Rev Esp Cardiol. 2023;**76(8)**:645–654

Review article

Machine learning for decision-making in cardiology: a narrative review to aid navigating the new landscape

John W. Pickering[a,b]

[a] *Christchurch Heart Institute, Department of Medicine, University of Otago Christchurch, New Zealand*
[b] *Emergency Care Foundation, Christchurch Hospital, New Zealand*

Article history:
Received 9 February 2023
Accepted 9 February 2023
Available online 9 March 2023

Keywords:
Machine learning
Artificial intelligence
Narrative review
Cardiology

ABSTRACT

Machine learning in cardiology is becoming more commonplace in the medical literature; however, machine learning models have yet to result in a widespread change in practice. This is partly due to the language used to describe machine, which is derived from computer science and may be unfamiliar to readers of clinical journals. In this narrative review, we provide some guidance on how to read machine learning journals and additional guidance for investigators considering instigating a study using machine learning. Finally, we illustrate the current state of the art with brief summaries of 5 articles describing models that range from the very simple to the highly sophisticated.

Aprendizaje automático para la toma de decisiones en cardiología: una revisión narrativa para ayudar a recorrer el nuevo panorama

Palabras clave:
Aprendizaje automático
Inteligencia artificial
Revisión narrativa
Cardiología

RESUMEN

El aprendizaje automático *(machine learning)* en cardiología es cada vez más frecuente en la literatura médica, pero los modelos de aprendizaje automático aún no han producido un cambio generalizado de la práctica clínica. En parte esto se debe a que el lenguaje utilizado para describir el aprendizaje automático procede de la informática y resulta menos familiar a los lectores de revistas clínicas. En esta revisión narrativa se proporcionan, en primer lugar, algunas orientaciones sobre cómo leer las revistas de aprendizaje automático y, a continuación, orientaciones adicionales para quienes se plantean iniciar un estudio utilizando el aprendizaje automático. Por último, se ilustra el estado actual de la técnica con breves resúmenes de 5 artículos que van desde un modelo de aprendizaje automático muy sencillo hasta otros muy sofisticados.

INTRODUCTION

This narrative review considers the use of machine learning (ML) or artificial intelligence (AI) in cardiology publications with an emphasis on diagnostic models. In 2022, 1 in 60 cardiovascular disease publications in PubMed had ML or AI as a medical subject heading. Despite the abundance of articles, few diagnostic models have been validated or translated into practice. While this may be no worse than with other research avenues, there were high expectations that AI would revolutionize medicine. So, when will the revolution arrive?

As with many new technologies, ML in cardiology will go through the Gartner hype cycle—currently it is probably in the “Trough of disillusionment” and the field now needs to move up the “Slope of enlightenment” to the “Plateau of productivity”.[1] To reach these next stages, 3 things need to happen. First, clinicians need to understand the language of AI and ML and what is being said, and not said, in the articles they read. Second, scientists and clinicians need improve on current study designs. Third, robust validation studies and translational research is needed to identify where and how ML will make daily medical practice more precise and productive.

This manuscript has been written from the perspective of a scientist most interested in how ML can improve diagnosis. It focuses on diagnostics, though much of the content could translate to other uses of ML.

AI is strictly a machine which continues to learn through a flow of data after the development of the initial model. This is extremely rare. ML and statistical models are those developed on a bounded data set. Almost all cardiology articles using the term, AI, are strictly ML or statistical models. At their core, whatever we call the

E-mail address: John.Pickering@otago.ac.nz
@KiwiskiNZ@mastodon.nz

https://doi.org/10.1016/j.rec.2023.02.009

Abbreviations

AUC: area under the curve
AI: artificial intelligence
ML: machine learning
NPV: negative predictive value
PPV: positive predictive value
ROC: receiver operator characteristic

method, they are attempting to improve prediction or classification.[2] The appearance of ML and AI in the medical literature means clinicians and medical researchers are interacting with the culture and language of computer science even while the language of statistics is evolving. In time the languages may converge, but for now they confuse—both those reading the literature and those moving between disciplines. For the purposes of this article, a single term, ML will be used to encompass all that others call AI, ML, or statistics.

