Understanding and interpreting artificial intelligence, machine learning and deep learning in Emergency Medicine

S Ramlakhan[1], R Saatchi[2], L Sabir[1], Y Singh[3], R Hughes[4], O Shobayo[2], D Ventour[3]

[1]Emergency Department, Sheffield Children's Hospital, Sheffield UK

[2]Electronics & Computer Engineering Research Institute, Sheffield Hallam University, Sheffield UK

[3]Faculty of Medical Sciences, University of the West Indies, Trinidad & Tobago

[4]Advanced Forming Research Centre, University of Strathclyde, Sheffield, UK

Word Count 2239

Keywords: Artificial Intelligence, Machine Learning, Neural Networks, Methods, Emergency Medicine

Contributorship - Conceptualization, Writing Original Draft – SR; Writing - Review & Editing - all other authors

Funding – none

Competing interests - SR is a decision editor for the EMJ

Data Sharing/Data availability – n/a

Ethics Approval Statements – n/a

Introduction

The field of Artificial Intelligence (AI) has been developing more prominently for over half a century. Innovations in computer processing power and analytical capabilities coupled with the availability of huge amounts of routinely collected data has meant that AI research and technology development has grown exponentially in recent years. The results of this growth can be seen in Emergency Medicine (EM) – with the FDA approving the first AI software as a medical device for wrist fracture detection in 2018. As of 2021, several more have been approved - for triage, x-ray identification of pneumothorax and notification and triage software for CT images. (1)

Between 2015 and 2021, there were over 500 publications indexed in MEDLINE involving AI in acute and emergency care, with more than half of these published within the last 2 years alone. There is recognition that AI technology can potentially play an important role in Emergency Department (ED) decision making, workflow and operations. (2–4) However, concerns with unstructured and often opaque reporting, inappropriate algorithm selection, proxy bias, data privacy and safety have led to calls for better standards for undertaking and reporting of research involving AI. (5–9) For practising ED clinicians, this will facilitate interpretation and understanding of AI research prior to model deployment or generalisation.

The aim of this paper is to serve as a primer for clinicians and researchers in understanding common AI methods as they relate to EM, and to provide a framework for interpreting AI research. A companion paper provides a more detailed exploration of the AI model building pipeline in an EM context.

What is the promise of AI for EM?

AI technology is seen in multiple aspects of day-to-day living – from e-mail spam filters, voice activated devices, suggestions from entertainment streaming services and social media to self-driving cars – all are powered by AI of varying complexities. The natural extension into healthcare, and EM in particular, is expected given the generalist, public-facing nature of the specialty.

AI has the potential to influence and improve ED triage and outcome prediction (10,11), forecasting and operations (3), diagnosis (2) and assessment of prognosis (12).

In addition, AI is facilitating the harnessing of new technology suitable for ED applications and research, such as natural language processing (13), radiomics (14) and machine vision (15). Large data repositories are being curated and leveraged to explore correlations between patient variables and urgent care outcomes. (10,16,17) Examples of recent studies with a reasonable rationale for using AI methods are summarised in Box 1.

However, the promise of AI must be tempered with acknowledgement of its evolving nature. This must be coupled with a realistic assessment of the acceptability and quality of current EM applications of the technology, especially compared with those which have been conventionally derived using traditional statistical methods. Certainly, in other specialties, AI models have not been shown to be superior to those derived by traditional logistic regression. (18)

The vast majority of AI EM diagnostic and prognostic studies use retrospective data, i.e., where the data was not collected specifically for the AI application (which also raises questions about the use of personal data without explicit consent). Importantly, less than 20% are externally validated in a traditional manner (19), with less than half of those externally validated according to AI standards. (8,20,21) Randomised trials are even rarer.

To compound matters, AI models can involve deep neural networks or similarly complex algorithms. The way that these models arrive at a decision cannot be interrogated in the clinical setting, making transparency and acceptability challenging to clinicians, patients and regulators. (22,23)

Although most published research compares AI with a clinician (19,24), this is arguably largely irrelevant. In EM practice, AI solutions have potential for secondary benefits by facilitating effective use of clinician time and reducing cognitive load. This can be through automating non-clinical tasks, allocating resources or prioritising workflows for efficiency. Rather than replacing clinicians or clinical tasks, AI's true promise lies in supplementing or improving on clinicians' care by harnessing the most appropriate human and machine abilities.

