Title

Guidelines for Conducting Ethical Artificial Intelligence Research in Neurology: A Systematic Approach for Clinicians and Researchers.

Permalink
https://escholarship.org/uc/item/2h49k7jr

Journal
Neurology, 97(13)

ISSN
0028-3878

Authors
Chiang, Sharon
Picard, Rosalind W
Chiong, Winston
et al.

Publication Date
2021-09-01

DOI
10.1212/wnl.0000000000012570

Peer reviewed

Guidelines for Conducting Ethical Artificial Intelligence Research in Neurology: A Systematic Approach for Clinicians and Researchers

Sharon Chiang, Rosalind W. Picard, Winston Chiong, Robert Moss, Gregory A. Worrell, Vikram R. Rao, Daniel M. Goldenholz

ABSTRACT

Pre-emptive recognition of the ethical implications of study design and algorithm choices in artificial intelligence (AI) research is an important but challenging process. AI applications have begun to transition from a promising future to clinical reality in neurology. As the clinical management of neurology is often concerned with discrete, often unpredictable, and highly consequential events linked to multimodal data streams over long timescales, forthcoming advances in AI have great potential to transform care for patients. However, critical ethical questions have been raised with implementation of the first AI applications in clinical practice. Clearly, AI will have far-reaching potential to promote, but also to endanger, ethical clinical practice. This article employs an anticipatory ethics approach to scrutinize how researchers in neurology can methodically identify ethical ramifications of design choices early in the research and development process, with a goal of pre-empting unintended consequences that may violate

principles of ethical clinical care. First, we discuss the use of a systematic framework for researchers to identify ethical ramifications of various study design and algorithm choices. Second, using epilepsy as a paradigmatic example, anticipatory clinical scenarios that illustrate unintended ethical consequences are discussed, and failure points in each scenario evaluated. Third, we provide practical recommendations for understanding and addressing ethical ramifications early in methods development stages. Awareness of the ethical implications of study design and algorithm choices that may unintentionally enter AI is crucial to ensuring that incorporation of AI into neurology care leads to patient benefit rather than harm.

KEYWORDS: Machine learning, Biomedical ethics; Anticipatory ethics; Anticipatory governance; Scenario analysis; Technology foresight

1. INTRODUCTION

Artificial intelligence (AI) applications have begun to transition from a promising future to clinical reality in neurology. AI can transform neurological clinical practice, with impacts on quality, cost, and access to care.[1] Epilepsy serves as a paradigmatic case of AI's potential in neurology, for which the breadth and depth of emerging AI applications spans a rapidly expanding number of diagnostic, therapeutic, and prognostic applications[2-10].

Critical ethical concerns have begun to arise with AI incorporation into clinical practice. These include maximization of patient benefit while avoiding harm, risks to patient privacy, perpetuation of bias, and tradeoffs between competing ethical goals. A fundamental question that the epilepsy and broader neurology community must answer in coming years is the degree of responsibility each party (researchers, industry, clinicians, regulatory agencies) carries in the AI pipeline, in order to facilitate a common goal of ensuring that AI promotes rather than endangers ethical clinical practice.[11]

With the rapid proliferation of AI in healthcare, there have been a number of broad initiatives[12-14] that provide general guidance on ethical values that AI should promote in healthcare. There is currently little consensus in the neurologic community on where responsibility lies in the AI pipeline for ensuring that benefit outweighs harm. The Food and Drug Administration (FDA)'s action plan to increase oversight over AI-based medical software[15] is anticipated to help guide safe usage of AI in later market development stages. However, consideration of the ethical implications of AI research is important starting from early development/validation stages. First, from the viewpoint of promoting safe AI, many fundamental study design and algorithm choices are made during early development/validation stages that have downstream ethical implications.

Second, from the viewpoint of direct benefit to early-stage researchers, early adoption of appropriate practices will decrease workload later when the system goes to market. It may be in researchers' interests to consider these factors early rather than engage in post-hoc consideration that may require data retraining or recollection.

