Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making

Hassan, Nehal; Slight, Robert; Morgan, Graham; Bates, David W.; Gallier, Suzy; Sapey, Elizabeth; Slight, Sarah

DOI:
10.1136/bmjhci-2023-100784

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Hassan, N, Slight, R, Morgan, G, Bates, DW, Gallier, S, Sapey, E & Slight, S 2023, 'Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making', *BMJ Health and Care Informatics*, vol. 30, no. 1, e100784. https://doi.org/10.1136/bmjhci-2023-100784

BMJ Health & Care Informatics

Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making

Nehal Hassan ,[1,2] Robert Slight,[2,3] Graham Morgan,[4] David W Bates,[5] Suzy Gallier,[6,7] Elizabeth Sapey ,[6,7] Sarah Slight[1,2]

To cite: Hassan N, Slight R, Morgan G, *et al*. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making. *BMJ Health Care Inform* 2023;**30**:e100784. doi:10.1136/bmjhci-2023-100784

Received 12 April 2023
Accepted 24 July 2023

[1]School of Pharmacy, Newcastle University School of Pharmacy, Newcastle Upon Tyne, UK
[2]Faculty of Medical Sciences, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK
[3]Freeman Hospital, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, UK
[4]School of Computing, Newcastle University, Newcastle upon Tyne, UK
[5]Department of General Internal Medicine, Harvard Medical School, Boston, Massachusetts, USA
[6]PIONEER Health Data Research Hub, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[7]Department of Health Informatics, PIONEER Health Data Research Hub, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Correspondence to
Sarah Slight;
Sarah.Slight@newcastle.ac.uk

ABSTRACT

Background Predictive models have been used in clinical care for decades. They can determine the risk of a patient developing a particular condition or complication and inform the shared decision-making process. Developing artificial intelligence (AI) predictive models for use in clinical practice is challenging; even if they have good predictive performance, this does not guarantee that they will be used or enhance decision-making. We describe nine stages of developing and evaluating a predictive AI model, recognising the challenges that clinicians might face at each stage and providing practical tips to help manage them.

Findings The nine stages included clarifying the clinical question or outcome(s) of interest (output), identifying appropriate predictors (features selection), choosing relevant datasets, developing the AI predictive model, validating and testing the developed model, presenting and interpreting the model prediction(s), licensing and maintaining the AI predictive model and evaluating the impact of the AI predictive model. The introduction of an AI prediction model into clinical practice usually consists of multiple interacting components, including the accuracy of the model predictions, physician and patient understanding and use of these probabilities, expected effectiveness of subsequent actions or interventions and adherence to these. Much of the difference in whether benefits are realised relates to whether the predictions are given to clinicians in a timely way that enables them to take an appropriate action.

Conclusion The downstream effects on processes and outcomes of AI prediction models vary widely, and it is essential to evaluate the use in clinical practice using an appropriate study design.

INTRODUCTION

Healthcare systems worldwide generate enormous amounts of patient-related health data, much of which is electronic in developed countries. There is growing interest among clinicians and healthcare staff in how they could use these data to support patient care.[1] Much of medicine is about anticipating and reducing risk, based on current and historical experiences. Predictive analytics in healthcare can help determine the risk of a patient developing a particular condition or complication, which can inform the shared decision-making process between clinicians and patients and improve patient satisfaction with their overall medical care.[2–7] With the new era of artificial intelligence (AI), clinical prediction tools can help personalise treatment and management decisions.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) framework was published to guide developing multivariate predictive models,[8] outlining what should be reported (eg, data sources, modelling techniques) when written up for publication.[9] However, a recent systematic review highlighted how these models' reporting has been rather poor since its publication.[10] TRIPOD also only focused on regression-based prediction models (although it can be applied to AI-generated approaches) and highlighted the need for more 'practical methods' for developing models more commonly used in healthcare (ie, supervised learning techniques).[11] The Consolidated Standards of Reporting Trials–AI guidelines were published in 2020 to help readers conceive studies with AI interventions; however, there was limited guidance on how these AI predictive models could be developed and usefully applied in clinical practice[12]; clinicians have sought further information on this.[1 13] Even if a newly developed AI model has a good predictive performance, this does not guarantee that it will be used in clinical practice or enhance clinical decision-making, let alone improve health outcomes.[14] The quality criteria important for evaluating AI predictive models were described in a recent scoping review; however, little information was provided on how such tools affect the clinical routine of physicians, which may vary per physician.[15]

The nine stages for developing and evaluating predictive AI models

Stage 1: clarifying the clinical question or outcome(s) of interest (output).

