Delft University of Technology

Developing, implementing and governing artificial intelligence in medicine
a step-by-step approach to prevent an artificial intelligence winter

Van De Sande, Davy; Van Genderen, Michel E.; Smit, Jim M.; Huiskens, Joost; Visser, Jacob J.; Veen, Robert E.R.; Van Unen, Edwin; Ba, Oliver Hilgers; Gommers, Diederik; Bommel, Jasper van

DOI
10.1136/bmjhci-2021-100495

Publication date
2022

Document Version
Final published version

Published in
BMJ Health and Care Informatics

Citation (APA)
Van De Sande, D., Van Genderen, M. E., Smit, J. M., Huiskens, J., Visser, J. J., Veen, R. E. R., Van Unen, E., Ba, O. H., Gommers, D., & Bommel, J. V. (2022). Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health and Care Informatics*, *29*(1), Article e100495. https://doi.org/10.1136/bmjhci-2021-100495

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

BMJ Health & Care Informatics

Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter

Davy van de Sande ,[1] Michel E Van Genderen ,[1] Jim M. Smit,[1,2] Joost Huiskens,[3] Jacob J. Visser,[4,5] Robert E. R. Veen,[6] Edwin van Unen,[3] Oliver Hilgers BA,[7] Diederik Gommers,[1] Jasper van Bommel[1]

To cite: van de Sande D, Van Genderen ME, Smit JM, *et al.* Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Inform* 2022;**29**:e100495. doi:10.1136/bmjhci-2021-100495

Received 06 October 2021
Accepted 24 January 2022

© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Michel E Van Genderen; m.vangenderen@erasmusmc.nl

SUMMARY

Objective Although the role of artificial intelligence (AI) in medicine is increasingly studied, most patients do not benefit because the majority of AI models remain in the testing and prototyping environment. The development and implementation trajectory of clinical AI models are complex and a structured overview is missing. We therefore propose a step-by-step overview to enhance clinicians' understanding and to promote quality of medical AI research.

Methods We summarised key elements (such as current guidelines, challenges, regulatory documents and good practices) that are needed to develop and safely implement AI in medicine.

Conclusion This overview complements other frameworks in a way that it is accessible to stakeholders without prior AI knowledge and as such provides a step-by-step approach incorporating all the key elements and current guidelines that are essential for implementation, and can thereby help to move AI from bytes to bedside.

INTRODUCTION

Over the past few years, the number of medical artificial intelligence (AI) studies has grown at an unprecedented rate (figure 1). AI-related technology has the potential to transform and improve healthcare delivery on multiple aspects, for example, by predicting optimal treatment strategies, optimising care processes or making risk predictions.[1 2] Nonetheless, studies in the intensive care unit (ICU) and radiology demonstrated that 90%–94% of the published AI studies remain within the testing and prototyping environment and have poor study quality.[3 4] Also in other specialties, clinical benefits fall short of the high set expectations.[2 5] This lack of clinical AI penetration is daunting and increases the risk of a period in which the AI hype will be tempered and reach a point of disillusionment expectations, that is, an 'AI winter'.[6]

To prevent such a winter, new initiatives must successfully mitigate AI-related risks on multiple levels (eg, data, technology, process and people) that impede development and might threaten safe clinical implementation.[2 3 7 8] This is especially important since the development and implementation of new technologies in medicine, and in particular AI, is complex and requires an interdisciplinary approach to engagement of multiple stakeholders.[9] A parallel can be drawn between the development of new drugs for which the US Food and Drug Administration (FDA) developed a specific mandatory process before clinical application.[10–12] Because the delivery of AI to patients is in need of a similar structured approach to ensure safe clinical application, the FDA proposed a regulatory framework for (medical) AI.[13–16] In addition, the European Commission proposed a similar framework but does not provide details concerning medical AI.[17] Besides regulatory progress, guidelines have emerged to promote quality and replicability of clinical AI research.[18]

Despite the increasing availability of such guidelines, expert knowledge, good practices, position papers and regulatory documents, the medical AI landscape is still fragmented and a step-by-step overview incorporating all the key elements for implementation is lacking. We have therefore summarised several steps and elements (figure 2) that are required to structurally develop and implement AI in medicine (table 1). We hope that our step-by-step approach improves quality, safety and transparency of AI research, helps to increase clinicians' understanding of these technologies, and improves clinical implementation and usability.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

