Guidance for the Use of Artificial Intelligence in Clinical Education Research (CERG-AI)

University of Cambridge Clinical Education Researchers' Group

1. Introduction

Artificial Intelligence (AI) is emerging as a transformative innovation in the field of clinical education research, offering the potential to improve learning, assessment, curriculum design, and research. As AI becomes increasingly embedded across healthcare, it is essential that its use in clinical education research is accompanied by human oversight and co-production, safeguards against data and record-keeping risks, and attention to existing inequalities in digital access.

Most AI systems operate as advanced forms of machine-learned pattern recognition and share many characteristics with other statistical or digital tools already governed by existing University and Clinical School policies. What often distinguishes current large-scale models, particularly large language models (LLMs), is not necessarily their fundamental analytical approach but the computational scale and reliance on external, often multinational, providers. This raises considerations around data transfer, privacy, intellectual property and recourse in the event of a breach, as well as the environmental and resource costs of maintaining the AI models. It is sometimes worth considering whether a locally hosted model without connectivity outside the University (e.g. a locally hosted "small language model") would be a viable alternative. Externally hosted LLMs are digital resources that require the storage and transfer of research data, potentially personal, and the policies of the Clinical School and University remain relevant in the same way as for any other digital resource. The transfer of data into and out of the University requires legal and contractual oversight.[1] Further information about research ethics requirements is available in the Cambridge Psychology Research Ethics Committee Handbook.[2]

In clinical education praxis and research it is expected that the use of any innovation, including AI, is founded on acknowledged educational theory and empirical research in line with an educational ethos and evidence-based praxis where possible. Whilst there has been a surge of interest in the application of AI to clinical education, there is a notable shortage of existing empirical research in this area. As clinical education research contributes directly to continuous improvement of evidence-based educational praxis, the responsible and rigorous use of AI in education research should serve as a model for how AI can be translated into clinical education practice. Researchers remain responsible for the accuracy, validity, and ethical use of outputs generated or influenced by AI when they are used in the context of research, even if the AI software is provided by a private partner. Relevant aspects of University Information Compliance should be consulted in all cases. For example, the University of Cambridge produces guidance on the administrative use of AI.[3] University Information Compliance guidance also covers AI for administrative tasks or drafting manuscripts, which is not covered by the present document.

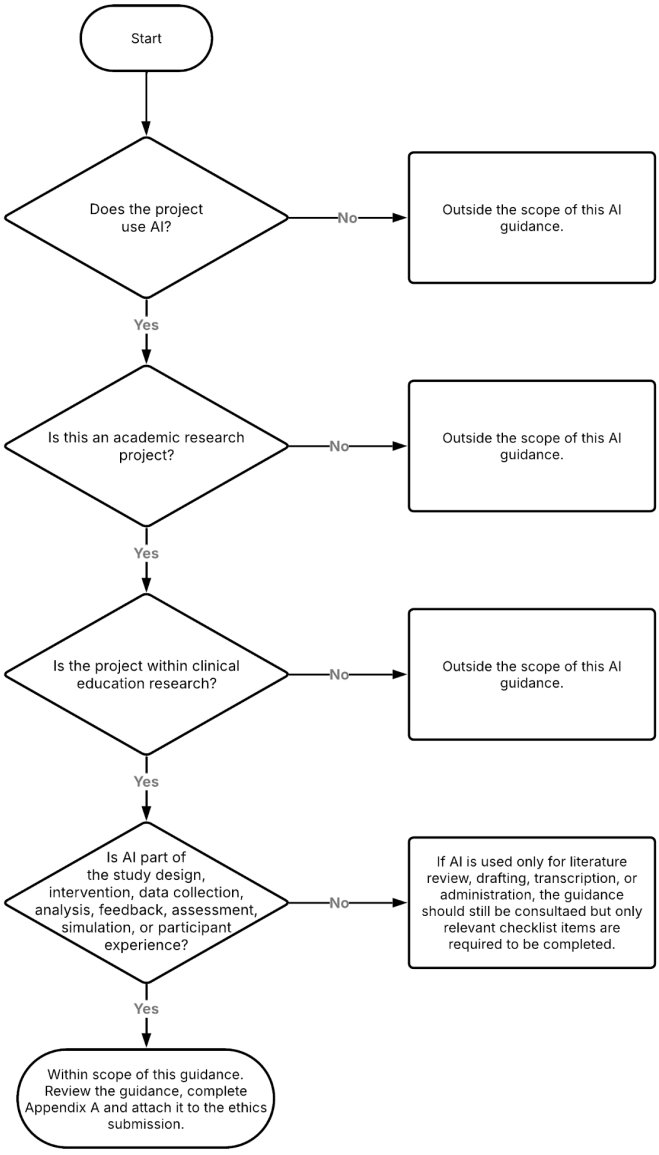
Transparency in the use of AI is essential to maintaining academic and professional integrity. Any use of AI in research, education, or related clinical-education activities should be clearly acknowledged, including wherever this prevents ambiguity regarding authorship or originality. Because AI systems generate outputs derived from data created by others, researchers must ensure a clear distinction between their own work and content that has been supplemented by AI. Ethical oversight of AI use should be undertaken as frequently as necessary to provide assurance to learners, educators, curriculum validators, clinicians, and the public that AI is being used responsibly and in accordance with established academic standards.