While literature exists on good statistical practices to improve rigor and reproducibility from a technical standpoint[16], as well as a number of documents on guiding values for AI,[12-14] there is limited practical guidance to researchers/clinicians on how to systematically identify ethical implications of design choices in emerging AI research. Given the domain-specificity of AI data streams, patient vulnerabilities, and outpatient/inpatient differences in neurological subspecialties, it may be helpful to consider distinct neurological subfields individually. The purpose of the present work is to discuss a practical framework that researchers, peer-reviewers, and clinicians may find helpful when evaluating the potential ethical ramifications of emerging AI research in the field of neurology, focusing on epilepsy as a paradigmatic case.

2. FIVE KEY ETHICAL PRINCIPLES FOR AI AND SYSTEMATIC FRAMEWORK

The four core principles of bioethics—*respect for patient autonomy, beneficence, nonmaleficence*, and *justice*[17]—are pertinent also in AI. In addition, there is consensus that a fifth key essential principle arises when evaluating AI: *explicability*, or transparency of process[14,18] (eTable 1, available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8). Although the community has established that an "ethical AI" should enhance these five principles,[1] there are limited guidelines on *how* to implement these principles in practice when conducting or evaluating AI research. Consideration of impact on these five ethical principles by developers is generally ad hoc.

It may be useful to contextualize ethical concerns in neurological AI research by breaking down the AI development pipeline into five stages (eFigure 1, available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8)[19]: *conception*; *development*, during which data collection, algorithm development, training, and testing take place; *calibration,* during which performance is evaluated; *implementation* in clinical practice; and *monitoring*, or maintenance in the clinical environment. Below, we focus on the initial three stages during which many fundamental AI choices are made: conception, development, and calibration. We decompose each stage into the various design/algorithm choices made in that stage and discuss the implications of each choice for the five key ethical principles of AI.

3. RECOMMENDATIONS FOR ETHICAL CONSIDERATIONS BY STAGE

We followed published guidelines for development of health research reporting guidelines[20] for recommendations development. Anticipatory case analysis was first conducted with a focus on examples from the field of epilepsy. eAppendix 1 (available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8) shows case scenarios in epilepsy, some hypothetical and some based on real cases, that illustrate unintended consequences of AI applications that may endanger rather than promote ethical values. Each of these cases motivates the recommendations developed in this document, illustrates potential failure points, and raises discussion of checkpoints that the neurological community can take in AI development.

Anticipatory case analysis was then combined with systematic literature review and modified Delphi methodology (Figure 1, eAppendix 2 [available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8]) to develop a set of 15 operational recommendations for conducting AI research in neurology (Tables 1-3).

A. CONSIDERATIONS IN STAGE 1 (ALGORITHM CONCEPTUALIZATION)

Q1. *To what extent were key stakeholders directly or indirectly involved in conceptualization/design phase of the AI application?* Stakeholders who may benefit or be affected by the AI application should be defined at conceptualization. Most commonly, stakeholders of AI research in neurology include the users of AI applications, and may include patients, healthcare providers, patient families, and/or care providers. It is helpful to include key stakeholders early in AI development to understand how likely a specific data type is to be accepted for collection or reliably acquired by the patient community. Patient concerns about privacy and data security are a major public concern that may limit a system’s implementation.[21,22] For example, specific data streams (such as video cameras, motion detectors, or chronic outpatient EEG) may impact patient privacy or fundamental rights;[23] others, such as mobile phones or smart glasses, may be turned off or not worn constantly. Understanding use cases early will help researchers understand anticipated limitations of potential data streams when deciding which variables to incorporate into algorithms. It is important to acknowledge when citing use cases that stakeholders may not be able to imagine all possible use cases of a technology that does not exist yet. Focus groups are often time-intensive to conduct and it may not be feasible to incorporate key stakeholders directly in the conceptualization/design phase. If prior research has been conducted describing the needs and concerns of key stakeholders with regards to the AI application, references to this literature should be provided. This step helps promote transparency about intentions, provides a means for fundamental rights and privacy assessment early in development, and helps ensure that patient perspectives are included early in the design phase.

