



Universiteit
Leiden
The Netherlands

Leveraging AI-based prediction in perioperative and critical care: from model development to clinical implementation

Meijden, S.L. van der

Citation

Meijden, S. L. van der. (2025, May 6). *Leveraging AI-based prediction in perioperative and critical care: from model development to clinical implementation*. Retrieved from <https://hdl.handle.net/1887/4245255>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4245255>

Note: To cite this publication please use the final published version (if applicable).

Chapter 1

General Introduction

Over the past few years, Artificial Intelligence (AI) has become an increasingly integral part of our daily lives. It has transformed the way we listen to music, drive our cars, and search for information online [1,2]. AI mimics ‘intelligent behavior’ and sometimes even exceeds human capabilities when performing specific tasks [3]. Beyond the applications of AI in daily life, there is the hope that AI may contribute to resolving the healthcare crisis of soaring costs and staff shortages [4]. To this extent and to improve patient outcomes, many AI applications have been developed in the last 20 years, ranging from reducing administrative workload to supporting decision-making and diagnosing diseases [5,6]. Despite the rise in academic publications on the development of AI tools for clinical purposes [7], the number of applications implemented in clinical practice is small [8]. The lack of AI tool implementations also applies to the clinical domains of perioperative and critical care medicine, where patients face a high risk of complications and a large proportion of hospital expenses are made [9, 10].

Accurately predicting patient outcomes using AI may reduce the impact and costs of these complications by allowing more timely diagnosis and intervention, ultimately being of value to clinicians, patients, and society. This dissertation therefore focuses on the clinical applicability of AI-based prediction tools to impact perioperative and critical care. To contribute to the safe and effective implementation of AI in perioperative and critical care, the challenges related to human factors, data heterogeneity, model validity and performance, and ethical considerations are studied throughout the model development and implementation lifecycle.

1.1. Artificial Intelligence and Machine Learning in Healthcare

Artificial Intelligence is the umbrella term for a wide range of mathematical algorithms, including Machine Learning models. These models learn from various data types to perform specific tasks. Such tasks range from image segmentation for tumor treatment planning and providing patient conversation summaries to predicting patient outcomes [11-13]. This dissertation focuses on clinical outcome prediction (e.g., the risk of intensive care unit readmission or postoperative infection) to support healthcare professionals in decision-making. To perform this prediction task, the model is fed with thousands of historical patient records for which it is known whether the clinical outcome of interest occurred. As a result of this so-called ‘supervised learning’, the model can predict the probability of the outcome occurring as a prognostic tool for new incoming patients.

In perioperative care, postoperative infections greatly impact other adverse patient outcomes, hospital length of stay, quality of life, and costs [14]. To be able to identify patients with postoperative infections sooner, predicting the risk of infection after surgery may result in earlier diagnosis, and treatment, potentially resulting in reduced hospital length of stay and improved patient outcomes. The AI tool PERISCOPE was therefore developed by the start-up Healthplus.ai in collaboration with the Leiden University Medical Center. PERISCOPE provides a 7-day and 30-day risk prediction of infection directly after surgery and summarizes all relevant infection-related patient factors to support physicians and nurses in identifying patients at risk of infection.