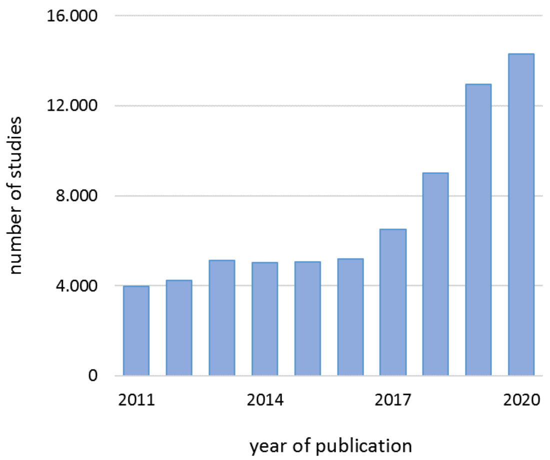

Figure 1 Global evolution of research in artificial intelligence in medicine. The number of AI papers in humans on PubMed.com was arranged by year, 2011–2020. The blue bars represent the number of studies. The following search was performed: ("artificial intelligence"[MeSH Terms] OR ("artificial"[All Fields) and "intelligence"[All Fields]) OR "artificial intelligence"[All Fields]) OR ("machine learning"[MeSH Terms] OR ("machine"[All Fields] AND "learning"[All Fields]) OR "machine learning"[All Fields]) OR ("deep learning"[MeSH Terms] OR ("deep"[All Fields] AND "learning"[All Fields]) OR "deep learning"[All Fields]).

IDENTIFYING KEY DOCUMENTS IN THE AI LITERATURE

Publications were identified through a literature search of PubMed, Embase and Google Scholar from January 2010 to June 2021. The following terms were used as index terms or free-text words: "artificial intelligence", "deep learning", "machine learning " in combination with "regulations", "framework", "review", and "guidelines" to identify eligible studies. Articles were also identified through searches of the authors' own files. Only papers published in English were reviewed. Regulatory documents were identified by searching the official web pages of the FDA, European Medicines Agency, European Commission and International Medical Device Regulators Forum (IMDRF). Since it was beyond our scope to provide a systematic overview of the AI literature, no quantitative synthesis was conducted.

PHASE 0: PREPARATIONS PRIOR TO AI MODEL DEVELOPMENT

Define the clinical problem and engage stakeholders

AI models should improve care and address clinically relevant problems. Not only should they be developed to predict illnesses, such as sepsis, but they also should produce actionable output directly or indirectly linked to clinical decision-making.[19] Defining the clinical problem and its relevance before initiating model development is therefore important.[20]

Varying skills and expertise are required to develop and implement an AI model, and formation of an interdisciplinary team is key. The core team should at least consist of knowledge experts, decision-makers and even users (figure 2).[9] While each of them are essential to make the initiative succeed, depending on the required skills

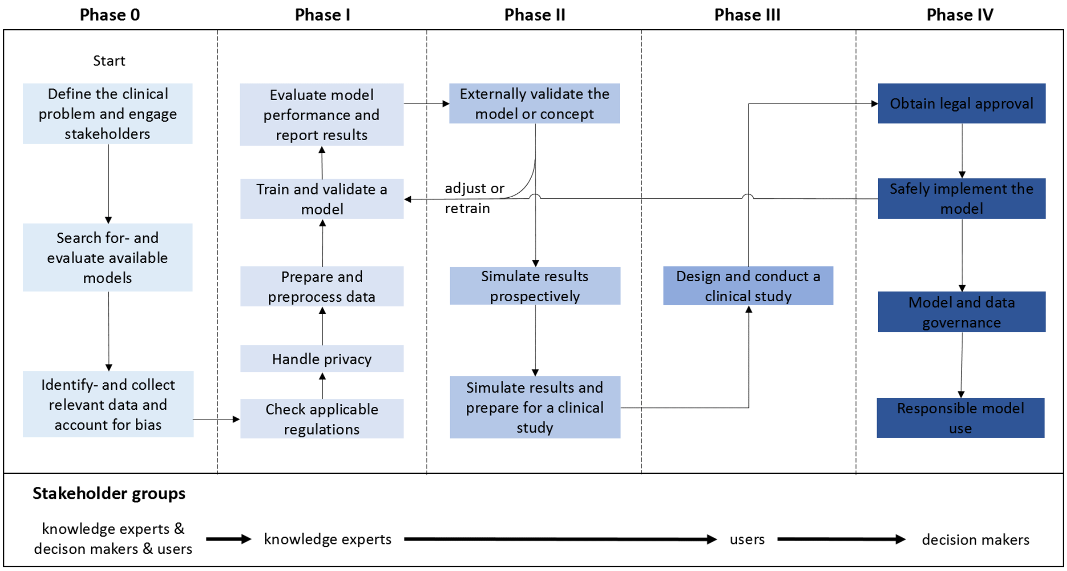
Phase 0
Phase I
Phase II
Phase III
Phase IV
Start
Define the clinical problem and engage stakeholders
Search for- and evaluate available models
Identify- and collect relevant data and account for bias
Evaluate model performance and report results
Train and validate a model
Prepare and preprocess data
Handle privacy
Check applicable regulations
Externally validate the model or concept
adjust or retrain
Simulate results prospectively
Simulate results and prepare for a clinical study
Design and conduct a clinical study
Obtain legal approval
Safely implement the model
Model and data governance
Responsible model use
Stakeholder groups
knowledge experts & decison makers & users
knowledge experts
users
decision makers

