Possibilities and challenges for artificial intelligence and machine learning in perioperative care

Meijden, S.L. van der; Arbous, M.S.; Geerts, B.F.

ELSEVIER

Revalidation FOR ANAESTHETISTS RCA Revalidation matrix
Matrix codes: 1G01, 2A03, 2A04, 3A03, 3I00, 3J01

BJA EDUCATION · RESEARCH · DISCOVERY · 1923-2023 · BRITISH JOURNAL OF ANAESTHESIA

BJA Education, 23(8): 288–294 (2023)

doi: 10.1016/j.bjae.2023.04.003
Advance Access Publication Date: 21 June 2023

Possibilities and challenges for artificial intelligence and machine learning in perioperative care

S.L. van der Meijden[1,2], M.S. Arbous[2] and B.F. Geerts[1,*]

[1]Healthplus.ai-R&D B.V., Amsterdam, The Netherlands and [2]Intensive Care Unit, Leiden University Medical Centre, Leiden, The Netherlands

*Corresponding author: bart@healthplus.ai

Keywords: artificial intelligence; decision support systems; perioperative care

Learning objectives

By reading this article, you should be able to:

- Explain the basic concepts of artificial intelligence and machine learning and their applications in perioperative care.
- Assess the validity and clinical applicability of artificial intelligence and machine learning applications and the available literature.
- Discuss the current challenges and limitations of developing and using artificial intelligence and machine learning in clinical practice.

Key points

- Artificial intelligence (AI) is the field of computer science where algorithms are learned to perform cognitive tasks similar to humans.
- Risk prediction and image or video processing are the predominant high-potential usages of AI in perioperative care.
- When evaluating the use of an AI solution, first assess the discriminative performance, calibration properties and decision-curve analysis.
- Next, evaluate whether impact studies have been performed on clinical and economic outcomes.
- Barriers to the safe implementation of AI include variations in study quality resulting in potential biases and reduced generalisability; technological, regulatory and data-related challenges; and a lack of explainability of AI systems.

With the increase of digitalised working in healthcare, increased computing power and an increase in data availability over the last few years, there has been a growing interest in the application of artificial intelligence (AI) in medicine.

Artificial intelligence is a field within computer science that aims to allow computers and algorithms to perform cognitive tasks similar to humans by learning and recognising patterns in data. In medicine, where there are increasing amounts of healthcare data, the potential of AI is to aid repetitive tasks, diagnosis, prediction, drug discovery, personalised diagnosis and treatment, and decision support. As perioperative medicine accounts for a large part of hospital care and costs, and generates a large amount of data, AI has a high potential to be of value in this field.[1] The aim of this article is to provide an introduction to AI, an overview of its potential applications in perioperative care, an introduction to the assessment of the validity and clinical applicability of AI systems and to provide an overview of current challenges and pitfalls.

Bart F. Geerts MD PhD MSc MBA *is CEO at Healthplus.ai. Bart is a trained anaesthetist, intensivist and clinical pharmacologist. He is chair of the A.I. workgroup for the Dutch Federation of Medical Specialists. With Healthplus.ai, Bart is working on an EHR-integrated AI-system to predict postoperative infections.*

M. Sesmu Arbous MD PhD MSc *is an anaesthesiologist-intensivist and clinical epidemiologist at Leiden University Medical Center. She is a board member of the NICE foundation (National Intensive Care Evaluation).*

Siri L. van der Meijden MSc *is technical physician, a senior PhD researcher and data scientist at LUMC and at Healthplus.ai. Her research is focused on the clinical applicability of AI in perioperative and critical care, covering the full scope from development to implementation.*

Accepted: 21 April 2023

What is AI?

