Review

Important Ethical, Technical, and Epidemiological Considerations in an AI Tool Production (ETEPAI): Scoping Review

Boon How Chew[1,2], MD, MMed Fam Med, PhD; Kee Yuan Ngiam[3,4], MBBS, MRCS, MMed Surgery, FRCS

[1]Department of Family Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[2]Family Medicine Specialist Clinic, Hospital Sultan Abdul Aziz Shah, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[3]Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[4]Division of General Surgery (Thyroid and Endocrine Surgery), Department of Surgery, National University of Singapore, Singapore, Singapore

Corresponding Author:

Boon How Chew, MD, MMed Fam Med, PhD
Family Medicine Specialist Clinic, Hospital Sultan Abdul Aziz Shah
Universiti Putra Malaysia
Serdang, Selangor 43400 UPM
Malaysia
Phone: 60397692538
Email: chewboonhow@upm.edu.my

Abstract

Background: Artificial intelligence (AI) tools are being developed in a rapidly evolving technology. The convergence of ethical, technical, and research methods' considerations is crucial for multidisciplinary teams aiming to produce effective AI tools. The success of these tools postdeployment hinges on the intricate interplay between the AI system's development on its output through rigorous decision-making processes and stakeholders' capacity to act on the AI's recommendations.

Objective: This paper synthesizes ethical, technical, and epidemiological considerations for all involved in artificial intelligence tool production (ETEPAI), based on established guidelines, checklists, and frameworks.

Methods: Relevant guidelines, checklists, frameworks, and expert recommendations were systematically identified and synthesized into ETEPAI, an ethical, technical, and epidemiological framework for AI tool development in health care.

Results: From 30 reviewed frameworks, ETEPAI integrates critical considerations across 4 stages (design, development, deployment, and postdeployment) and 3 domains (ethics, technical, and epidemiological), providing a compact yet comprehensive guide. It includes probing questions, key indicators, and common pitfalls to support high-quality, ethically sound, and clinically relevant AI tools. ETEPAI aligns with European Union trustworthiness standards and is supported by a research proposal template and supplementary references to aid implementation and adoption. We present probing questions and critical pointers across 4 stages from the design, development, deployment, and postdeployment, highlighting their relevance in health care settings. The designing stage aligns with epidemiologic research methodologies, while the development stage emphasizes transparent project execution. Deployment and postdeployment stages focus on real-world implementation. Additionally included are common pitfalls and challenges to emphasize the importance of due attention to the importance of ETEPAI considerations to avoid serious consequences.

Conclusions: Applying ETEPAI ensures comprehensive, complete, compact, and crisp consideration from conception to execution, promoting high-quality, ethically sound, and clinically relevant AI tools. The brevity and conciseness of ETEPAI might be adequate for trained personnel and serve as clear signposts to unprepared stakeholders.

JMIR AI 2026;5:e80340; doi: 10.2196/80340

Keywords: artificial intelligence; ethics; epidemiology; clinical decision support systems; practice guidelines

Introduction

In the rapidly evolving landscape of artificial intelligence (AI) development, the convergence of ethical, technical, and research methods considerations is paramount for multidisciplinary teams aiming to achieve success in the production of AI tools [1]. The impact of AI tools after deployment depends on the intricate interplay between the AI system's development, the decision-making based on its output, and the capacity of the stakeholders involved to take the necessary subsequent actions. Foreseeing, estimating, and designing the AI tools that meet all the challenges before deployment are essential to bridge the chasm between AI tools development and achievable benefits [2] with acceptable risks [3].

However, there is a proliferation of guidelines, checklists, assessments, frameworks, and recommendations [4,5] which could lead to confusion, inconsistency, and inefficiency in adherence and application [6]. Overwhelmed by the many similar yet different and lengthy referent materials, developers and stakeholders may overlook the most relevant and essential practices, rendering many good referent materials less effective, stalling improvement in the AI tools production process, or risking the perpetuation of the poor conduct and reporting of such studies [7]. These include a lack of data and code availability, an absence of or small human comparator groups, and a shortage of real-world clinical relevance, transparency, and inappropriate conclusions, causing overall high risk of bias [8], leading to more research waste [9-11]. To address these issues, there is a pressing need to integrate the various referent materials into a unified, thorough, concise, and clear approach across different study designs to inform about the best practices and ethical standards that accentuate the clinical relevance of AI tools and systems production that integrate into clinical workflow for sustainability throughout the whole developmental process.

For the purposes of this review, we use the term "artificial intelligence" as a broad umbrella to encompass the spectrum of machine learning (ML) models commonly used for clinical tasks, from traditional regression to deep learning. While the technical complexity of these algorithms varies, the fundamental principles of ethical oversight, rigorous epidemiological validation, and postdeployment governance required for safe clinical integration remain consistent.

Methods

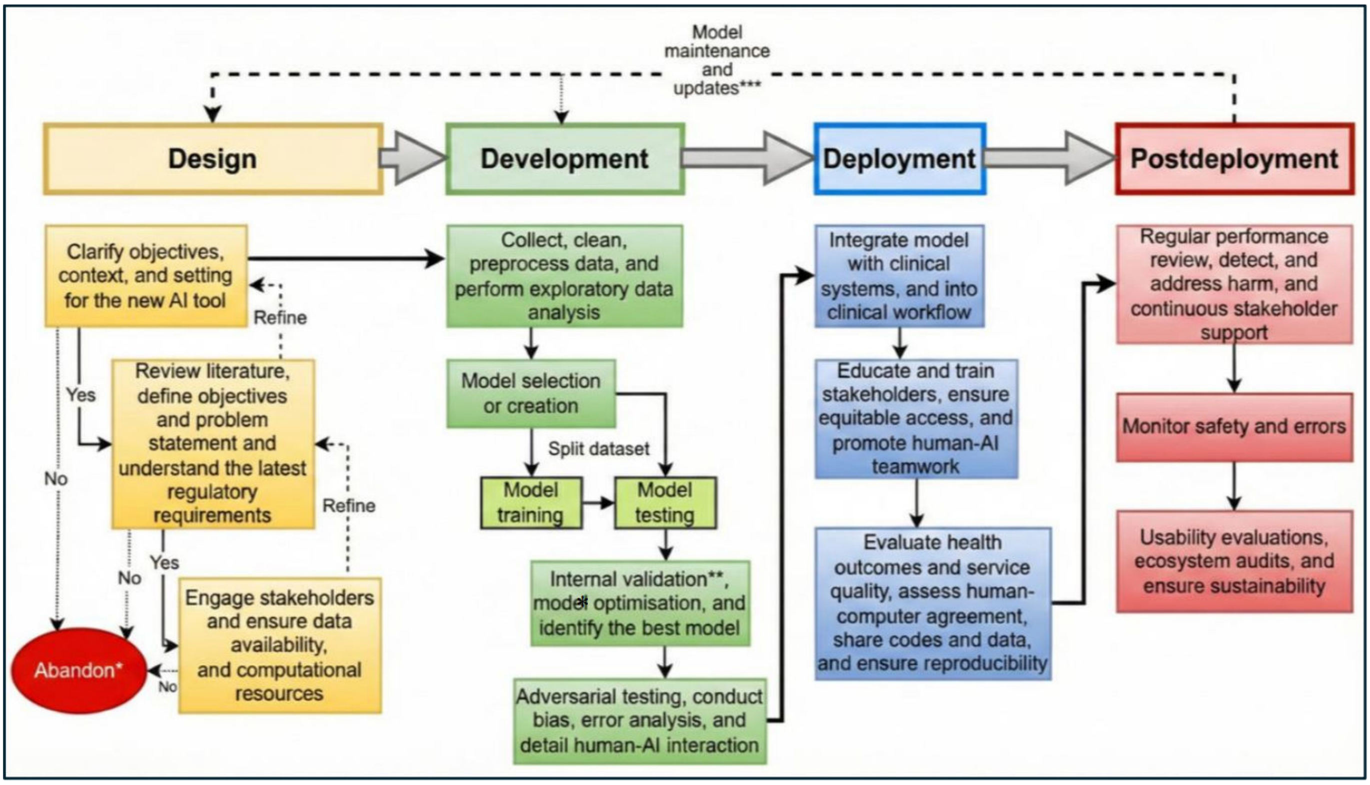
Search Strategy and Selection Criteria

Searches for relevant recommendations were conducted through a multipronged search strategy combining systematic database searches with supplementary methods to capture a broad range of relevant recommendations. First, a systematic search was conducted in key academic databases, including PubMed, Google Scholar, and Semantic Scholar. The search strategy used a combination of keywords and Boolean operators to identify relevant literature. A representative search string was: ("artificial intelligence" OR "machine learning") AND ("guideline" OR "framework" OR "checklist" OR "recommendation" OR "best practice") AND ("healthcare" OR "clinical" OR "medicine"). Second, this initial search was supplemented by several methods. We conducted a targeted review of regulatory documents from prominent health and technology organizations, published policy, and gray literature. We also performed citation searching by manually reviewing the reference lists of highly relevant guidelines and systematic reviews to identify eligible papers. This process was further enriched through consultation with domain experts to ensure seminal works were not overlooked. Additionally, we used AI-powered academic search tools such as Elicit, SciSpace, and Perplexity to identify relevant preprints and newly published papers that may not have been fully indexed in traditional databases. Recommendations that were selected must be developed by groups of experts through a systematic and scientific process, provide explicit, actionable recommendations for the design, development, validation, or deployment of clinical AI tools, and must be published in English.

Data Extraction and Synthesis

The included recommended guidelines [3,12-16], checklists [17-23] and the AIPA (Artificial Intelligence Prediction Algorithm) for medical sectors [24], the Stanford's FURM (Fair, Useful, the Reliable Artificial Intelligence Model) assessment [2], the SUDO (pseudo-label discrepancy) framework [25], and the medical algorithmic audit framework [26], and the STANDING (Standards for Data Diversity, Inclusivity, and Generalisability) Recommendations [27] are summarized with the checklists of the referenced materials provided via links on the tables in the Multimedia Appendix 1 [3,12,13,16-21,24,26-51], and excluded guidelines were discussed further and described in Multimedia Appendix 2 [52-66]. We assimilated the referenced materials into important ethical, technical, and epidemiological considerations for all involved in artificial intelligence tool production (ETEPAI) as critical pointers according to 4 coherent stages of design, development, deployment, and postdeployment in 3 domains of ethics, technical, and epidemiological principles (Figure 1). While these 4 stages are analogous to a standard software development life cycle to ensure familiarity and ease of integration, the critical pointers within each stage are specifically tailored to the unique challenges of developing safe and effective AI tools for health care.

Figure 1. Key stages of AI tools from design to postdeployment. *No responses that lead to an abandonment include nonethical conduct, superior alternatives, infeasibility of whatever causes, nonacceptance by users, strategic or value misalignment to the institutions involved, financial viability, and economic factors to sustain. **May include external validation, decision curve analysis, (early) health technology assessments, and impact studies such as randomized controlled clinical trials. ***Model maintenance and updates may require a revisit from the beginning or to the tuning in the development stage. AI: artificial intelligence.