2. Scope

This document provides guidance for clinical education researchers who intend to incorporate AI into their academic research. This document should be consulted prior to submitting a project proposal for ethical approval, in addition to existing guidance available for all research projects. This guidance applies to the use of externally hosted commercial AI platforms as well as locally hosted and open-source models.

[1] https://www.research-services.admin.cam.ac.uk/manage-risk/contracts
[2] https://www.bio.cam.ac.uk/files/prec_handbook.pdf
[3] https://www.information-compliance.admin.cam.ac.uk/data-protection/guidance/ai-guidance

The checklist in Appendix A should be completed and attached to the ethics review board submission for the project. This guidance does not pertain to the use of AI in research outside the field of clinical education and does not pertain to clinical education research where AI is not part of the study design. The guidance is particularly focussed on the use of generative AI (e.g. large language models) rather than rule-based, deterministic, or non-generative AI techniques, which are less prevalent in clinical education research. The scope of this guidance is illustrated below. Where AI is being used solely for administrative tasks, including transcription, only relevant aspects of the guidance need to be reported in the checklist. General use of AI for administrative tasks or manuscript preparation outside the study design is specifically covered by University Information Compliance guidance. The University currently recommends using Copilot, Gemini and NotebookLM where possible under University licenses, though up-to-date guidance should be consulted at the time of study design.[3] Other tools are allowable and features from this guidance could be used to consider which may be most appropriate in a clinical education research setting. Where any new Intellectual Property (IP) is expected to be generated from the study, relevant University Intellectual Property guidance should be consulted at the time of study design.[4,5]



[4] https://www.hr.admin.cam.ac.uk/hr-staff/information-staff/staff-guide/terms-employment/intellectual-property
[5] https://www.research-services.admin.cam.ac.uk/manage-risk/contracts

3. Information to Learners and Participants in Research

- Clearly disclose to all participants that AI is being used and explain in what capacity.
- Identify the specific models or platforms employed (e.g., GPT-X.X, Med-PaLM, local model etc.), their intended role in the study, and any relevant adaptations.
- Participants and researchers should be instructed not to input personal data into the AI interface. If the AI model is not participant-facing, the research team must ensure that no personal data of participants is inputted. This may be exempted in certain cases (e.g. where inputted data neither leaves local servers nor is used as training data for the model, or where there are specific risk mitigations in place).
- Where participants themselves are engaging directly with an externally hosted AI interface, a statement should be included in the Consent Form and Participant Information Sheet about the risk of inputting personal data, for example: *"Inputting data into *the AI model* could be considered equivalent to putting it into the public domain. Therefore, please do not input any personal data into *the AI model*."* This may be exempted in certain cases (e.g. where inputted data neither leaves local servers nor is used as training data for the model, or where there are specific risk mitigations in place).
- Where relevant, participants should be informed that any data, including anonymised data, entered into AI platforms may be stored, used for further training, and, where an externally hosted model is used, may be subject to the provider's terms of use and data-processing arrangements. Once incorporated, such data may not be retrievable, removable, or capable of being excluded from downstream use.
- If AI is used to simulate clinician or patient responses, participants should be instructed that these outputs may not reflect real-world behaviour or clinically validated reasoning and are at risk of 'hallucinations'.
- Deploying LLMs without informed consent, or using them deceptively in place of a human interlocutor, is not permitted except under tightly justified and approved circumstances.
- Researchers should be explicit about model limitations and training data sources. Where the training data sources of models are unavailable, this should be explicitly stated, e.g. with the following statement: "*AI models are limited by the data that they are trained on. We cannot reliably predict how they will behave if exposed to data that they have not seen before during model training. As we cannot be sure what data was used to train *the AI model*, we cannot be sure it will not 'hallucinate' or produce incorrect information whilst it is being used. We would therefore advise caution when interpreting its output."*