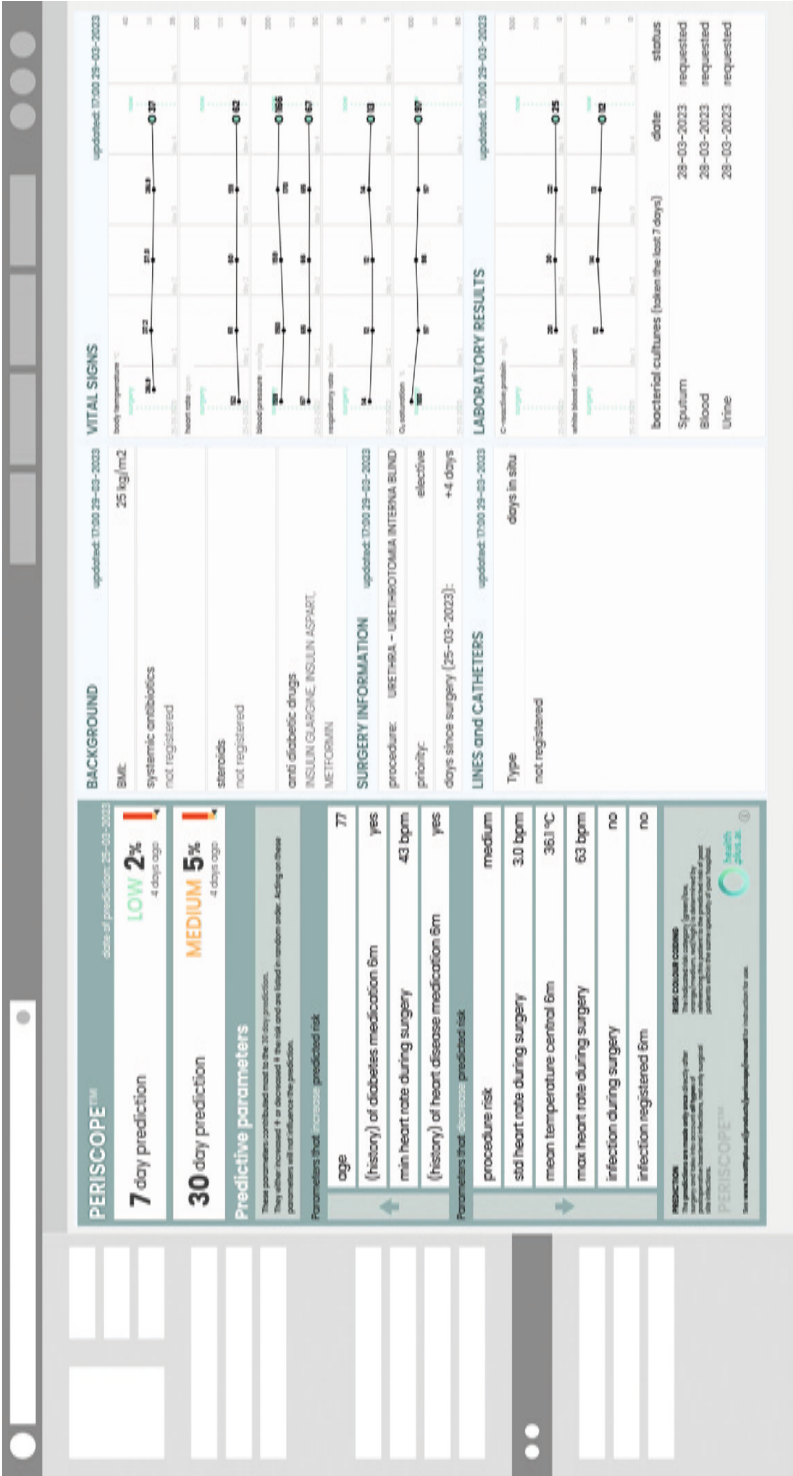


Figure 1: The PERISCOPE dashboard for a fictive patient integrated into the electronic health record.

To allow integration into the daily workflow, PERISCOPE is developed to be integrated into the local hospital electronic health record (EHR). See Figure 1 for an example PERISCOPE dashboard for a fictive patient. While the development of AI tools comparable to PERISCOPE is not novel, the road from development to implementation to be of clinical value to clinicians and patients needs further research. In this dissertation, the development and evaluation of PERISCOPE is a central theme, but the findings may often apply to other AI tools in the clinical domain.

1.2. The AI for healthcare development and evaluation trajectory

With the rise of AI for clinical applications, numerous frameworks have been established to guide the development and evaluation of safe and effective tools before implementation into clinical practice [7, 15]. The central aim of these frameworks is to ensure proper methodology steps for developing and validating AI tools. Ranging from data preparation and validation in different settings to software application development, clinical impact evaluation, implementation, regulatory compliance, and monitoring. Furthermore, the field of fair-AI focusses on identifying and mitigating biases in underlying data, predictions, and decision-making. Unfair use of AI mostly impacts minority groups and can increase healthcare disparities.