Q2. *Is the explainability of methods justified against potential harm in the case of erroneous predictions or unreliable human supervision*? Methods with technical explainability can help to improve AI safety through understanding of key assumptions, unintended biases, and cases where performance may be low.[14,15] However, explainability of AI decisions is not always possible (or adequate), particularly in the case of “black box” approaches, such as deep neural networks. Although some advocate for avoiding “black box” approaches[21,24], technical solutions such as post-hoc and hybrid approaches can increase explainability for different machine learning models (eTable 2, available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8).

Tradeoffs may also sometimes be necessary between increased accuracy (at the cost of explainability) and enhanced explainability (at the cost of accuracy). The level of expected technical explainability should also be balanced against the degree of explainability in the corresponding gold standard non-AI process; for example, if an AI process attempts to reproduce a human decision that is not explainable, the degree of technical explainability reasonably expected from the AI may not be as high. In all cases, but particularly in cases where explainability is reduced, including a discussion of other measures (e.g. limitations and generalizability of testing data, traceability, auditability, and transparent communication about system capabilities) is needed for internal verification, which also allows external reviewers to understand how unintended biases and model assumptions may affect performance[14]. When determining whether the provided level of explainability is appropriate for an emerging AI, the potential severity of consequences in case of inaccurate predictions or unreliable human supervision should be considered. For example, certain outcomes, such as identification of surgical candidates, seizure forecasting, and sudden unexpected death in epilepsy (SUDEP) prediction may have more severe consquences than others in case of inaccurate predictions.

Because of the often tacit assumption that AI predictions are closer to “truth” than patient report, it is important for clinicians to be cognizant of this latent assumption, especially in AI with low explainability, and evaluate when incorrect.

Q3. *Is the AI algorithm intended for locked or continuous learning*? Continuously learning applications automatically update using inputs during use, as opposed to locked applications, which do not change after initial training. Evaluating safety, efficiency, and equity for continuous and locked learning involves contain distinct challenges. “Distributional drift” is a phenomenon that occurs most frequently in locked learning when training data does not match ongoing testing data. Various types of distributional drift can occur, including covariate drift (where the input distribution changes), prior probability drift (where the outcome distribution changes), and concept drift (where the relationship between covariates and predicted outcome changes). Locked learning AI algorithms are particularly susceptible to distributional drift, which may lead to inaccurate conclusions. To ensure that AI operating in dynamic environments does not degrade over time, drift detection and performance re-evaluation when drift is suspected/detected are needed, particularly in locked learning. The rate of distributional drift will

vary on a case-by-case basis. If original test data are diverse and large, distributional shift may be gradual, while if original test data are non-representative, small, or if a major event occurs, distributional drift may occur quickly. Drift detection methods[25] are helpful for detecting when distributional drift occurs. If distributional drift is detected, performance estimates may no longer up-to-date and should be re-evaluated. Benefits of performance re-evaluation must be reasonably weighed against financial costs and computational time. Alternatively, continuous learning AI mitigates distributional shift, but unless performance estimates are published in realtime, it can lead to outdated and inaccurate performance estimates. The FDA does not currently have defined guidelines for monitoring changes in performance in continuously learning systems.

Q4. *Is the AI algorithm assistive or autonomous*? Assistive AI algorithms provide recommendations whereas autonomous AI algorithms operate autonomously without human supervision. Some devices and algorithms, such as responsive neurostimulation (RNS) or physiology-based smart watches,[3,26] can operate as either. For example, currently both FDAapproved/cleared versions of these systems run autonomously in their seizure detection capacities; however, the clinician (in the case of RNS) or patient (in the case of a smart watch) assists the AI in confirming detected events as true seizures or false alarms. If an algorithm intended solely for assistive capacity is misused in an autonomous capacity, this may lead to harm due to lack of appropriate human supervision. AI proposed for solely autonomous usage, such as closed-loop systems that operate outside of human without supervision, should be held to higher performance standards. There may be different implications for assumption of responsibility and liability in assistive versus autonomous AI systems, with some advocating for liability to be imposed on AI developers for autonomous systems[21,27].