Figure 2 Structured overview of the clinical AI development and implementation trajectory. Crucial steps within the five phases are presented along with stakeholder groups at the bottom that need to be engaged: knowledge experts (eg, clinical experts, data scientists and information technology experts), decision-makers (eg, hospital board members) and users (eg, physicians, nurses and patients). Each of the steps should be successfully addressed before proceeding to the next phase. The colour gradient from light blue to dark blue indicates AI model maturity, from concept to clinical implementation. The development of clinical AI models is an iterative process that may need to be (partially) repeated before successful implementation is achieved. Therefore, a model could be adjusted or retrained (ie, return to phase I) at several moments during the process (eg, after external validation or after implementation). AI, artificial intelligence.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 1 Crucial steps and key documents per phase throughout the trajectory

Phase	Guidelines, position papers and regulatory documents
0: preparations prior to AI model development	
1. Define the clinical problem and engage stakeholders.	Wiens *et al*[9]
2. Search for and evaluate available models.	Benjamens *et al*,[21] ECLAIR[22]
3. Identify and collect relevant data and account for bias.	FHIR,[26] FAIR,[28] Riley *et al*[23] Wolff *et al*[25]
4. Handle privacy.	HIPAA[30] and GDPR[31]
I: AI model development	
5. Check applicable regulations.	'Proposed regulatory framework' (FDA),[13] 'Harmonised rules on AI' (EU)[17]
6. Prepare and preprocess the data.	Ferrão *et al*[40]
7. Train and validate a model.	Juarez-Orozco *et al*[42]
8. Evaluate model performance and report results.	Park and Han,[50] TRIPOD,[51] TRIPOD-ML* [52]
II: assessment of AI performance and reliability	
9. Externally validate the model or concept.	Ramspek *et al*,[53] Riley *et al*,[54] Futoma *et al*[55]
10. Simulate results and prepare for a clinical study.	DECIDE-AI* [59]
III: clinically testing AI	
11. Design and conduct a clinical study.	SPIRIT-AI,[63] Barda *et al*,[65] CONSORT-AI[66]
IV: implementing and governing AI	
12. Obtain legal approval.	Muehlematter *et al*[35]
13. Safely implement the model.	TAM,[70] Sendak *et al*[72]
14. Model and data governance.	FAIR,[28] 'SaMD: clinical evaluation' (FDA),[79] 'Application of Quality Management System'(IMDRF)[78]
15. Responsible model use.	Martinez-Martin *et al*[19]

Based on emerging themes in medical AI literature, important steps have been highlighted and categorised in five phases analogous to the phases of drug research. For each phase, the crucial steps are noted on the left and the corresponding key documents are noted on the right.
Standard protocol items: recommendations for interventional trials.
*Guidelines are currently under construction.
AI, artificial intelligence; CONSORT-AI, Consolidated Standards of Reporting Trials–Artificial Intelligence; DECIDE-AI, Developmental and Exploratory Clinical Investigation of Decision-Support Systems Driven by Artificial Intelligence; ECLAIR, Evaluating Commercial AI Solutions in Radiology; EU, European Union; FAIR, Findable, Accessible, Interoperable and Reusable; FDA, Food and Drug Administration; FHIR, Fast Healthcare Interoperability Resources; GDPR, General Data Protection Regulation; HIPAA, Health Insurance Portability and Accountability Act; IMDRF, International Medical Device Regulators Forum; ML, machine learning; SaMD, software as a medical device; SPIRIT-AI, Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence; TAM, technology acceptance model; TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis.

for each step, some will play a more important role than others.

Search for and evaluate available models

Numerous AI models have already been published, so it is knowledgeable to search for readily available models when encountering a clinical problem (https://medicalfuturist.com/fda-approved-ai-based-algorithms/)[21] and to evaluate such models using the 'Evaluating Commercial AI Solutions in Radiology' guideline.[22] Although the latter guideline was developed for radiology purposes, it can be extrapolated to other specialties.