Phase	Key elements	PERISCOPE	
0 Preparations prior to model development	<ul style="list-style-type: none"> ➤ In collaboration with (clinical) stakeholders, define clinical problem, corresponding outcome to address, and the appropriateness to address this problem in with the use of AI ➤ Search for available models to address clinical problems ➤ Collect relevant observational patient data, accounting for potential biases ➤ Determine relevant regulations 	PERISCOPE was initiated in collaboration with surgeons, infectious disease specialists, and intensive care physicians to tackle one of the most impactful complications after surgery (postoperative infection). A broad outcome definition was established to identify postoperative infections based on readily available electronic health record data. Patient data was collected from different hospitals. PERISCOPE falls under (most importantly) the MDR and GDPR in Europe.	
1 AI model development	<ul style="list-style-type: none"> ➤ Prepare and preprocess data ➤ Development (i.e., training and hyperparameter tuning (optimization)) of AI model ➤ Internal validation, reporting on results. ➤ Software application development 	Observational EHR data was preprocessed and checked for missingness, biases etc. An XGBoost (tree boosting) model outperformed others and was optimized. Internal validation took place at the development site (The Leiden University Medical Center). The software application was developed, including an on-premise (i.e., local model integrated at the hospital) or cloud option, and the user interface (dashboard).	
2 Assessment of AI performance and reliability	<ul style="list-style-type: none"> ➤ External validation in geographical and temporal settings (with or without model updating), addressing model fairness. ➤ Prospective evaluation of model performance without showing the predictions to the user ➤ Usability testing with user interface ➤ Prepare for a clinical study ➤ Cost-effectiveness simulation study ➤ Proceed certification processes 	PERISCOPE was externally and prospectively validated in two geographically distinct hospitals, as well as temporarily to evaluate stable performance over time. Model updating was performed to ensure optimal model performance and to account for data changes. Usability testing was performed with end-users (physicians and nurses). Preparations are performed for clinical trials and cost-effectiveness studies. The validation and usability results served as clinical evaluation sources for the CE certification process.	
3 Clinical testing of AI	<ul style="list-style-type: none"> ➤ Compare to state-of-the-art (e.g., to physicians' predictive performance) ➤ Investigate the impact on clinical outcomes in clinical trials 	The predictions of PERISCOPE were compared to those of surgeons (in training) directly after surgery. Clinical trials are planned.	
4 Implementing and governing AI	<ul style="list-style-type: none"> ➤ Obtain regulatory approval or certification ➤ Safely integrate in workflow and EHR ➤ Monitor changes in input data and model performance to account for biases ➤ Investigate long-term impact on patient outcomes and costs 	PERISCOPE obtained CE-certification under the MDR. Preparations into the workflow and EHR, including user training, are completed as well as model and data monitoring strategies.	

Figure 2: AI development and evaluation trajectory key elements, specified for PERISCOPE, adapted from [7]. Bold items in the 'key elements' list are the specific parts covered in this dissertation. EHR = electronic health record, GDPR = general data protection regulation, MDR = medical device regulation.

This dissertation is mostly centered around developing the AI tool PERISCOPE. We aimed to cover the phases analogous to those of drug research for PERISCOPE, to enable safe implementation into clinical practice (Figure 2) [7]. As PERISCOPE is a commercial tool that needs to comply with the European Medical Device Regulation (MDR), we identified several gaps that could be added to the framework. These gaps include cost-effectiveness studies to evaluate the (potential) impact on patient outcomes and costs, software application development, user interface testing through usability research, integration into the EHR, and long-term evaluation of patient outcomes and costs. While all the key elements are conducted or planned for PERISCOPE, this dissertation does not cover all parts of the trajectory in full.

1.3. AI model evaluation metrics

Different evaluation metrics are used to quantify the performance of AI tools for predicting patient outcomes in phase I and II (Figure 2). To evaluate model performance in a population, the predictions are compared to a 'ground truth', i.e., the label. In PERISCOPE, this implies that the predictions of postoperative infections are compared to whether a patient developed an infection after surgery according to the established diagnostic criteria. Predictive models are usually evaluated amongst three domains (Figure 3): 1) discrimination (main metric: Area under the receiver operating characteristic curve (AUROC), 2) calibration (main metric: calibration slope and intercept), and 3) clinical utility (main metric: net benefit) [16].