Statistics is a young discipline (it is only 120 years since the concept of a statistical hypothesis was introduced),[3] and computer science is even younger. In the last 20 years, the processing power of computers has become sufficient for ML to enter mainstream and to drive the new discipline of data science (an amalgam of computer science and statistics). As with any young discipline, we should expect not only changes in language, but a turnover in key concepts. This should engender caution within disciplines, such as cardiology, which make use of statistics and computer science concepts. Difficulty arises when concepts transfer from one discipline to another and become axiomatic—an example is the concept of statistical significance, which is ubiquitous in medicine, but is now considered poor methodology by academic statisticians.[4]

The first section is aimed at all who read the medical literature. Differences in language between statistics and computer science used to describe diagnostic model or risk score development are discussed. In the second section, for those looking to incorporate ML in their research, additional concepts that need to be considered before investing the time and effort to develop new models are introduced. Finally, a few exemplar articles on ML in the cardiology literature are evaluated.

PART 1: HOW TO EVALUATE ML PAPERS

To understand an ML article, one must be familiar with the terms used. Table 1 translates computer science terms to statistics terms. The table is not exhaustive and many ML glossaries are available online.[8] Understanding the specific methods used is important, but recognizing good science and good algorithm development methodology is also crucial, regardless of the ML techniques applied.

Understanding ML methods

Q1: What is the purpose of the study?

ML articles can be grouped into 3 categories: identifying associations between input variables and an outcome, illustrating the potential of a technique, or developing a model to change practice. Sometimes, articles combine these goals. An article may claim to develop a diagnostic model but also discuss causal relationships, which is not necessary for the model to be clinically useful. Often, there is no attempt to show that the developed model is better than current methods. This does not invalidate the study, but it should be recognized that the model is in an early stage of development and is not ready for implementation.

Q2: What is the cohort/from where does the data come?

If the study population does not reflect the population in which the model will be applied, the algorithm may be biased. This is known as spectrum bias (see Tseng et al.[9] for a cardiology example). Readers should pay special attention to how race, age, and ethnicity are dealt with. For example, the American Heart Association risk score for mortality in heart patients assigns lower risk to black patients,[10] which may lead to less intervention, and this can embed systematic inequities in society and health care systems.

Q3. What is the outcome?

When evaluating the usefulness of a model, it is important to consider if it is prognosticating something clinically meaningful and if its outcome would affect treatment or investigation decisions. Additionally, consider if the model is diagnosing something that is otherwise difficult to diagnose *at the time it is expected to be applied*. These are the important “so what?” questions.

Additionally, note if the model classifies (assign to a class, eg, whether a patient has heart failure or not) or provides a probability of a patient being in a particular class (eg, a patient has a 31.4% probability of having heart failure). A prediction may then turn into a classification by application of a threshold. An example of a classifier is a decision tree, and a method capable of giving probabilities is logistic regression. The ML literature does not always distinguish between the 2 types and may, for example, call logistic regression a classifier by applying a threshold (often arbitrarily at a probability of 50%).

Q4. What is the machine learning method used and why?

The choice of ML method often appears arbitrary and is often not justified. Recent successful methods for similar problems may be the reason for choosing a particular method. Where several methods are considered, a process using a training and/or testing data set may be used to create the best ML model (figure 1). The exact process can vary, and the reader will come across descriptions which may be quite complex, often relegated to a supplement, such as feature selection, cross validation (K-fold, leave-one-out, bootstrapping), and hyperparameter tuning to control the learning process. A difficulty that a clinical reader may face is not knowing how well the article has been refereed from a computer science (or statistical) methodological perspective. Editors could help by indicating the types of reviewers of an article.

Q5. What metrics are being used to determine validity?

The most crucial metrics for evaluating ML are those for the validation cohort with the chosen “best” model. Readers may also encounter metrics used to compare models. Some of these are discussed in the section for model developers below.