B. CONSIDERATIONS IN STAGE 2 (ALGORITHM DEVELOPMENT)

Q5. *How well are latent biases in training data and sources of missingness assessed and mitigated?* Training datasets often include demographic inequalities, historical bias, and incompleteness. There are at least two major types of bias that can be present in training data. First, training data may not reflect accurately the epidemiology within a given demographic; for example, underdiagnosis of dissociative seizures in areas without access to tertiary epilepsy centers, underrepresentation of low socio-economic status or rural populations in top research centers, or off-label use of medications/devices. Second, training data may undersample specific subgroups. For example, patients with rare epilepsies, children, and elderly patients are often underrepresented in training and testing data. Missing data can lead similarly to unrecognized systematic undersampling; for example, members of groups that have historically faced discrimination or other disadvantages may be more reluctant to provide personal information that could be used against them. These sources of bias may result in decreased accuracy in undersampled subgroups, perpetuate discrimination/marginalization, or lead to over/underestimation of risk in specific populations.[28] Identifiable sources of bias should be acknowledged and removed in the data collection phase when possible.

Strategies can include recruiting from diverse backgrounds, training the algorithm exclusively on the cohort alone, or training on data evenly distributed across cohorts. To promote transparency in potential latent biases, demographic characteristics should be reported in training/testing data for all AI algorithms. There are several key demographics in epilepsy (age, sex, socio-economic status, intellectual/developmental disability) which can be useful to report in epilepsy due to common demographic inequities. Depending on the application, other characteristics may be relevant as well, such as race, seizure frequency, height/weight, comorbidities, medications, and epilepsy etiology. Whenever demographic subpopulations are underrepresented, latent bias should be acknowledged as a limitation.

Q6. *Are proxy outcomes used and what are sources of measurement error?* Use of proxy outcomes and measurement error may lead to algorithmic bias against patient groups. If proxy outcomes are used, a careful evaluation is warranted of cases where the proxy outcome will not reflect the desired outcome and consideration of how differences may result in algorithmic bias against patient subgroups. Examples of proxy outcomes include the use of hospital visits as a proxy for illness,[29] heart rate escalations as a proxy for seizures,[30] and sustained detection of epileptiform activity as a proxy for electrographic seizures.[23] Measurement error can independently be present in variables themselves, which can also result in algorithmic bias. For example, measurement error in counting self-reported seizures may be higher in seizures with loss of consciousness, which may result in algorithmic bias against patients with focal dyscognitive or generalized seizures.[31]

Q7. *Could the AI lead to self-fulfilling prophecy and perpetuate disparities present in training data?* Training algorithms on real-world data will reflect disparities present in data and may result in perpetuation of AI bias and deescalation of care, violating non-maleficence through selffulfilling prophecy. For example, if clinicians deescalate anti-seizure medications (ASMs) early for patients predicted to be at high likelihood for failure from a particular ASM, then further training the AI on data reflecting these clinical decisions will likely classify these patients as likely to fail the ASM, resulting in even higher likelihood of early ASM de-escalation. To mitigate this, sources of training bias should be acknowleged and attempts made to decrease effects of bias. Algorithms trained on real-world data, which is at greater risk for being subject to this bias, can also be trained on randomized clinical trial data to reduce bias. However, if algorithms are trained only on randomized trial data, this omits clinically important sources of knowledge present in real-world data. Samples studied in randomized clinical trials may also be outliers to the broader epilepsy community, including generally higher seizure frequencies than typical patients.[32] Therefore, in data subject to bias from self-fulfilling prophecy, training on both real-world data and randomized clinical data is ultimately needed, along with acknowledgment of known bias sources and attempts for mitigation.