Box 1. Emergency medicine-relevant studies using artificial intelligence (AI)/machine learning (ML)

Fracture recognition (2)

Problem: Misinterpretation of limb x-rays by less experienced ED clinicians is not uncommon. Because contemporaneous radiologist reporting is not widely available, a system which automatically highlights abnormalities on an image may improve fracture detection.

Why AI? Deep neural networks (DNN) can be trained to identify areas (pixels) on an image where a fracture is likely. A DNN trained on a large number of images could improve the diagnostic accuracy of a clinician by providing a visual 'opinion' of the presence of a fracture. This principle can be extended to most imaging modalities.

Issues: Although the reported DNN-aided accuracy was better than the clinician-only interpretation, this was assessed using images on which the DNN had been trained. The DNN would therefore have had an advantage, as it would have 'seen' the images before, and biasing direct comparison with human interpretation. In addition, the clinical significance of a missed or delayed fracture diagnoses must be considered - with the actual patient benefit of any diagnostic improvement from AI clearly shown.

Waiting time estimation(17)

Problem: Knowledge of expected waiting times(WT) are helpful for ED patients, families and providers. Traditional triage or estimation based on historical time/day WTs are considered crude, and often do not reflect actual time spent waiting.

Why AI? Traditional statistical models derived from small datasets/single sites may not capture important predictors and are prone to overfitting. AI algorithms can look for important variables in large multi-centre routine datasets and derive predictive models which should theoretically give accurate estimates of WTs across a network or at single sites. This approach would be more resource efficient than each site developing their own model.

Issues: Even within a geographically defined area using a large dataset, estimates of WT from AI models were only accurate within 30 minutes of the actual wait time between 40% and 63% of the time. This may not be sufficient for operational benefit but may be for patient information purposes. External validation of models developed for one setting also appear to be less accurate at others, emphasising the importance of robust external evaluation and the need for site-specific customisation.

Predicting severe sepsis (25)

Problem: Early recognition and treatment of sepsis improves outcomes. Several sepsis scoring systems (EWS, qSOFA, SIRS) are used to predict severe sepsis and outcomes. However, these generally cannot leverage trends or correlate individual measures and have moderate predictive value. Automated Electronic Patient Record (EPR) sepsis alerting systems using these scores suffer from low specificity and false positives.

Why AI? Large datasets with physiological and outcome variables are available using EPR repositories. Conventional regression methods may not capture the interdependencies of individual parameters at specific times as well as relevant time-trends. The complexity of rationalising multiple variable combinations and the volume of data is challenging for human-designed analysis; however, gradient boosted ensemble ML algorithms can predict time-series events, and select important parameters/parameter combinations while handling class imbalance. They can therefore predict an outcome at multiple given time points, making it particularly useful in the ED where patients may present at different stages in disease progression. It also allows prediction, in this case, at up to 48 hours before the onset of severe sepsis.

Issues: The model demonstrated impressively high AUROCs, which was stable on external validation up to 6 hrs before onset of severe sepsis. However, performance at 48 hrs deteriorated markedly on external validation, suggesting overfitting (overtraining so it fits too closely to training data and is unable to generalise well for new data). For imbalanced datasets, AUROC is not an ideal metric for head-to-head comparison (given that the FPR changes minimally due to the high number of true negatives – see Fig 5). Whether predicting the onset of severe sepsis translates into improved outcomes still requires evaluation – with further external validation and/or a randomised trial.

Adverse outcomes from Covid-19 (26)

Problem: The rapid spread and variable disease progression of SARS-CoV-2 infections have meant that clinicians are working with an unprecedented lack of data, knowledge base or decision tools for predicting outcomes.

Why AI? The combination of individual patient variables which determine outcome are largely unknown. There is a paucity of quality data which considers disease outcomes at variable time points and at different stages of illness. In order to leverage available, good quality datasets to explore multiple predictor variables and their relative importance, an interpretable ensemble algorithm (Random Forest) can be used. This allows various models to be constructed with iterations of setting, clinical features, laboratory data and temporal measures, thus allowing flexibility in deployment and interpretability. In this case, the availability of a relatively small amount of good quality data to train and externally validate a model is preferable to larger poor-quality datasets from a single institution or region.