Diagnostic metrics

For diagnostic and risk classification, the receiver operator characteristic (ROC) plot and the associated area under the

Table 1
Common terms with a “translation” between computer science and statistical language.

The language of computer science (AI/ML)	The language of statistics commonly used in the medical literature	Comment
Algorithm/inducer/learner	Method (eg, logistic regression)	The program that learns from the data to produce a model
Model/network	Model	The learned program that maps inputs to predictions/classes
Inputs	Predictor variable data	_
Feature	Independent (explanatory/ predictor) variable or covariate	_
Feature selection	Variable selection	This is often an automated process to attempt to choose relevant variables. Most familiar in the medical literature are stepwise selection. These techniques are not robust and prespecification of variables by domain experts is preferred[5,6]
Feature engineering	Domain knowledge variable selection	The selection of variables based on domain knowledge by experts. It may involve a further step such as principal components analysis to reduce the dimension of the dataset
Label/outcome/response/class	Outcome/event/dependent variable	_
Optimization	Model fitting (regression)	_
Supervised learning	Prediction or regression	_
Classification	_	Rather than a prediction (continuous scale), simply outputs a predicted class; eg, predicts heart failure or not heart failure
Weights	Parameters (eg, beta coefficients in logistic regression models)	Often converted to the likes of odds ratios or hazard ratios
Confusion matrix of outcome verse classification	N x N (often 2 x 2) matrix	While in the medical literature convention appears to have the “true” outcomes in the order of positive, negative, as columns and the test outcomes as rows, this is not always the case with ML
Data sets		
Training	Development/derivation	The data set used to train a model (arrive at the models weights/ parameters)
Validation/testing (sometimes)	_	Sometimes used to choose the best of several models or to optimize ML algorithm. This may be wrapped up with the training set and implied in the method used – eg, k-fold cross validation
Testing/hold out	Validation/generalisability	Application of the (best) of the trained models in a data set that was set-aside for this purpose. The most important performance metrics in a article are those of the performance within this data set
Data analysis		
Prevalence	Prevalence	(TP + FN)/n
Recall	_	Proportion of all in a class that are predicted to be in that class
Recall (for binary outcome)/TP rate	Sensitivity (TP rate)	TP/(TP + FN)
	Specificity	TN/(TN + FP)
FP rate	1-specificity (FP rate)	FP/(FP + TN)
Precision	_	Proportion of all predicted to be in a class that are truly in that class
Precision (for binary outcome)	Positive predictive value	TP/(TP + FP) [prevalence dependent]
_	Negative predictive value	TN/(TN+FN) [prevalence dependent]
Accuracy	_	(TP + TN)/n [prevalence dependent]
_	Negative likelihood ratio (LR-)	Probability that someone with the event tests negative/probability that someone without the event tests negative (< 1 is diagnostic)
	Positive likelihood ratio (LR+)	Probability that someone with the event tests positive/probability that someone without the event tests positive (the larger the better)
F1 score	_	2TP/(2TP + FP + FN) [harmonic mean of precision and recall] (the larger the better; prevalence agnostic)
Calibration/reliability	Calibration	For prediction of diagnosis states this is a graph of actual proportion diagnosed with the disease to predicted proportion with the disease
Precision/recall curve	_	Precision on the y-axis vs recall on the x-axis
ROC curve	ROC curve	ROC curve. A curve formed by plotting sensitivity vs 1-specificity
UC	AUC	Area under the ROC curve
Additional		
Class imbalance	_	When the proportion of patients in each class is not equal
Up-sample/oversampling	_	A method used for development of some classification algorithms when the classes are imbalanced. This process may degrade performance[7]

AUC, area under the curve; FN, false negative; FP, false positive; n, sum of TP, TN, FP, FN; ROC curve, receiver operator characteristic curve; TN, true negative; TP, true positive.

