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

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

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. 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. Also in other specialties, clinical benefits fall short of the high set expectations. 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’.

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. 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. 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. 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. In addition, the European Commission proposed a similar framework but does not provide details concerning medical AI. Besides regulatory progress, guidelines have emerged to promote quality and replicability of clinical AI research.

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.
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. Defining the clinical problem and its relevance before initiating model development is therefore important. 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). While each of them are essential to make the initiative succeed, depending on the required skills
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://medical-futurist.com/fda-approved-ai-based-algorithms/) and to evaluate such models using the ‘Evaluating Commercial AI Solutions in Radiology’ guideline. 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 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. In order to mitigate the risk of bias and to collect representative training data, tools such as the Prediction Model Risk of Bias Assessment System (IMDRF) can be of help. 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 Findable, Accessible, Interoperable and Reusable (FAIR) guidelines. 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. 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.

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

<table>
<thead>
<tr>
<th>Phase</th>
<th>Guidelines, position papers and regulatory documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: preparations prior to AI model development</td>
<td></td>
</tr>
</tbody>
</table>
1. Define the clinical problem and engage stakeholders. Wiens et al  
2. Search for and evaluate available models. Benjamens et al; ECLAIR  
3. Identify and collect relevant data and account for bias. FAIR, FAIR, Riley et al, Wolff et al  
4. Handle privacy. HIPAA and GDPR  |
| I: AI model development |  
5. Check applicable regulations. ‘Proposed regulatory framework’ (FDA), ‘Harmonised rules on AI’ (EU)  
6. Prepare and preprocess the data. Ferrão et al  
7. Train and validate a model. Juarez-Orozco et al  
8. Evaluate model performance and report results. Park and Han, TRIPOD, TRIPOD-ML  |
| II: assessment of AI performance and reliability |  
9. Externally validate the model or concept. Ramspek et al, Riley et al, Futoma et al  
10. Simulate results and prepare for a clinical study. DECIDE-AI  |
| III: clinically testing AI |  
11. Design and conduct a clinical study. SPIRIT-AI, Barda et al, CONSORT-AI  
12. Obtain legal approval. Muehlematter et al  
13. Safely implement the model. TAM, Sendak et al  
15. Responsible model use. Martinez-Martin et al  |
| IV: implementing and governing AI |  
for each step, some will play a more important role than others. |
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). 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. 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. In general, the higher the risk, the higher the requirements to obtain legal approval. A recent review by Muehlematter et al 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. Therefore, these data must be prepared and preprocessed prior to AI model development. 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.

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. In a thoughtful review, Juarez-Orozco et al 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.

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.

Evaluate model performance and report results

Clinical implementation of inaccurate or poorly calibrated AI models can lead to unsafe situations. 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. A guideline by Park and Han 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). Since the TRIPOD statement was intended for conventional prediction models, a specific machine learning extension has recently been announced.

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. However, it is commonly accepted that poor generalisability should be avoided prior to implementation; it is argued that broad generalisability is probably impossible since ‘practicespecific information is often highly predictive’ and models should thus be locally trained whenever possible, that is, site-specific training. 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...
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. 57,68 Nowadays, some models already received a CE mark 55 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. 70–72 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). 19,75–79 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). 19,62–84 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.
The importance of such a framework to transparently develop and implement clinical AI models has been highlighted by a study of Wong et al.87; 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**
Department of Adult Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands
Pattern Recognition and Bioinformatics group, EEMCS, Delft University of Technology, Delft, The Netherlands
SAS Institute Inc, Health, Huizen, The Netherlands
Department of Radiology and Nuclear Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
Department of Information Technology, Chief Medical Information Officer, Erasmus Medical Center, Rotterdam, The Netherlands
Department of Information Technology, theme Research Suite, Erasmus Medical Center, Rotterdam, The Netherlands
Department of Information Technology, theme Research Suite, Erasmus Medical Center, Rotterdam, The Netherlands
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 Novaulang (medical advisory board 2015–2018). JH currently works as industry expert healthcare at SAS Institute. Evl 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**


79 Administration USFDA. Software as a medical device (SaMD): clinical evaluation. In: Health USDHasFDAACIDaR, 2016.


82 Liu VX. The future of AI in critical care is augmented, not artificial, intelligence. *Crit Care* 2020;24:673.