Issues: As with any AI model, deterministic capability is lacking. This is compounded by similar lack of pathophysiological knowledge of poor outcomes in Covid-19; for example - is obesity associated with poor outcomes due to immunomodulation or simply reduced respiratory capacity and is there therefore covariance and imputation issues which impacts on the model? A relatively small dataset makes the model potentially unstable, however external validation is reassuring other than for predicting ICU admission (likely due to the small numbers). As additional data becomes available, the model performance and parameters may change, and validation in more diverse settings would be warranted.

Artificial Intelligence, Machine Learning and Deep Neural Networks.

Much of the terminology relating to AI can be confusing as they are often used interchangeably – at least partly due to popular usage stemming from science fiction references. Even within the fields of computing and data science (not to mention philosophy!), there is no clear definition of AI, and this is at least partly attributable to the evolving nature of the technology. Box 2 contains a glossary to aid in understanding AI terminology and descriptions uses in this paper. From a clinician's perspective, *Artificial Intelligence* can be considered any human-like intelligence displayed by a machine (computer). This definition is valid insofar as humans demonstrate intelligence by learning from experience or observations, then use this knowledge to recognise, interpret and take autonomous actions when faced with similar situations.

Machine Learning (ML) is a subset of AI (Figure 1), and involves the use of algorithms which can identify patterns in data, learn from these patterns, improve with experience and come to conclusions when faced with new data – all without being explicitly programmed (27).

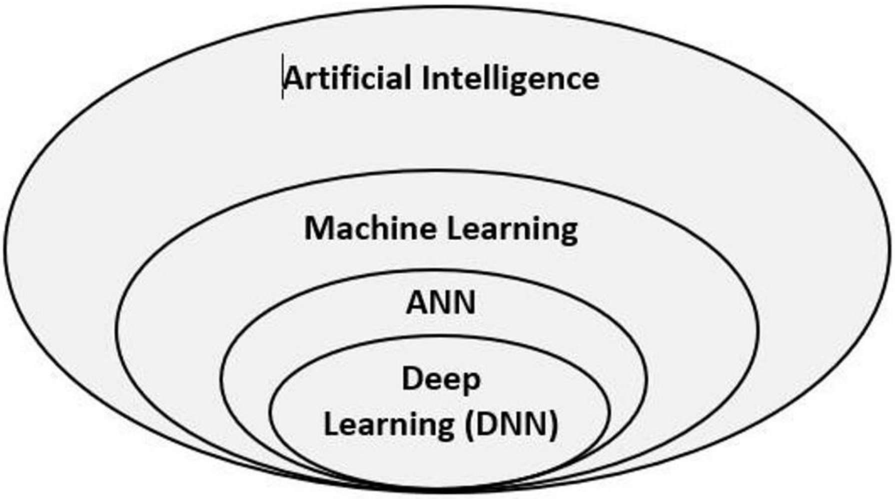


Figure 1 Relationship between artificial intelligence, machine learning, artificial neural networks (ANN) and deep learning (deep neural networks (DNN)).

In EM, this data is generally from images, electronic patient records and ED or population databases, but can reasonably include any type of data which can be interpreted by an algorithm. These algorithms usually consist of programming code (written in programming languages such as Python), mathematical formulae or combinations of these represented as pseudocode. There are several ML algorithm repositories/libraries where researchers can access “off-the-shelf” algorithms which have functions relevant to specific tasks. An AI algorithm which has ‘learnt’ from data is referred to as a model.

The term *Artificial Neural Network* (ANN) is used to describe an algorithm which consists of an input, output and often intermediate (hidden) layer(s). Data (input variables, also called features) is fed into the algorithm, which then weighs various combinations of these variables in the intermediate layer, and feeds forward an output which is dependent on being activated at set thresholds. As the number of hidden (intermediate) layers increase, this layer becomes deeper – hence *Deep Learning (DL) or* Deep *Neural Networks (DNN*) - and so does its ability to undertake more complex learning tasks. Output is passed from one layer to the next and is dependent on multiple inputs into multiple layers. DL models can also correct their own errors by backpropagation of data in order to refine and improve on performance.