Stage 2: identifying appropriate predictors (features selection).

Stage 3: choosing relevant datasets.

Stage 4: developing the AI predictive model.

Stage 5: validating and testing the developed model.

Stage 5: presenting and interpreting the model prediction(s).

Stage 7: licensing the AI predictive model.

Stage 8: maintaining the AI predictive model.

Stage 9: ongoing evaluation of the impact of the AI predictive model.

It is vital to seek the input of a multidisciplinary team early when developing AI predictive models. This includes clinical specialists when deciding how the model could potentially enhance clinical decision-making and computing scientists when selecting the most appropriate algorithm(s).[16] Patients and providers should also be involved in deciding if the recommendations will be presented to them, including what, how and when information might be usefully presented (ie, content and alerts).[2 7 17] Taking each of these stages in turn.

STAGE 1: CLARIFYING THE CLINICAL QUESTION OR OUTCOME(S) OF INTEREST (OUTPUT)

The clinical question or outcome(s) of interest should be clearly defined from the onset. An example of a clinical question might be 'what is the likelihood of a patient developing type 2 diabetes mellitus (T2DM)?' to modify some of the patient's potential risk factors through lifestyle changes and/or prescribing medication.[18] It is essential to consider how we define T2DM here. Kopitar *et al* defined it as a fasting plasma glucose level of 6.1 mmol/L or higher without diabetes symptoms.[18] This definition makes the model a prognostic rather than diagnostic predictive model, given that it focuses on predicting a future health outcome. It is worth mentioning that this definition varies from those presented in different clinical guidelines[18] and can also change over time, highlighting the importance of model upgrading and maintenance. Another example of a clinical question could be 'what is the likelihood of a patient developing an infection and subsequent sepsis as an inpatient?'. Again, multiple definitions of sepsis could be used,[19–21] each varying in how closely aligned it is with the systemic effects of sepsis syndrome (see figure 1).[19 20] However, the choice of definition here is critical as it can directly influence the model performance measures, particularly specificity, which we will discuss later.[22] Clinicians should decide on the most accurate clinical definition for the predicted output, with the model upgraded to reflect any future changes to this definition.

STAGE 2: IDENTIFYING APPROPRIATE PREDICTORS (FEATURE SELECTION)

The second step involves identifying appropriate clinical predictors (features) related to the outcome of interest. Thus, if we take our sepsis-3 definition (figure 1), the next question relates to 'what clinical variables should we use for predicting sepsis?'. These clinical predictors will again depend on whether you want to develop a prognostic predictive model (which predicts the likelihood of sepsis occurring before the systemic inflammation process begins)[23] or a diagnostic predictive model (which early detects the likelihood of sepsis but after the inflammation process has already begun).[24] A review of the medical literature can help identify potential predictors that might be worth considering; 194 clinical predictors have been previously used to train machine learning algorithms for sepsis prediction, 13 of which were used across all 17 newly developed algorithms.[22] These 13 predictors contained a blend of non-modifiable (eg, age, gender) and modifiable (eg, blood glucose levels, blood pressure) predictors, the latter potentially increasing the applicability of the model in clinical practice.[22] It is important to consider here how these predictors have been defined and selected in previous studies, their source (ie, retrospective or real-time data) and whether any were excluded, thus recognising any inherent bias.[14 25] In terms of predictor type, numerical predictors should be given preference over categorical predictors, whenever possible.[8 26–28] A classic example is blood pressure, which can be recorded as a numerical (eg, 110 mm Hg) or categorical (eg, high,