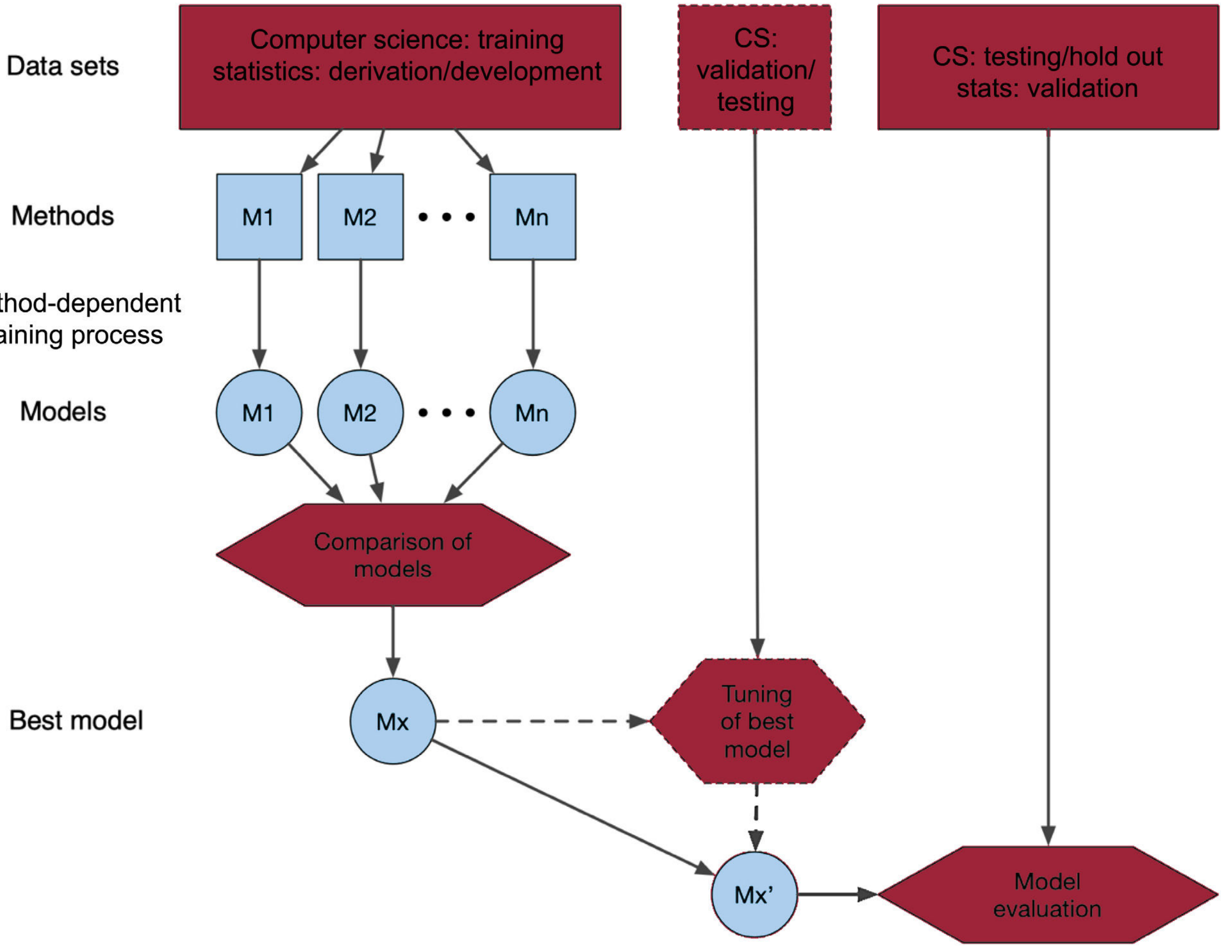

Figure 1. Central illustration. Schematic of process of developing and validating a machine learning model. One or several models are trained/derived and compared using different methods. The best model, Mx, may then undergo some further validation or tuning of parameters. The final model is validated in a data set that may be external (from other sources), temporally different from the other data set(s), or randomly chosen from the same data source as the other data set(s). CS, computer science; M1, model 1; M2, model 2; Mn, model n.

ROC curve (AUC) are the most commonly used metrics. However, they should not be used on their own or without further interpretation.