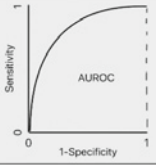
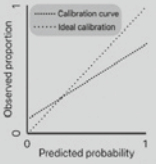
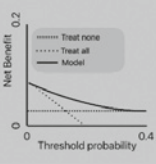
Domain	Main outcome measure(s)	Explanation	Example graph
<i>Discrimination</i>	Area under the receiver operating characteristic curve (AUROC)	Ability of the model to distinguish between patients with or without the outcome of interest. The curve is the function of sensitivity against 1-specificity over all classification thresholds. Area under the curve (AUROC) of 1 implies perfect discrimination and 0.5 not better than random guessing.	
<i>Calibration</i>	Calibration curve with: Calibration slope Calibration intercept	Curve representing the agreement between predicted probabilities and observed outcomes. E.g., in 100 patients with 10% risk estimates, 10 should be positive for the outcome. The slope is ideally 1, where > 1 indicates too moderate predictions and < 1 too extreme predictions. The intercept is ideally 0, < 0 means estimated risks are too low and > 0 too high.	
<i>Clinical utility</i>	Net Benefit	Decision analytic measure that puts benefits and harms on the same scale, resulting in a decision curve where net benefit is plotted as a function of threshold (i.e., decision) probabilities. At higher threshold probabilities, one is willing to accept less false positives ('worried about the intervention') than at lower threshold probabilities ('worried about the disease').	

Figure 3: Quantitative evaluation domains for clinical prediction tools with main outcome measure(s) and example graphs [19, 20].

A wide array of other model performance metrics has been determined [17, 18], this dissertation is centered around these three metrics as they are established to be most important for informing decisions with prediction scores between 0-100%. A specific classification threshold is chosen to determine when a prediction is positive or negative for classification models resulting in a binary outcome (e.g., to prescribe antibiotics or not). The choice of the classification threshold determines for example the sensitivity and specificity of the model as determined by the confusion matrix including the true and false positive and negative predictions [19].

1.4. Research questions

This dissertation's overall aim is to investigate the potential of AI in predicting outcomes in perioperative and critical care. The three specific research questions are as follows:

1. What are the human and data-related challenges associated with developing and integrating AI-based prediction models into clinical practice?
2. How can we ensure the validity and fairness of AI-based prediction models across various clinical settings?
3. How effective is AI in predicting patient outcomes, and how does its performance compare to physicians?

1.5. Data sources

Various data sources are used throughout this dissertation, including observational EHR data from three hospitals in the Netherlands and Belgium, including 253,010 surgical patient records in total, and the open-source EHR database MIMIC-IV [21]. The focus is on observational, routinely collected EHR data in tabular format. Furthermore, questionnaire data was collected amongst intensive care unit physicians, surgeons, and surgeons in training.

Table 1: Used data sources throughout this dissertation. EHR = Electronic health record.

Chapter	Data type	Source	Total dataset size
3	Questionnaires	Leiden University Medical Center and the Amsterdam University Medical Center	64 questionnaires
5	Observational EHR data	Leiden University Medical Center	59,106 procedures
6	Observational EHR data	Leiden University Medical Center, Radboud University Medical Center, Hospital Oost Limburg (Belgium)	253,010 procedures
7	Observational EHR data and questionnaires	Leiden University Medical Center	2,280 procedures and 544 questionnaires
8	Observational EHR data	Leiden University Medical Center, MIMIC-IV [12]	15,482 procedures

1.6. Outline: The AI development lifecycle

The chapters of this dissertation align with the phases identified in Figure 2. Initially, an introduction to AI in perioperative care with current applications, challenges, and opportunities is presented in **chapter 2**.

Phase 0: Preparations prior to model development

Before starting to develop AI-based prediction models for clinical application, the clinical problem should be defined in collaboration with relevant stakeholders, leading to a predicted outcome of interest. Whether AI is the appropriate technology to address the clinical problem, both data challenges as well as end-user acceptance of this new technology should be studied. As part of ‘phase 0: Preparations prior to model development’, **chapter 3** covers a pre-implementation survey study among intensive care unit physicians to investigate their perspectives on AI and their current clinical decision-making behavior. In **chapter 4**, an essential challenge of using AI in clinical practice is covered i.e., the difficulty in identifying patients for the outcome of interest in readily available EHR data. This chapter is focused on the different methods used to identify patients with (postoperative) infections in prediction and surveillance studies.