Identify and collect relevant data and account for bias

Adequate datasets are required to train AI models. These datasets need to be of sufficient quality and quantity to achieve high model performance; Riley *et al*[23] therefore proposed a method to calculate a required sample size similar to traditional studies. Information on the outcome of interest (model output) as well as potential predictor variables (model input) need to be collected while accounting for potential bias. Unlike bias in traditional studies (eg, selection bias), bias in AI models can additionally be categorised in algorithmic and social bias which can arise from factors such as gender, race or measurement errors, leading to suboptimal outcomes for particular groups.[24] In order to mitigate the risk of bias and to collect representative training data, tools such as the Prediction Model Risk of Bias Assessment Tool can be of help.[24,25] Nonetheless, these clinical data are often underused since they are siloed in a multitude of medical information systems complicating fast and uniform extraction, emphasising the importance of adopting unified data formats such as the Fast Healthcare Interoperability Resources.[26,27] To enhance usability and sharing of such data, it must be findable, accessible, interoperable and reusable as described in the Findable, Accessible, Interoperable and Reusable (FAIR) guideline.[28] In this phase, developers should also look beyond interoperability of resources within institutions; namely, if AI models are to be used at scale, compatibility between hospitals' information systems may be challenging as well.[29]

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Handle privacy

Regarding privacy, special care should be taken when handling such patient data (particularly when sharing data between institutions to combine datasets). A risk-based iterative data deidentification strategy for the purposes of the US Health Insurance Portability and Accountability Act as well as the European General Data Protection Regulation should therefore be taken into account. Such a strategy was recently applied to an openly available ICU database in the Netherlands.[30–32]

PHASE I: AI MODEL DEVELOPMENT

Check applicable regulations

Although medical device regulations are important in effectively implementing and scaling up newly developed models (phase IV), developers should be aware of it early on. AI models are qualified as a 'software as a medical device' (SaMD), when intended to diagnose, treat or prevent health problems (eg, decision support software that can automatically interpret electrocardiograms or advise sepsis treatment).[33] These devices should be scrutinised to avoid unintended (harmful) consequences, and as such, the FDA and the European Commission have been working on regulatory frameworks.[2 13 17] The IMDRF uses a risk-based approach to categorise these SaMDs into different categories reflecting the risk associated with the clinical situation and device use.[34] In general, the higher the risk, the higher the requirements to obtain legal approval. A recent review by Muehlematter *et al*[35] summarises the applicable regulating pathways for the USA and Europe.

Prepare and preprocess the data

Raw data extracted directly from hospital information systems are prone to measurement/sensing errors, particularly monitoring data, which increases the risk of bias.[36 37] Therefore, these data must be prepared and preprocessed prior to AI model development.[38 39] Data preparation consists of steps such as joining data from separate files, labelling the outcome of interest for supervised learning approaches (eg, sepsis and mortality), filtering inaccurate data and calculating additional variables. On the other hand, data preprocessing consists of more analytical data manipulations (specifically used for model training) such as smart imputations of missing values (eg, multiple imputation), variable selection (ie, selecting those highly predictive variables) and others to create a so called 'data preprocessing pipeline'. An example of such a data preprocessing framework has been described in more detail by Ferrão *et al*.[40]

Train and validate a model

To address the clinical problem, different AI models can be used. Herein, a distinction can be made between traditional statistical models such as logistic regression and AI models such as neural networks.[41] In a thoughtful review, Juarez-Orozco *et al*[42] provided an overview of advantages and disadvantages of multiple AI models and categorised them according to their learning type (broadly categorised as supervised, unsupervised and reinforcement learning) and purpose (eg, classification and regression). When selecting a model, trade-offs exist between model sophistication and AI explainability; the latter refers to the degree AI models can be interpreted and should not be overlooked.[43]

To determine whether AI models are reliable on unseen data, they are usually validated on a so-called 'test dataset' (ie, internal validation). Several internal validation methods can be used. For example, by randomly splitting the total dataset into subsets (train, validation and test dataset) either once or multiple times (which is known in literature as k-fold cross-validation) in order to evaluate model performance on the test dataset such as that demonstrated by Steyerberg *et al*.[44]

Evaluate model performance and report results

Clinical implementation of inaccurate or poorly calibrated AI models can lead to unsafe situations.[45] As no single performance metric captures all desirable model properties, multiple metrics such as area under the receiver operating characteristics, accuracy, sensitivity, specificity, positive predictive value, negative predictive value and calibration should be evaluated.[41 46–49] A guideline by Park and Han[50] can assist model performance evaluation. Afterwards, study results should be reported transparently, following transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).[51] Since the TRIPOD statement was intended for conventional prediction models, a specific machine learning extension has recently been announced.[52]