The ROC plot is a curve created by evaluating diagnostic performance at all possible outcome thresholds (probabilities for ML). The sensitivity and specificity are calculated at each threshold. The ROC plot is the curve formed on a sensitivity vs 1-specificity grid, and the AUC is the area under that curve. It is often assumed that the main diagonal represents a coin toss and that only values on the curve above that diagonal are useful. This, though, is incorrect.[11] Only the point in the middle of the diagonal (0.5, 0.5) is equivalent to a fair coin toss. Other points may contain diagnostically useful information depending on the prevalence. For example, the point 0.0 at the bottom left is equivalent to saying all diagnostic test results are negative, which in a low prevalence population has a probability much greater than 0.5 (equal to 1-prevalence).

The AUC may be interpreted as the probability that if one randomly drew the model outputs for a patient who had the outcome of interest and for another who did not, that the patients who had the outcome of interest had a greater output value than the patient who did not. In this light, it is difficult to understand why the AUC has become such a popular metric to describe the model performance. Figure 2A illustrates how 2 models with identical AUCs (0.94) can have different ROC curves. If the clinical situation is such that the cost of missing a diagnosis is very high, then the curve with the highest sensitivity at high specificity would be the preferred model.

Figure 2 illustrates some less common diagnostic plots that may appear in the medical literature. The precision/recall (positive predictive value [PPV]/sensitivity) is common in computer science. It can be used to choose between models at desired levels of PPV or sensitivity (figure 2B). Figure 2C is a violin plot. In this case, it shows that the baseline model has more very high probabilities for those with the event. The net benefit curve (decision curve), figure 2D, shows that, for the situation where false negatives and false positives are evenly weighted, there is net benefit in using the new model only where a prediction threshold of < 0.1 is applied. The risk assessment plot, figure 2E, illustrates that the difference in performance between the baseline and new model is a result of a small improvement in the lowering of the probability of those without the event (the solid teal, ‘‘1-specificity’’, curve for the new model moves nearer the bottom left compared with the dotted baseline model), but a large, inappropriate, lowering of probability for those with the event (the solid sensitivity curve moves toward the lower left for the new model, rather than the top right, compared with the baseline model).