4. Risks and Mitigations

- Faculty and researchers must not upload identifiable student, patient, or faculty data to externally hosted AI platforms without prior approval from the Clinical School (Research Governance Team).
- Where AI models are being used to provide feedback to learners, feedback should be qualitative in nature, typically providing constructive positive and negative feedback. Feedback should usually not be in the form of quantitative marking or 'Pass/Fail' marking unless marking is checked by a human prior to release to participants.
- AI models should not typically be used as a marking tool for assessment of curriculum competences except in cases where the primary research objective of the study is to assess the ability of AI to assess competences. In such cases, marking should be generally checked by a human prior to release to participants.
- [Optional] For higher risk projects, AI models could be 'stress-tested' prior to use with real participants, in the form of pilot testing (this could be conducted within the study team). The summary results of pilot testing, if conducted, should be included in the submission to the Research Ethics Committee.
- If AI-generated feedback is being provided to learners, researchers should assess the potential psychological impacts of incorrect or overly confident outputs and signpost support mechanisms in place for the individual project (e.g. speaking to a supervisor).
- The protocol should describe a process for reviewing the quality of AI outputs, proportionate to their role and potential impact in the study. For example, outputs used directly in the study, such as those presented to learners or used to generate results, may require closer monitoring than those used only for preparatory or internal purposes.
- Where relevant, researchers should disclose in the study protocol any potential algorithmic bias in the AI model. This could be based on previous research, pilot studies or hypothetical biases. Biases may

arise through subject selection and amplification of existing biases, which should be addressed or disclosed in the study protocol. Where relevant, mitigation strategies for identified biases should be documented.

- Early public and patient involvement, including designing projects in collaboration with potential participants and/or patients is advised, particularly if patients will be participating in the study. Where the study design involves patients engaging with AI software, it is expected that some public and patient involvement will have been sought prior to submitting the study proposal for review to the Research Ethics Committee. This should be summarised in the ethics submission and in eventual research outputs.
- Standard processes for early termination of studies apply, including notification requirements. Where relevant, AI-specific circumstances that may warrant pausing, amending, or terminating the study include, but are not limited to:
    - A pattern of AI outputs that are factually incorrect, unsafe, clinically inappropriate, or potentially harmful to participants, identified through routine quality monitoring or participant report.
    - Evidence of systematic bias in AI outputs affecting a participant subgroup, where this was not anticipated and cannot be mitigated in a timely manner.
    - A material change in the AI model or platform, including a change in provider terms, loss of institutional access, or substantive change in model performance, that alters the study conditions in a way that may affect data integrity, study validity, participant safety or consent.
    - Unanticipated psychological distress in participants attributable to AI-generated content.

5. Regulatory Considerations

- Where AI is being used to process patient data, support clinical decision-making, or perform a function that may fall within the definition of Software as a Medical Device, the protocol should include a statement on its regulatory status. This should include, where applicable, whether the software is UKCA/CE marked; whether any required MHRA registration has been completed; the manufacturer's stated intended purpose; whether the proposed study use falls within that intended purpose; and its medical-device risk classification under the applicable UK or EU Medical Device Regulations. Stating this information helps reviewers assess the extent of regulatory oversight the AI tool has undergone, particularly where it is being used in relation to clinical or patient data. Software that is appropriately UKCA/CE marked and used within its stated intended purpose would generally be considered relatively lower risk.
- In the context of clinical education research, it is expected that most AI use cases will not necessarily fall under the definition of a Software as a Medical Device. Medical device registration and UKCA/CE markings are **not** prerequisites for ethics approval in most clinical education research. In many cases, the following statement can be used as justification for not requiring medical device registration or UKCA/CE marking: *"No patient data is being processed by the AI software, and the proposed use does not fall within the scope of Software as a Medical Device"*. However, if the AI software is processing patient data or simulating clinically relevant functions, researchers should provide a rationale for why such registration and regulatory markers are not necessary for their specific use case. If the AI software is not yet appropriately UKCA/CE marked but the proposed study will be contributing to the evidence required to achieve registration, this may be an acceptable justification. However, ethics approval alone should not be treated as a substitute for applicable medical-device regulatory requirements, including any MHRA clinical investigation notification requirements. Researchers should consult current MHRA (Medicines and Healthcare products Regulatory Agency) guidance where regulatory status is uncertain, as UK medical-device requirements may change.
- Researchers should confirm that the proposed processing of personal data using the AI system complies with UK GDPR as amended by the Data (Use and Access) Act 2025.

6. Data Storage and Intellectual Property

- As with other externally hosted digital research tools used within the University, appropriate licensing and governance arrangements must be in place before the system is used to process research data.
- Any identifiable raw results data produced by the study should be pseudonymised prior to analysis.

- Researchers should maintain documentation of prompt design, fine-tuning, custom instructions, parameters, and modified system settings, particularly when using externally hosted generative AI models.
- Identifiable results data and/or pseudonymisation keys should be retained in keeping with relevant departmental, University or funder policies. Only anonymous data should be retained thereafter.
- Where descriptive statistics of participants have been collected in the study, results for each subgroup should be reported, except where the sample size of a subgroup presents a disclosure risk.
- Proprietary materials (e.g., unpublished research data, teaching cases, exams) must not be uploaded to AI platforms without appropriate approval, even if this is anonymised.
- Where any new Intellectual Property (IP) is expected to be generated from the study (e.g. in a partnership with a start-up company seeking to test a product), this should be discussed with the Principal Investigator and Head of Department in the first instance. The University's IP Guidance will apply.