The fundamental principle of the subtypes of AI is the requirement for learning from data. In ML, this is generally through either supervised or unsupervised learning. There are other types such as semi-supervised or reinforcement learning, which are less common in EM applications. Supervised learning is used in most ML applications and involves the use of labelled data (usually a type of variable as well as outcome of interest) fed to the algorithm so that it can learn the relationships and differences between the variables and outcome(s). Conversely, in unsupervised learning, unlabelled data is provided to the algorithm, and it is allowed to determine patterns and features of the data on its own, without human input. Unsupervised learning is used primarily in exploratory (clustering) algorithms, and generally requires much more data to train than supervised models.

Box 2. Glossary of AI/ML terms

Glossary

Algorithm - programming code/pseudocode or formulae which has been developed for a certain task or process

Model – a representation of a trained algorithm i.e., an algorithm which has learned from data

Ensemble – combinations of diverse simpler algorithms to improve overall performance.

Data Leakage – where data from outside the training set manages to leak into the model building (training) process, hence inappropriately influencing the model and its validation.

Feature – a measurable characteristic of the data or population of interest (commonly a variable)

Training set- data used by an algorithm to create or fit a model

Validation set- data used to evaluate the model fit while tuning hyperparameters, or to select features

Test set- data used to assess a model's performance on unseen data.

Ground Truth – analogous to Gold Standard

Underfitting – where the model does not learn adequately, and performs poorly in training and testing. It has a high bias and low variance.

Overfitting – where the model has learnt the training set too well, and thus performs well in training but poorly in testing. It has a low bias and high variance.

Bias – the difference between predicted and actual values

Variance – how varied the predictions are between different sets of input

Parameter- a variable whose value is calculated from the data and forms part of the model itself. It is independent of the analyst and cannot be manually adjusted.

Hyperparameter- a setting or weight whose value cannot be calculated from the data and is external to the model. It can be tuned by the analyst.

Loss function- a measure of how the predicted output of each training instance differs from the actual outcome.

Regularisation –modifications to a learning algorithm intended to reduce its generalization error but not its training error. (28) This is usually by penalising or limiting weights or early stopping of training for example. This can be achieved by algorithms using least absolute shrinkage and selection operator (LASSO) or Ridge regression for example.

Decision Tree – a type of algorithm which splits data based on probabilities or attributes in a branching pattern, until an output is determined.

Support Vector Machine – a classification algorithm which separates classes by finding a (hyper)plane of separability between them.

Naïve Bayes – uses Bayes theorem i.e., how pre-event variables influences post-event outcome probability. It is naïve as it assumes all variables are independent.

K-Nearest Neighbours – an algorithm which predicts an outcome by finding the k-nearest instances of the input variable and averaging their outcomes.

Random Forest – a decision-tree based bagging **ensemble** model. It *Random*ly selects variables/data and builds a *Forest* of multiple decision trees.

AI or Conventional methods?

ML/AI uses some biostatistical methods and measures that will be familiar to Emergency Physicians. Some algorithms utilise methods based on linear regression, decision trees and

Bayesian reasoning, amongst others and these can usually be represented mathematically. Significantly, AI/ML can go further, by using algorithms to leverage previously unrecognised variables (or features) which are different to the human-determined variables used in traditional statistics. In addition, models built on these algorithms can harness non-linear relationships between variables and outcomes, which make them well suited to identifying complex or unapparent data inter-relationships. They are therefore better at exploratory predictive (regression or classification) modelling with large or computationally heavy datasets, whereas classic statistical approaches are arguably better for confirmatory, hypotheses-light analyses with relatively smaller datasets. Of note, the process of evaluation and comparison of different models is almost always based on conventional statistical methods. These methods can often be adapted for ML models, commonly using programming code written specifically to simplify some types of analysis.

AI models learn from the data they are given – the algorithms can incorporate multiple data inputs and make predictions based on their analysis of the relationships between inputs and outputs. Just as an experienced emergency physician (EP) who has looked at thousands of ECGs can interpret an ECG at a glance by what would be considered tacit knowledge, an AI/ML model should theoretically be able to similarly interpret specific ECGs - provided that it has learned to do so by being trained on a sufficient number of similar ECGs for it to discern the relationships between the labelled ECG 'variables' and diagnosis. It can do this without explicitly learning – for example, it will not 'learn' the voltage criteria for LBBB, the electro-pathological reasons for the pattern or how it affects STEMI diagnosis. It may however, still be able to diagnose LBBB or a STEMI. This analogy highlights an often-ignored aspect of AI – it has no deterministic capability and cannot discern physical, physiological or pathological determinants or constraints. It can therefore find spurious relationships which may lead to inappropriate predictions. With large datasets with multiple combinations of input variables, it is not difficult to see how 'statistically' significant relationships can be derived purely by chance - so called 'p-hacking'.