Phase 1: AI model development

The next step is to develop and optimize an AI model on high-quality EHR data. Preparing patient data and evaluating data quality requires a lot of data exploration, mapping, and cleaning. Furthermore, the predicted outcome of interest (in the PERISCOPE case that is ‘postoperative infections’) should be identified in the data where we face the challenge of under-registration of complications in EHRs. ‘Phase I: AI model development’ aligns with **chapter 5**, where a feasibility study of PERISCOPE is presented with internal validation results and a validation of the label definition of postoperative infections.

Phase 2: Assessment of AI performance and reliability

After the development of the AI model, extensive validation should be performed to assess AI performance and reliability in different settings. Traditionally, external validation of prediction models is performed by directly applying the model to new, unseen data. However, data heterogeneity and differences in patients and protocols may require updating of AI models to achieve comparable performance to internal validation results. Therefore, the external validation and updating of PERISCOPE (Phase II: Assessment of AI performance and reliability) is presented in **chapter 6** in a multicenter validation setting.

Phase 3: Clinical testing of AI

Whether the AI model outperforms the ‘state-of-the-art’, i.e., the established way of working including other predictive modeling tools is studied in the phase of clinical testing of AI. Ultimately, the impact of AI tools on patient outcomes and costs is investigated in clinical trials. While no clinical trials have yet been for PERISCOPE, ‘phase III: Clinical testing AI’ is touched upon in **chapter 7**, where the performance of PERISCOPE is compared to those of physicians.

Phase 4: Implementing and governing AI

The implementation of AI in clinical practice requires specific considerations in terms of e.g., regulatory approvals, integration in the workflow, and monitoring of model performance. Furthermore, while AI aims to improve patient outcomes, careful consideration of its ethical implications and governance is essential to prevent exacerbating health disparities. The fair application of AI (phase IV: implementing and governing AI) is discussed in **chapter 8** where two distinct use cases are presented along the theoretical frameworks provided for AI fairness evaluation.

1.7. Terminology

The field of AI is bothered with an abundance of definitions and terminology. Important definitions with synonyms throughout this dissertation are provided in Table 2.

Table 2: Definitions and synonyms

Definition	Synonym	Explanation
AI		Artificial Intelligence (AI) is the broad field of advanced algorithms that mimic human thinking and behavior (e.g., pattern recognition). AI can be used for many applications, but this dissertation focuses on predictive, supervised models that aim to predict a certain outcome of interest.
AI tool	AI-based clinical decision support system, AI system	An AI application aimed to be used to support human decision-making. Includes the AI model, as well as the software and user interface. Can be either used as a clinical decision support system to classify patients to support a specific decision, or to inform decision-making by providing a risk score between 0-100%.
Bias		In the context of AI, bias refers to systematic errors in the data or algorithms that can lead to inaccurate predictions or decisions.
CE-certification		CE certification indicates that a product meets the European Union's health, safety, and environmental protection requirements. The "CE" marking is mandatory for a wide range of products sold within the European Economic Area, including medical devices using AI software (Software as a Medical Device).
EHR		Electronic health records are real-time digital patient records. Information is only made available to authenticated users. They typically contain a broad view of a patient's medical history and are used in daily clinical decision-making.
Fairness		There is no consensus on the definition of AI fairness (see chapter 8). In general, the field of fair AI strives to reduce health disparities in minority patient groups.
Label		The annotation of data instances (e.g., patients) to be positive or negative for the outcome of interest (e.g., postoperative infection). Labels are necessary for a model to learn the relation between input data and the outcome of interest, as well as to validate model performance.