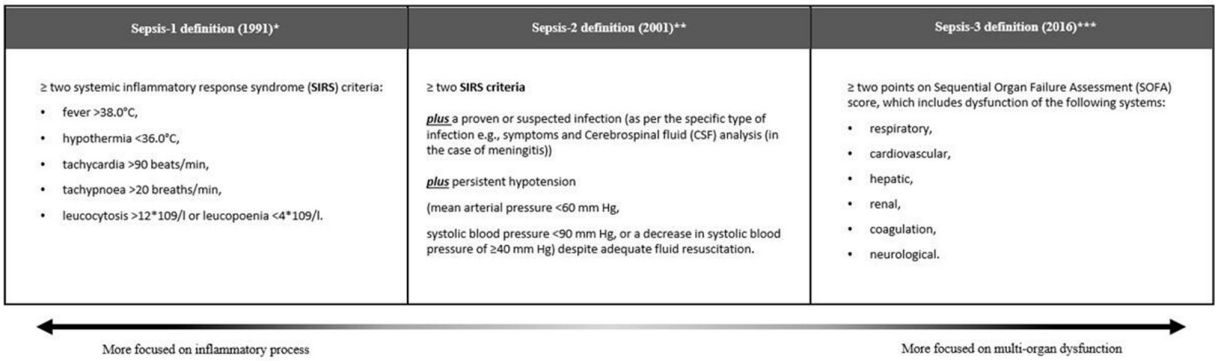

Figure 1 Different definitions of sepsis and their related clinical predictors. *Note that SIRS criteria are non-specific on the type of infection. **Note that suspected infection became a requirement to define sepsis. ***Note that clinical parameters are more specific to the systemic mechanism of sepsis.

normal, low) value. The latter assumes that a patient with systolic blood pressure of 110 mm Hg has the same level of hypotension as another patient with systolic blood pressure below 90 mm Hg, which is more characteristic of sepsis. In the T2DM example mentioned above, Kopitar *et al* screened the electronic health records (EHRs) of patients who went on to develop T2DM to identify potentially modifiable (eg, total cholesterol) and non-modifiable (eg, age) predictors.[18] EHR data can also allow exploring variables with predictive potentials that might not have been considered.[18]

The potential clinical predictors are then correlated to the model's outcome of interest (output) using either statistical methods or machine learning techniques.[29] Some predictors are likely to correlate strongly to the output but may be more suitable for a diagnostic rather than a prognostic predictive model. For example, the Sequential Oragn Failure Assessment Score (SOFA) Score, which reflects multiorgan dysfunction, will have a strong correlation with the sepsis diagnosis and would be more suitable for developing a diagnostic predictive model, whereas lipid profile will have a strong correlation to the diabetes prognosis and would be more suitable for developing a prognostic predictive model; this is because patients with established diabetes are likely to have hypercholesterolaemia.[18] We suggested using a 'blended approach' for predictor selection, where the predictors are correlated to the model's output and clinical input is also obtained on the choice to support its clinical application.[19 22 30]

STAGE 3: CHOOSING RELEVANT DATASETS

The existence, choice and access to relevant datasets often represent a limiting step for developing predictive AI models.[1 31] Thousands of organisations hold health datasets in the UK, so it can be difficult for clinicians, researchers and innovators to discover what datasets already exist.[32–34] Developers should first look at the relevance, data size and diversity of potential datasets; the proposed dataset should ideally represent the targeted population where the AI model is intended to be used to reduce the risk of inherent bias.[35] If the key outcome(s) of interest is unidentified, developers may have to decide how these available variables are used to define the key outcome. Researchers and innovators can search and request access to UK health-related datasets through 'the Gateway', a common entry point established by the Health Research Authority for nine UK-based health data research (HDR) hubs across the country.[33] These hubs include DATAMIND (mental health data), PIONEER (acute care data) and Discover-Now (primary care data), the latter being one of the largest primary care datasets in Europe. The UK HDR Alliance is also an independent alliance of leading healthcare and research organisations united to establish best practices for the ethical use of UK health data for research at scale.[34] In the UK, patients' information is protected by the General Data Protection Regulation and patients can refuse to permit their confidential data to be used through the national data opt-out service. Deidentification can be challenging, specifically with demographic variables, some of which can be important predictors when training the model. Removing them can potentially risk the efficiency of the model performance. A trusted research environment with anonymised patient data can be prepared for the clinician or researcher, once all the necessary ethical approvals have been obtained and the required training on data use and security completed.[36–39] Alternatively, data can be processed in a safe environment either at a hospital or university site; however, checks will need to be made on the safety of these environments and these data not approved for release if they do not meet the HDR UK five safes (safe people, safe projects, safe settings, safe outputs and safe data).[34] The diabetes risk prediction model mentioned above was developed using anonymised data collected from 10 diabetes screening clinics pooled in a single database.[18] Internationally, the Medical Information Mart for Intensive Care (MIMIC) database has clinical information from more than 40 000 patients admitted to critical care units at one tertiary centre (Beth Israel Deaconess Medical Centre, Boston, Massachusetts, USA). Similarly, healthcare professionals can freely access the dataset after completing appropriate data use and security training and signing a data usage agreement.[36 40] An important consideration is how these data have been collected and recorded. Numerical variables in the chosen dataset should ideally be collected and recorded synchronously.[37] The MIMIC database developers recognised this as a potential limitation of their dataset, with vital signs like heart rate and blood pressure recorded at different time points, thus potentially impacting the accuracy of the model.[36] Clinicians should help decide which dataset best represents the patient population that this model is intended to be used in.