Q8. *How is data ownership/access defined?* It is important to define data ownership and patients' and researchers' rights to access data up-front. There is an open debate about these choices.[33] Stakeholders claiming ownership may include patients, researchers, institutions, industry, and funding agencies. Data at present are often owned by the entity collecting the data, such as industry, funding agencies, or institutions. Patients may also seek access to their own research data.[23] Although considerations of autonomy suggest that patients should be provided direct access to research data, unregulated access may also lead to harm when there are no guidelines available for interpretation of raw data. Allowing open access is beneficial to the scientific community; however, doing so may decrease the competitive advantage of entities that have invested significant resources into data collection, who will likely incur additional costs to ensure data privacy protection and appropriate sharing. Establishing how these issues will be handled early on can help avoid downstream issues, as each choice has different implications on autonomy, non-maleficence, and beneficence.

C. CONSIDERATIONS IN STAGE 3 (ALGORITHM CALIBRATION)

Q9. *How comprehensive is the performance testing?* Due to the context-specific nature of AI applications in epilepsy, traditional testing on simulated data and a single real-world case example carries limited generalizability. Models tested on one clinical dataset may poorly generalize to datasets with other patient groups. Testing conducted by both internal and independent external parties increases auditability. Multi-institutional datasets, adversarial testing to “break” the system, incentive competitions for external developers, and AI self-play can be considered (eTable 2, available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8).

Q10. *Are practices employed that may lead to overly optimistic performance estimates?* Practices leading to overly optimistic performance estimates of AI systems include failure to compare with null models or the gold standard, comparison only to sub-par competitors; improper separation of training and testing data; overfitting; and poor quality or biased labeling practices.[34,35]

Q11. *When optimizing performance testing metrics, was optimization tailored toward metrics most valued by the target population?* At minimum, all AI should report performance metrics that are standard in statistical practice. Approriate performance testing practices are not within the scope of this article and is discussed elsewhere.[34] Accuracy, sensitivity, and specificity should not be reported in isolation. For example, when predicting whether an event will occur (such as seizure forecasting), one can achieve 100% sensitivity by always predicting that the event will occur. Similarly, since accuracy is the weighted average of sensitivity and specificity, it should never be reported in isolation; one can easily achieve near perfect accuracy even in the presence of low specificity if there is a high prevalence of events. As different patient populations may weight false positives and negatives differently, it is helpful to conduct an analysis or literature review prior to algorithm development of which populations are most likely to utilize the algorithm, and to gain an understanding of the relative importance of false positives and negatives to the populations at greatest probability of usage. eTable 2 (available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8) highlights several examples.

Q12. *Is there equity in performance testing?* Ideally, estimates of performance should be provided for multiple patient subgroups in the intended use population. Subgroups evaluated should be of sufficient size for valid performance estimation. As additional data collection incurs cost to research/development, benefits must be balanced against cost practicalities.

Q13. *Is there a reasonable plan for periodic reevaluation of AI performance?* The safety, efficiency, and equity distribution of AI performance will change over time as clinical contexts change. A reasonable plan for re-evaluation is needed in both locked and continuous learning AI systems.

Q14. *Is the AI's performance level justified against the potential cost to patients in case of AI error?* AI errors can include incorrect predictions, induced human complacency, and data/device failure modes (e.g., unreliable data collection or supervision). Minimum standards for reasonable performance and reliability should be weighed against the potential for patient, stakeholder, and societal harm and realistic worst case scenarios for harm caused by AI errors.

Q15. *What is the ecological impact?* AI systems with large computational costs may have an ecological impact in terms of carbon footprint. For example, natural language processing (NLP) models generally incur high computational costs and may leave a greater carbon footprint than, for example, models built on seizure counts, which have far fewer events and categories to classify than natural language[36]. However, computational cost and ecologic impact must be balanced against performance and reproducibility, as the benefit of better performance or greater reproducibility may or may not outweigh the ecological cost of a greater carbon footprint.

Strategies such as transfer learning and variational inference can help reduce resource consumption. In cases where equal performance and high reproducibility can be attained, more efficient approaches are preferred.

4. CONCLUSIONS

Awareness of the potential ethical implications of study design and algorithm choices that may unintentionally enter AI research is crucial to ensuring that the impact of AI in neurology leads to patient benefit rather than harm. Concrete steps in the early stages of research can help preempt inherent structural issues contributing to later biases and unintended consequences.