The term 'Artificial Intelligence' was first introduced by John McCarthy in 1955, but its application in medicine has taken hold in the last two decades since the rapid advancements in computing technology and cloud storage.[2] The definition of AI is 'a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behaviour, and with the creation of artefacts that exhibit such behaviour'.[3] Today, AI is used as the umbrella term for all algorithms and systems that can perform cognitive tasks such as prediction, clustering, classification, speech recognition and decision-making. Artificial intelligence is part of the field of data science which encompasses not only these advanced algorithms, but also disciplines such as statistics and data mining.[4]

Artificial intelligence also incorporates machine learning (ML), which includes deep learning (DL), where computers are trained to apply obtained knowledge from input data to newly presented data without being explicitly programmed to do so (Fig. 1). Most AI applications in healthcare are 'supervised' ML or DL algorithms, where the algorithm learns to predict or diagnose an outcome based on available data of patients for which the outcome is known using retrospective data. The algorithm learns to recognise (non-linear) patterns between the input data, for example, a chest X-ray image, and an outcome, for example, COVID-19. Supervised algorithms require a development dataset for which the outcome is known, to be able to learn relations between input and output variables. 'Unsupervised' learning is the field of AI where the outcome is unknown, and the algorithm is trained to cluster, for example, patient groups based on characteristics in order to discover new patterns in groups of patients or to reduce dimensionality. In the remainder of this article, we use the term AI to describe all types of self-learning algorithms (i.e. supervised and unsupervised).

Supervised ML, also known as prediction or classification algorithms, is one of the most often studied and implemented subforms of AI in medicine. Classification algorithms may perform numerous tasks on different types of data, such as tabular patient data stored in the electronic health record, image data from MRI or CT images to diagnose diseases or perform image segmentation, time-series data from EEG signals to predict delirium, or textual data on which natural language processing (NLP) is applied to automatically summarise or perform diagnosis. Besides informing physicians or patients on diagnosis and prognosis, AI is also used to guide treatment decisions in clinical decision support systems, for example, to prescribe medication. However, the use of clinical decision support systems is not exclusive to AI as many rule-based systems exist.[5]

It is debatable where the field of classical statistics ends and AI begins. A possible distinction between AI and statistics is that ML models can directly and automatically learn from