Model maintenance and updates***
Design
Development
Deployment
Postdeployment
Clarify objectives, context, and setting for the new AI tool
Refine
Yes
No
Review literature, define objectives and problem statement and understand the latest regulatory requirements
Refine
Yes
No
Engage stakeholders and ensure data availability, and computational resources
No
Abandon*
Collect, clean, preprocess data, and perform exploratory data analysis
Model selection or creation
Split dataset
Model training
Model testing
Internal validation**, model optimisation, and identify the best model
Adversarial testing, conduct bias, error analysis, and detail human-AI interaction
Integrate model with clinical systems, and into clinical workflow
Educate and train stakeholders, ensure equitable access, and promote human-AI teamwork
Evaluate health outcomes and service quality, assess human-computer agreement, share codes and data, and ensure reproducibility
Regular performance review, detect, and address harm, and continuous stakeholder support
Monitor safety and errors
Usability evaluations, ecosystem audits, and ensure sustainability

Important concepts and must-have indicators from the referenced materials are deemed sufficient and included in the list of items in ETEPAI as they represent well the rest of the indicators in their respective material. To prime users at the start of using ETEPAI are probing questions that aim to enhance readiness (Textbox 1). Added to these guides on best practices is a section on the common pitfalls and prevailing challenges to highlight and re-emphasize the importance of due consideration to every step in the AI tools development in order to avoid serious consequences. In this approach, ETEPAI is more inclusive and thorough compared to any of the referenced materials. ETEPAI considerations are comprehensive, complete, compact, and crisp to use to guide high-quality, ethically sound, and clinically relevant AI tools from idea conception to plan execution, and from promoting to monitoring. As a result, we also provided an AI research proposal template in Multimedia Appendix 3. Thus, ETEPAI serves as annotated content pages of books that signpost authors and users to the necessity of each step in the process but would require further reading in order to gain further understanding if this did not happen before. Similarly, readers and users of this ETEPAI are advised to refer to the original references for full explanation (Multimedia Appendix 1), and to other more elaborative literatures [67], on rigorous external model validation [16,68-70] and for clinical trial and economic evaluation [21,23,28], or for additional help in complete reporting guidelines [5,71], in organizational capabilities strengthening for AI adoption [72], in sociotechnical dimensions to consider when integrating AI tools in complex adaptive health care systems [73], and to identify technological solutions for achieving large-scale sustained adoption [74]. The scale and brevity of ETEPAI may be sufficient for trained personnel and serve as a sure sign to unmotivated stakeholders, as a standard evaluation and as materials to equip organizations to be AI-capable [72]. Further elaboration on its uses is given below.

ETEPAI's product-centric approach could effectively align with European Union (EU)–defined trustworthiness standards [29,75] by emphasizing safety, robustness, and ethical considerations at every stage of the AI life cycle, from design through deployment. With its focus on rigorous processes, techniques, and verifiable methods, ETEPAI promotes comprehensive risk identification and mitigation strategies that address both product-level and system-level impacts on users and society. This focus directly supports compliance with the EU's safety, health, and fundamental rights protections, ensuring AI systems are both reliable and ethically responsible. While ETEPAI aligns well with the principles of the EU AI Act, its fundamental focus on data privacy, responsible data handling, and accountability ensures its principles are compatible with and supportive of other major regulatory frameworks, such as HIPAA (Health Insurance Portability and Accountability Act) in the United States [76].

Textbox 1. Probing questions in the 4 stages of artificial intelligence (AI) tools design, development, deployment, and postdeployment.

Design

- What problem does this AI model seek to solve?
- Is the AI tool necessary and appropriate?
- How will the AI tool be used?
- Is the context in which the AI tool will be used appropriate?
- When and where will it be used or not used?
- Should a health care provider use the AI tool? Who else will use it?
- Will there be secondary (indirect) users of the AI tool?
- What are the main functions of the AI tool?
- What are the expected outcomes, potential secondary, and unexpected outcomes?
- How will the output of the model be used?
- How might this model impact patients, personnel, and the health care system?
- What would be the impact and consequences of the unexpected outcomes?
- What approaches would mitigate risks arising from the use of the AI tool?
- Given the cost and work capacity involved, what net benefits could be realized?
- Would the AI model-guided workflow be financially sustainable?

Development

- What are the available resources, data sources, and potential trade-offs for the AI system and technology?
- How should the objectives and functions of the AI tool be prioritized according to the available resources?
- Reuse an existing model or learn a new one?
- How to get the best and fairest model within the required time?

Deployment

- Is the deployment process the simplest possible for the acceptable iteration speed?
- How to get the outputs back into the clinical workflow?
- Are the validity and efficiency of the AI tool limited over time?
- How long can the results or the technology that supports the AI tool be used?

Postdeployment

- Does the AI model have the intended impact?
- How often should the AI tool or system be updated?
- Who is responsible for updating the AI system?
- Is governance in place to monitor the AI tool or system for its safety and errors, to audit the ecosystem within the components of larger societal systems?

Results

Critical Pointers and Domains

Figure 2 shows the flow diagram [77]. Figure 1 illustrates the key stages of AI tools from development to postdeployment. Textbox 1 lists probing questions in the 4 stages of design, development, deployment, and postdeployment stages according to their relevance and importance when thinking about having an AI tool in a health care setting. These would alert and engage users' attention to the scopes and contexts of ETEPAI. Similarly, Table 1 provides critical pointers in AI tools production arranged in 3 domains of ethics, technical, and epidemiological principles under the same 4 stages. Although the probing questions and the critical pointers are grouped into different stages and domains, they are all to be considered when any AI tool is being conceived by an individual or discussed by a team of people, which should either result in a comprehensive research project proposal or abandonment if it is found to be unsound or a nonfeasible idea. Please see Checklist 1 for PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) documentation.

Figure 2. Flow diagram of sources screened and assessed for eligibility.

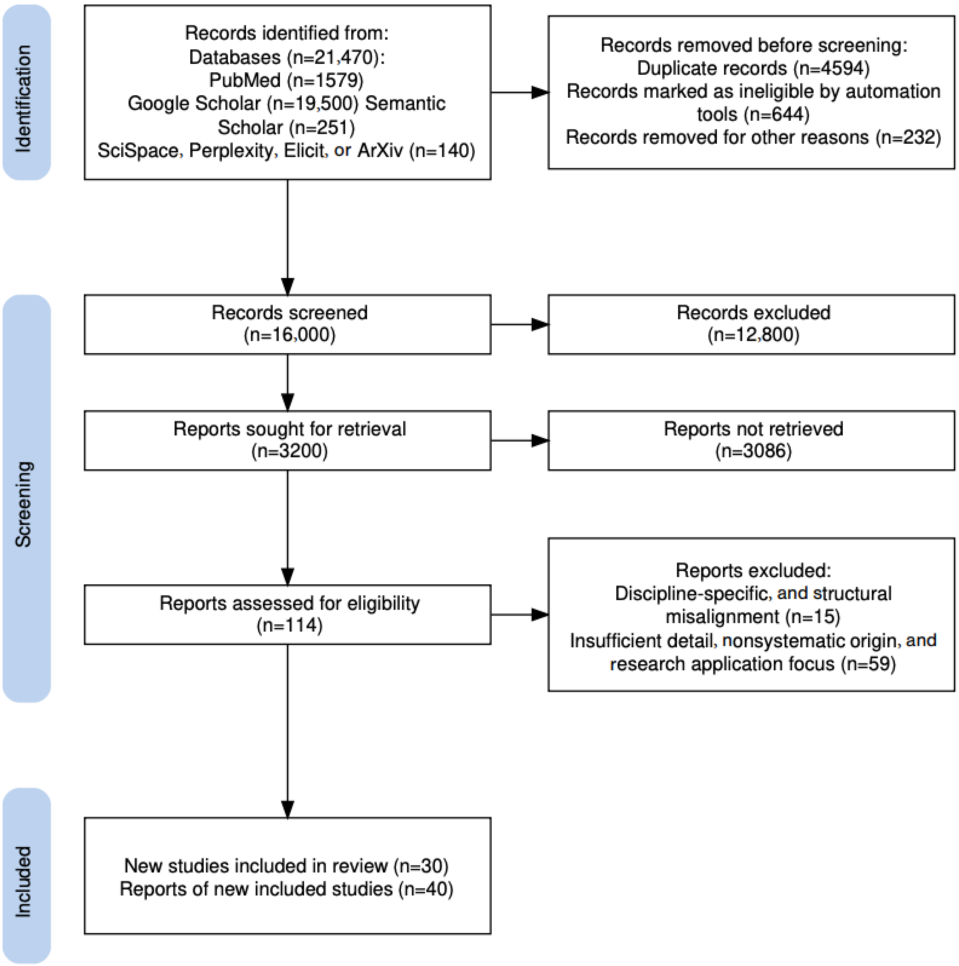

Identification
Records identified from:
Databases (n=21,470):
PubMed (n=1579)
Google Scholar (n=19,500) Semantic Scholar (n=251)
SciSpace, Perplexity, Elicit, or ArXiv (n=140)
Records removed before screening:
Duplicate records (n=4594)
Records marked as ineligible by automation tools (n=644)
Records removed for other reasons (n=232)
Screening
Records screened (n=16,000)
Records excluded (n=12,800)
Reports sought for retrieval (n=3200)
Reports not retrieved (n=3086)
Reports assessed for eligibility (n=114)
Reports excluded:
Discipline-specific, and structural misalignment (n=15)
Insufficient detail, nonsystematic origin, and research application focus (n=59)
Included
New studies included in review (n=30)
Reports of new included studies (n=40)

Table 1. Critical pointers in 3 domains of ethics, technical, and epidemiological principles when considering an AI[a] tool.

	Design	Development	Deployment	Postdeployment
Core concept	• Clarify the objectives, context, and setting for the AI tool. • Adopt standards and best practices to guide the whole design, development, and deployment processes to ensure compliance and interoperability of the AI technology with the health systems.	• Conduct according to the proposed designs and approaches, which are made publicly available as a register on an open platform, a preprint, or a published journal paper.	• Engage and educate multiple stakeholders for deployment and maintenance.	• Evaluate the impact and improve performance of the deployed AI tool and system.
Ethical principles	• Define relevant ethical issues through consultation, assess risk, and address biases. • Prioritize safety in high-risk decisions and procedures. • Promote transparency to assign responsibility, ensure trust in target users and the public, and protect patient rights. • Address bias in the datasets for mitigation and not replication[b].	• Preserve and enhance human autonomy in the use of the AI tool, where informed consent and the decision to refuse are allowed in the AI system. • Maintain privacy and confidentiality in the collection and use of patient data, and uphold transparency of the AI system and its training data.	• Reliability proven from testing in a similar setting. • Nonmaleficence as harms and worst-case scenarios were considered. • Ensure equitable access to the AI tool and related health care technologies and services.	• Governance is in place to detect harm, redress plans, with a mechanism for humans to roll back, disengage, or deactivate the AI model.