PHASE II: ASSESSMENT OF AI PERFORMANCE AND RELIABILITY

Externally validate the model or concept

Unlike medical devices, such as mechanical ventilators, AI models do not operate based on a universal set of preprogrammed rules but instead provide patient-specific predictions. They might work perfectly in one setting and terribly in others. After local model development, AI models should undergo external validation to determine their generalisability and safety.[53 54] However, it is commonly accepted that poor generalisability should be avoided prior to implementation; it is argued that broad generalisability is probably impossible since 'practice-specific information is often highly predictive' and models should thus be locally trained whenever possible, that is, site-specific training.[55] Therefore, the AI concept (ie, the concept based on the specific variables and outcomes) may need to be validated rather than the exact model. Whether validating the exact model or concept, it is always important to evaluate whether the training and validation population are comparable in order to compare results appropriately. In case external validation demonstrates

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

inconsistencies with previous results, the model may need to be adjusted or retrained.[56]

Simulate results and prepare for a clinical study

In order to safely test an AI model at bedside, potential pitfalls should be timely identified. It has been suggested that model predictions can be generated prospectively without exposing the clinical staff to the results, that is, temporal validation.[57] Such a step is pivotal to evaluate model performance on real-world clinical data and is used to ensure availability of all required data (ie, data required to generate model predictions) for which a real-time data infrastructure should be established.[58] Because variation across local practices and subpopulations exists and clinical trials can be expensive, the Developmental and Exploratory Clinical Investigation of Decision-Support Systems Driven by Artificial Intelligence is being developed to decrease the gap to clinical testing.[59]

PHASE III: CLINICALLY TESTING AI

Design and conduct a clinical study

To date, only 2% of AI studies in the ICU were clinically tested while it is an important step to determine clinical utility and usability.[3] Clinical AI studies preferably need to be carried out in a randomised setting where steps are described in detail to enhance replication by others.[60–62] Such studies can have different designs similar to traditional studies, and the same considerations need to be made (eg, randomised versus non-randomised, monocentric versus multicentric, blinded versus non-blinded). At all times, the Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence guideline should be followed.[63] Since AI models are primarily developed to improve care by providing actionable output, it is important that the output is appropriately conveyed to the end users; that is, output should be both useful and actionable. For example, Wijnberge *et al*[64] clinically tested a hypotension prediction model during surgery and provided the clinicians the output via a specific display. A recent framework can help to design such user-centred AI displays, and reporting via the Consolidated Standards of Reporting Rrials–Artificial Intelligence guideline can promote quality, transparency and completeness of study results.[65 66]

PHASE IV: IMPLEMENTING AND GOVERNING OF AI

Obtain legal approval

Regulatory aspects (as described in phase I), data governance and model governance play an important role in the clinical implementation and should be addressed appropriately. Before widespread clinical implementation is possible, AI models must be submitted to the FDA in the USA and in Europe, they need to obtain a Conformité Européenne (CE) mark from accredited companies (these can be found on https://ec.europa.eu/growth/tools-databases/nando/), unless exempted by the pathway for health institutions.[67 68] Nowadays, some models already received a CE mark[35] or FDA approval.[21]

Safely implement the model

If an AI model is not accepted by the users, it will not influence clinical decision-making.[69] Factors such as usefulness and ease of use, which are described in the technology acceptance model, are demonstrated to improve the likelihood of successful implementation and should therefore be taken into account.[70 71] Furthermore, implementation efforts should be accompanied by clear and standardised communication of AI model information towards end users to promote transparency and trust, for example, by providing an 'AI model facts label'.[72] To ensure that AI models will be safely used once they are implemented, users (eg, physicians, nurses and patients) should be properly educated, particularly on how to use them without jeopardising the clinician–patient relationship.[19 73 74] Specific AI education programmes can help and have already been introduced.[75 76]

Model and data governance

After implementation, hospitals should implement a dedicated quality management system and monitor AI model performance during the entire life span, enabling timely identification of worsening model performance, and react whenever necessary (eg, retire, retrain, adjust or switch to an alternative model).[49 77–79] Governance of the required data and AI model deserves special consideration. Data governance covers items such as data security, data quality, data access and overall data accountability (see also the FAIR guideline).[19 28] On the other hand, model governance covers aspects such as model adjustability, model version control and model accountability. Besides timely identifying declining model performance, governing AI models is also vital to gain patients' trust.[80] Once a model is retired, the corresponding assets such as documentation and results should be stored for 15 years (although no consensus on terms has been reached yet), similar to clinical trials.[81]