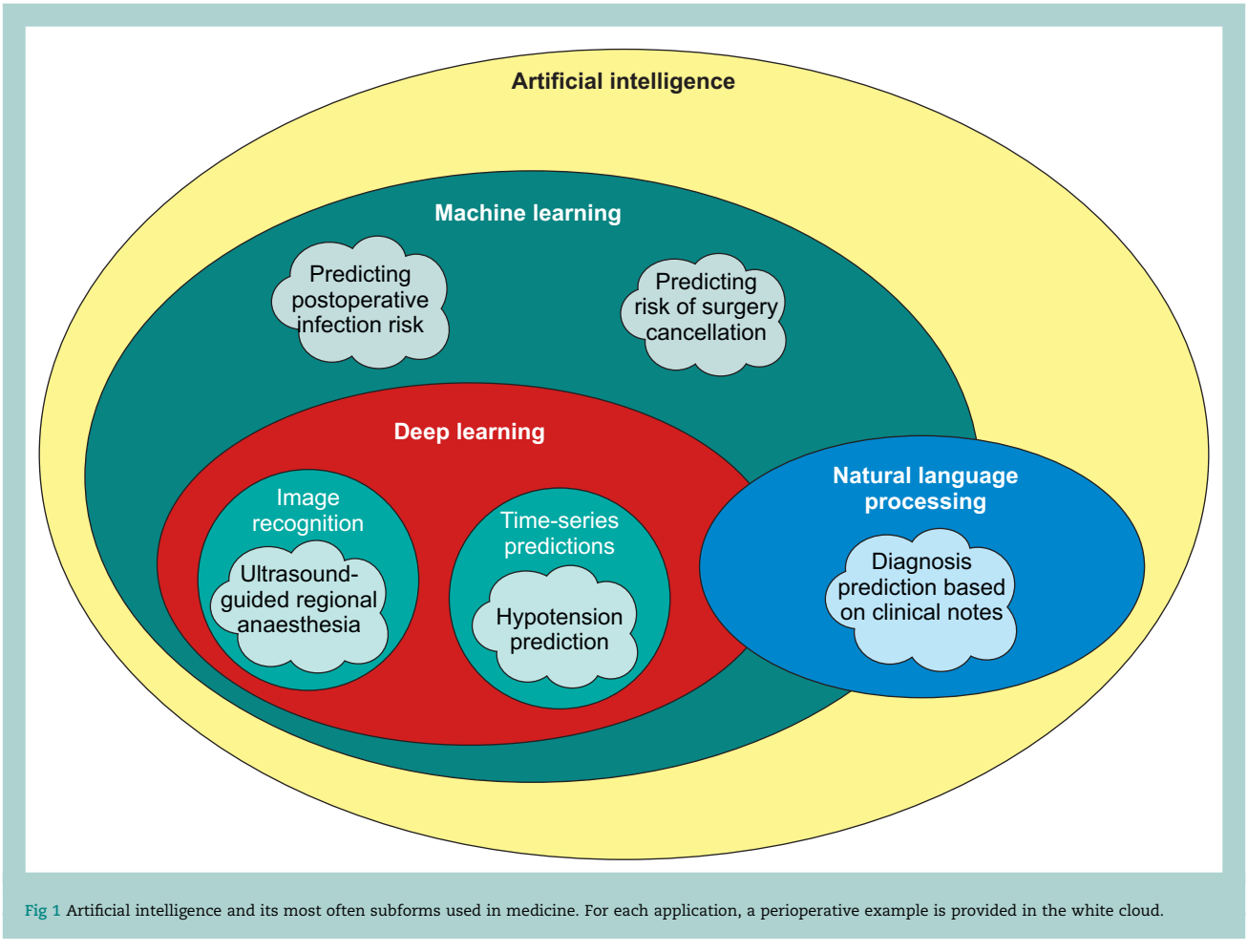

Fig 1 Artificial intelligence and its most often subforms used in medicine. For each application, a perioperative example is provided in the white cloud.

data, whereas regression models are more subject to human intervention and knowledge in model specification. With a more complex algorithmic structure, AI models are better at discovering non-linear relationships in data than classical statistics, but they require large amounts of data to be superior to classical statistics and to avoid overfitting. Overfitting is the phenomenon where models are not generalisable when applied to different datasets of comparable populations. In general, using AI in prediction models has not shown superior performance over logistic regression where they have been compared.[6] However, it must be noted that these findings of AI not outperforming logistic regression were on relatively small datasets with a limited number of predictors, whereas AI could outperform logistic regression when applied to larger datasets, in terms of, for example, establishing non-linear patterns. For more complex data structures such as images or textual data, this does not apply as traditional statistical models cannot be used. With increasing data availability, improved data quality and the potential use of other data sources such as imaging and textual data, AI may exceed statistical models in the near future. Compared with classical statistics, AI models are better suited for non-tabular data such as free-text clinical notes using NLP or image recognition using DL.

AI in perioperative care

The first attempts to use algorithms to aid the practice of anaesthesia date back to the 1950s when maintenance of anaesthesia was controlled using an EEG-guided closed-loop setting. For several decades, research into applied AI in the perioperative setting has been scarce, until recently. We distinguish the following perioperative domains in which AI applications may be of value: anaesthesia; surgery; prediction of the risk of surgical complications; operating room (OR) organisation; and nursing practice.

Anaesthesia

Several AI applications have been developed and become widely available recently in anaesthesia. An assistive AI device has proved to improve ultrasound image acquisition and interpretation for ultrasound-guided regional anaesthesia, which could result in the technology becoming more widely available as it may be applied by non-experts.[7] A more widely studied subject is the prediction of intraoperative hypotension, also known as the Hypotension Prediction Index (HPI) and Hemodynamic Stability Index (HSI).[8,9] A systematic review identified that HPI has the potential to improve haemodynamic management in terms of reduction in occurrence, duration, and severity of intraoperative hypotension compared with standard care, but more high-quality evidence is needed to prove this finding.[8]

The automation of the management of anaesthesia may benefit from airway evaluation and closed-loop anaesthesia assisted by AI devices.[10] Classical closed-loop systems for haemodynamic and pharmacological monitoring do not require AI systems, but using AI may result in fewer fluctuations, as future states may be predicted more accurately and more timely, and more input variables and closed-loop systems may be combined.[11] In the field of pharmacokinetics/pharmacodynamics (PK/PD) modelling, AI is more frequently used as it may combine different sources of input variables, but is compared with traditional PK/PD modelling more prone to overfitting and less interpretable.[12]

Surgery

With the increasing availability of sensor technology, robots and intraoperative imaging in the OR, more high-quality data are available for use by AI algorithms. AI may potentially benefit robot-assisted surgery by performing motion analysis to assess surgical skills, and in the recognition and classification of sutures and other surgical tasks. However, because of methodological caveats (limited data size and no external validation) in published studies, there is no proof that AI is already of benefit for robot-assisted surgery.[13] Other AI systems have been developed to perform video analysis or spectral light analysis to allow detection of cancer, tool detection, surgical phase recognition, workflow recognition and endoscopic guidance.[14]