	Design	Development	Deployment	Postdeployment
	• Safeguard the privacy of the datasets from any form of reidentification, and implement a dataset retention or removal plan. • Responsibility and accountability among the model developer, IT[c] staff at the deploying organization, and clinical staff are clearly defined. • Limit the environmental impacts.	• Traceability and auditable AI methods[d].		• Institute regular challenges and reviews to monitor performance of the AI tool and system[e].
Technical principles	• Define problem statement and project scope, identify regulatory requirements, and policy framework[f]. • Engage multiple stakeholders and understand contexts including geographical scope, users' background, main languages, and digital skills[f]. • Designs that are inclusive of all related stakeholders, effective preparation for source and learned code sharing. • Data availability, quality[b], and data procurement plans for data protection, high-quality annotation, and leakage prevention. • Ensure sufficient computational resources[g]. • Strive to achieve model interpretability and explainability[h]. • Design the AI system with scalability and performance to handle growing data volumes and user demands[i].	• Conduct data collection as planned for all types and sources of data, data governance, and access[b]. • Clean and preprocess the data to handle missing values, outliers, noise, and inconsistencies[b]. • Perform exploratory data analysis to gain insights into the data and understand its characteristics[b]. • Feature (variables) engineering to improve the performance of the model[j]. • Model architecture and algorithm selection, and training that fit the project objectives[k]. • Evaluation and validation to assess the performance, accuracy, and generalization capabilities of the model[l]. • Hyperparameter tuning and optimization to fine-tune the model's parameters and improve performance[m]. • Design, data, and process audits[d]. • Use adversarial attacks and red-teaming to identify worst-case behaviors[n].	• Clearly delineate responsibility for what to do, when, and how. • Train stakeholders in why, how, and when to use the tool, including the main objectives, functions, and features, and differences among usage scenarios, when applicable[o]. • Prepare the AI model for deployment in a production environment, integrate the model with the target system or application, and deploy it to production servers or cloud platforms. • Promote effective human-AI teaming[h]. • Implement mechanisms for model versioning, monitoring, and rollback to manage the deployment process effectively. • Codes and data are shared with the community.	• Documentation , knowledge transfer, and engage continuously with stakeholders and support users[p]. • Establish a regular technical review to determine[e] whether the AI tool is having the intended impact, is filling a gap in need, and is improving health care[q]. • Maintenance, incorporating, verifying, and validating changes to the tool or system[e]. • Ecosystem audit[r], health economic evaluation, and/or full algorithmic audit[d].
Epidemiological methods	• Background: review existing relevant literature exploring AI models for the problem being addressed. • Objective and problem: clearly state what the proposed AI tool aims to address with respect	• Model specification: specify the final panel of features included, ensure the independence between training and test sets[w], and hyperparameters tuned[m]. The cohorts (training and test sets)	• Model explanation and interpretability to improve acceptability, uptake, and sustainability[h].	• Safety and errors: monitor and report any risks to patient safety or instances of harm. Provide a

	Design	Development	Deployment	Postdeployment
	to the study setting, population, and outcome. Define the research question clearly[f]. • Eligibility criteria for patients and features or input data, and rationale[s]. A fair distribution of the severity of disease and alternative diagnoses is required (spectrum bias). Time interval between index test and reference standard is within an appropriate interval[f]. • Ground truth or referent standard: define the ground truth of interest, conditions, and outcome events, and rationale (if alternatives exist)[t]. Describe how it will be collected and encoded[u]. Define the test positivity cutoffs, distinguishing the prespecified from the exploratory. • Source of data: describe how the dataset will be obtained and the study period[b,f]. • Data abstraction, cleaning, and preparation to develop the final dataset[v,w]. • Data splitting: specify how the data is to be divided into the training and testing cohorts[l]. • Sample size estimation: provide rationale for sample size required for model development (eg, based on power calculation)[x]. • Baseline model: describe the baseline model that will serve as a comparison for the AI model. • Model description and evaluation: describe the software libraries to be investigated, the evaluation metrics to assess performance and calibration, transformation and optimization strategy[t,u,v]. Include clinical utility assessment, bias assessment, and error analysis (justify why not) beside the statistical methods to analyze the primary and secondary outcomes, subgroup analyses, and their rationale.	are shown to be representative of real-world clinical settings (to be discussed if not), and missingness is addressed: reported, imputed, or corrected[y]. • Clinical utility assessment: use appropriate metrics for the risk or benefit trade-offs at the specified decision threshold[z]. • Validation or efficacy: nonclinical and clinical research for validation or to estimate the AI tool's clinical effect in the routine clinical workflow, usability to those involved, and effects on clinical outcomes[z]. • Bias assessment: compare evaluation metrics for the AI tool and reference standard, including stratification by patient- and task-specific subgroups, or subtyping[t,u,v,w]. • Error analysis: analyze predictive errors to identify characteristics that are more prone to inaccurate predictions[x,t,u,v]. Determine if there are any surprise errors. • Hyperparameter tuning: specify all model hyperparameters that are optimized, the search space for hyperparameter tuning, and evaluation metrics used to optimize parameters[m]. • Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users[h]. Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice[h].	• Adoption: implement the AI system within the intended clinical workflow or care pathway[e]. • Effectiveness: evaluate all patients' health outcomes and service quality indicators, before and after deployment[aa]. • Human-computer agreement: evaluate and report any instances of and reasons for user variation from the AI system's recommendations, and if applicable, users changing their mind based on the AI system's recommendations[aa]. • Reproducibility: share the data, source code, or release an application that runs the code. A data dictionary involves providing descriptions of all features and ground truth.	description of how significant errors or malfunctions were defined and identified[ab] . • Maintenance or sustainability: conduct usability evaluation, according to recognized standards or frameworks, the user learning curves evaluation, stakeholder and patients' acceptability of the AI tools or systems. Support for the intended use of the AI system in clinical settings.

[a]AI: artificial intelligence.

[b]Ensure data adequacy by using representative, diverse, and sufficient data, properly labeled and curated to reduce bias. When using pre-existing cohorts for case-control sampling, make necessary sampling adjustments (eg, reweighting) to ensure correct calibration. Maintain process transparency by documenting the entire data pipeline, including collection methods, dataset purpose, and handling of missing or complex data. Finally, detail all known biases, disparate outcomes, or data shifts; describe mitigation procedures; and transparently report on any formal bias and fairness assessments (STANDING Together 2023 [Standards for Data Diversity, Inclusivity, and Generalisability Together 2023] [27], AIPA [Artificial Intelligence Prediction Algorithm] [24], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence

Studies in Healthcare] [17], CODE-EHR [Clinical Outcomes in Digital Enterprise - Electronic Health Records] [19], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], the medical algorithmic audit [26], and UN Resolution on AI [United Nations Resolution on Artificial Intelligence] 2024 [31]).

[c]IT: information technology.

[d]Artificial intelligence auditing is a critical mechanism for accountability in system decisions, with several key approaches: data audits evaluate training data; process audits scrutinize development documentation; and ecosystem audits assess human-AI interactions. This practice of algorithmovigilance, best conducted with all stakeholders, proactively identifies vulnerabilities to mitigate risk, guide critical thinking on system acceptability, and inform future model improvements. Audits should be complemented by related impact assessments and health economic evaluations to estimate cost-effectiveness (three key questions [78], the medical algorithmic audit [26], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence] [32], UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33], and CHEERS-AI [Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence] [28,34]).

[e]Establish a maintenance plan to regularly monitor model performance, data quality, and user feedback. This is critical for addressing performance degradation caused by factors such as data distribution shift (concept drift) or inconsistencies from new devices, and for guiding how to iterate on the artificial intelligence system with necessary updates based on evolving requirements (APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence] [32], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33], and STANDING Together 2023 [Standards for Data Diversity, Inclusivity, and Generalisability Together 2023] [27]).

[f]Clearly define the artificial intelligence application’s context by specifying its medical indication, target population, intended end user (eg, specialist or patient), and the health care process it aims to improve with its expected benefits. The application’s timing of use within the clinical workflow and its type (eg, diagnostic, prognostic, and monitoring), including any prediction horizon, must also be defined. Crucially, ensure active stakeholder engagement to align with user needs, build trust, and ensure usability (AIPA [Artificial Intelligence Prediction Algorithm] [24], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], and CODE-EHR [Clinical Outcomes in Digital Enterprise - Electronic Health Records] [19]).

[g]Evaluate the computational resources required for training and deploying the artificial intelligence model. Consider factors such as the complexity of the model, the size of the dataset, and the computational power needed for training, inference, and scaling (ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], and UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33]).

[h]Consider the importance of model interpretability and explainability to enhance transparency into how the artificial intelligence model makes decisions. Various methods can provide insights, such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) for local and global feature importance, partial dependence plots to illustrate feature-outcome relationships, or saliency maps to highlight influential areas in images (AIPA [Artificial Intelligence Prediction Algorithm] [24], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence] [32], The medical algorithmic audit [26], and UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33]).

[i]Design the artificial intelligence system with scalability and performance in mind to handle growing data volumes and user demands. Consider distributed computing, parallel processing, and optimization techniques to improve efficiency and scalability (ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30] and UN Resolution on AI 2024 [United Nations Resolution on Artificial Intelligence] [31]).

[j]Apply feature engineering techniques (eg, dimensionality reduction, scaling, and transformation) to improve model performance, and meticulously document all analytical and modeling procedures in sufficient detail to allow a third party to accurately reproduce the results (AIPA [Artificial Intelligence Prediction Algorithm] [24], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], and FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence] [32]).

[k]Choose appropriate algorithms (eg, supervised and deep learning) based on project objectives, split the data into training, validation, and test sets, and clearly document all modeling techniques and stages from model selection to parameter tuning and calibration or recalibration to ensure reproducibility (AIPA [Artificial Intelligence Prediction Algorithm] [24], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], and the medical algorithmic audit [26]).

[l]Validate the model against predefined criteria using appropriate internal validation methods (eg, bootstrapping and cross-validation), and assess its performance using a comprehensive suite of performance metrics, such as those for accuracy and precision, quantitatively and qualitatively [79] (AIPA [Artificial Intelligence Prediction Algorithm] [24], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], and TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16]).

[m]Explore techniques such as grid search, random search, or Bayesian optimization to search for optimal hyperparameters efficiently (APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17] and TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16]).

[n]Before real-world deployment, conduct adversarial attacks and red-teaming exercises where teams actively try to exploit vulnerabilities to proactively identify worst-case behaviors, potential for malicious use, and unexpected failures that fixed benchmarks might miss (APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting

of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence] [32], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], the medical algorithmic audit [26], and UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33]).

[o] Ensure transparency and usability by providing a comprehensive model card. This should detail the model's purpose, intended use, performance metrics, and methodologies, with information tailored to the specific needs of different end users, such as providing implementation details for clinicians and clear explanations of impact for patients (AIPA [Artificial Intelligence Prediction Algorithm] [24], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], and UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33]).

[p] For high-stakes tools on data without ground-truth labels, advanced techniques such as the SUDO (pseudo-label discrepancy) framework can identify unreliable predictions and bias. This should be part of a comprehensive documentation of the entire development process, with training materials provided to ensure effective knowledge transfer to stakeholders (APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis - Artificial Intelligence] [16], FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness and Explainability - Artificial Intelligence] [32], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], and UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33]).

[q] Conduct thorough testing (eg, unit, integration, and system) to verify functionality and detect performance shifts caused by real-world dataset shifts (eg, population and prevalence). Pay close attention to performance gaps in specific subgroups (hidden stratification) and systematic algorithmic errors or failure modes [26] (APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence [32]], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], and the medical algorithmic audit [26]).

[r] Postdeployment analysis allows AI systems to be studied as components of larger societal systems; monitoring real-world usage, including emergent threats such as jailbreaks or deepfakes, is crucial for shaping scientific research on mitigating harms (ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], the medical algorithmic audit [26], and UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33]).

