PHI / identifiable data

Confirm that the specific service is institutionally approved for PHI and that the study follows the conditions of that approval. IRB review generally should not repeat technical security review already completed by Information Security.

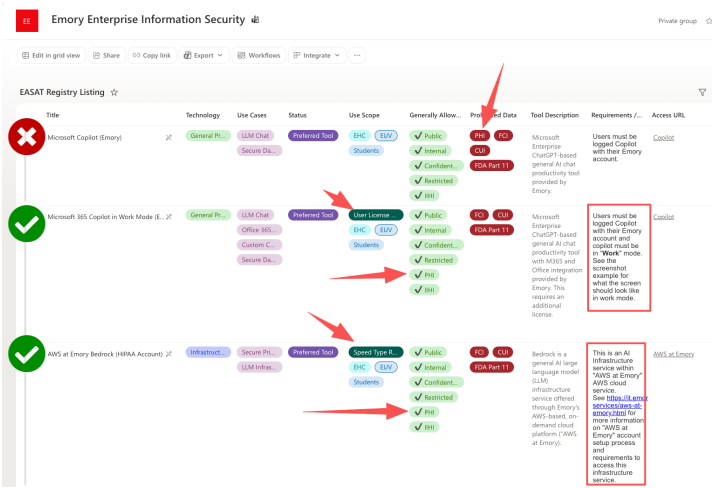
Credibly de-identified data

If the data are credibly de-identified and the outputs do not affect participants or care, an external LLM or API is usually low AI risk. Details about hosting, API architecture, or vendor identity generally are unnecessary unless they change the risk assessment.

General advice

- **Safer default:** Use an institutionally approved environment, even for data believed to be de-identified.
- **Other services:** May be appropriate only when the data are fully and credibly de-identified.
- **Expect questions:** Reviewers may ask how de-identification was established, especially for MRI images, ZIP codes or other granular location data, and AI-assisted de-identification.

Institutionally Approved AI Environments for PHI



Source: Emory EASAT Registry Listing

When Is AI Expert Review Usually Needed?

AI expert review is generally appropriate when:

- Participants directly interact with generative AI that can respond without a fixed script
- AI influences diagnosis, treatment, eligibility, safety, or other participant-level decisions
- AI can take autonomous actions or access clinical systems or sensitive data in a way that could affect participants, expose PHI, or bypass meaningful human oversight
- A model trained on identifiable information will be distributed and may disclose participant information

Usually No Expert Review

Retrospective analysis of de-identified data with a fixed model.

Expert Review More Likely

A participant-facing chatbot giving unscripted health guidance.

Take-Home for Study Teams

Make six things explicit in the AI/ML submission:

AI system

- **Model and task:** Name the model or method, provider/version, and research function.
- **Hosting:** State whether it is local/self-hosted or external/cloud-hosted and who operates the environment.
- **Changes:** Distinguish fixed use, prompt/template or retrieval/tool changes, fine-tuning/weight updates, and adaptive learning.

Use and protections

- **Data flow:** State what participant information enters or leaves, whether PHI is processed by a fixed model or used for training/updating, and relevant retention/sharing.
- **Outputs:** Identify who receives them and whether they affect recruitment, eligibility, monitoring, diagnosis, treatment, or care.
- **Protections and lifecycle:** Cover approved environments, human oversight, multi-site transfers, model retention/reuse/distribution, and verification of AI-assisted de-identification.

Focus on details that could change participant risk or what the IRB needs to require.

Case Study — Example C: A Complete Applicant Response

A GPT-5.5-based model is fine-tuned on identifiable clinical notes to predict Alzheimer's risk. A complete response could state:

AI system

- **Model and task:** A GPT-5.5-based model will be fine-tuned once to generate research-only Alzheimer's-risk scores from clinical notes.
- **Hosting:** The model runs on an institutionally managed local server. Access is limited to authorized study staff using individual password-protected accounts and multifactor authentication.
- **Changes:** The model version, fine-tuned weights, and prompt template are locked before analysis. The model does not learn or adapt from new study data.

Use and protections

- **Data flow:** Identifiable notes are used for fine-tuning—not only inference. They remain within the institution-controlled environment and are removed from the training workspace after fine-tuning; logs follow approved retention terms.
- **Outputs:** Risk scores are available only to the research team and are not returned, placed in the EHR, or used for eligibility, diagnosis, treatment, or care.
- **Protections and lifecycle:** Role-based access, encryption, and audit logging protect the local server. The model is not distributed; any reuse requires privacy review for memorization or disclosure risk.

This description distinguishes processing PHI from training on PHI and makes clear that the retained model is fixed, not adaptive.

Discussion: Withdrawal After Model Training

Problem

Example C: A participant requests that their clinical notes be removed *after* the LLM has been fine-tuned to predict Alzheimer's risk.

Discussion: What Applicants Should Address

Problem

Example C: A participant requests that their clinical notes be removed *after* the LLM has been fine-tuned to predict Alzheimer's risk.

What applicants should address

- **Consent and withdrawal language:** Explain what can be removed after training and clearly state any limits.
- **Data versus model:** Distinguish deleting source notes or derived data from modifying or retraining the trained model. Do not promise retraining unless it is feasible.
- **Model retention and privacy:** State whether the model or weights will be retained, shared, or reused and address any realistic disclosure or memorization risk.
- **Response process:** Describe how withdrawal requests will be handled and when the study team will seek IRB or other institutional guidance.