Prediction of the risks of anaesthesia and postoperative surgical complications

Most AI applications for perioperative care have been developed in the field of predicting the risk of intraoperative complications, postoperative surgical and anaesthesia-related outcomes. The prediction of a certain outcome requires the ability to identify which patients had the outcome to train the prediction model on. This process is called 'labelling' and requires the outcome to be either identifiable based on electronic healthcare record (EHR) data or performed by manual review of EHR systems. Few outcomes are stored with high accuracy and completeness in EHR systems, limiting the application of prediction models based on large datasets. One often predicted outcome is the risk of mortality, which is accurately recorded in most EHR systems but has less clinical relevance in terms of being amenable to action by clinicians. Outcomes such as respiratory complications, cardiovascular complications, renal complications, sepsis, venous thromboembolism, hypotension after induction of anaesthesia, hypoxia after intubation, postoperative nausea and vomiting, and postoperative delirium can be predicted with high accuracy (often area under the receiver operating characteristic curve [AUROC] >0.90) based on clinical variables stored in HER systems.[1] Accurately predicting these outcomes will allow personalised treatments, monitoring and useful deployment of equipment. However, to be of assistance to physicians, AI systems should go beyond predicting clinical outcomes by providing recommendations for management that are tailored to an individual patient. The authors are currently working on an AI system that aims to tailor the management of postoperative infections.[15]

Organisation of the operating theatre

Apart from assisting physicians in clinical care, AI has the potential to improve hospital logistics. Predicting surgery duration and surgeries with a high risk of cancellation has been shown to be of benefit in operating theatre scheduling and usage.[1,14] Improving efficiency and scheduling will result in cost savings and decreased waiting times. Beyond the operating theatre, predicting postoperative hospital length of stay, risk of admission or readmission to critical care and readmission to hospital after discharge may benefit hospital logistics and costs.[1]

Nursing practice

Most AI applications in perioperative care are aimed at physicians, but AI is potentially equally of benefit for nursing

practice. As there is a high documentation load for nurses, AI could play a role in making these tasks more efficient. For perioperative nurses specifically, management of capnographic and false alarm management, clinical decision support systems for nursing diagnoses, faster detection of a patient's physiological changes and automated assessment of postoperative pain are examples of applications where AI plays a role.[16]

How to assess the validity and clinical applicability of an AI system

From the aforementioned examples of AI applications in perioperative medicine, it is clear that there are many potential applications for this technology. Only limited applications have been implemented in clinical practice, and it is important for a clinician to assess the validity and applicability before using AI in clinical practice.

Development and validation of AI systems

The development and validation of an AI algorithm are performed in different subsequent steps. The development dataset often consists of data from one hospital or one department. First, these data on which the algorithm is trained (i.e. the process to recognise patterns in the data to be able to make a prediction or classification on a new patient or dataset) need to be prepared. This data 'preprocessing' is often the most time-consuming step, as the data must be collected, cleaned for outliers, imputed for missing values and labelled for the predicted outcome. After preprocessing, the development dataset is split into a training dataset, for example 80% of all patients, and a test dataset, for example 20% of all patients. The model is trained and optimised over the training dataset, using for example cross-validation, and final performance is evaluated on the unseen patients in the test dataset. This process is called internal validation. The next step is to perform external validation on unseen data that may be prospectively collected from the same site (temporal validation), but ideally, the model is applied to another hospital's dataset (geographical validation). Performance often decreases in other settings, requiring retraining or recalibrating on data from the new site, or both. During validation, the model should ideally be compared with a 'baseline' or benchmark, for instance, a clinical risk score, or comparing predictions from physicians to those of the model.