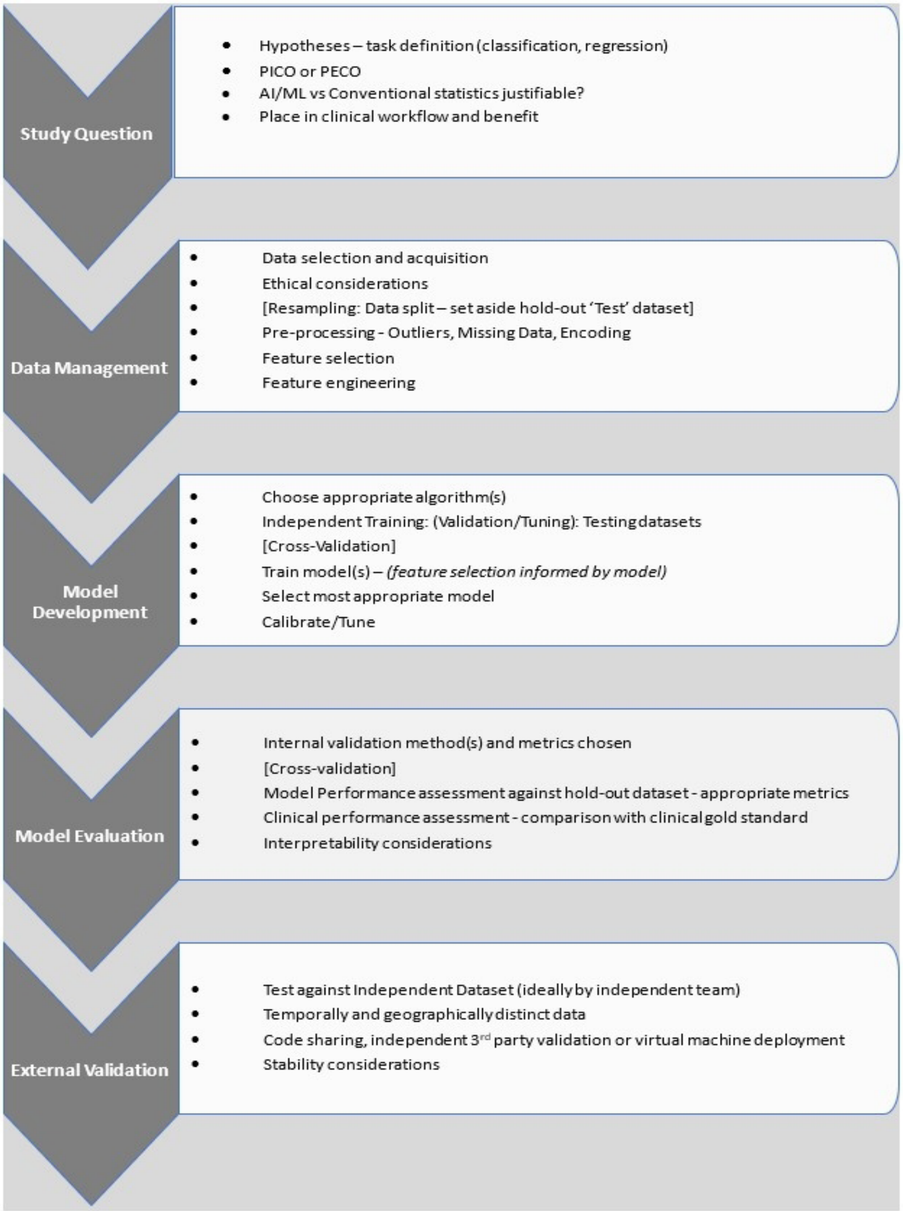
Another driver for using AI/ML stems from the type of data that is being made available. Besides large-scale routinely collected data, other types of data which were previously relatively untapped due to their complexity are well suited for AI algorithmic analysis. These include diagnostic imaging, laboratory or physiological data and free text from clinical notes.