This work is intended to provide AI developers and researchers with an operational set of guidelines for conducting ethical AI research in neurology, and to provide clinicians and peerreviewers with a systematic approach to evaluating the potential ethical consequences of emerging AI research. While we focus on epilepsy as a paradigmatic case, similar approaches may be followed in other subfields of neurology: for example, considerations of assistive/autonomous usage, locked/continuous learning, and potential for self-fulfilling prophecy in automated detection of stroke and intracerebral hemorrhage[37]; explicability and patient privacy in deep learning to predict Alzheimer’s disease[38]; patient privacy and latent bias in AI-based systems to predict diabetic neuropathy using facial recognition from home cameras[39].

Adopting a systematic approach to considering the ethical ramifications of emerging research on the principles of beneficence, non-maleficence, patient autonomy, justice, and explicability can help ensure that the patient’s benefit remains at the forefront of the neurological community’s efforts.

5. LIMITATIONS

The field of AI is fluid, and there are several caveats to these recommendations. (1) The proposed recommendations are intended for AI research in development/validation stages in neurology. There are various other ethical issues and questions that arise in later stages of AI development (e.g., implementation and maintenance in the clinical environment), and by other stakeholders, including end-users and deployers, which have been addressed by other experts[40]. Regulatory guidance from the FDA is needed at later stages[15]. (2) These recommendations are intended only to address ethical considerations related to AI use in neurology, and not its technical quality, which is addressed in other resources. (3) Issues already covered by Institutional Review Board or FDA requirements, such as data privacy and protection, usability testing, regulations for off-label indications, informed consent, and liability/redress are not addressed here.

TABLES

Table 1. Checklist of ethical considerations when conducting or evaluating AI research in epilepsy during algorithm conceptualization.

Question	**Recommendation**	**Ethical principle(s)**
1. To what extent were key stakeholders directly or indirectly involved in conceptualization/design phase of the AI application?	State whether key stakeholders were directly or indirectly involved in conceptualization of the algorithm. If key stakeholders were not directly involved, include references to existing literature on stakeholder needs (e.g., priorities, privacy, fundamental rights considerations). In either case, provide the characteristics of participating stakeholders, paying careful attention to which groups may not have been fully represented. An effort should be made to ensure that the participants involved in shaping the technology include anticipated potential future user groups, including lower socio-economic status, poor health, and pediatric and elderly populations.	Beneficence, non maleficence, autonomy
2. *Is the explicability of methods justified against potential harm in the case of erroneous predictions or unreliable human*	Employ methods with high technical explainability or technical solutions for increasing explainability when possible (eTable 2, available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8). Include	Explicability

supervision?	a technical as well as a non-technical explanation adapted to readers with a non-statistical background to increase transparency; details that help increase explicability include sources of data, data collection procedure, cleaning/transformations, data labeling (including the background of persons labeling data), algorithm, model assumptions and parameter settings, sensitivity analysis, evaluated test cases, and limitations. Limitations should include a description of examples of the types of errors that may occur with the technology in language understandable to all users. Weigh the expected degree of algorithm explicability against the severity of consequences in the case of erroneous predictions or unreliable human supervision.	
3. *Is the AI algorithm intended for locked or continuous learning*?	Explicitly state whether the AI is locked or continuously learning. If locked learning is used, recognize as a limitation that distributional drift may occur and performance re-evaluation may be needed. Consider distributional shift monitoring methods to guide timing for performance re evaluation. Discuss role of out-of-sample testing to re-evaluate algorithm performance.	Non maleficence, justice, explicability
4. *Is the AI algorithm assistive or autonomous*?	State whether the AI is assistive, autonomous, or both. Pay particular attention to moments of trade off between the human and the AI, e.g., where the human may assume the AI is doing more than it accurately can, or the AI is handing off a task to a human unable to take appropriate or timely action. If the AI is developed for use in an assistive capacity, exercise caution prior to use in an autonomous capacity.	Non maleficence, explicability

Table 2. Checklist of ethical considerations when conducting or evaluating AI research in epilepsy during algorithm development.