After evaluating how the model performs on different datasets, the clinical and economic impact should be evaluated, but there are currently few impact studies compared with the number of model development studies, indicating a gap between the development and implementation of these systems.[17]

Performance evaluation

Classically, the evaluation of statistical and AI prediction models is based on their discriminative performance and calibration properties.[18] Discriminative performance is the ability of the model to distinguish between subjects with and without the outcome. Classification models give a binary outcome that is dependent on a probability threshold above which the prediction is deemed to be positive or negative. Using a lower threshold will increase sensitivity, at the cost of the positive predictive value. Conversely, using a higher threshold will increase specificity, at the cost of the negative predictive value. The overall measure of discriminative performance, which is independent of the chosen threshold as it plots the sensitivity against the false positive rate for all thresholds between 0 and 1, is the area under the receiving operating characteristic curve (concordance statistic or AUROC). An AUROC of 1.0 is a perfect classifier and an AUC of 0.5 not better than chance. As a rule of thumb, AUROC >0.7 is seen as acceptable and >0.8 is seen as a good classifier, but there are no clear cut-offs for AUROC to determine the clinical utility of the model.[19] It is important to note that other, traditional measures such as accuracy are not useful when the predicted outcome is 'imbalanced'. For example, if only 2% of patients develop hypotension and a hypotension prediction algorithm predicts with a 98% accuracy, it may be that the model predicts 'no hypotension' for all patients.

Often, AI systems evaluate and display the predicted probability between 0% and 100%, and perform a classification of potential outcomes. Calibration is the agreement between the predicted probability of the outcome and the actual proportion of patients that had the outcome with a certain prediction. An example of good calibration is that in 100 patients with a 10% predicted probability of mortality, 10 patients will actually die. Calibration is evaluated in calibration plots, where a slope of 1.0 and an intercept of 0 is optimal.

A measure beyond discrimination and calibration, as a measure of potential clinical utility of AI models, is the decision-curve analysis, where the net benefit of the model is calculated taking into account the number of false positives clinicians are willing to accept to find one true positive.[20] Calculating the net benefit of a model allows us to compare AI systems to standard practice and to traditional risk scores in a more clinically relevant setting.

Guidelines

There are no strict guidelines stating when an AI system is 'good enough' for use in clinical practice, as it may depend on the use case what sufficient performance is. However, it is important to assess whether the methodology of coming to certain outcome metrics is scientifically sound. The EQUATOR Network provides several guidelines for reporting and assessing the quality of AI-based diagnostic and prognostic prediction modelling studies, including TRIPOD-AI for development and validation studies (currently in the making), PROBAST-AI for risk of bias assessment, DECIDE-AI for early-stage clinical trials with AI systems and SPIRIT-AI and CONSORT-AI for clinical trial reporting on AI systems.[21] The number of guidelines for AI reporting and assessment is rapidly increasing, and different guidelines are available for different phases of AI prediction model development and implementation.[22]

It is not possible to cover all aspects of evaluating AI systems in this review, but we want to highlight some important aspects. First, determine whether the AI system was externally validated in a temporal or geographic setting to determine the robustness and generalisability to new settings and over time, or both. It must be noted that retraining or

recalibrating AI models may be necessary when entering new hospitals.[23] This implies that retrospective clinical data should be made available to the AI system manufacturer to validate, and if needed, retrain or recalibrate the model. Second, determine whether the AI model in question has been shown to be of additional value against a benchmark, state-of-the-art risk prediction models or healthcare professionals. Third, determine whether subgroup analyses have been performed to account for biases in (minority) patient populations. Fourth, evaluate published studies on the AI system in question according to the EQUATOR-guidelines.

Challenges and pitfalls of AI

Despite the increasing interest in AI in research and perioperative clinical practice, there are still several challenges and pitfalls in this emerging field. Considering the large number of publications regarding the development of AI systems, the number of clinical and economic impact studies on AI applications is still surprisingly low.[24] One of the most widely implemented clinical AI prediction systems, the Epic Sepsis Model, showed poor external performance and poor label quality, emphasising the need to incorporate domain knowledge and perform external validation and retraining of AI models.[25] The quality of research on clinical AI needs to be improved, as there is currently a high risk of bias as a result of small sample sizes, lack of comparison groups, lack of model, input and output variables, transparency and incomplete performance reporting.[24]

Aside from the shortcomings in study methodology, other challenges for AI are the 'black box' nature of the algorithms, which limits the explainability of the algorithms concerning the underlying factors of certain predictions. The field of explainable AI aims to improve explainability, but as a result of the complex structure and numerous input variables, full transparency is often not feasible. As the explainability of AI may have legal, medical and ethical consequences, one should be aware of the limitations of both '*post hoc*' explainability of AI systems that aim to uncover which input factors were of impact to the predictions and inherent explainability.[26] *Post hoc* explanations should not be used to assess whether a model predicted the outcome correctly, but may be used to identify biases.

