Implementing AI models in clinical workflows: a roadmap

Fei Wang ,[1] Ashley Beecy[1,2]

10.1136/bmjebm-2023-112727

[1]Weill Cornell Medical College, New York, New York, USA
[2]NewYork-Presbyterian Hospital, New York, New York, USA

Correspondence to:
Dr Fei Wang, Weill Cornell Medical College, New York, New York, USA; few2001@med.cornell.edu

To cite: Wang F, Beecy A. *BMJ Evidence-Based Medicine* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjebm-2023-112727

Artificial Intelligence (AI) aims at mimicking human intelligence through computer programmes. Machine learning (ML), especially deep learning technologies, aiming at inferring insights from complex data through mathematical modelling, offers an effective way of achieving AI and has achieved great success in many disciplines, such as computer vision and natural language processing. Over the past decade, many ML models have also been developed with the goal of improving healthcare, such as predicting the risk of sepsis shock for patients in critical care,[1] identifying patients who are at high risk of developing postpartum depression from their historical clinical records[2] and screening patients who are infected by SARS-CoV-2 according to their routine blood test results.[3]

Real-world clinical trials are essential for proving that AI applications are safe, effective and fit for use in healthcare by assessing their performance across diverse conditions and populations, ensuring regulatory compliance and addressing ethical concerns. Despite the need for clinical trials and the promising results reported in research papers, the ratio of these models that have been implemented in real-world clinical workflows is relatively small. One of the inherent reasons is the complex interactions among multiple stakeholders in the healthcare system including patients, providers, policymakers and insurance companies. In a recent review, Li *et al*[4] identified 19 technical/algorithm, stakeholder and social levels barriers to the application of AI in healthcare and called for future endeavours to address them. With this demand, there has been more and more efforts focusing on particular aspects of these barriers[5 6] or exemplar implementations in different disease contexts,[1 2 7] but guidelines for the holistic process of implementating AI models in clinical workflows are still sporadic.

To fill in this gap, in this perspective, we provide an AI model implementation roadmap in clinical workflows, including three main phases: pre-implementation, peri-implementation and post-implementation. Key modules of each phase and how they are interconnected to impact the overall outcome of the entire solution are discussed, with the goal of providing a comprehensive picture on the lifecycle of AI model implementation. Figure 1 summarises these different stages and the critical components that we will discuss as follows:

Pre-implementation

Pre-implementation refers to the stage when the model has been developed and demonstrated strong promise during retrospective analysis. Before we integrate the model into the actual clinical workflow, we need to make sure of the following items:

Model performance. The model's performance needs to be extensively evaluated before it can be deployed. In a recent paper, Wong *et al*[8] reported a significant drop of the performance of the sepsis risk prediction model integrated in the Epic system. Finlayson *et al*[6] stated that 'this was a case in which the dataset shift fundamentally altered the relationship between fevers and bacterial sepsis'. Although external validation has been emphasised as an important step to ensure the generalisability of the developed model, researchers have recently argued that such external validation could be unrealistic due to various reasons including population and measurement differences, and it has been suggested to conduct repeated local validation instead. Therefore, retrospective evaluation using the local data from the site that the model will be deployed to is critical. During localisation, the operating characteristics and threshold determination can be made based on the specific use case.

Data and infrastructure. After the model is developed and appropriately evaluated for performance and bias, we need to map out the entire data flow of the model deployment cycle and understand where the data will be fed into the model and how the model output will be demonstrated to the end user. For example, a clinical risk prediction model can be implemented within the electronic health record (EHR) system, such as Epic, through their provided applied programming interfaces. During this process, the model developers need to work closely with the information technology service (ITS) team to build appropriate connectors (eg, through the Fast Healthcare Interoperability Resources) so that the EHR data can be fed into the model and model outputs can be transmitted back to the EHR system. We also need to consider where the model will be stored and how frequently the model inference will be needed. Costs and resources to complete this work should be incorporated into the value assessment of the tool.