Question	**Recommendation**	**Ethical principle(s)**
1. *How well are latent biases in training data*	Perform analysis to identify underrepresented patient subpopulations, non-representative training	Non maleficence,

and sources of missingness assessed and mitigated?	sets, and potential latent biases present in training/testing data. Describe anticipated sources/mechanisms of missingness. Report demographic characteristics of training/testing data, including age, sex, socio-economic status, intellectual/developmental disability, and any other relevant demographic subgroups for whom performance or needs may be anticipated to vary. If demographic subpopulations are suspected to be underrepresented in training data, reasonable attempts should be made to train the algorithm on a representative dataset sampled evenly across demographics or specifically on underrepresented cohorts. Provide clear statements as to which populations the performance estimates apply, and acknowledge that performance estimates may differ in non-represented or underrepresented populations.	justice
2. *Are proxy outcomes used and what are sources of measurement error?*	State whether proxy outcomes are used and evaluate potential cases where the proxy may not reflect the desired outcome. Discuss subgroups for which measurement error is anticipated to be greater. Consider how this may result in algorithmic bias.	Non maleficence, justice
3. *Could the AI lead to self-fulfilling prophecy and perpetuate disparities present in training data?*	Evaluate potential disparities present in training data, and seek to identify potential scenarios in which bias may be perpetuated or lead to de escalation of care. If present, training on both real world and randomized trial data can help mitigate this bias. Sources of training bias and attempts toward mitigation should be acknowledged within the limitations section. A clear clinical strategy should be outlined to prevent clinical de-escalation of care due to self-fulfilling prophecy.	Non maleficence, justice
4. *How is data ownership/access defined?*	Questions of data ownership and patient/researcher access to their own data should be clarified up front. Unless it is stated from the start that changes can be made after an agreement is established, data ownership and access should not be made more restrictive to patients/researchers without explicit permission.	Beneficence, non maleficence, autonomy

Table 3. Checklist of ethical considerations when conducting or evaluating AI

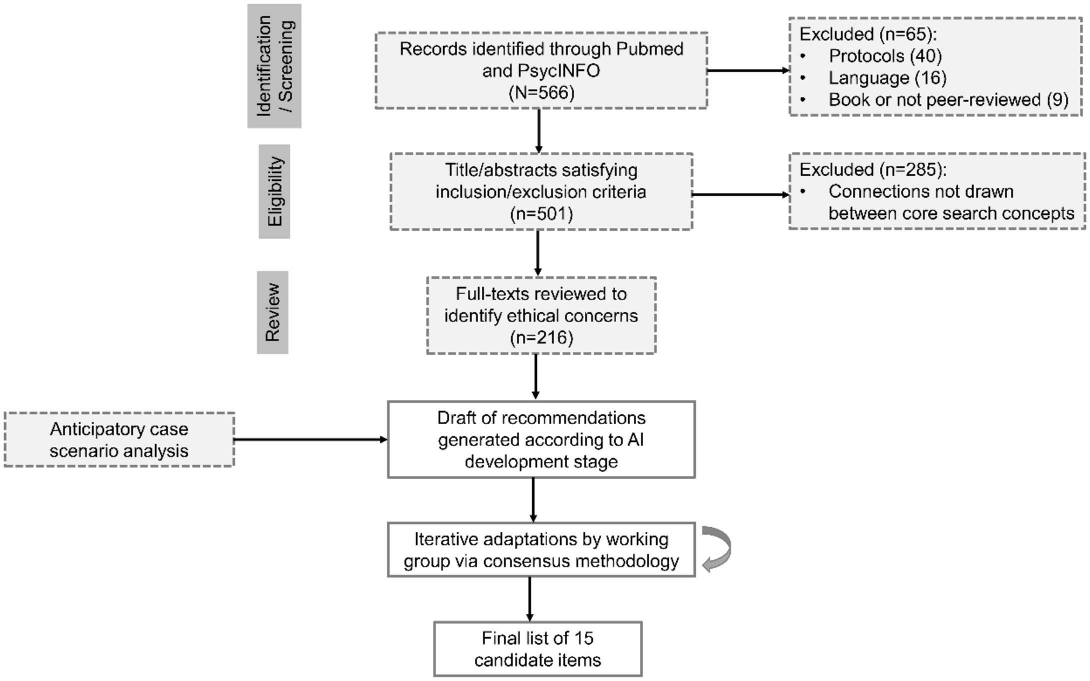
research in
epilepsy during algorithm calibration.