The field of AI fairness studies how biases in AI research may be identified and how to best mitigate them.[22] Biases in clinical AI predictions may occur when the development dataset does not reflect all patient groups for which the system will be used. This may result in suboptimal model performance in minority and vulnerable patient groups such as ethnic minorities. Therefore, model performance should be evaluated on large datasets with sufficient sample size in different patient groups and one should be aware of the population on which the model was trained. This emphasises the need for transparent model reporting.

ChatGPT is one of the most groundbreaking and now well-known applications of AI; it is a large language model based on NLP, that will inevitably be used in the healthcare domain, for example, to automatically summarise EHRs and reduce administrative burdens. However, clinicians should be aware of the potential limitations and biases that may occur with this technology: answers produced by ChatGPT may be incorrect and should be interpreted with caution.[27]

As AI systems are built on data from hospitals and patients, one must be aware of the technological and regulatory challenges involved in developing, implementing and scaling AI systems. Lack of data standardisation and differences in coding of patient-related variables across hospitals and countries limit the scalability and generalisability to new settings. Additive technologies such as AI are dependent on the quality of the input data, as for example visual recognition may be limited by the image quality. Furthermore, data format and coding may change over time, requiring (post-market) monitoring of the system. Hospitals and manufacturers of AI systems often lack the resources that are needed to uniform EHR datasets and to enable the information technology (IT) infrastructure needed to deploy AI systems in a real-world setting. Along with these technological challenges for deployment, governance is needed to oversee safe deployment of AI in healthcare, including fairness, transparency, trustworthiness and accountability.[28] The high costs involved and lack of expertise in technical integration and governance are two burdens for the implementation of AI systems in clinical care.

The certification of AI systems under the Medical Device Regulation (MDR) in Europe and the Food and Drug Administration (FDA) in the USA is focused on the effective and safe deployment of AI systems in clinical practice.[29] For the UK market, AI systems must be registered with the Medicines and Healthcare products Regulatory Agency (MHRA). Next to the certification of AI systems as a medical device, the General Data Protection Regulation (GDPR) protects patients from their data usage without consent (with certain exceptions) and prohibits AI systems from performing stand-alone decision-making. Clinicians involved in developing AI systems should be aware of the regulatory requirements involved.

The focus of clinically applicable AI has long been on technological challenges, model performance, reporting, explainability and biases, but many other factors are of importance beyond the model itself.[30] Besides data and technology-related challenges, human factors play an important role in the successful integration of AI systems. Therefore, a multidisciplinary team should develop AI tools with technological, medical and methodological expertise.[22] As stated in the Topol review, successful implementation of AI in healthcare requires investment in people and technology, and co-development should take place with clinicians, patients and technologists from initial design to final implementation.[30] Good interaction between AI systems and their end-users is essential for safe and effective use and should be evaluated in early-stage clinical trials or usability studies.[31] Understanding the human factors involved in utilising AI systems and focusing on organisational change management is seen as one of the steps to close the gap towards the implementation of AI systems.

Summary

Different types of AI systems are increasingly being developed and studied in perioperative medicine, but implementation remains scarce. Artificial intelligence applications for perioperative medicine may be divided into clinical risk prediction models and decision support tools, image recognition, robot-assisted surgery, advanced closed-loop monitoring and predicting OR and hospital logistical outcomes such as

readmission and surgery cancellation. Before utilising an AI system in clinical practice, the performance of the model should be evaluated in external validation studies in terms of discriminative performance, calibration properties and clinical usefulness to properly assess safety and risk of bias, the need for recalibration and clinical utility. The risk of bias as a result of insufficient sample size, inappropriate choice of input variables and insufficient representation of minority groups in the training dataset asks for a critical appraisal of study results and subgroup analyses with data from multiple sites. Regulatory requirements from the GDPR, MDR, FDA or MHRA, or a combination, must be met for the safe and effective implementation of AI systems in clinical practice. Scaling AI systems to other hospital settings requires data standardisation and investment in technological infrastructure and governance. Beyond these regulatory and technological challenges, there is a need to study the clinical benefit of AI systems in decision-making and with respect to clinical and economic outcomes, but also to investigate what is needed for end-users to effectively use these new technologies. Therefore, human factors research, usability studies and clinical trials will need to be performed to bridge the gap towards implementation.