[continued on next page]

Table 2: *[continued]*

Definition	Synonym	Explanation
ML		Machine Learning (ML) is a subform of AI and includes for example tree-boosting and deep learning models. Instead of being explicitly programmed to perform specific tasks, ML algorithms identify patterns and relationships within data, allowing the system to improve its performance over time through experience.
Model	(Predictive) algorithm	Mathematical structures that are used to understand and make predictions based on data. It uses historical data to identify patterns and relationships to make predictions or decisions about new, unseen data. Can be a statistical or artificial intelligence-based (i.e., machine learning) model.
PERISCOPE		PERISCOPE is an AI tool developed by Healthplus.ai to predict the risk of postoperative infections within 7- and 30-days of surgery using routinely collected electronic health record data.
Updating		Model updating strategies adapt existing models to improve performance in new patient populations. This may include retraining the model as well as recalibrating it.
Validation		Model validation is performed to assess the performance (in terms of discrimination, calibration, and clinical utility (figure 3) in the development dataset (internal validation) and external datasets (external validation). External validation is performed to assess the generalizability of the model in a distinct temporal, geographical, or domain setting [21].

REFERENCES

1. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. ChatGPT: Exploring the Use of Artificial Intelligence in Ophthalmology by Comparing the Accuracy, Safety, and Readability of Responses to Frequently Asked Patient Questions Regarding Cataracts and Cataract Surgery. *Semin Ophthalmol*. 2024 Mar 22:1-8.
2. Oeding JF, Lu AZ, Mazzucco M, Fu MC, Taylor SA, Dines DM, Warren RF, Gulotta LV, Dines JS, Kunze KN. ChatGPT-4 Performs Clinical Information Retrieval Tasks Utilizing Consistently More Trustworthy Resources Than Does Google Search for Queries Concerning the Latarjet Procedure. *Arthroscopy*. 2024 Jun 25
3. Buttazzo G. Rise of artificial general intelligence: risks and opportunities. *Front Artif Intell*. 2023 Aug 25;6:1226990.
4. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*. 2020:25–60.
5. Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV, Al Muhanna D, Al-Muhanna FA. A Review of the Role of Artificial Intelligence in Healthcare. *J Pers Med*. 2023 Jun 5;13(6):951.
6. Secinaro, S., Calandra, D., Secinaro, A. et al. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis Mak* 21, 125 (2021).
7. 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 Feb;29(1).
8. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JY. Randomized Controlled Trials of Artificial Intelligence in Clinical Practice: Systematic Review. *J Med Internet Res*. 2022 Aug 25;24(8)
9. Kaye DR, Luckenbaugh AN, Oerline M, Hollenbeck BK, Herrel LA, Dimick JB, Hollingsworth JM. Understanding the Costs Associated With Surgical Care Delivery in the Medicare Population. *Ann Surg*. 2020 Jan;271(1):23-28.
10. Cecconi M, Spies CD, Moreno R. Economic sustainability of intensive care in Europe. *Intensive Care Med*. 2024 Jan;50(1):136-140.
11. Khalighi S, Reddy K, Midya A et al. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *npj Precis. Onc*. 8, 80 (2024).
12. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV. PhysioNet. Available from: <https://doi.org/10.13026/s6n6-xd98> (2021).
13. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, Gupta R, Blecker SB, Feldman J. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Netw Open*. 2024 Mar 4;7(3).
14. Toner A, Hamilton M. The long-term effects of postoperative complications. *Curr Opin Crit Care*. 2013;19(4):364-368.
15. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med*. 2022 Jan 10;5(1):2.
16. Erickson BJ, Kitamura F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. *Radiol Artif Intell*. 2021 May 12;3(3)
17. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022 Apr 8;12(1):5979.
18. de Hond AAH, Shah VB, Kant IMJ, et al. Perspectives on validation of clinical predictive algorithms. *NPJ Digit Med*. 2023 May 6;6(1):86.
19. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014 Aug 1;35(29):1925-31.
20. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006 Nov-Dec;26(6):565-74.
21. de Hond AAH, Shah VB, Kant IMJ, Van Calster B, Steyerberg EW, Hernandez-Boussard T. Perspectives on validation of clinical predictive algorithms. *NPJ Digit Med*. 2023 May 6;6(1):86. doi: 10.1038/s41746-023-00832-9. PMID: 37149704