Question	**Recommendation**	**Ethical principle(s)**
1. *How comprehensive is the performance testing?*	Seek to identify and state all limitations on comprehensiveness of performance testing, e.g. single institutional dataset, multi-institutional data, simulated testing, adversarial testing, self play, labelers' training backgrounds, and method for how "true" labels were adjudicated. Consider mechanisms to allow testing by external and internal parties to increase auditability.	Justice
2. *Are practices employed that may lead to overly optimistic performance estimates?*	Performance comparison should be made for at least three use cases: 1) a model for uninformed guessing, such as a majority class predictor or rate-matched forecast; 2) a standard statistical approach; and 3) the current preferred standard clinical method. Statistical rigor to avoid improper separation of training/testing data, overfitting, and biased or poor-quality labeling is essential.	Beneficence, explicability
3. *When optimizing performance testing metrics, was optimization tailored toward metrics most valued by the target population?*	Report at minimum performance metrics that are standard in statistical practice. Perform sensitivity analysis to assess how performance may vary. Conduct analysis or literature review of patient populations most likely to utilize the algorithm and the relative importance of false positives, false negatives, and tolerable levels of accuracy to the populations at greatest probability of usage. If estimates of accuracy or sensitivity (specificity) are provided, the specificity (sensitivity) should also be reported. When possible, show how the AI algorithm can be adapted to change its weights to accommodate different cost tradeoffs.	Beneficence, non-maleficence
4. *Is there equity in performance testing?*	Test performance for patient subgroups in the intended use population, and state for which subgroups the performance measures are reported. Acknowledge limitations in extrapolating to subgroups with lower performance.	Beneficence, non-maleficence, autonomy, justice, explicability
5. *Is there a reasonable plan for periodic reevaluation of AI performance?*	Each AI in clinical use should undergo re assessment of safety, efficiency, and equity periodically to determine if these endpoints deviate from the prior performance standards.	Beneficence, non-maleficence, justice
6. *Is the AI's performance level justified against the*	Even in the early stages of AI development/validation, consider potential harmful consequences to patients in the case of	Non-maleficence

potential cost to patients in case of AI error?	erroneous predictions, indirect effects on inducing human complacency, or possible failure modes, such as unreliable data collection or unreliable supervision. Consider holding AI with greater potential for patient or societal harm to greater minimum performance/reliability standards. Costs to patients, stakeholders, and society to consider include financial, psychological, legal, morbidity/mortality, and exacerbation of disparities.	
7. *What is the ecological impact?*	State the total expected computational requirements of training, testing and expected implementation, being mindful of computationally intensive processes that increase carbon footprint. Adopt strategies to minimize computational cost in cases where similar performance can be obtained via more efficient or less computational expensive algorithms.	Non maleficence, justice

FIGURE LEGENDS

Figure 1. Flowchart demonstrating steps used to generate recommendations. Details are in
eAppendix 2 (available on Dryad: https://doi.org/10.5061/dryad.9zw3r22f8).


Identification / Screening
Records identified through Pubmed and PsycINFO (N=566)
Excluded (n=65):
• Protocols (40)
• Language (16)
• Book or not peer-reviewed (9)
Eligibility
Title/abstracts satisfying inclusion/exclusion criteria (n=501)
Excluded (n=285):
• Connections not drawn between core search concepts
Review
Full-texts reviewed to identify ethical concerns (n=216)
Anticipatory case scenario analysis
Draft of recommendations generated according to AI development stage
Iterative adaptations by working group via consensus methodology
Final list of 15 candidate items