Understanding the differences between AI and traditional statistical methods also comes from appreciating the focus of the main proponents of AI development – usually in data rich technology and marketing industries - where the performance of a model takes priority over describing how it works. ML/AI is focussed on real life predictions or decisions based on new or dynamic data, whereas traditional statistics is more focussed on understanding and articulating data relationships using established statistical theories and assumptions. This difference is critical – the point of creating an AI model is not to show how it performs in vitro, but to deploy it in the real world (clinically or operationally). The ultimate goal therefore, is developing a model which can be confidently generalised.

Stages in developing an AI (Machine Learning) Model

EM research in AI is almost entirely focussed on ML models, and therefore the focus henceforth will be on ML, although the framework and principles will also apply to almost any AI model. There is significant variability in the reporting and conduct of ML research, which can make interpretation challenging – particularly when some of the models used are by their nature ‘black boxes’. (22) A Machine Learning model should be developed systematically, with the same structure and clinical focus as any traditional EM clinical research. The steps in model development from conception to deployment are outlined in Figure 2.

An understanding of model development is useful when interpreting an AI research paper. More detail on this process is discussed in a linked companion paper (reference to second paper).



Study Question

- Hypotheses – task definition (classification, regression)
- PICO or PECO
- AI/ML vs Conventional statistics justifiable?
- Place in clinical workflow and benefit

Data Management

- Data selection and acquisition
- Ethical considerations
- [Resampling: Data split – set aside hold-out 'Test' dataset]
- Pre-processing - Outliers, Missing Data, Encoding
- Feature selection
- Feature engineering

Model Development

- Choose appropriate algorithm(s)
- Independent Training: (Validation/Tuning): Testing datasets
- [Cross-Validation]
- Train model(s) – *(feature selection informed by model)*
- Select most appropriate model
- Calibrate/Tune

Model Evaluation

- Internal validation method(s) and metrics chosen
- [Cross-validation]
- Model Performance assessment against hold-out dataset - appropriate metrics
- Clinical performance assessment - comparison with clinical gold standard
- Interpretability considerations

External Validation

- Test against Independent Dataset (ideally by independent team)
- Temporally and geographically distinct data
- Code sharing, independent 3rd party validation or virtual machine deployment
- Stability considerations

Figure 2. Steps in artificial intelligence (AI) model development. Square brackets denote steps which are sometimes described during data management, model development or evaluation. Feedback and iterations of preceding steps are expected. ML, machine learning. PICO/PECO - Population/ Intervention (Exposure)/Comparison/Outcome

Interpreting an AI/ML paper

The ability to interpret and understand ML methodology is vitally important to EM clinicians. Key considerations when appraising an EM AI/ML paper are summarised in Box 3.

There are concerns that AI in clinical medicine is overhyped and suffers from poor methodology, reporting and transparency. (9) This is reflected in the findings of systematic reviews which have highlighted incomplete and non-standardised methods and reporting in ML studies. (6,18,19,24,29) This has led to calls for reporting standards to address these shortcomings, with extensions to current reporting guidelines (8,30,31) developed along with suggested general (20,32) and specialist (21) frameworks and checklists being proposed.

There are several ML repositories and libraries which allow researchers to use ML techniques out-of-box using accessible platforms such as R(r-project.org). As with any novel tool, there is a risk of inappropriate or clinically disconnected use. In addition, internal validation processes are heterogeneous with heuristics and analyst preference regarding algorithms, procedures, metrics and validation being common. Because of the interdependencies of various steps in model development, bias or poor choice of processes at any part of data selection, preparation or model development will have a knock-on effect, and impact negatively on the performance of the final model(s). This and other model building considerations are explored and expanded upon in a linked paper, along with discussion of the of the pitfalls to be aware of when interpreting AI model performance.