STAGE 4: DEVELOPING THE AI PREDICTIVE MODEL

There are four major types of machine learning algorithms: supervised learning, unsupervised learning, semisupervised learning and reinforcement learning.[41] The choice of machine learning algorithm will depend on some factors, including the outcome of interest (ie, numerical or discrete value); the number of predictors; the 'shape' of the dataset (ie, size, completeness, uniformity); and the performance measures of the algorithm (ie, sensitivity, specificity, accuracy, area under the curve).[30] In the case of the latter, a number of algorithms may need to be tried first before finally deciding on the most suitable one or combination (ensemble model).[41] Supervised learning is commonly used for predictive models and can be subclassified into regression (ie, numerical output) or classification (ie, discrete output) algorithms.[42] The higher the number of predictors used, the more computational power needed to train the model and the higher the potential risk of overfitting.[42] An overfitting model is

a model that has high accuracy during the training phase, but lower accuracy during the validation and testing phase; potential ways to overcome this are described below.[26 42 43] It is important to remember, however, that strong computational correlations primarily depend on the entry values (eg, non-extreme vs extreme) and amount of missing data. Missing data can be potentially managed by statistical methods (ie, multiple imputations) or machine learning algorithms (ie, K-nearest neighbours), the choice of which will usually depend on the type and extent of missing information.[41 44 45]

Deep learning and artificial neural networks can perform better than conventional machine learning techniques. These networks act as a net of neurons that can identify patterns and correlations in a dataset so the model can self-learn from these patterns. The 'deep' refers to the depth of layers in a neural network and the performance measures of a deep learning model are directly correlated to the data size (ie, the larger the dataset, the better the model performance).[46 47] However, this can be challenging with rare diseases.[42 46 48]

Python is one of the most common programming languages for developing AI predictive models and is freely available.[49] After importing the dataset into programming software, you usually divide it into two portions: training the algorithm (70%) and internal validation (30%).[41 43] As described in stage 2 above, each predictor is then correlated to the outcome of interest (feature selection) using the training set and the performance measures of the algorithm calculated. This includes the specificity, sensitivity, receiver operating characteristic (ROC) curve and the area under the ROC curve (AUROC curve). The AUROC curve measures the distinctive ability of the algorithm to predict the outcome, with a value of >0.9 considered excellent.[22 50] AI systems learn to make decisions based on these training data, which may reflect human biases or social inequities, even if predictors such as race or gender have been removed.[51] It is beneficial to have the input of a programming specialist when preparing/revising the codes and judging the performance measures of any resulting models.

STAGE 5: VALIDATING AND TESTING THE AI PREDICTIVE MODEL

After developing the model, its predictive accuracy is reassessed using a validation dataset (internal validation) and again in a completely new, unseen dataset (ie, externally validated), ideally from another site. This comparison of performance measures is important for evaluating the risk of over/underfitting and widening the generalisability of the model, considering the diversity and representation of the patient population.[52] The testing phase usually involves running the model in a silent clinical environment, where the output is not shared with clinicians but compared with conventional clinical judgement and diagnosis. The T2DM prediction model was tested in a silent clinical environment over 6 months to assess its performance, before 'going live' to support clinical decision-making.[18] It is important to recognise that not all data are equal in quality; laboratory values may be coded differently or missing for some or all of an entire predictor in validation and training datasets. Complete case analysis is a method that can handle missing data and involves removing all missing patient cases; however, this requires a large sample size and may introduce selection bias. Alternatively, mean imputation can be used for missing numerical predictors, but can be sensitive to outliers (ie, extreme values).[53]