Declaration of interests

BG is currently CEO and majority shareholder of Healthplus.ai B.V. and subsidiaries. BG has also consulted for and received research grants from Philips NV and Edwards Lifesciences LLC. SvdM works as a data scientist and PhD at Healthplus.ai and LUMC. SvdM owns share options in Healthplus.ai.

MCQs

The associated MCQs (to support CME/CPD activity) will be accessible at www.bjaed.org/cme/home by subscribers to *BJA Education*.

References

1. Bellini V, Valente M, Bertorelli G *et al*. Machine learning in perioperative medicine: a systematic review. *J Anesth Analg Crit Care* 2022; **2**: 2
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25**: 44–56
3. Shapiro SC. Artificial intelligence. In: Shapiro SC, editor. *Encyclopedia of artificial intelligence*. vol. 1, 2nd Edn. New York: Wiley; 1992
4. Sarker IH. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Comput Sci* 2021; **2**: 377
5. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020; **6**: 3–17
6. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; **110**: 12–22
7. Bowness JS, Macfarlane AJR, Burckett-St Laurent D *et al*. Evaluation of the impact of assistive artificial intelligence on ultrasound scanning for regional anaesthesia. *Br J Anaesth* 2023; **130**: 226–33
8. Li W, Hu Z, Yuan Y, Liu J, Li K. Effect of hypotension prediction index in the prevention of intraoperative hypotension during noncardiac surgery: a systematic review. *J Clin Anesth* 2022; **83**, 110981
9. Rahman A, Chang Y, Dong J *et al*. Early prediction of hemodynamic interventions in the intensive care unit using machine learning. *Crit Care* 2021; **25**: 388
10. Gambus PL, Jaramillo S. Machine learning in anaesthesia: reactive, proactive predictive. *Br J Anaesth* 2019; **123**: 401–3
11. Wingert T, Lee C, Cannesson M. Machine learning, deep learning, and closed loop devices-anesthesia delivery. *Anesthesiol Clin* 2021; **39**: 565–81
12. Janssen A, Bennis FC, Mathôt RAA. Adoption of machine learning in pharmacometrics: an overview of recent implementations and their considerations. *Pharmaceutics* 2022; **14**: 1814
13. Moglia A, Georgiou K, Georgiou E, Satava RM, Cuschieri A. A systematic review on artificial intelligence in robot-assisted surgery. *Int J Surg* 2021; **95**, 106151
14. Birkhoff DC, van Dalen ASHM, Schijven MP. A review on the current applications of artificial intelligence in the operating room. *Surg Innov* 2021; **28**: 611–9
15. van der Meijden SL, van Boekel AM, de Boer MGJ *et al*. *Towards proactive surgical infection management: development and external validation of an AI-based prediction tool*. Dallas, TX: Paper presented at: 41st Annual Meeting of the Surgical Infection Society; 2022
16. Seibert K, Domhoff D, Bruch D *et al*. Application scenarios for artificial intelligence in nursing care: rapid review. *J Med Internet Res* 2021; **23**, e26522
17. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med* 2021; **47**: 750–60
18. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; **35**: 1925–31
19. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. *Lancet Digit Health* 2022; **4**: e853–5
20. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**: i6
21. Equator Network. Available from https://www.equator-network.org/(accessed 27 December 2022).
22. 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; **5**: 2
23. de Hond AAH, Kant IMJ, Fornasa M *et al*. Predicting readmission or death after discharge from the ICU: external validation and retraining of a machine learning model. *Crit Care Med* 2023; **51**: 291–300
24. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021; **23**, e25759
25. Habib AR, Lin AL, Grant RW. The epic sepsis model falls short—the importance of external validation. *JAMA Intern Med* 2021; **181**: 1040–1

26. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3**: e745–50
27. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023; **47**: 33
28. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020; **27**: 491–7
29. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; **17**: 195
30. Topol E. *The Topol review: preparing the healthcare workforce to deliver the digital future*. Feb 2019. Available from: https://topol.hee.nhs.uk/. [Accessed 10 March 2023]
31. Vasey B, Nagendran M, Campbell B *et al*. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: decide-ai. *Nat Med* 2022; **28**: 924–33