Model integration. In addition to the technical aspects involving model, data and infrastructure, incentives for the integration of the solution should be aligned as the stakeholder who made the request may not be the same as those that will be responsible for acting on the results. It is imperative to understand the current and future state care delivery process

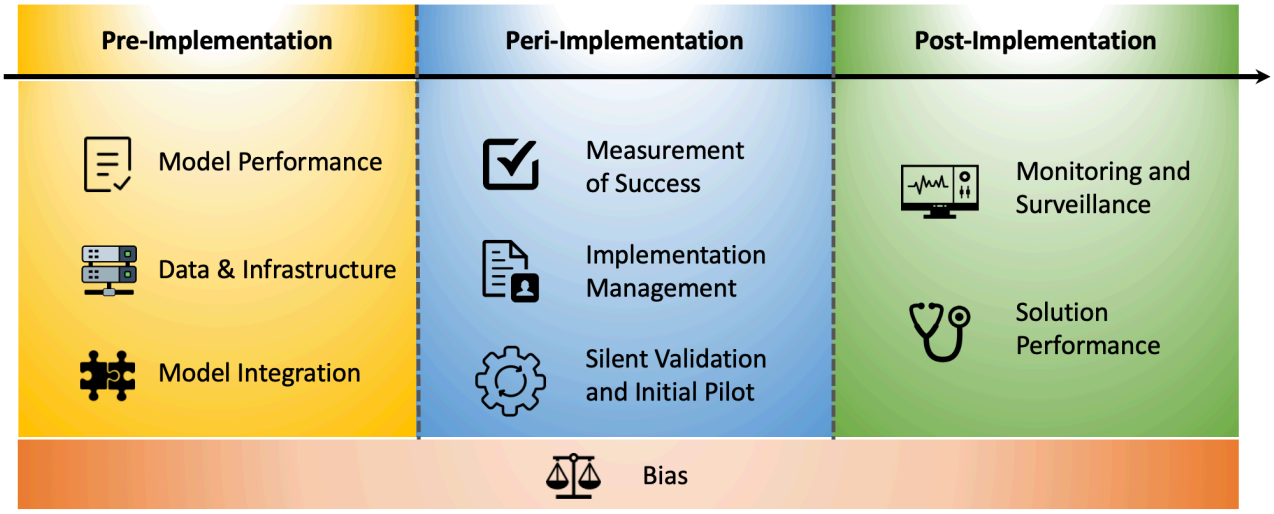

Figure 1 Stages across the deployment of artificial intelligence models in clinical workflows.

as adoption of the tool will be dependent on its fit into a given workflow. The five rights of clinical decision support can be used as a guide: the right person, information, time, context and channel.[9] A user-centred design approach should be taken, and an effector arm should be implemented.[10] Patient and provider input provides valuable insights into the user-friendliness, effectiveness and overall impact of AI applications on care. At this stage, it is appropriate to consider engaging the community for feedback through groups such as a patient advisory council.

Peri-implementation

Peri-implementation refers to the stage right before and during the model is implemented in the clinical workflow. During this phase, the following items are critical.

Measurement of success. It is critical to define the measurement of success during model deployment and ensure the data to quantify this measurement is captured during implementation. Typically, such measurement is not directly the model performance, but it is derived from the model's inference. For example, Adams *et al*[1] used mortality reduction to measure the effectiveness of a sepsis shock prediction algorithm, where the doctors who act on the best practice advisory alert would prescribe antibiotics earlier and may improve patient outcomes such as mortality. In clinical operations, metrics in the EHR, such as Epic's 'Pyjama Time', are used to track interventions aimed at reducing physician administrative burden.[11] The measurement of success should be compared against the pre-deployment standard of care to understand the impact of the tool.

Implementation management. The oversight of medical AI is crucial to ensure its safety and effectiveness, not only on a centralised scale, like the US Food and Drug Administration, but also at the local level to address variations in care, patients and system performance.[12] A clear local governance structure is needed during the model deployment process, as this will involve coordination and collaborations across multiple teams. These teams may include information technology, informatics, data science, health equity, legal, compliance, and information security. An efficient and effective communication mechanism is also required across these teams and with the leadership and end-users. A well-organised documentation structure is needed so that problems and troubles can be resolved in time.