STAGE 6: PRESENTING AND INTERPRETING THE MODEL PREDICTION(S)

It is essential to consider how the model prediction(s) is presented to target users (patients/clinicians) and whether a recommendation accompanies it. The predicted probability (output) can be presented to users without any corresponding recommendations; this assistive presentation format allows clinicians to combine these predictions with clinical judgement.[54 55] In contrast, a directive prediction model provides the physician with a recommendation in addition to the predicted probability; this, in turn, can potentially increase the ease of use of the AI prediction model, especially if integrated into the electronic ordering system.[56 57] Clinicians should be informed of the underlying assumptions of the model, including which predictors were included and why, any inherent bias (eg, if groups are over-represented or under-represented in the training data) and how patients with specific outcome risk profiles might be affected by different recommendations.[14] For example, the inclusion of health costs as a proxy for health needs could potentially introduce racial bias, as less money is spent on black patients who have the same level of need in the USA; in other words, the algorithm could falsely conclude that black patients are healthier than equally sick white patients.[58] There is some evidence that clinicians in English-speaking countries have felt more legally supported when using decision support tools because they can provide documented evidence for the rationale behind their decisions.[59] Chua *et al* proposed an AI–human interface, where clinicians identify which patients might be eligible to use the tool, and the algorithm identifies (more accurately) which patients have serious illness communication needs and promotes upstream data collection.[7] Target users should contribute to the design of the model interface, ensuring that it is user-friendly, and any outputs and recommendations are easy to understand.

STAGE 7: LICENSING THE AI PREDICTIVE MODEL

In the UK, AI-based tools are classified as medical devices and therefore need the Medicines and Healthcare products Regulatory Agency (MHRA) approval. Before Brexit, approved tools required either the 'United Kingdom Conformity Assessed' (UKCA) or 'Conformité Européenne' logo to be marketed in Europe.[60] However,

from July 2023, only tools with the UKCA logo will be allowed to be marketed in the UK.[61] In Europe, AI-based software and tools are regulated by the EU Medical Device Regulation (EU MDR),[31 62 63] whereas in the USA, AI-based tools are regulated by the Food and Drug Administration.[64]

To licence an AI predictive tool in the UK, the MHRA must ensure that it complies with certain 'conformity assessment' standards, described by the National Institute for Health and Care Excellence (NICE) in 2018 and updated in 2021.[65] It is worth mentioning that NICE framework is designed for AI tools with fixed algorithms (ie, no change over time) rather than AI tools with adaptive algorithms (ie, continually and automatically change)[65]; the latter are covered by separate standards (including principle 7 of the code of conduct for data-driven health and care technology).[65] Higher-risk AI tools are classified as those that either target vulnerable patient populations, have serious consequences with errors or system failure, are used solely by patients without healthcare professionals' support or require a change in clinical workflow.[65] For EU-approved tools, the tool should comply with the general safety and performance requirements stated by the EU MDR.[66 67] Clinicians should be aware of the appropriate approvals that need to be obtained, especially with the growing adoption of these tools.

STAGE 8: MAINTAINING THE AI PREDICTIVE MODEL

Maintenance of the model and knowledge management are critical.[68] It may be necessary to update the model as populations, diseases and treatments change and include an expiry date.[68] In the UK, NICE data framework recommends a regression test be done when the model is updated to ensure that any new changes do not have a negative impact on its performance, reliability and functionality.[65] Model developers should also keep users (clinicians and patients) informed when releasing new model versions. In the USA, model recertification is needed when AI predictive models are updated,[15] although the US FDA is currently working on a framework that allows repeated updating of an AI predictive model without recertification through a change control plan.[69]