Responsible model use

Importantly, one must be aware that AI models can be used in biased ways when real-world data do not resemble the training data due to changing care/illness specific paradigms (ie, data shift).[19 62 82–84] Clinicians always need to determine how much weight they give to AI models' output in clinical decision-making in order to safely use these technologies.[82 85]

DISCUSSION

We believe that this review complements other referenced frameworks by providing a complete overview of this complex trajectory. Also, stakeholders without prior AI knowledge should now better grasp what is needed from AI model development to implementation.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

The importance of such a framework to transparently develop and implement clinical AI models has been highlighted by a study of Wong *et al*[86]; they externally validated a proprietary sepsis prediction model which has already been widely implemented by hundreds of hospitals in the USA despite no independent validations having been published yet. The authors found that the prediction model missed two-thirds of the patients with sepsis (ie, low sensitivity), while clinicians had to evaluate eight patients to identify a patient with sepsis (ie, high false alarm rate).[86] It is important to question why such prediction models can be widely implemented while they may be harmful to patients and may negatively affect the clinical workflow; they may, for example, lead to overtreatment (eg, antibiotics) of false-positive patients, undertreatment of false-negative patients and alarm fatigue among clinicians.

The main challenges to deliver impact with clinical AI models are interdisciplinary and include challenges that are intrinsic to the fields of data science, implementation science and health research, which we have addressed throughout the different phases in this review. Although it was outside the scope of this review to provide a comprehensive overview of the ethical issues related to clinical AI, they are of major concern to the development as well as clinical implementation and hence are an important topic on the AI research agenda.[87] Some examples are protecting human autonomy, ensuring transparency and explainability, ensuring inclusiveness, and equity, which are described in a recent guidance document on AI ethics by the WHO.[88]

In an attempt to prevent an AI winter, we invite other researchers, stakeholders and policy makers to comment on the current approach and to openly discuss how to safely develop and implement AI in medicine. By combining our visions and thoughts, we may be able to propel the field of medical AI forward, step-by-step.

CONCLUSION

This review is a result of an interdisciplinary collaboration (clinical experts, information technology experts, data scientists and regulations experts) and contributes to the current medical AI literature by unifying current guidelines, challenges, regulatory documents and good practices that are essential to medical AI development. Additionally, we propose a structured step-by-step approach to promote AI development and to guide the road towards safe clinical implementation. Importantly, the interdisciplinary research teams should carry out these consecutive steps in compliance with applicable regulations and publish their findings transparently, whereby the referenced guidelines and good practices can help.

Still, future discussions are needed to answer several questions such as the following: what is considered as adequate clinical model performance? how do we know whether predictions remain reliable over time? who is responsible in case of AI model failure? and how long must model data be stored for auditing purposes?

Author affiliations
[1]Department of Adult Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands
[2]Pattern Recognition and Bioinformatics group, EEMCS, Delft University of Technology, Delft, The Netherlands
[3]SAS Institute Inc, Health, Huizen, The Netherlands
[4]Department of Radiology and Nuclear Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
[5]Department of Information Technology, Chief Medical Information Officer, Erasmus Medical Center, Rotterdam, The Netherlands
[6]Department of Information Technology, theme Research Suite, Erasmus Medical Center, Rotterdam, The Netherlands
[7]Active Medical Devices/Medical Device Software, CE Plus GmbH, Badenweiler, Germany

Twitter Davy van de Sande @davy_sande

Contributors DvdS, MEVG and JH searched the literature.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests DG received speaker's fees and travel expenses from Dräger, GE Healthcare (medical advisory board 2009–2012), Maquet, and Novalung (medical advisory board 2015–2018). JH currently works as industry expert healthcare at SAS Institute. EvU currently works as principal analytics consultant at SAS Institute. No financial relationships exist that could be construed as a potential conflict of interest. All other authors declare no competing interests.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing is not applicable as no datasets were generated and/or analysed for this study. Not applicable.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Davy van de Sande http://orcid.org/0000-0003-4484-0995
Michel E Van Genderen http://orcid.org/0000-0001-5668-3435

REFERENCES

1 Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309:1351–2.
2 He J, Baxter SL, Xu J, *et al*. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30–6.
3 van de Sande D, van Genderen ME, Huiskens J, *et al*. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med* 2021;47:750-760.
4 Kim DW, Jang HY, Kim KW, *et al*. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol* 2019;20:405-410.
5 Wilkinson J, Arnold KF, Murray EJ, *et al*. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* 2020;2:e677–80.
6 Floridi L. Ai and its new winter: from myths to realities. *Philos Technol* 2020;33:1–3.
7 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
8 Roski J, Maier EJ, Vigilante K, *et al*. Enhancing trust in AI through industry self-governance. *J Am Med Inform Assoc* 2021;28:1582–90.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

9 Wiens J, Saria S, Sendak M. Do no harm: a roadmap for responsible machine learning for health care (vol 25, PG 1337, 2019). *Nat Med* 2019;25:1627–27.