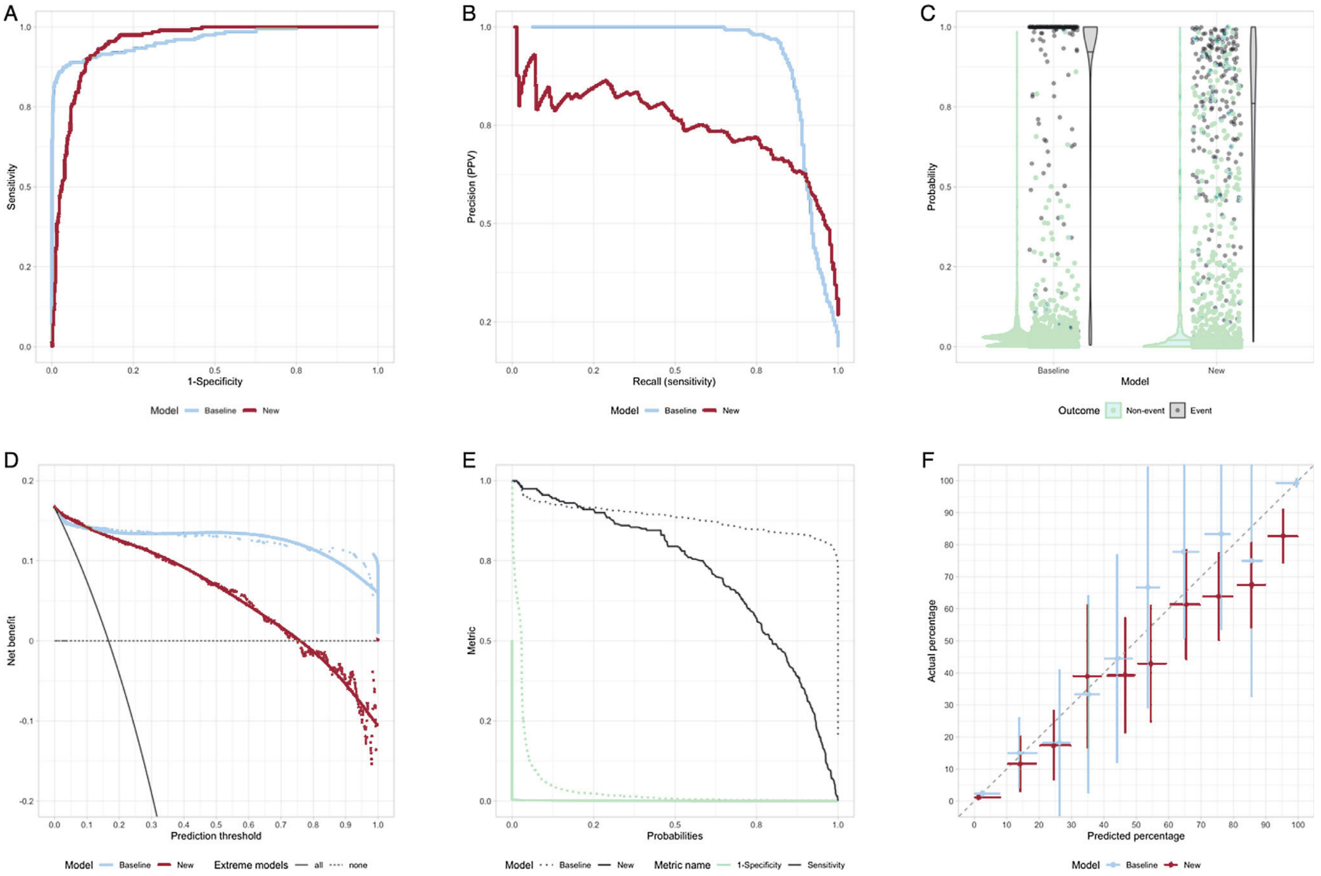

Figure 2. A: receiver operator characteristic (ROC) curves of 2 models with identical areas under the ROC curve. B: precision-recall curve. The baseline model is best at highest precision, but for recall > 0.92 the new model has best precision. C: scatter and violin plots. The bar on the violin plot is at the median probability for the model. An alternative is a box-and-whiskers plot. D: decision curve/net benefit curve. The greater the net benefit the better. Usually illustrate only over prediction thresholds of clinical relevance. In this case, the baseline model is better than the new model except at low (< 0.15) probabilities. E: risk assessment plot. The closer the teal curves are to the bottom left, the better the model is at assigning low probabilities to those without the outcome. The closer the black curves to the top right, the better the model is at assigning high probabilities to those with the outcome. In this case, the new model improves on the baseline for those without the outcome, but for probabilities > 0.2 it is worse than the baseline model. F: calibration plot. Ideally all points are along the diagonal line, indicating that predicted risk accurately reflects the actual risk. 95% confidence intervals are shown.

Some ML models report accuracy, the proportion of true positives and true negatives in the cohort, but this is not a useful measure in typical clinical situations where prevalence is low. For example, one could stand at the door of a hospital and turn away everyone with chest pain. The accuracy for diagnosis of myocardial infarction is likely to be 90% or more. Unfortunately, the sensitivity is 0% and one would be out of a job, in jail, or worse.

For classification models, common metrics include sensitivity, negative predictive value (NPV), specificity and PPV (see table 1). To generate these metrics, it is important to choose, by consensus among clinicians, clinically meaningful thresholds.[12] When there are differences in prevalence between training, test, and validation data sets, NPV and PPV should not be compared as these vary with prevalence. Accompanying these may be the negative and positive likelihood ratios (LR-, LR+) which are robust against prevalence. These ratios indicate if the test adds any diagnostic value. Data scientists may also report the F1 score.