STAGE 9: ONGOING EVALUATION OF THE IMPACT OF THE AI PREDICTIVE MODEL

Introducing an AI prediction model into clinical practice can be considered a complex intervention; it usually consists of multiple interacting components including the accuracy of the model predictions, physician and patient understanding and use of these probabilities, expected effectiveness of subsequent actions or interventions, and adherence to these. A new framework has now replaced the UK Medical Research Council's guidance for developing and evaluating complex interventions. It focuses on recent developments in methods and the need to optimise the efficiency, use and impact of research.[70]

The downstream effects on patient outcomes of using an AI prediction model are not always predictable. For example, Kappen *et al* described no decrease in the incidence of postoperative nausea and vomiting, despite an increase in the administration of prophylactic antiemetics in the cluster-randomised trial of the AI prediction model (using an assistive presentation format).[56] This may indicate that either the predictive performance of the model was insufficient, the impact on physician decision-making was too small (eg, too few prophylactic drugs were administered despite high predicted probabilities), the antiemetic drugs were not as effective as thought, and/or patients chose not to take them.[56] Collecting additional data (observations and interviews) may help improve our understanding of these study results.

When designing an impact study before applying to licensing, a clinician needs to consider whether the complex intervention will have an individual effect on patients or whether it induces a more group-like effect.[56] A prediction model often aims to affect the clinical routine of a physician, which may vary per physician; this could lead to clustering of the effect per physician or practice (hospital) when the AI model use is compared across providers or practices.[19 31 56] After repeated exposure to the predictions, clinicians may also become better at estimating the probability in subsequent similar patients, even when those patients are in the control group.[19 31 56] This likely dilutes the effectiveness and thus impact of the model use.[48 56] As Kappen *et al* highlights, the effects of a learning curve may be minimised, though not completely prevented, by randomisation at a cluster level, for example, physicians or hospitals.[52 56]

CONCLUSION

We have provided a road map which clinicians and others developing algorithms can use to develop and evaluate AI predictive models to inform clinical decision-making. We described the nine stages, recognising the challenges that clinicians might face at each stage and practical tips to manage them. A 'blended approach' should be considered for clinical predictor selection, and the proposed dataset clearly represents the targeted population where the AI model is intended to be used. Comparing performance measures between the different training, validation and unseen clinical datasets are important for evaluating the risk of over/underfitting and widening the generalisability of the model. The format of the predictive model (assistive or directive) should be carefully chosen and designed. The maintenance of the model is important as populations, diseases and treatments change. The downstream effects on patient outcomes of using an AI prediction model are not always predictable, and it is important to evaluate its use in clinical practice using an appropriate study design.

Twitter Nehal Hassan @Nehal_Hassan_

Contributors RS, SS and NH conceived this article. NH, RS and SS led the writing of this manuscript, with all other coauthors (GM, DWB, SG and ES) commenting on subsequent drafts. All authors gave their approval for the final version to be published. NH is an early career researcher.

Funding The first author (NH) was awarded NUORS scholarship by Newcastle University, which covers tuition fees; this project is a part of a doctorate degree.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Nehal Hassan http://orcid.org/0000-0002-8302-5769
Elizabeth Sapey http://orcid.org/0000-0003-3454-5482