This guidance will be updated based on feedback and reflections during the course of its usage. This guidance was developed by members of the Clinical Education Researchers' Group at the University of Cambridge, with contributions from:

- Dr Anmol Arora, *Artificial Intelligence, Technology & Simulation Lead, Public Health Education Group*
- Dr Charlotte Tulinius, *Medical Education Research Lead, School of Clinical Medicine*
- Professor Paul Wilkinson, *Clinical Dean, School of Clinical Medicine*
- Professor Niels Peek, *Professor of Data Science and Healthcare Improvement*
- Dr Mark Lillicrap, *Clinical Professor of Medical Education Practice*
- Dr Vanessa Nash, *Lead for Clinical Reasoning and Communication Skills, School of Clinical Medicine*
- Dr Simon Duke, *Medical Education Technology Lead, School of Clinical Medicine*
- Professor Raj Jena, *Professor of Artificial Intelligence in Radiation Oncology*
- Stephanie Smith, *Senior Tutor in Clinical Reasoning and Communication Skills (CRCS), Cambridge University Hospitals Foundation Trust, School of Clinical Medicine*
- Professor John Suckling, *Chair, Cambridge Psychology Research Ethics Committee*
- NIHR Cambridge BRC Public and Patient Representative Panel

APPENDIX A: CERG-AI PROTOCOL CHECKLIST

This checklist should be completed and uploaded as an Appendix to any ethics review submission.

Project Title:	
Principal Investigator (and email):	
Contact Author (and email):	

Item	Checklist item	Page # where reported	Additional Notes (optional)
	STUDY DESIGN		
1	Describe the rationale for including AI in the clinical education study		
2	Introduce previous research using AI in this context, including limitations of prior studies		
3	Identify the specific AI models or platforms employed (with any relevant adaptations) and specify their intended role in the study		
4	List all data that will be uploaded to the AI model.		
5	[Optional] Where relevant, describe the results of any pilot study/'stress testing' that has been performed to help identify specific risks associated with this study, including 'hallucinations'		
6	List all descriptive statistics of participants that will be collected ("Year of Study", "Age", "Sex", etc.) and identify for which variables subgroup reports will be generated when reporting results		
7	Document prompt design, fine-tuning and engineering techniques used to adapt the model (if relevant)		
8	If AI-generated feedback is being provided to learners, describe the potential psychological impacts of incorrect or overly confident feedback, and the support mechanisms that will be signposted		
9	Describe the process for reviewing the quality of AI outputs, proportionate to their role and potential impact in the study		
10	Specify how data will be stored once collected as part of the study		
11	Specify the data retention period		
12	Where relevant, provide a statement on how IP generated from the study will be treated		
13	Provide detail regarding the regulatory status of the AI software, including details on medical device risk classification where appropriate		
14	Provide a statement of compliance to UK GDPR as amended by the Data (Use and Access) Act 2025		
15	Disclose any identified risks of algorithmic bias and mitigation strategies		
16	Describe any co-design with participants/public in developing this study		
	INFORMATION TO PARTICIPANTS		
17	Clearly disclose to all participants that AI is being used and explain in what capacity		
18	Identify the specific models or platforms employed (e.g., GPT-X.X, Med-PaLM, local model etc.)		
19	Where relevant, briefly describe to what extent participants/public co-design has been included in the study		
20	If participants will be directly using an externally hosted AI model, participants should be informed not to input any personally identifiable information into the AI interface, e.g. with the following statement: *"Inputting data into *the AI model* could be considered equivalent to putting it into the public domain. Therefore, please do not input any personal data into *the AI model*"* This may be exempted in certain cases (e.g. where inputted data does not leave local servers or where there are specific risk mitigations in place).		
21	Describe model limitations and training data sources. Where the training data sources of a model are unavailable, this should be explicitly stated, e.g. with the following statement: *"AI models are limited by the data that they are trained on. We cannot reliably predict how they will behave if exposed to data that they have not seen before during model training. As we cannot be sure what data was used to train *the AI model*, we cannot be sure it will not 'hallucinate' or produce incorrect information whilst it is being used. We would therefore advise caution when interpreting its output."*		
22	Where relevant, participants should be informed that any data, including anonymised data, entered into AI platforms may be stored, used for further training, and, where an externally hosted model is used, may be subject to the provider's terms of use and data-processing arrangements. Once incorporated, such data may not be retrievable, removable, or capable of being excluded from downstream use.		
23	Clearly disclose that outputs of AI may not reflect real-world behaviour or clinically validated reasoning if AI is used to simulate clinician or patient responses		