All metrics should be reported with a confidence interval and it is important to consider both bounds when interpreting results.[13] The point estimate is only one of many possible values the underlying population could take. The null value may lie within the confidence interval, but, contrary to what is often stated, this does not mean the test is not useful. For example, if the LR- (95% confidence interval) is 0.80 (0.55-1.05), and if LR- < 0.9 is considered clinically meaningful then, in this case, values from 0.6 to 0.9 are all more likely than the null. For diagnostic algorithms, a particular bound may be important from a safety or utility perspective; eg, the lower bound for sensitivity is most important for assessing safety.

Importance of calibration

Calibration is important, yet often missing in analysis.[14] Goodness of fit tests (eg, the Hosmer-Lemonshow test), are much less informative than a calibration plot of the actual proportion with the outcome vs the predicted portion with the outcome, figure 2F. Models which discriminate may have poor calibration, which limits their usefulness. Check calibration curves at the probabilities of interest. For example, when the predicted rate is 3%, but the actual rate is 6%, then this could result in a clinically relevant underestimation of risk in a very large proportion of people in whom it is applied.

Q6. How are bias and equity handled?

In addition to assessing how the article has handled possible spectrum bias, consider if evidence is provided on the performance

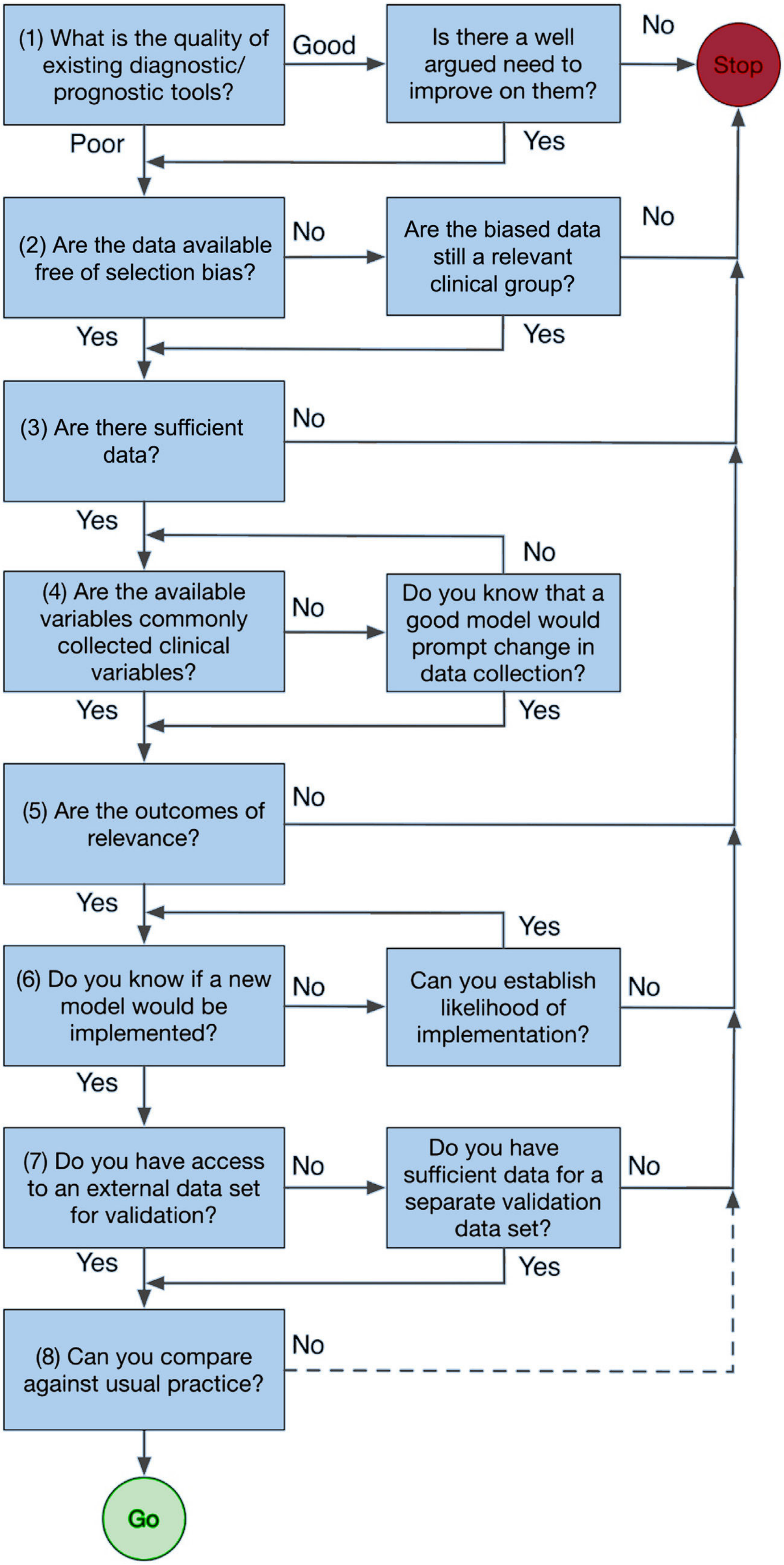
(1) What is the quality of existing diagnostic/ prognostic tools?
Good
Is there a well argued need to improve on them?
No
Stop
Poor
Yes
(2) Are the data available free of selection bias?
No
Are the biased data still a relevant clinical group?
No
Yes
Yes
(3) Are there sufficient data?
No
Yes
No
(4) Are the available variables commonly collected clinical variables?
No
Do you know that a good model would prompt change in data collection?
Yes
Yes
(5) Are the outcomes of relevance?
No
Yes
Yes
(6) Do you know if a new model would be implemented?
No
Can you establish likelihood of implementation?
No
Yes
(7) Do you have access to an external data set for validation?
No
Do you have sufficient data for a separate validation data set?
No
Yes
Yes
(8) Can you compare against usual practice?
No
Go