Box 3. Key considerations when appraising an EM AI/ML paper

		Key Considerations
INTERNAL VALIDITY	Study question	i. Is the study question and aim clearly stated? ii. Is a comparison with the current EM practice baseline and rationale for improvement explicitly stated? iii. Is the use of ML justifiable as opposed to traditional statistical methods? iv. Is the strategy for model development, validation and evaluation clearly described in a clinical EM context?
	Baseline Data	i. Is sufficient good quality data available? Has a sample size assessment been made? ii. Is data representative of the population and setting in which the model is being deployed? iii. Has data pre-processing been clearly described and handled in an EM applicable manner? Is this reproducible? iv. Have redundant (noisy or collinear) variables been identified and addressed? v. Has any baseline class imbalance been addressed, and is the method appropriate to the intended EM task? vi. Is Ground Truth valid and reflects the EM “gold standard”
	Data split/Resampling	i. Has the approach to data resampling been clearly described? *(is a data flow diagram provided)* ii. Have internal and external validation test sets been clearly identified, and are they as independent as possible? iii. Is the data split clear, rationalised and stratified if appropriate? iv. Has data leakage been avoided?
	Algorithm selection	i. Has the rationale for selecting candidate algorithms been explained in relation to the study aims? Has overfitting been explicitly addressed? ii. Are candidate algorithms appropriate to EM and representative of a range of complexities? iii. Has a variable selection procedure been described? Are chosen variables plausible in the EM context?

		iv. Have the metrics for model evaluation been defined and rationalised?
	Model Validation	i. Are the model evaluation metrics appropriate to the task and clinical question? If not, has model performance been reported with a clinically applicable (EM) metric? ii. Are estimates of model variance reported (from cross validation/bootstrapping)? iii. Has tuning of model hyperparameters been undertaken, and have the settings been reported? iv. Have model calibration and discrimination been reported, and have these been tuned?
EXTERNAL VALIDITY	Reproducibility	i. Has a statement regarding code (and data) availability been included? ii. Are all steps in the model development described in sufficient detail to allow independent replication of the model pipeline? iii. Has interpretability been explicitly and reasonably addressed (e.g., by model visualisations)?
	Generalisability	i. Has external validation on a geographically (and temporally) independent test set been undertaken? ii. Are performance metrics consistent with the estimates of spread obtained during internal validation? iii. Have the reasons for good or poor performance on external validation been objectively explored? iv. Is the final model suitable for deployment in the ED? Does it have high computational or resource requirements? v. Has an assessment been made of the potential for exacerbation of bias by the deployed model?

This paper should provide the reader with conceptual insight on how AI models are developed and a framework to interpret AI applications as it relates to EM practice and research. There is scope for synergism between AI and EM clinicians to maximise efficiency within EDs by utilizing these methods in workflow, triage processes, automating certain tasks, and diagnostic applications. There is likely to be an ongoing increase in the number of studies describing these methods. Therefore the ability to critically appraise these papers and careful awareness that despite exciting promise, AI models still have to be methodologically sound and applicable to real world practice. The true benefit of AI is likely in assisting clinicians rather than replacing them.

Twitter Shammi Ramlakhan @shammi_ram
Contributors Conceptualisation, writing original draft: SLR. Writing—review and editing: all other authors.
Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.
Competing interests SLR is a Decision Editor for the EMJ.
Patient consent for publication Not applicable.
Ethics approval This study does not involve human participants.
Provenance and peer review Not commissioned; externally peer reviewed.
ORCID iDs
Shammi Ramlakhan http://orcid.org/0000-0003-1792-0842
Lisa Sabir http://orcid.org/0000-0001-6488-3314

References

1. 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 Sep 11;3(1):1–8.

2. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci. 2018 Nov 6;115(45):11591–6.

3. Nas S, Koyuncu M. Emergency Department Capacity Planning: A Recurrent Neural Network and Simulation Approach [Internet]. Vol. 2019, Computational and Mathematical Methods in Medicine. Hindawi; 2019 [cited 2021 Jan 30]. p. e4359719. Available from: https://www.hindawi.com/journals/cmmm/2019/4359719/

4. Shafaf N, Malek H. Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. Arch Acad Emerg Med [Internet]. 2019 Jun 3 [cited 2021 Feb 4];7(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6732202/

5. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019 Mar;28(3):231–7.

6. Yusuf M, Atal I, Li J, Smith P, Ravaud P, Fergie M, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. BMJ Open. 2020 Mar 1;10(3):e034568.

7. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ [Internet]. 2020 Mar 20 [cited 2020 Dec 1];368. Available from: https://www.bmj.com/content/368/bmj.l6927

8. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet. 2019 Apr 20;393(10181):1577–9.

9. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. N Engl J Med. 2017 Jun 29;376(26):2507–9.

10. Kwon J, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. PLOS ONE. 2018 Oct 15;13(10):e0205836.

11. Shung DL, Au B, Taylor RA, Tay JK, Laursen SB, Stanley AJ, et al. Validation of a Machine Learning Model That Outperforms Clinical Risk Scoring Systems for Upper Gastrointestinal Bleeding. Gastroenterology. 2020 Jan;158(1):160–7.

Formatted: Italian (Italy)

12. D'Ascenzo F, Filippo OD, Gallone G, Mittone G, Deriu MA, Iannaccone M, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. The Lancet. 2021 Jan 16;397(10270):199–207.

13. López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui F (Rich). Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J Biomed Inform. 2015 Dec 1;58:60–9.

Formatted: Italian (Italy)

14. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp [Internet]. 2018 Nov 14 [cited 2021 Jan 30];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6234198/

15. Haque A, Guo M, Alahi A, Yeung S, Luo Z, Rege A, et al. Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance. ArXiv170800163 Cs [Internet]. 2018 Apr 24 [cited 2021 Jan 30]; Available from: http://arxiv.org/abs/1708.00163

16. Walker KJ, Jiarpakdee J, Loupis A, Tantithamthavorn C, Joe K, Ben-Meir M, et al. Predicting Ambulance Patient Wait Times: A Multicenter Derivation and Validation Study. Ann Emerg Med. 2021 Jul 1;78(1):113–22.

17. Walker K, Jiarpakdee J, Loupis A, Tantithamthavorn C, Joe K, Ben-Meir M, et al. Emergency medicine patient wait time multivariable prediction models: a multicentre derivation and validation study. Emerg Med J [Internet]. 2021 Aug 25 [cited 2022 Jan 27]; Available from: https://emj.bmj.com/content/early/2021/08/24/emermed-2020-211000

18. 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 Jun 1;110:12–22.

19. Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review. Acad Emerg Med [Internet]. [cited 2021 Feb 7];n/a(n/a). Available from: https://www.onlinelibrary.wiley.com/doi/abs/10.1111/acem.14190

20. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020 Sep;26(9):1320–4.

21. Sengupta PP, Shrestha S, Berthon B, Messas E, Donal E, Tison GH, et al. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A

Checklist: Reviewed by the American College of Cardiology Healthcare Innovation Council. JACC Cardiovasc Imaging. 2020 Sep 1;13(9):2017–35.

22. Price WN. Big Data and Black-Box Medical Algorithms. Sci Transl Med [Internet]. 2018 Dec 12 [cited 2021 Jan 20];10(471). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6345162/

23. Sartor G, European Parliament, European Parliamentary Research Service, Scientific Foresight Unit. The impact of the General Data Protection Regulation (GDPR) on artificial intelligence: study [Internet]. 2020 [cited 2021 Jan 30]. Available from: http://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

24. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019 Oct 1;1(6):e271–97.

Formatted: German (Germany)

25. Burdick H, Pino E, Gabel-Comeau D, Gu C, Roberts J, Le S, et al. Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals. BMC Med Inform Decis Mak. 2020 Dec;20(1):276.

Formatted: German (Germany)

26. Jimenez-Solem E, Petersen TS, Hansen C, Hansen C, Lioma C, Igel C, et al. Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients. Sci Rep. 2021 Feb 5;11(1):3246.

Formatted: German (Germany)

27. Koza JR, Bennett FH, Andre D, Keane MA. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero JS, Sudweeks F, editors. Artificial Intelligence in Design '96 [Internet]. Dordrecht: Springer Netherlands; 1996 [cited 2021 Jan 19]. p. 151–70. Available from: https://doi.org/10.1007/978-94-009-0279-4_9

28. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006. 738 p.

29. Király FJ, Mateen B, Sonabend R. NIPS - Not Even Wrong? A Systematic Review of Empirically Complete Demonstrations of Algorithmic Effectiveness in the Machine Learning and Artificial Intelligence Literature. ArXiv181207519 Cs Stat [Internet]. 2018 Dec 18 [cited 2020 Dec 1]; Available from: http://arxiv.org/abs/1812.07519

30. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020 Sep;26(9):1351–63.

31. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74.

32. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016;18(12):e323.