Silent validation and initial pilot. Before the model is integrated in the actual clinical workflow, a silent validation and a pilot study are needed to check production data feeds and understand how such a model will impact the clinical workflow. Here 'silent' validation means the end-users do not have access to the model results, with the goal of recording information on the data input and the model output to ensure it is in line with the retrospective evaluation. A subsequent pilot study, typically in a smaller subset of the final intended population, allows assessment of education materials, communication plan, user interface and potential effector arm.

Post-implementation

AI model deployment is not a one-stop procedure. After deploying the model, its performance and the impact to the entire workflow should be closely monitored. Necessary actions, such as model updating, re-training and even decommissioning, should be taken when the model's behaviour deviates from its original intention or becomes harmful to patients.

Monitoring and surveillance. Most of the disease conditions progress over time, and thus, the model trained using patient data collected from a certain period may not work in the future. For example, with COVID-19, the different SARS-CoV-2 variant waves have been associated with different acute infection outcomes. Therefore, a clinical risk prediction model built during the first wave, which is associated with the most severe clinical outcomes in the acute phase for patients who were infected, may not work for later waves. In addition, public health policies and resource abundance may also impact model performance. For instance, Yang *et al*[2] created a COVID-19 risk prediction model using patient blood test results collected during the first wave of the pandemic in the New York city area. During that time, resources needed for conducting the reverse transcription polymerase chain reaction (RT-PCR) test—the golden standard for confirming a patient is infected by SARS-CoV-2—is limited. Consequently, patients could only take the test if they had relevant symptoms such as fever and cough, which led to a high positive rate (close to 50%). However, after the first wave, such resources became much more available, and the policy also changed so anyone can take the test if they

wanted, which reduced the positive rate to around 2%. Yang *et al*[13] found that the routine blood test profile distributions of the patients who took the RT-PCR test had changed significantly, and the model performance was drastically decreased. Therefore, the model performance needs to be closely monitored, and appropriate actions are needed when there are abnormal observations.

Solution performance. After model deployment, its behaviour will interact with clinicians' practice, which may impact the model's performance, and further model tuning or retraining is needed. Vaid *et al*[14] systematically studied this problem in a simulation framework and found that such model adjustment would further deteriorate the model's performance and lead to unintended consequences. Therefore, it is critical to carefully log all details of the model deployment process, including when the model was deployed, how it interacted with clinicians and how the model performance was changing over time. Liu *et al*[15] proposed a medical algorithmic audit framework to better understand the mechanism of the AI model failure and encourage feedback between the end-user, model developer and ITS team, which can better ensure a safe model deployment process.

Bias

Evaluation of bias should be done at each phase of model deployment to ensure that the model does not introduce or perpetuate healthcare inequities. During retrospective evaluation, model developers should review the training data to ensure that patients represented in the data match the intended target population.[16] If race or inputs from other protected classes are used as features, then the rationale for inclusion of that input should be clearly understood and communicated. The use of surrogate variables for inputs or outcome labels should be reviewed.[17] Model performance should be measured across demographics retrospectively and prospectively to identify potential disparate performance across groups, which could lead to the introduction or perpetuation of bias. Lastly, the favourable outcome (eg, resource, intervention) should be identified, and during the post-implementation period, the distribution of the favourable outcome should be measured to determine whether the model interventions are equitable or as expected. Xu *et al*[18] summarised various potential causes of biased decisions made by algorithms. To deal with this challenge, researchers have developed different checklists for potential algorithmic bias. For example, Finlayson *et al*[6] developed an 'AI safety checklist' to recognise and mitigate dataset shifts in AI models. Wolff *et al*[19] created the Prediction model Risk Of Bias ASsessment Tool for assessing the risk of bias of the predictive models. These checklists and tools should be used as references for assessing the potential bias in AI algorithms.