10 Komorowski M. Clinical management of sepsis can be improved by artificial intelligence: Yes. *Intensive Care Med* 2020;46:375–7.

11 Park SH, Do K-H, Choi J-I, *et al*. Principles for evaluating the clinical implementation of novel digital healthcare devices. *J Korean Med Assoc* 2018;61:765–75.

12 Komorowski M. Artificial intelligence in intensive care: are we there yet? *Intensive Care Med* 2019;45:1298–300.

13 Administration FaD. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback*. Food and Drug Administration, 2019.

14 FaD A. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. Action Plan: Food and Drug Administration, 2021.

15 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58.

16 Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA* 2020;324:1397–8.

17 Commission E. *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts*. Brussels: European Commission, 2021.

18 The Lancet Digital Health . Walking the tightrope of artificial intelligence guidelines in clinical practice. *Lancet Digit Health* 2019;1:e100.

19 Martinez-Martin N, Luo Z, Kaushal A, *et al*. Ethical issues in using ambient intelligence in health-care settings. *Lancet Digit Health* 2021;3:e115–23.

20 Gutierrez G. Artificial intelligence in the intensive care unit. *Crit Care* 2020;24:101.

21 Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020;3:118.

22 Omoumi P, Ducarouge A, Tournier A, *et al*. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol* 2021;31:3786–96.

23 Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.

24 Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019;322:2377–8.

25 Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51–8.

26 Mandel JC, Kreda DA, Mandl KD, *et al*. Smart on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc* 2016;23:899–908.

27 Ghassemi M, Naumann T, Schulam P, *et al*. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health* 2019;1:e157–9.

28 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.

29 Lehne M, Sass J, Essenwanger A, *et al*. Why digital medicine depends on interoperability. *NPJ Digit Med* 2019;2:79.

30 OOTA S, Evaluation FPA. Health Insurance Portability and Accountability Act of 1966: U.S. Department of Health & Human Services; 08/21/1996. Available: https://aspe.hhs.gov/report/health-insurance-portability-and-accountability-act-1996

31 Commission E. Data protection in the EU: European Union, 2016. Available: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en

32 Thoral PJ, Peppink JM, Driessen RH, *et al*. Sharing ICU patient data Responsibly under the Society of critical care Medicine/European Society of intensive care medicine joint data science collaboration: the Amsterdam University medical centers database (AmsterdamUMCdb) example. *Crit Care Med* 2021;49:e563-e577.

33 Group ISW. *Software as a medical device (SaMD): key definitions international medical device regulators forum*, 2013.

34 Group ISaaMDSW. "*Software as a Medical Device*": *Possible Framework for Risk Categorization and Corresponding Considerations: International Medical Device Regulators Forum*, 2014.

35 Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. *Lancet Digit Health* 2021;3:e195-e203.

36 Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013;51:S30–7.

37 Maslove DM, Dubin JA, Shrivats A, *et al*. Errors, omissions, and outliers in hourly vital signs measurements in intensive care. *Crit Care Med* 2016;44:e1021–30.

38 Miller DD. The medical AI insurgency: what physicians must know about data to practice with intelligent machines. *NPJ Digit Med* 2019;2:62.

39 Johnson AEW, Ghassemi MM, Nemati S, *et al*. Machine learning and decision support in critical care. *Proc IEEE Inst Electr Electron Eng* 2016;104:444–66.

40 Ferrão JC, Oliveira MD, Janela F, *et al*. Preprocessing structured clinical data for predictive modeling and decision support. A roadmap to tackle the challenges. *Appl Clin Inform* 2016;7:1135–53.

41 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.

42 Juarez-Orozco LE, Martinez-Manzanera O, Nesterov SV. The machine learning horizon in cardiac hybrid imaging. *Eur J Hybrid Imaging* 2018:1–15.

43 Amann J, Blasimme A, Vayena E, *et al*. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020;20:310.

44 Steyerberg EW, Harrell FE, Borsboom GJ, *et al*. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774–81.

45 Gottesman O, Johansson F, Komorowski M, *et al*. Guidelines for reinforcement learning in healthcare. *Nat Med* 2019;25:16–18.

46 Shillan D, Sterne JAC, Champneys A, *et al*. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Crit Care* 2019;23:284.