[s] Inappropriate participant inclusion or exclusion criteria can harm generalizability, as the training data must be representative of the target population. Artificial intelligence systems perform well on in-distribution data (interpolation) but poorly on out-of-distribution data that requires extrapolation (STANDING Together 2023 [Standards for Data Diversity, Inclusivity, and Generalisability Together 2023] [27], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], and the medical algorithmic audit [26]).

[t] Bias can arise from flawed methods of outcome determination, such as using suboptimal or inconsistently applied criteria, incorrect timing, or knowledge of predictors influencing the assessment. A key issue is incorporation bias, where a predictor is part of the outcome definition, leading to overly optimistic performance estimates (APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], CODE-EHR [Clinical Outcomes in Digital Enterprise - Electronic Health Records] [19], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], and the medical algorithmic audit [26]).

[u] Bias can also be introduced by flawed predictor definition and measurement, such as when predictors are defined inconsistently across participants or when knowledge of the outcome influences their assessment (APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], and STANDING Together 2023 [Standards for Data Diversity, Inclusivity, and Generalisability Together 2023] [27]).

[v] For model development, ensure sufficient sample size, typically an events per variable of ≥20, and often >200 for artificial intelligence models to minimize overfitting and avoid selecting predictors based on univariable analysis. For model validation, a minimum of 100 participants with the outcome is recommended, and unlike for development, a priori sample size calculations are generally feasible (TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16]).

[w] Select predictors based on existing knowledge and clinical credibility, not univariable analysis; retain important predictors regardless of statistical significance. Handle continuous predictors appropriately: avoid dichotomization, which loses information and risks bias, and instead model them continuously while examining for nonlinearity (eg, using fractional polynomials or splines; TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16] and STANDING Together 2023 [Standards for Data Diversity, Inclusivity, and Generalisability Together 2023] [27]).

[x] While larger datasets are generally better, especially for complex models or class imbalance, this must be balanced with ethical data minimization. For models where a priori sample size formulas are unavailable, use a posteriori methods such as learning curves to assess data sufficiency and minimize overfitting (AIPA [Artificial Intelligence Prediction Algorithm] [24] and ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30]).

[y] For a robust analysis, include all participants, handling missing data with appropriate methods such as multiple imputation. Evaluate both calibration and discrimination, and use internal validation techniques (eg, bootstrapping and cross-validation) to adjust performance estimates for model optimism (TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17], STANDING Together 2023 [Standards for Data Diversity, Inclusivity, and Generalisability Together 2023] [27], FUTURE-AI [Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence] [32], and the medical algorithmic audit [26]).

[z] Evaluate and document the model's expected benefits using methods such as decision curve analysis, which assesses net benefit on decision-making, or a more comprehensive early health technology assessment to evaluate medical, economic, and social implications (AIPA [Artificial Intelligence Prediction Algorithm] [24], TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence] [16], and APPRAISE-AI [Tool for Adapting Practice Parameters for Reporting Artificial Intelligence Studies in Healthcare] [17]).

[aa]Evaluate potential outcomes by aligning with the quadruple aim (patient experience, cost, population health, and provider perception), considering a range of process, health, and societal impacts. A comparative study is essential to assess the tool's added benefit, ideally a randomized comparative design, though alternatives such as controlled before and after studies may be used when randomization is not feasible (AIPA [Artificial Intelligence Prediction Algorithm] [24], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33], and TRIPOD-AI [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis r Diagnosis - Artificial Intelligence] [16]).

[ab]For effective implementation, continuous monitoring is essential. This includes tracking performance and errors (eg, miscalibrations, false positives or negatives, and technical failures) and assessing fairness by analyzing outcome disparities across different populations to detect and mitigate bias (AIPA [Artificial Intelligence Prediction Algorithm] [24], ALTAI [Assessment List for Trustworthy Artificial Intelligence] [13,30], and UNESCO Recommendation on the Ethics of Artificial Intelligence 2022 [United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence 2022] [33]).

ETEPAI Tool Production and Use

The designing stage is aligned with the epidemiologic research methodologies comprising the theoretical design, data collection design, and statistical analysis design [80,81]. The theoretical design involves identifying the real need for the AI tool, and defining the research question by having a clear target population, the exposure (or independent variable) of interest or interventional program, context or control, and the outcome measured at specified time point and setting (the PE/ICOTS [Population or Participants, Exposure/Intervention, Comparator, Outcomes, Timing, and Setting] acronym) [35,80]. The data collection is usually on already available datasets to be further selected and cleansed according to eligibility criteria. Prospective sampling is instructed by some guidelines. The statistical analysis includes the AI-specific models' development or selection, evaluation, and validation procedures, estimating the best-performing AI-based prediction model. This is often summarized in terms of calibration and discrimination, without overoptimism, and with internal (cross-validation), external validation, decision curve analysis, (early) health technology assessments, and impact studies (such as via randomized clinical trials) [1]. The development stage is the conduct of the project according to the published and openly available project or research proposal. This allows public scrutiny of the project conduct from the proposed designs. Therefore, justification as a clear explanation must be given for every deviation in the actual conduct of the project from the proposal. The deployment and postdeployment stages are similar to the implementation research [82,83], where the tested AI tools are used in the routine practice. Deployment cost and organizational effort required to integrate the AI systems into a clinical workflow should be estimated to facilitate sustainability [78]. Most pointers in Table 1 have explanatory footnotes with further explanations in the Multimedia Appendix 4, and informative referenced materials given in parentheses. While the ethics domain has its referenced materials in Table S1 in Multimedia Appendix 1. This keeps the table within a readable limit and with adequate explanatory reference information. The explanatory footnotes are taken from the reference materials that are combined for maximal practical and educational clarity.

Common Pitfalls and Prevailing Challenges

Afterthoughts and neglecting the aforementioned crucial considerations for AI tools development would inevitably cause serious problems and costly consequences. These pitfalls include inadequate data quality, biased algorithms, and misaligned AI capabilities that often emerge despite rigorous planning [84]. To re-emphasize the importance of due consideration of ETEPAI at every step throughout the development process to mitigate these risks [84], this section shares some potential pitfalls and appropriate preventive and corrective measures (Table 2) that can lead to models that perform effectively in real-world settings. These notable pitfalls and veracious challenges during AI tools development may vary according to different tools and settings. Further guides on best practices to remedy pitfalls, challenges, errors, and biases are available elsewhere [85,86].

Table 2. Potential pitfalls and possible best practices that could prevent and correct them. The content of this table is summarized from extracted materials in the study by Aliferis et al [86].

Number	Pitfalls	Best practices and measures
1	• Using hard-to-reproduce, nonstandardized data input steps. • Using nonrepresentative datasets or unusual populations and making wrong inferences.	• Use high-quality and representative datasets and appropriate populations, and make claims based on appropriate datasets. • Use dense time series data, leverage population models, and address abrupt distribution shifts in patient models.
2	• Using normalization or data transforms requiring the entire sample, affecting test independence.	• Use normalization or data transforms that do not require the entire sample (or confine such within discovery/training and validation datasets independently).
3	• Not conducting power sample analysis or understanding sample size effects on modeling.	• Perform reanalysis with improved protocols and domain knowledge.

Number	Pitfalls	Best practices and measures
4	- Models are designed with excessive complexity relative to the data and sample size.	- Manage model complexity using regularization, dimensionality reduction, feature selection, Bayesian methods, and model selection. - Explore all relevant learning method families for better predictivity. Conduct power sample analysis and characterize sample size effects on modeling.
5	- Failing to coordinate analysis across teams and datasets using unbiased, collaborative protocols.	- Coordinate analysis across teams using unbiased protocols.
6	- Not correcting for multiple statistical hypothesis tests.	- Correct for multiple statistical hypotheses' tests. - Anticipate and incorporate multiple modeling stages to avoid overfitting.
7	- Using biased estimators or introducing bias into otherwise unbiased ones.	- Systematically explore the hyperparameter space.
8	- Allowing uncontrolled iterative modeling, leading to analysis creep[a].	- Conduct iterative or sequential modeling using unbiased protocols. - Follow theoretical and empirical specifications of reference methods.
9	- Inappropriately modeling individual patients and overinterpreting generalizability.	- Combine individual patient modeling with population modeling where possible. - Use dense time series data, leverage population models, and address abrupt distribution shifts in patient models.
10	- Not examining the stability of models and parameters, nor investigating unstable findings.	- Examine the stability of models and parameters, and investigate unstable findings.
11	- Allowing models to learn incorrect patterns due to spurious co-occurrence or uncontrolled structural relations.	- Prevent models from learning incorrect patterns through proper control and domain knowledge.
12	- Controlling only some factors contributing to overconfidence.	- Control all relevant factors contributing to overconfidence via nested model selection.
13	- Ignoring the statistical uncertainty of strong performance estimates.	- Deploy and explore all relevant data preparation steps for the domain and task.
14	- Insufficient evaluation of scalability and generalizability for bespoke models.	- Use label reshuffling and independent dataset validation cautiously.
15	- Reporting only the strongest models or results.	- Fully report all procedures used to obtain models for independent verification. - Inform analyses with methods literature to explore both best-known and novel methods.

[a] Analysis creep is the gradual introduction of biases or distortions into a model due to uncontrolled or repeated adjustments, often resulting in overfitting or overconfidence in the results.

Common Pitfalls to Caution

In the rapidly evolving field of AI, developers and data scientists face a myriad of challenges that can significantly impact the effectiveness and reliability of their models. While much emphasis is placed on the theoretical foundations and potential applications of AI, it is equally important to be aware of the common pitfalls that can undermine these efforts. This section highlights several notable pitfalls that practitioners must navigate to ensure robust and accurate outcomes, from data preparation to model selection and beyond.

One major challenge in AI development is the narrow selection of models and methods [87]. A common pitfall is failing to explore the appropriate model family during model selection, such as using only linear regression models when the data-generating function is nonlinear and discontinuous [88]. Additionally, data scientists or vendors often have strong preferences for a limited set of methods or technologies, even when they are not the most suitable for the task [89]. This narrow approach can significantly hinder model performance and limit the potential of AI applications.

Another common issue is the tendency to rely on single-stage modeling attempts without refinement. Initial models typically require iterative enhancements, as first attempts rarely meet performance goals in complex problems. A single-stage approach lacks the iterative understanding necessary to optimize model interaction with data [90]. Furthermore, ignoring established best methods or misapplying robust methods outside of their intended use can result in underperforming models [91]. Adhering to proven methods and following established guidelines are critical for maintaining model integrity and achieving reliable outcomes.

Insufficient exploration of hyperparameters and data preparation is also a critical issue. Even when the right model families are chosen, inadequate tuning of hyperparameters can lead to suboptimal outcomes [92]. Similarly, improper data preparation, such as neglecting feature construction, selection, normalization, or discretization, can greatly impair a model's effectiveness [93,94]. Thorough attention to both hyperparameter optimization and data preparation is essential for achieving robust model performance.

Data contamination and bias are additional challenges that can compromise model accuracy. When data processing steps such as normalization or discretization are conducted across both training and test sets, it can introduce bias and lead to inflated performance metrics [95]. This issue is exacerbated when nonrepresentative datasets or unusual populations are used, leading to models that do not generalize well. In some cases, models may produce misleading results by focusing on correlated factors rather than underlying causes, which can result in ineffective or incorrect conclusions.