REFERENCES

1 Fontana G, Ghafur S, Torne L, *et al*. Ensuring that the NHS realises fair financial value from its data. *Lancet Digit Health* 2020;2:e10–2.
2 Hassan N, Slight RD, Bimpong K, *et al*. Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review. *Lancet* 2021;398:S80.
3 Flynn D, Nesbitt DJ, Ford GA, *et al*. Development of a computerised decision aid for thrombolysis in acute stroke care. *BMC Med Inform Decis Mak* 2015;15:6.
4 Wasson JH, Sox HC, Neff RK, *et al*. Clinical prediction rules. applications and methodological standards. *N Engl J Med* 1985;313:793–9.
5 Silvestrin TM, Steenrod AW, Coyne KS, *et al*. An approach to improve the care of mid-life women through the implementation of a women's health assessment tool/clinical decision support Toolkit. *Womens Health (Lond)* 2016;12:456–64.
6 Motorny S, Sarnikar S, Noteboom C. Design of an intelligent patient decision aid based on individual decision-making styles and information need preferences. *Inf Syst Front* 2022;24:1249–64.
7 Chua IS, Ritchie CS, Bates DW. Enhancing serious illness communication using artificial intelligence. *NPJ Digit Med* 2022;5:14.
8 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 2015;102:148–58.
9 Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, *et al*. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open* 2020;10:e041537.
10 Andaur Navarro CL, Damen JAA, Takada T, *et al*. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *Epidemiology* [Preprint] 2021.
11 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577–9.
12 Liu X, Rivera SC, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020;370:m3164.
13 Hercheui M, Mech G. Factors affecting the adoption of artificial intelligence in healthcare. *Global Journal of Business Research* 2021;15:77–88.
14 Kappen TH, van Klei WA, van Wolfswinkel L, *et al*. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagn Progn Res* 2018;2:11.
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;5:2.
16 Romero-Brufau S, Wyatt KD, Boyum P, *et al*. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. *Int J Med Inform* 2020;137:104072.
17 Nanji KC, Garabedian PM, Shaikh SD, *et al*. Development of a perioperative medication-related clinical decision support tool to prevent medication errors: an analysis of user feedback. *Appl Clin Inform* 2021;12:984–95.
18 Kopitar L, Kocbek P, Cilar L, *et al*. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep* 2020;10:11981.
19 Nemati S, Holder A, Razmi F, *et al*. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med* 2018;46:547–53.
20 Seymour CW, Liu VX, Iwashyna TJ, *et al*. Assessment of clinical criteria for sepsis: for the third International consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 2016;315:762–74.
21 Marik PE, Taeb AM. SIRS, qSOFA and new sepsis definition. *J Thorac Dis* 2017;9:943–5.
22 Hassan N, Slight R, Weiand D, *et al*. Preventing sepsis; how can artificial intelligence inform the clinical decision-making process? A systematic review. *Int J Med Inform* 2021;150:104457.
23 Faisal M, Scally A, Richardson D, *et al*. Development and external validation of an automated computer-aided risk score for predicting sepsis in emergency medical admissions using the patient's first electronically recorded vital signs and blood test results*. *Crit Care Med* 2018;46:612–8.
24 Nachimuthu SK, Haug PJ. Early detection of sepsis in the emergency department using dynamic Bayesian networks. AMIA Annual Symposium Proceedings 2012; American Medical Informatics Association, 2012:653
25 Riley RD, Moons KGM, Snell KIE, *et al*. A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ* 2019;364:k4597.
26 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Ann Intern Med* 2015;162:735–6.
27 Gupta A, Liu T, Shepherd S. Clinical decision support system to assess the risk of sepsis using tree augmented Bayesian networks and electronic medical record data. *Health Informatics J* 2020;26:841–61.
28 Khojandi A, Tansakul V, Li X, *et al*. Prediction of sepsis and in-hospital mortality using electronic health records. *Methods Inf Med* 2018;57:185–93.
29 Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 2019;19:64.
30 D'Ascenzo F, De Filippo O, Gallone G, *et al*. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *Lancet* 2021;397:199–207.
31 Beckers R, Kwade Z, Zanca F. The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. *Phys Med* 2021;83:1–8.
32 Reilly G, Varma S. Health data research innovation gateway. *ITNOW* 2021;63:60–3. 10.1093/itnow/bwab061 Available: https://doi.org/10.1093/itnow/bwab061
33 Health data research innovation gateway. Available: https://www.hdruk.ac.uk/access-to-health-data/health-data-research-innovation-gateway/ [Accessed 18 Mar 2022].
34 Gallier S, Price G, Pandya H, *et al*. Infrastructure and operating processes of PIONEER, the HDR-UK data Hub in acute care and the workings of the data trust committee: a protocol paper. *BMJ Health Care Inform* 2021;28:e100294.
35 Mehrabi N, Morstatter F, Saxena N, *et al*. A survey on bias and fairness in machine learning. *ACM Comput Surv* 2022;54:1–35.
36 Clifford GD, Scott DJ, Villarroel M. *User guide and documentation for the MIMIC II Database; MIMIC-II Database Version; Physionet.org*. Cambridge, MA, USA, 2009.
37 Liang Y, Abbott D, Howard N, *et al*. How effective is pulse arrival time for evaluating blood pressure? Challenges and recommendations from a study using the MIMIC database. *J Clin Med* 2019;8:337.
38 Symons JD, Ashrafian H, Dunscombe R, *et al*. From EHR to PHR: let's get the record straight. *BMJ Open* 2019;9:e029582.
39 Budhdeo S, Weerasuriya CK, Zhang J, *et al*. Interoperability in NHS acute trusts within england: a situation and capability analysis using freedom of information requests. *Health Informatics* [Preprint].
40 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
41 Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci* 2021;2:160.
42 Huang J-C, Ko K-M, Shu M-H, *et al*. Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Comput & Applic* 2020;32:5461–9.
43 Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.