47 Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997;30:1145–59.

48 Collins GS, de Groot JA, Dutton S, *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.

49 Bouwmeester W, Zuithoff NPA, Mallett S, *et al*. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1–12.

50 Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800–9.

51 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.

52 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577–9.

53 Ramspek CL, Jager KJ, Dekker FW, *et al*. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2021;14:49–58.

54 Riley RD, Ensor J, Snell KIE, *et al*. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140.

55 Futoma J, Simons M, Panch T, *et al*. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020;2:E489–92.

56 Davis SE, Greevy RA, Fonnesbeck C, *et al*. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inform Assoc* 2019;26:1448–57.

57 Kelly CJ, Karthikesalingam A, Suleyman M, *et al*. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195.

58 van de Sande D, Van Genderen ME, Huiskens J, *et al*. Generating insights in uncharted territories: real-time learning from data in critically ill patients-an implementer report. *BMJ Health Care Inform* 2021;28.

59 DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;27:186-187.

60 Sibbald B, Roland M. Understanding controlled trials. why are randomised controlled trials important? *BMJ* 1998;316:201.

61 Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol* 1995;48:23–40.

62 Colak E, Moreland R, Ghassemi M. Five principles for the intelligent use of AI in medical imaging. *Intensive Care Med* 2021;47:154-156.

63 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;2:e549–60.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

64 Wijnberge M, Geerts BF, Hol L, *et al*. Effect of a machine Learning-Derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the hype randomized clinical trial. *JAMA* 2020;323:1052–60.
65 Barda AJ, Horvat CM, Hochheiser H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med Inform Decis Mak* 2020;20:257.
66 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health* 2020;2:e537–48.
67 Union E. Notified bodies Nando: European Commission. Available: https://ec.europa.eu/growth/tools-databases/nando/
68 Union E. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance)Text with EEA relevance: European Union, 2017. Available: https://eur-lex.europa.eu/eli/reg/2017/745/2017-05-05
69 Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020;22:e15154.
70 Davis FD, Bagozzi RP, PR W. User acceptance of computer technology: a comparison of two theoretical models. *Manage Sci* 1989;35:982–1003.
71 Jauk S, Kramer D, Avian A, *et al*. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. *J Med Syst* 2021;45:48.
72 Sendak MP, Gao M, Brajer N, *et al*. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med* 2020;3:41.
73 Keane PA, Topol EJ. AI-facilitated health care requires education of clinicians. *Lancet* 2021;397:1254.
74 Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. *Lancet Digit Health* 2020;2:e111–2.
75 Paranjape K, Schinkel M, Nannan Panday R, *et al*. Introducing artificial intelligence training in medical education. *JMIR Med Educ* 2019;5:e16048.
76 Coalition DA. The National AI-healthcare course (in Dutch: de nationale AI-zorg cursus), 2021. Available: https://zorg.ai-cursus.nl/home
77 Lee CS, Lee AY. Clinical applications of continual learning machine learning. *Lancet Digit Health* 2020;2:e279–81.
78 Group ISW. Software as a Medical Device (SaMD): Application of Quality Management System. In: *International medical device regulators forum*. Forum IMDR, 2015.
79 Administration USFaD. *Software as a medical device (SAMD): clinical evaluation*. In: Healt USDoHaHSFaDACfDaR, 2016.
80 Falco G, Shneiderman B, Badger J, *et al*. Governing AI safety through independent audits. *Nat Mach Intell* 2021;3:566–71.
81 Commission E. *Amending directive 2001/83/EC of the European Parliament and of the Council on the community code relating to medicinal products for human use*. Brussels: Commission E, 2003.
82 Liu VX. The future of AI in critical care is augmented, not artificial, intelligence. *Crit Care* 2020;24:673.
83 Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018;15:e1002689.
84 MBAM BN, Chauhan G, Naumann T, *et al*. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. machine learning for health (ML4H). *NeurIPS* 2018.
85 Shaw JA, Sethi N, Block BL. Five things every clinician should know about AI ethics in intensive care. *Intensive Care Med* 2021;47:157-159.
86 Wong A, Otles E, Donnelly JP, *et al*. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181:1065-1070.
87 Gibney E. The battle for ethical AI at the world's biggest machine-learning conference. *Nature* 2020;577:609.
88 Governance WsHEa, Health uitdoRfHatdoD. *Ethics and governance of artificial intelligence for health*. Geneva: World Health Organization, 2021.

BMJ Health & Care Informatics: first published as 10.1136/bmjhci-2021-100495 on 20 February 2022. Downloaded from https://informatics.bmj.com on 24 February 2025 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.