In summary, we provided an overview of the lifecycle of implementing AI models in clinical workflows. Different from existing studies focusing on model development or a particular phase of the model implementation process, we provided a complete picture of the aspects at its different phases and how they are interconnected to impact the outcome of the overall solution, which aligns well with the real-world scenario when we actually implement these models. We hope our paper can provide a roadmap and trigger holistic thinking in our communities.

Contributors FW conceptualised the manuscript. FW and AB drafted and proofread the whole manuscript.

Funding FW would like to acknowledge the support from NIH awards R01MH124740, RF1AG072449, R01AG080991, R01AG080624, R01AG076448, R01AG076234 and NSF award 1750326 and 2212175.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iD

Fei Wang http://orcid.org/0000-0001-9459-9461

References

1 **Adams R**, Henry KE, Sridharan A, *et al*. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med* 2022;28:1455–60.
2 **Liu Y**, Joly R, Reading Turchioe M, *et al*. Preparing for the bedside—optimizing a postpartum depression risk prediction model for clinical implementation in a health system. *J Am Med Inform Assoc* 2024;31:1258–67.
3 **Yang HS**, Hou Y, Vasovic LV, *et al*. Routine laboratory blood tests predict SARS-Cov-2 infection using machine learning. *Clin Chem* 2020;66:1396–404.
4 **Li LT**, Haley LC, Boyd AK, *et al*. Technical/algorithm, Stakeholder, and society (TASS) barriers to the application of artificial intelligence in medicine: a systematic review. *J Biomed Inform* 2023;147:104531.
5 **Reddy S**, Rogers W, Makinen V-P, *et al*. Evaluation framework to guide implementation of AI systems into Healthcare settings. *BMJ Health Care Inform* 2021;28.
6 **Finlayson SG**, Subbaswamy A, Singh K, *et al*. The clinician and Dataset shift in artificial intelligence. *N Engl J Med* 2021;385:283–6.
7 **Boag W**, Hasan A, Kim JY, *et al*. The algorithm journey map: a tangible approach to implementing AI solutions in Healthcare. *NPJ Digit Med* 2024;7:87.
8 **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;181:1065–70.
9 **Osheroff JA**, Teich JM, Middleton B, *et al*. A roadmap for national action on clinical decision support. *J Am Med Inform Assoc* 2007;14:141–5.
10 **Martinez VA**, Betts RK, Scruth EA, *et al*. The Kaiser Permanente northern California advance alert monitor program: an automated early warning system for adults at risk for in-hospital clinical deterioration. *Jt Comm J Qual Patient Saf* 2022;48:370–5.
11 **Arndt BG**, Micek MA, Rule A, *et al*. Refining vendor-defined measures to accurately quantify EHR workload outside time scheduled with patients. *Ann Fam Med* 2023;21:264–8.
12 **Price WN**, Sendak M, Balu S, *et al*. Enabling collaborative governance of medical AI. *Nat Mach Intell* 2023;5:821–3.
13 **Yang HS**, Hou Y, Zhang H, *et al*. Machine learning highlights Downtrending of COVID-19 patients with a distinct laboratory profile. *Health Data Sci* 2021;2021:7574903.
14 **Vaid A**, Sawant A, Suarez-Farinas M, *et al*. Implications of the use of artificial intelligence predictive models in health care settings: A simulation study. *Ann Intern Med* 2023;176:1358–69.
15 **Liu X**, Glocker B, McCradden MM, *et al*. The medical Algorithmic audit. *Lancet Digit Health* 2022;4:e384–97.
16 **Jamali H**, Castillo LT, Morgan CC, *et al*. Racial disparity in oxygen saturation measurements by pulse Oximetry: evidence and implications. *Ann Am Thorac Soc* 2022;19:1951–64.
17 **Obermeyer Z**, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
18 **Xu J**, Xiao Y, Wang WH, *et al*. Algorithmic fairness in computational medicine. *EBioMedicine* 2022;84:104250.
19 **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.