Overconfidence in models, particularly in the context of overfitting and underfitting, poses a significant risk. Models may appear effective during development but fail in real-world applications due to high-dimensional data, small sample sizes, or overly complex learners [87]. This issue is closely related to the challenge of power-sample calculation in AI, which is more complex than in traditional statistical analysis. Learning curves, which describe the generalization error as a function of sample size, are often unknown, making it difficult to determine the required sample size for reliable model performance.

Addressing common pitfalls is essential for the effective development of AI models, but it is equally important to navigate the more complex challenges that can emerge. By adopting a balanced approach that includes both careful model selection and iterative refinement while also considering algorithmic transparency, regulatory impacts, and complete process reporting, developers can better ensure the reliability and ethical deployment of AI systems. These considerations are crucial for fostering innovation without compromising on responsibility, as AI continues to play an increasingly significant role in various fields.

Veracious Challenges to Circumspect

Navigating the complex landscape of AI development requires a careful balance between transparency, innovation, and responsible implementation. While algorithmic transparency is critical for validating methods, it must be handled with caution to avoid unintended consequences and misuse. Similarly, overengineering and overregulation can stifle innovation, making it essential to strike a balance that promotes both progress and safety. Complete transparency in reporting is also vital, ensuring that all analytical processes are fully disclosed and accessible, enabling thorough evaluations of model performance and potential biases.

Algorithmic transparency is crucial for the validation of methods and tools [96]. However, it also introduces risks, such as the potential for misuse, unintended consequences, or manipulation of models and systems. Black box methods, while generally undesirable when they fail to meet expected safety and performance standards, can be advantageous in scenarios where securing the system against tampering is necessary. Moreover, as argued by some AI literature [97,98], if a well-validated black box model demonstrates a substantial statistical advantage over the best transparent model, it may be both impractical and ethically questionable to disregard its superior performance and use [98]. Navigating these trade-offs requires careful consideration and expertise.

Overengineering and overregulating pose challenges in science and technology [99]. When best practices are enforced in rigid, bureaucratic ways, there is a risk of stifling innovation and slowing progress. Additionally, disguising decision models that guide user actions as mere advisory tools is a persistent issue in health AI [12]. Regulation must address these practices, as they can render regulation ineffective and distort performance and safety requirements during the design of AI systems. Ensuring safety and performance through best practices is essential, but it must be balanced against the risks of delaying the deployment of valuable AI/ML applications in health care and health sciences. Achieving this balance is crucial for fostering both innovation and responsible implementation.

It is essential to practice complete reporting of the full process, especially all the analyses and modeling procedures applied to the data [100,101]. This would enable a thorough evaluation of the robustness and any potential issues of bias. Providing complete access to the algorithms and data, with a detailed disclosure of the entire analytical process, including all model selection and error estimation steps, would definitely help the comprehensive evaluation of model performance and reliability, and its associated aspects, such as overfitting or overconfidence.

Discussion

Principal Findings and Framework Synthesis

This work results in the ETEPAI, which is structured around 4 key stages: design, development, deployment, and postdeployment. It is a wholesome consideration of an AI tool production from designing, development, deployment, and postdeployment stages to be executed in a sound approach and with good adherence to the ethical principles, solid preparation, and execution that meet expected achievable usefulness, information technology feasibility, and crucial technical requirements including deployment strategy, monitoring and evaluation plan, impact assessment, financial projections for sustainability, and take into account what is cohesive to the epidemiologic research methodology of rigorous validation and complete reporting [15]. This is achieved by integrating multiple frameworks and guidelines to ensure a broad and comprehensive approach to AI tool development in health care. This corroborates the declaration of Innsbruck that underscores the evaluation of information systems in health care throughout design and implementation stages [102], and the multistep process design framework of iterative evaluation and system assessment as central to overall development planning [103], and also matches well the key challenges posed by AI in a real-world context of care and services according to the health technology assessment core model [104].

Comparison With Existing Guidelines

Since ETEPAI is essentially a synthesized summary of existing robust recommendations, it is an evidence-based guidance on the tasks it is meant for, which may include using it as an evaluation tool in a systematic review of papers reporting on AI tools development, which would ease the task of using multiple similar checklists [105]. ETEPAI is different from other similar tools, but does not form its referenced base. Multimedia Appendix 2 tabulates the characteristics of these guidelines, checklists, and frameworks. The methodological guidelines tool reviewed the methodologies applied in 134 selected studies and developed its checklist to provide a systematic framework for estimating population-based health indicators using linked data and ML techniques [52], but lacked in certain emphasis, such as on error analysis, guidance on deployment, and postdeployment strategies. Another was based on theoretical knowledge (Clinical Artificial Intelligence Research checklist) [53], one that is discipline-specific that has yet to be developed through the proper processes (Model for Assessing the Value of Artificial Intelligence in Medical Imaging) [54], another translational-to-practice focused evaluation framework (Translational Evaluation of Health Care Artificial Intelligence framework) [55], another that is a publication appraising tool in radiology [56] to complement an existing one (Data Algorithm Training Output method) [57], another for setting specific "verification paradigms" of AI tools in clinical decision-making [58], one a function-specific (algorithm selection) [59], another that is ethics-focused [60], one forcedly matched to the sequential phases of experimental testing and clinical research for drugs and medical devices [61], while another is a reporting guideline developed through the Delphi method of 11 researchers from 3 institutions on 3 different continents, providing suggestions for every section of a journal paper from title to discussion on limitations [62]. There is another reporting guideline specific to radiomic research (Checklist for Evaluation of Radiomics Research) [63] for study planning, paper writing, and evaluation during the review process. There is one radiology AI software specification, classification, and performance evaluation framework till postdeployment, which is comprehensive but presented as fragmented parts and might complicate its application [64]. Another radiology framework (Radiology Artificial Intelligence Deployment and Assessment Rubric) emphasizes proper study designs in the assessment of its 7 preidentified function domains named as hierarchical levels of efficacy [106], and not along a clear development process of AI tools. A recent checklist, OPTICA (Organizational Perspective Checklist for Artificial Intelligence Solutions Adoption), was developed by a single institution primarily to assess completed AI solutions that are to be adopted in health care organizations [65]. It comprises 13 chapters of 3 to 12 items in each chapter to be completed by 5 main stakeholders who are assumed to be educated and able to mark off the checklist items competently and sequentially by a single identified stakeholder. This is different from ETEPAI, where full consideration is to be given by the project team at the beginning, and adhered to or referenced during the process till the end. Although OPTICA does not specifically address ethics, the 77 probing questions are indeed noteworthy to all stakeholders during adoption consideration. This is unlike ETEPAI, which mandates early and continuous adherence to ethical and procedural considerations throughout the entire AI project life cycle in general. TRIPOD-LLM (Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis-Large Language Models) [66] focuses on text-only large language model (LLM) projects and standardized reporting to enhance clarity, transparency, and accountability. While ETEPAI is embedded in project workflows, TRIPOD-LLM aids in appraising study quality and facilitating reproducibility with less emphasis on guiding processes from initiation to completion if used as a post hoc assessment tool. Another, which is named as Awesome AI Guidelines on GitHub, is a rich compilation of many policy documents, guidelines, and checklists related to AI, with the majority of them on ethics principles; only a few are dated beyond the year 2020 [107]. Although every evaluation framework has its respective strengths and should be used where appropriate, ETEPAI is believed to have all the pivotal aspects required of AI tools production, with an emphasis on methodology.

Addressing Implementation and Sustainability Challenges

Recent studies highlighted that current regulatory approvals for AI models often did not ensure fairness, with bias evaluation and mitigation typically occurring post-deployment [108]. Integrating AI models into clinical practice involves significant challenges, including information technology system integration, local fine-tuning, and addressing unintended consequences such as automation bias [108]. Robust human-machine collaboration and continuous monitoring are essential to improve performance and prevent failures caused by changes in software or equipment. Postdeployment maintenance and sustainability of AI models in clinical settings present significant ongoing challenges, including model drift, data drift, concept drift, harmful feedback loops, adversarial attacks, and potential biases that may emerge in real-world applications, retraining schedules, and managing overall AI tools maintenance [109]. Additionally, usability evaluation faces obstacles such as the lack of transparency and explainability in AI systems, often referred to as "black boxes" or "black box inside a black box" nature of commercial AI models [108], which can hinder user trust and satisfaction. To address these issues, it is essential to conduct usability evaluations according to recognized standards, assess user learning curves to inform training programs, evaluate stakeholder and patient acceptability to ensure alignment with user expectations, and redesign AI systems tailored for their intended clinical use, focusing on integration into workflows, reliability, and regulatory compliance [55].

Strategic Value and Generalizability

The ETEPAI framework is structured to proactively address these multifaceted challenges. By mandating that critical issues such as fairness, bias, and workflow integration be considered from the initial design stage, it shifts mitigation from a reactive, postdeployment afterthought to a foundational project requirement, including agentic AI uses [110]. Its dedicated postdeployment stage provides a clear roadmap for the continuous monitoring and usability evaluations essential for long-term sustainability and trust. This comprehensive, life cycle–based approach distinguishes ETEPAI from other valuable but more specialized guidelines. While built upon excellent existing frameworks such as TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence, a reporting guideline), FUTURE-AI (Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence, imaging-focused), and ALTAI (Assessment List for Trustworthy Artificial Intelligence, an ethics checklist), ETEPAI's novelty lies in its integration of ethical, technical, and epidemiological considerations into a single, cohesive structure. Consequently, this end-to-end structure establishes its practicality as an actionable guide for multidisciplinary teams from project inception, rather than as a post hoc reporting or evaluation checklist.

Consequently, ETEPAI is designed to be model-agnostic. Whether the underlying technology is a traditional logistic regression, a convolutional neural network, or an LLM, the necessity for ethical oversight, rigorous epidemiological study design (such as external validation), and postdeployment monitoring remains constant. While the specific performance metrics may differ between model types, such as calibration plots for regression vs red-teaming for generative artificial intelligence (GenAI), the governance checkpoints provided by ETEPAI are universally applicable steps in the translational pathway.

Application Nuances

Notwithstanding the comprehensive and process-centric evaluation of ETEPAI, the degree of relevance and importance of ETEPAI's items within the 4 stages of design, development, deployment, and postdeployment, and the 3 domains of ethics, technical, and epidemiological principles may differ in different AI projects of tools or systems. The different objectives, settings, and complexities of solutions of each AI tool would demand different attention on different items in ETEPAI. Nevertheless, every item requires consideration and must be justified if bypassed. When any of the items, stages, or domains raises concern, for proper understanding and evaluation, we are to refer to the original guideline or checklist (as given in the parentheses in the footnote to Table 1).

There are a few potential drawbacks of ETEPAI arising from the recommendations it is built on and the evolving techniques and technology of AI. Some of these include a lack of detail on the class imbalance mitigation strategy, approaches to address fairness, heterogeneity in estimates of model parameter values, and model performance. Additionally, further guidance will be needed on specific AI areas such as AI tools to influence social and health behaviors at scale in social media and different platforms, artificial general intelligence in health care, and other guides that were developed for specific purposes, as alluded to in this discussion, such as the OPTICA tool in assessing AI solutions in health care settings [65]. This is because ETEPAI is mainly for sign-posting the important considerations as necessary indicators and not for instructing the proper means of achieving them, which are dependent on the dataset's quality and distribution. However, ETEPAI reminds developers and stakeholders to provide a clear explanation and reporting of all decisions taken during the whole process, from development to postdeployment. Guidance on advanced evaluation methods, such as handling Pareto frontiers for multiobjective optimization (balancing fairness and accuracy) [111,112], would require more elaborative and extensive explanation beyond the foundational structure ETEPAI aims to provide.