44 Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods* 2018;15:233–4.
45 Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest? In: *International workshop on machine learning and data mining in pattern recognition*. Berlin, Heidelberg: Springer, 2012: 154–68.
46 Sarker IH, Colman A, Han J, *et al*. A behavioral decision tree learning to build user-centric context-aware predictive model. *Mobile Netw Appl* 2020;25:1151–61.
47 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. reply. *N Engl J Med* 2019;380:2589–90.
48 Battineni G, Sagaro GG, Chinatalapudi N, *et al*. Applications of machine learning predictive models in the chronic disease diagnosis. *J Pers Med* 2020;10:21.
49 Downloading python. Available: https://www.python.org/downloads/ [Accessed 7 May 2022].
50 Bloch E, Rotem T, Cohen J, *et al*. Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction. *J Healthc Eng* 2019;2019:5930379.
51 What do we do about the biases in AI? Available: https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai [Accessed 14 Jun 2022].
52 Ramspek CL, Jager KJ, Dekker FW, *et al*. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2021;14:49–58.
53 Tsvetanova A, Sperrin M, Peek N, *et al*. Missing data was handled inconsistently in UK prediction models: a review of method used. *J Clin Epidemiol* 2021;140:149–58.
54 Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. *BMJ* 2004;328:343–5.
55 Lobach D, Sanders GD, Bright TJ, *et al*. Enabling health care decision making through clinical decision support and knowledge management. *Evid Rep Technol Assess (Full Rep)* 2012:1–784.
56 Kappen TH, van Loon K, Kappen MAM, *et al*. Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol* 2016;70:136–45.
57 Bates DW, Auerbach A, Schulam P, *et al*. Reporting and implementing interventions involving machine learning and artificial intelligence. *Ann Intern Med* 2020;172:S137–44.
58 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.
59 Kennedy G, Gallego B. Clinical prediction rules: a systematic review of Healthcare provider opinions and preferences. *Int J Med Inform* 2019;123:1–10.
60 National Health Service. *Accelerating AI in health and care: results from a State of the Nation Survey*. London, United Kingdom: Department of Health and Social Service, 2018.
61 Guidance: medical device stand-alone software including apps (including Ivdmds) V1.08. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/999908/Software_flow_chart_Ed_1-08b-IVD.pdf [Accessed 5 Jun 2022].
62 Scheibner J, Sleigh J, Ienca M, *et al*. Benefits, challenges, and contributors to success for national eHealth systems implementation: a scoping review. *J Am Med Inform Assoc* 2021;28:2039–49.
63 Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020;3:118.
64 Larson DB, Harvey H, Rubin DL, *et al*. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. *J Am Coll Radiol* 2021;18:413–24.
65 Evidence standards framework for digital health technologies. Available: https://www.nice.org.uk/corporate/ecd7/resources/evidence-standards-framework-for-digital-health-technologies-pdf-1124017457605 [Accessed 2 May 2022].
66 Pesapane F, Volonté C, Codari M, *et al*. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* 2018;9:745–53.
67 Petersen E, Potdevin Y, Mohammadi E, *et al*. Responsible and regulatory conform machine learning for medicine: a survey of technical challenges and solutions. *IEEE Access* 2021;10:58375–418.
68 Solomonides AE, Koski E, Atabaki SM, *et al*. Defining AMIA's artificial intelligence principles. *J Am Med Inform Assoc* 2022;29:585–91.
69 Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (Samd). Available: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf [Accessed 10 Jun 2022].
70 Skivington K, Matthews L, Simpson SA, *et al*. A new framework for developing and evaluating complex interventions: update of medical research council guidance. *BMJ* 2021;374:n2061.