Limitations

While comprehensive, the ETEPAI may not fully address all aspects of certain types of AI models, especially newer or more complex models such as GenAI or agentic AI where multiple AI tools are developed and deployed to complement collaboratively [113], or in other sectors than health care such as education [114,115] and research enterprise [116,117], which are evolving and with robust evaluation frameworks that are yet available [118]. The POLARIS-GM (Partnership for Oversight, Leadership, and Accountability in Regulating Intelligent Systems–Generative Models in Medicine) [118] is an ongoing initiative focused on developing granular, scenario-based regulatory guidance for the unique challenges of GenAI and LLMs in medicine, using a multiphase, consensus-driven approach to generate new recommendations for postimplementation controls. In contrast, ETEPAI serves as a broad, comprehensive guide and evaluation tool for general AI tool production, synthesizing existing recommendations into practical pointers. We acknowledge that certain technical specifics of traditional ML do not map directly to GenAI, such as the standard calibration metrics, discrimination measures such as receiver operating characteristic area under curve, or the evaluation of feature stability, which are often replaced in GenAI by semantic similarity scores and human-preference alignment, and shift toward detecting toxicity and protecting against jailbreaking attempts, respectively. However, ETEPAI's core principles remain applicable if adapted appropriately. These and the stages of considerations are still relevant to GenAI and agentic AI across their life cycle [119], from defining a safe operational domain in the design stage, and conducting adversarial "red-teaming" and error analysis to evaluate hallucination rates and consistency in the development stage, to mitigating automation bias through stakeholder education during the deployment stage and continuously monitoring for emergent harms in the postdeployment stage. While ETEPAI provides the overarching governance structure, developers must supplement it with specific, evolving guidelines for generative models, which incorporate initiatives such as the TRIPOD-LLM and the POLARIS-GM, to address their unique technical and ethical challenges.

When the referenced materials are used, the interpretation of criteria for items relies on expert ratings, which can introduce subjectivity and potential bias, leading to inconsistencies in the evaluation of AI tools across different studies and users. As this study is primarily on assimilating existing recommended materials to produce one overarching guidance and not on evaluating the quality of the referenced materials, there is neither analysis of their processes and products, nor were comparisons made among them. However, ETEPAI uses a consensus-driven approach by using the collective judgment of multiple experts, which balances individual biases and ensures a more reliable and balanced assessment. Additionally, it emphasizes transparency by documenting the criteria and rationale behind the ratings. This openness allows for thorough scrutiny and review, helping to identify and address potential biases or inconsistencies in the evaluation process.

All care was taken to screen through eligible guides and checklists at the time of publication of ETEPAI. There are many similarities among the included and excluded guides, but only those that are considered to be robust in development and focus on the process from design to postdeployment formed the basis of ETEPAI. However, the content of ETEPAI is noted to be inclusive of almost all important considerations in the excluded guides, as discussed. The similarity of ETEPAI to many guidelines and checklists is intended as per the method used to create it. As ETEPAI draws its validity from 30 robust, consensus-derived guidelines it synthesizes, this provides a strong evidence base for ETEPAI's applicability across diverse AI models. Although the acceptability, applicability, and effectiveness of ETEPAI in real-world uses have not been tested, it is expected to be similar to those that were robustly tested. Although ETEPAI's content validity is established through its systematic synthesis of existing, evidence-based, and widely accepted guidelines, making it conceptually sound upon proposal, it would require subsequent empirical work to confirm them. This may include methods such as (1) expert consensus: conducting a formal Delphi study with international experts in clinical AI, ethics, and epidemiology to refine and weigh the framework's items; (2) pilot testing: implementing the ETEPAI framework within one or more real-world AI development projects to gather qualitative feedback on its practicality and utility; and (3) retrospective application: using ETEPAI as an evaluation tool for a cohort of published AI studies to assess its applicability and identify areas for improvement. While ETEPAI as a unified entity awaits prospective empirical testing, its component "data points," the individual requirements for fairness, robustness, and reporting, are already established standards in the scientific community.

Despite its comprehensive nature, the ETEPAI may still be challenging for some stakeholders to fully grasp without sufficient training or motivation. This is evidenced by the tool's expansive supplementary materials for full explanations, which may limit its immediate usability, as users need to refer to additional resources for a lucid understanding. Many pointers in the footnotes are cross-disciplinary and could be beyond any project team member. For example, the emphasis on diversity in the datasets and caution on using existing datasets that might limit the representativeness or comprehensiveness necessary for some AI tool development scenarios are to be appreciated by all stakeholders and upheld by methodologists or epidemiologists to ensure compliance throughout the whole production process. Although the framework highlights the need for rigorous external validation and adherence to reporting guidelines, these processes can be resource-intensive and may not be feasible for all projects. Furthermore, the rapidly evolving nature of AI technology, AI models updating methods [120], and health care practices may necessitate frequent updates to the ETEPAI framework to maintain its relevance and effectiveness. Public scrutiny and transparency are important. Publishing project protocols or making them available on public platforms can help ensure that the best practices recommended by ETEPAI are followed. It also makes it easier to explain and document any changes or deviations from the proposed design.

Conclusions

The ETEPAI is not just another guideline, but it combines existing robustly developed policies, guidelines, and checklists in one place that map out the ideal route, navigational aids, and risky areas throughout the whole AI tools development process. By focusing on its core function as a guide, this synthesized guideline encourages thoughtful application, ensuring its adaptability to diverse settings while avoiding misuse as a rigid standard, evaluation benchmark, or risk assessment framework. The guideline provides flexible principles and recommendations that must be adapted to specific contexts rather than being followed rigidly or universally applied. While it promotes quality, it does not include structured criteria or metrics for evaluating performance or outcomes. The guideline does not focus on identifying, quantifying, or mitigating risks but may complement such tools when integrated appropriately. Therefore, ETEPAI should be more useful as a training reference manual than a quality standard reference manual. This is because to appreciate and to put it to immediate practice requires previous training on AI concepts and technologies. It is possible to use ETEPAI as a template for training purposes to impart AI-related clinical competencies in health care professionals [121]. It is both comprehensive and complete, containing almost all critical considerations, as well as briefly and crisply presenting those considerations in a table of probing questions and another table of critical pointers in 3 domains of ethics, technical, and epidemiological principles when considering an AI tool.

Applying ETEPAI ensures comprehensive guidance and encourages a multidisciplinary approach to AI tool development, from design through postdeployment. This framework promotes adherence to ethical standards, robust technical execution, and rigorous epidemiological research methodologies. By integrating these considerations, AI tools can achieve their intended impact in health care settings, offering reliable, effective, and ethically sound solutions. To ensure effective application of ETEPAI and meticulous attention to the aspects throughout the AI development life cycle, it is essential for users to understand each and every included guideline, checklist, assessment, and framework, and be familiar with the ethical and regulatory concepts. We highly advocate its use to ensure that developers and authors appropriately design and comprehensively report AI work leading to high-quality, practical AI applications that align with clinical needs and ethical imperatives. ETEPAI's product-centric approach may effectively align with EU trustworthiness benchmarks on safety, robustness, and ethics from design to deployment [13]. This helps the final AI products to be compliant with the EU's requirements for the protections of health, safety, fundamental rights, and comprehensive risk management that considers impacts on both products and systems [75]. Using ETEPAI should be easy and sufficient for experienced professionals and serve as clear signposts for less prepared stakeholders. As the field of AI continues to evolve with new technologies and algorithms, we anticipate ongoing updates to ETEPAI as new guidelines and reporting standards emerge to ensure high-quality scientific research and the ethical application of AI in medicine.

Acknowledgments

BHC is grateful to Dr. Judice and Dr. Mohammad Shaheryar Furqan for their hospitality when hosting him at the Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore (NUS), where this work was completed. BHC would like to express his appreciation to the Foundation for Advancing Family Medicine (FAFM) for the grant of the Besrour Centre Family Medicine Early Career Researcher Award 2022 that supported his attendance at the online courses of the European Information Technologies Certification Academy/Artificial Intelligence Academy. BHC greatly appreciates his colleagues in the Department of Family Medicine, UPM, who were kind enough to spare him and to stand in for him to make his sabbatical and research leaves possible. During the preparation of this work, the authors used ChatGPT 3.5, 4o, and o1 (OpenAI) in order to assist in the drafting and language editing of portions of this paper. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

BHC was supported and sponsored financially by Universiti Putra Malaysia (UPM) to attend some of the courses and for his sabbatical leave to National University of Singapore (NUS) for a period of 9 months, and using the department's trust account for the publication of this paper, with a few other works that were also produced besides this one.

Authors' Contributions

BHC contributed to the conceptualization, data curation, formal analysis, methodology, project administration, validation, visualization, and writing of the original draft. KYN also contributed to resources, supervision, and review and editing of the writing. All the authors read and approved the final version of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Table S1: recommended guidelines, checklists, assessment frameworks, and recommendations, and Table S2: AI ethics, safety, and dataset diversity policy frameworks (descending orders of relevancy and recentness). AI: artificial intelligence.
[PDF File (Adobe File), 422 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Table S3: characteristics of guidelines, checklists, and frameworks related to AI tools (according to the alphabetical and ascending year of the publication). AI: artificial intelligence.
[PDF File (Adobe File), 222 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Outline of an AI research proposal. AI: artificial intelligence.
[PDF File (Adobe File), 142 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Further footnotes to Table 2.
[PDF File (Adobe File), 224 KB-Multimedia Appendix 4]

Checklist 1

PRISMA-ScR checklist.
[PDF File (Adobe File), 180 KB-Checklist 1]

References

1. van Royen FS, Asselbergs FW, Alfonso F, Vardas P, van Smeden M. Five critical quality criteria for artificial intelligence-based prediction models. Eur Heart J. Dec 7, 2023;44(46):4831-4834. [doi: 10.1093/eurheartj/ehad727] [Medline: 37897346]
2. Callahan A, McElfresh D, Banda JM, et al. Standing on FURM ground: a framework for evaluating fair, useful, and reliable AI models in health care systems. NEJM Catalyst. Sep 18, 2024;5(10). [doi: 10.1056/CAT.24.0131]
3. International scientific report on the safety of advanced AI: interim report. GOV.UK. May 2024. URL: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai [Accessed 2026-01-22]
4. Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. Clin Exp Ophthalmol. Jul 2021;49(5):470-476. [doi: 10.1111/ceo.13943] [Medline: 33956386]
5. Klontzas ME, Gatti AA, Tejani AS, Kahn CE. AI reporting guidelines: how to select the best one for your research. Radiol Artif Intell. May 2023;5(3):e230055. [doi: 10.1148/ryai.230055] [Medline: 37293341]
6. Shiferaw KB, Roloff M, Balaur I, Welter D, Waltemath D, Zeleke AA. Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review. Health Inf. Preprint posted online on May 28, 2024. [doi: 10.1101/2024.05.27.24307991]
7. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. Oct 2019;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
8. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. Mar 25, 2020;368:m689. [doi: 10.1136/bmj.m689] [Medline: 32213531]
9. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. Lancet. Jan 2014;383(9912):156-165. [doi: 10.1016/S0140-6736(13)62229-1]
10. Macleod MR, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. Lancet. Jan 2014;383(9912):101-104. [doi: 10.1016/S0140-6736(13)62329-6]
11. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. Jan 2014;383(9912):166-175. [doi: 10.1016/S0140-6736(13)62227-8]
12. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021. URL: https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf?sequence=1 [Accessed 2026-01-22]
13. Ethics guidelines for trustworthy AI. European Commission. 2019. URL: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html [Accessed 2026-01-22]
14. Ethical impact assessment: a tool of the recommendation on the ethics of artificial intelligence. UNESCO; 2023. [doi: 10.54678/YTSA7796]
15. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (Minimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. Dec 9, 2020;27(12):2011-2015. [doi: 10.1093/jamia/ocaa088] [Medline: 32594179]
16. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
17. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw Open. Sep 5, 2023;6(9):e2335377. [doi: 10.1001/jamanetworkopen.2023.35377] [Medline: 37747733]
18. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. Sep 2020;26(9):1320-1324. [doi: 10.1038/s41591-020-1041-y] [Medline: 32908275]
19. Kotecha D, Asselbergs FW, Achenbach S, et al. CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research. Lancet Digit Health. Oct 2022;4(10):e757-e764. [doi: 10.1016/S2589-7500(22)00151-0] [Medline: 36050271]
20. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. May 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9]
21. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. Sep 2020;26(9):1351-1363. [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
22. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. Oct 28, 2015;351:h5527. [doi: 10.1136/bmj.h5527] [Medline: 26511519]

23. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ. Sep 9, 2020;370:m3164. [doi: 10.1136/bmj.m3164] [Medline: 32909959]
24. van Smeden M, Moons KGM, Hooft L, Chavannes NH, van Os HJA, Kant I. Guideline for high-quality diagnostic and prognostic applications of AI in healthcare. OSF; Oct 19, 2023. URL: https://doi.org/10.17605/OSF.IO/TNRJZ [Accessed 2026-01-22]
25. Kiyasseh D, Cohen A, Jiang C, Altieri N. A framework for evaluating clinical artificial intelligence systems without ground-truth annotations. Nat Commun. Feb 28, 2024;15(1):1808. [doi: 10.1038/s41467-024-46000-9] [Medline: 38418453]
26. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health. May 2022;4(5):e384-e397. [doi: 10.1016/S2589-7500(22)00003-6] [Medline: 35396183]
27. The STANDING Together collaboration. Recommendations for diversity, inclusivity, and generalisability in artificial intelligence health technologies and health datasets. Zenodo; Oct 30, 2023. [Accessed 2026-01-22] [doi: 10.5281/ZENODO.10048356]
28. Elvidge J, Hawksworth C, Avşar TS, et al. Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI). Value Health. Sep 2024;27(9):1196-1205. [doi: 10.1016/j.jval.2024.05.006] [Medline: 38795956]
29. Artificial Intelligence Act. The European Parliament; Mar 13, 2024. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf [Accessed 2026-01-22]
30. Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Front Digit Health. 2024;6:1267290. [doi: 10.3389/fdgth.2024.1267290] [Medline: 38455991]
31. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. United Nations; Mar 11, 2024. URL: https://digitallibrary.un.org/record/4043244?ln=en&v=pdf [Accessed 2026-01-28]
32. Lekadir K, Osuala R, Gallin C, et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv. Preprint posted online on Jul 22, 2024. [doi: 10.48550/ARXIV.2109.09658]
33. Recommendation on the ethics of artificial intelligence. UNESCO; Nov 23, 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en [Accessed 2026-01-22]
34. Hawksworth C, Elvidge J, Knies S, et al. Protocol for the development of an artificial intelligence extension to the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022. Health Econ. Preprint posted online on Jun 1, 2023. [doi: 10.1101/2023.05.31.23290788]
35. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. Jan 1, 2019;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
36. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. Jul 9, 2021;11(7):e048008. [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
37. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. Jan 7, 2015;350(jan07 4):g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]
38. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. Mar 24, 2025;388:e082505. [doi: 10.1136/bmj-2024-082505] [Medline: 40127903]
39. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open. Jun 28, 2021;11(6):e047709. [doi: 10.1136/bmjopen-2020-047709] [Medline: 34183345]
40. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. Oct 2025;31(10):3283-3289. [doi: 10.1038/s41591-025-03953-8] [Medline: 40954311]
41. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and2109.09658 reviewers. Radiol Artif Intell. Mar 2020;2(2):e200029. [doi: 10.1148/ryai.2020200029] [Medline: 33937821]
42. Bilbro NA, Hirst A, Paez A, et al. The IDEAL Reporting Guidelines. Ann Surg. 2021;273(1):82-85. [doi: 10.1097/SLA.0000000000004180]

43. McCulloch P, Altman DG, Campbell WB, et al. No surgical innovation without evaluation: the IDEAL recommendations. Lancet. Sep 26, 2009;374(9695):1105-1112. [doi: 10.1016/S0140-6736(09)61116-8] [Medline: 19782876]
44. Marcus HJ, Bennett A, Chari A, et al. IDEAL-D Framework for Device Innovation. Ann Surg. 2022;275(1):73-79. [doi: 10.1097/SLA.0000000000004907]
45. Standing on FURM ground -- a framework for evaluating fair, useful, and reliable AI models in healthcare systems. arXiv. Preprint posted online on Mar 14, 2024. [doi: 10.48550/arXiv.2403.07911]
46. Readiness assessment methodology. A tool of the recommendation on the ethics of artificial intelligence. UNESCO; 2023. [doi: 10.54678/YHAA4429]
47. Voluntary code of conduct on the responsible development and management of advanced generative AI systems. Government of Canada. Sep 2023. URL: https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems [Accessed 2026-01-22]
48. A guide to good practice for digital and data-driven health technologies. GOV.UK. Jan 19, 2021. URL: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology [Accessed 2026-01-22]
49. Safe, secure, and trustworthy development and use of artificial intelligence: Executive Order 14110 of October 30, 2023. Federal Register. Oct 30, 2023. URL: https://www.federalregister.gov/d/2023-24283 [Accessed 2026-01-22]
50. OECD AI principles overview. OECDAI. URL: https://oecd.ai/en/ai-principles [Accessed 2026-01-22]
51. Universal guidelines for AI. Center for AI and Digital Policy. URL: https://www.caidp.org/universal-guidelines-for-ai/ [Accessed 2026-01-22]
52. Haneef R, Tijhuis M, Thiébaut R, et al. Methodological guidelines to estimate population-based health indicators using linked data and/or machine learning techniques. Arch Public Health. Jan 4, 2022;80(1):9. [doi: 10.1186/s13690-021-00770-6] [Medline: 34983651]
53. Olczak J, Pavlopoulos J, Prijs J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop. Oct 2021;92(5):513-525. [doi: 10.1080/17453674.2021.1918389] [Medline: 33988081]
54. Fasterholdt I, Kjølhede T, Naghavi-Behzad M, et al. Model for Assessing the Value of Artificial Intelligence in Medical Imaging (MAS-AI). Int J Technol Assess Health Care. Oct 3, 2022;38(1):e74. [doi: 10.1017/S0266462322000551] [Medline: 36189821]
55. Reddy S, Rogers W, Makinen VP, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform. Oct 2021;28(1):e100444. [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]
56. Park SH, Sul AR, Ko Y, Jang HY, Lee JG. Radiologist’s guide to evaluating publications of clinical research on AI: how we do it. Radiology. Sep 1, 2023;308(3):e230288. [doi: 10.1148/radiol.230288]
57. Kelly BS, Judge C, Hoare S, Colleran G, Lawlor A, Killeen RP. How to apply Evidence-Based Practice to the Use of Artificial Intelligence in Radiology (EBRAI) using the Data Algorithm Training Output (DATO) method. Br J Radiol. Oct 2023;96(1150):20220215. [doi: 10.1259/bjr.20220215] [Medline: 37086062]
58. Bragazzi NL, Garbarino S. Toward clinical generative AI: conceptual framework. JMIR AI. Jun 7, 2024;3:e55957. [doi: 10.2196/55957] [Medline: 38875592]
59. Forghani R. A practical guide for AI algorithm selection for the radiology department. Semin Roentgenol. Apr 2023;58(2):208-213. [doi: 10.1053/j.ro.2023.02.006]
60. Chiang S, Picard RW, Chiong W, et al. Guidelines for conducting ethical artificial intelligence research in neurology: a systematic approach for clinicians and researchers. Neurology (ECronicon). Sep 28, 2021;97(13):632-640. [doi: 10.1212/WNL.0000000000012570]
61. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open. Oct 1, 2020;3(3):326-331. [doi: 10.1093/jamiaopen/ooaa033]
62. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. Dec 16, 2016;18(12):e323. [doi: 10.2196/jmir.5870]
63. Kocak B, Baessler B, Bakas S, et al. Checklist for Evaluation of Radiomics Research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging. May 4, 2023;14(1):75. [doi: 10.1186/s13244-023-01415-8] [Medline: 37142815]
64. Tanguay W, Acar P, Fine B, et al. Assessment of radiology artificial intelligence software: a validation and evaluation framework. Can Assoc Radiol J. May 2023;74(2):326-333. [doi: 10.1177/08465371221135760]
65. Dagan N, Devons-Sberro S, Paz Z, et al. Evaluation of AI solutions in health care organizations — the OPTICA tool. NEJM AI. Aug 22, 2024;1(9):AIcs2300269. [doi: 10.1056/AIcs2300269]

66. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
67. Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls. Springer International Publishing; 2024. [doi: 10.1007/978-3-031-39355-6] ISBN: 978-3-031-39354-9
68. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. Nat Med. Jun 2020;26(6):807-808. [doi: 10.1038/s41591-020-0941-1]
69. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. Jan 2021;14(1):49-58. [doi: 10.1093/ckj/sfaa188] [Medline: 33564405]
70. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. Feb 29, 2000;19(4):453-473. [doi: 10.1002/(sici)1097-0258(20000229)19:4<453::aid-sim350>3.0.co;2-5] [Medline: 10694730]
71. The EQUATOR Network. URL: https://www.equator-network.org [Accessed 2026-01-22]
72. Novak LL, Russell RG, Garvey K, et al. Clinical use of artificial intelligence requires AI-capable organizations. JAMIA Open. Jul 2023;6(2):ooad028. [doi: 10.1093/jamiaopen/ooad028] [Medline: 37152469]
73. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. Oct 2010;19 Suppl 3(Suppl 3):i68-74. [doi: 10.1136/qshc.2010.042085] [Medline: 20959322]
74. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res. Nov 1, 2017;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
75. Josep SG, Sarah DN, Elias B, et al. Harmonised standards for the European AI Act. European Commission; Oct 24, 2024. URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC139430 [Accessed 2026-01-22]
76. Edemekong PF, Annamaraju P, Afzal M, Haydel MJ. Health Insurance Portability and Accountability Act (HIPAA) Compliance. StatPearls Treasure Island (FL): StatPearls Publishing; 2025. [Medline: 29763195]
77. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Syst Rev. Jun 2022;18(2):e1230. [doi: 10.1002/cl2.1230] [Medline: 36911350]
78. Morse KE, Bagley SC, Shah NH. Estimate the hidden deployment cost of predictive models to improve patient care. Nat Med. Jan 2020;26(1):18-19. [doi: 10.1038/s41591-019-0651-8] [Medline: 31932778]
79. Teo ZL, Thirunavukarasu AJ, Elangovan K, et al. Generative artificial intelligence in medicine. Nat Med. Oct 2025;31(10):3270-3282. [doi: 10.1038/s41591-025-03983-2] [Medline: 41053447]
80. Chew BH. Planning and conducting clinical research: the whole process. Cureus. Feb 20, 2019;11(2):e4112. [doi: 10.7759/cureus.4112] [Medline: 31058006]
81. Grobbee DE, Hoes AW. Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. Jones & Bartlett Publishers; 2014. ISBN: 978-1-4496-7433-5
82. Pinnock H, Barwick M, Carpenter CR, et al. Standards for Reporting Implementation studies (StaRI) statement. BMJ. Mar 6, 2017;356:i6795. [doi: 10.1136/bmj.i6795] [Medline: 28264797]
83. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated consolidated framework for implementation research based on user feedback. Implement Sci. Oct 29, 2022;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
84. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull. Sep 10, 2021;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
85. Aliferis C, Ma S, Wang J, Simon G. Characterizing, diagnosing and managing the risk of error of ML & AI models in clinical and organizational application. In: Simon GJ, Aliferis C, editors. Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Springer International Publishing; 2024:607-622. [doi: 10.1007/978-3-031-39355-6_13] ISBN: 978-3-031-39354-9
86. Aliferis C, Simon G. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. In: Simon GJ, Aliferis C, editors. Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Springer International Publishing; 2024:477-524. [doi: 10.1007/978-3-031-39355-6_10] ISBN: 978-3-031-39354-9
87. Khoei TT, Slimane HO, Kaabouch N. Deep learning: systematic review, models, challenges, and research directions. Neural Comput Appl. Nov 2023;35(31):23103-23124. [doi: 10.1007/s00521-023-08957-4]
88. Chen D, Hu F, Nian G, Yang T. Deep residual learning for nonlinear regression. Entropy (Basel). Feb 7, 2020;22(2):193. [doi: 10.3390/e22020193] [Medline: 33285968]

89. Borrego-Díaz J, Galán-Páez J. Explainable artificial intelligence in data science: from foundational issues towards socio-technical considerations. Minds Mach. Sep 2022;32(3):485-531. [doi: 10.1007/s11023-022-09603-z]
90. Wynn DC, Eckert CM. Perspectives on iteration in design and development. Res Eng Design. Apr 2017;28(2):153-184. [doi: 10.1007/s00163-016-0226-3]
91. Guerraoui R, Gupta N, Pinot R. Robust Machine Learning: Distributed Methods for Safe AI. Springer Nature Singapore; 2024. [doi: 10.1007/978-981-97-0688-4] ISBN: 978-981-97-0687-7
92. Arnold C, Biedebach L, Küpfer A, Neunhoeffer M. The role of hyperparameters in machine learning models and how to tune them. PSRM. Oct 2024;12(4):841-848. [doi: 10.1017/psrm.2023.61]
93. Fernandes AAA, Koehler M, Konstantinou N, Pankin P, Paton NW, Sakellariou R. Data preparation: a technological perspective and review. SN Comput Sci. Jun 2, 2023;4(4):425. [doi: 10.1007/s42979-023-01828-8]
94. Shantal M, Othman Z, Bakar AA. A novel approach for data feature weighting using correlation coefficients and min–max normalization. Symmetry (Basel). 2023;15(12):2185. [doi: 10.3390/sym15122185]
95. Cabello-Solorzano K, Ortigosa De Araujo I, Peña M. The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis. In: 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023). Springer Nature Switzerland; 2023:344-353. [doi: 10.1007/978-3-031-42536-3_33] ISBN: 978-3-031-42535-6
96. Barbierato E, Gatti A. The challenges of machine learning: a critical review. Electronics (Basel). Jan 19, 2024;13(2):416. [doi: 10.3390/electronics13020416]
97. Holm EA. In defense of the black box. Science. Apr 5, 2019;364(6435):26-27. [doi: 10.1126/science.aax0162]
98. Miller K. Should AI models be explainable? that depends. Stanford University. Mar 16, 2021. URL: https://hai.stanford.edu/news/should-ai-models-be-explainable-depends [Accessed 2024-08-17]
99. Wheeler T. The three challenges of AI regulation. The Brookings Institution. Jun 15, 2023. URL: https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation/ [Accessed 2026-01-22]
100. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform. Aug 2021;28(1):e100385. [doi: 10.1136/bmjhci-2021-100385] [Medline: 34426417]
101. Brown OM, Curtis AB, Goodwin JA. Principles for evaluation of AI/ML model performance and robustness, revision 1. Massachusetts Institute of Technology; Jan 21, 2021. URL: https://www.ll.mit.edu/sites/default/files/publication/doc/principles-evaluation-aiml-model-performance-brown-md-62.pdf [Accessed 2026-01-22]
102. Talmon JL, Ammenwerth E. "The declaration of Innsbruck": some reflections. Stud Health Technol Inform. 2004;110(68–74):68-74. [Medline: 15853254]
103. Shortliffe EH. Role of evaluation throughout the life cycle of biomedical and health AI applications. BMJ Health Care Inform. Dec 11, 2023;30(1):e100925. [doi: 10.1136/bmjhci-2023-100925] [Medline: 38081766]
104. Alami H, Lehoux P, Auclair Y, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res. Jul 7, 2020;22(7):e17707. [doi: 10.2196/17707] [Medline: 32406850]
105. Di Bidino R, Piaggio D, Andellini M, et al. Scoping meta-review of methods used to assess artificial intelligence-based medical devices for heart failure. Bioengineering (Basel). Sep 22, 2023;10(10):1109. [doi: 10.3390/bioengineering10101109] [Medline: 37892839]
106. Boverhof BJ, Redekop WK, Bos D, et al. Radiology AI Deployment and Assessment Rubric (RADAR) to bring value-based AI into radiological practice. Insights Imaging. Feb 5, 2024;15(1):34. [doi: 10.1186/s13244-023-01599-z] [Medline: 38315288]
107. Saucedo A. Awesome artificial intelligence guidelines. GitHub. URL: https://github.com/EthicalML/awesome-artificial-intelligence-guidelines.git [Accessed 2026-01-22]
108. Gichoya JW, Thomas K, Celi LA, et al. AI pitfalls and what not to do: mitigating bias in AI. Br J Radiol. Oct 2023;96(1150):20230023. [doi: 10.1259/bjr.20230023] [Medline: 37698583]
109. Bhargava S, Singhal S. Challenges, solutions, and best practices in post deployment monitoring of machine learning models. IJCTT. 2024;72(11):63-71. [doi: 10.14445/22312803/IJCTT-V72I11P107]
110. Yee L, Chui M, Roberts R. One year of agentic AI: six lessons from the people doing the work. QuantumBlack AI McKinsey. Sep 12, 2025. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/one-year-of-agentic-ai-six-lessons-from-the-people-doing-the-work [Accessed 2026-01-22]
111. Nagpal R, Khan A, Borkar M, Gupta A. A multi-objective framework for balancing fairness and accuracy in debiasing machine learning models. MAKE. Sep 20, 2024;6(3):2130-2148. [doi: 10.3390/make6030105]
112. Li M, Chen T, Yao X. How to evaluate solutions in pareto-based search-based software engineering: a critical review and methodological guidance. IIEEE Trans Software Eng. May 1, 2022;48(5):1771-1799. [doi: 10.1109/TSE.2020.3036108]
113. Sapkota R, Roumeliotis KI, Karkee M. AI agents vs. agentic AI: a conceptual taxonomy, applications and challenges. SuperIntelligence - Rob - Saf Alignment. 2025;2(3). [doi: 10.70777/si.v2i3.15161]

114. Tolsgaard MG, Pusic MV, Sebok-Syer SS, et al. The fundamentals of artificial intelligence in medical education research: AMEE guide No. 156. Med Teach. Jun 2023;45(6):565-573. [doi: 10.1080/0142159X.2023.2180340] [Medline: 36862064]
115. Chaudhry MA, Kazim E. Artificial Intelligence in Education (AIEd): a high-level academic and industry note 2021. AI Ethics. 2022;2(1):157-165. [doi: 10.1007/s43681-021-00074-z] [Medline: 34790953]
116. Chawla DS. Should AI have a role in assessing research quality? Nat New Biol. Oct 14, 2022. [doi: 10.1038/d41586-022-03294-3]
117. Unlu O, Shin J, Mailly CJ, et al. Retrieval-augmented generation–enabled GPT-4 for clinical trial screening. NEJM AI. Jun 27, 2024;1(7). [doi: 10.1056/AIoa2400181]
118. Ong JCL, Ning Y, Collins GS, et al. International partnership for governing generative artificial intelligence models in medicine. Nat Med. Sep 2025;31(9):2836-2839. [doi: 10.1038/s41591-025-03787-4]
119. Wang Z, Wang H, Danek B, et al. A perspective for adapting generalist AI to specialized medical AI applications and their challenges. npj Digit Med. Jul 11, 2025;8(1):429. [doi: 10.1038/s41746-025-01789-7]
120. Meijerink LM, Dunias ZS, Leeuwenberg AM, et al. Updating methods for artificial intelligence–based clinical prediction models: a scoping review. J Clin Epidemiol. Feb 2025;178:111636. [doi: 10.1016/j.jclinepi.2024.111636]
121. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence–based tools by health care professionals. Acad Med. Mar 1, 2023;98(3):348-356. [doi: 10.1097/ACM.0000000000004963]

Abbreviations

AI: artificial intelligence
AIPA: Artificial Intelligence Prediction Algorithm
ALTAI: Assessment List for Trustworthy Artificial Intelligence
ETEPAI: ethical, technical, and epidemiological considerations for all involved in artificial intelligence tool production
EU: European Union
FURM: Fair, Useful, the Reliable Artificial Inelligence Model
FUTURE-AI: Fairness, Universality, Traceability, Usability, Robustness, and Explainability - Artificial Intelligence
GenAI: generative artificial intelligence
HIPAA : Health Insurance Portability and Accountability Act
LLM: large language model
ML: machine learning
OPTICA : Organizational Perspective Checklist for Artificial Intelligence Solutions Adoption
PE/ICOTS: Population or Participants, Exposure/Intervention, Comparator, Outcomes, Timing, and Setting
POLARIS-GM: Partnership for Oversight, Leadership, and Accountability in Regulating Intelligent Systems–Generative Models in Medicine
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
STANDING: Standards for Data Diversity, Inclusivity, and Generalisability
SUDO: pseudo-label discrepancy
TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Artificial Intelligence
TRIPOD-LLM: Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis - Large Language Models

Edited by Bradley Malin; peer-reviewed by Dip Bharatbhai Patel, Jitender Jain, Manan Wason, Nicolas Bievre, Priyanshu Sinha, Saritha Kondapally; submitted 09.Jul.2025; final revised version received 01.Jan.2026; accepted 05.Jan.2026; published 05.Mar.2026

Please cite as:
Chew BH, Ngiam KY
Important Ethical, Technical, and Epidemiological Considerations in an AI Tool Production (ETEPAI): Scoping Review
JMIR AI 2026;5:e80340
URL: https://ai.jmir.org/2026/1/e80340
doi: 10.2196/80340