Figure 3. Heuristic for deciding whether to undertake model development in the first place.

of the model in specific subgroups known to face inequity in the population. For example, are performance metrics given for an indigenous population? When looking at such data, be aware that confidence intervals will be wider where numbers are low and take care not to interpret the model as "not working" in specific populations because of this.

Further literature

Useful articles for understanding ML are those by Marteen van Smeden,[15] Sebastian Vollmer from the Alan Turing institute,[16] and Google Alphabet ML expert Yun Liu.[17] Van Smeeden et al. start with the question "Is a new prediction model really needed?" Vollmer's first question relates to patient benefit. Where the researchers have not been "cognisant of the path from development to implementation" the manuscript may not pass the "so what?" test. In Yun Liu's article on how to read ML articles they state "clinical gestalt plays a crucial role in evaluating whether the results are believable: because one of the biggest strengths of ML models is consistency and the lack of fatigue, a useful check for believable ML results is whether an experienced expert could reproduce the claimed accuracy given an abundance of time." Finally, if a formal tool is wanted for assessing risk of bias in a prediction model, then I suggest the PROBAST tool.[18,19]

PART 2: HOW TO UNDERTAKE RESEARCH WITH ML

The heuristic, figure 3, is a brief guide to reaching a decision on the viability of a study.

(1) Before developing a new model, assess the quality of existing tools and determine if they can answer your research question, then assess the availability of relevant data. To pass the "so what?" test, survey those most affected by the introduction of a new tool, including patients, clinical staff, and health system management. Perhaps adopt "co-design" principles. These go a step further than consultation and ensure those most affected receive a "product" meaningful to them. For example, in New Zealand codesign is being used to reduce inequities for Māori in the delivery of cardiac health. By going first to Māori communities with a very limited agenda, the researcher acknowledges the primacy of the community and individual to make choices about their own health. The research may, for example, find the community has little interest in the probabilities of the proposed model, but instead suggests something different. Notably, the United States' "Blueprint for an AI Bill of Rights" adopts a similar approach with its first principal, which states that "Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system".[20]

(2) Assess the suitability of available data. Large, routinely collected data sets may have limitations, such as missing clinically relevant material or clinical outcome data biased because it is primarily used for financial purposes. To avoid selection bias, consider the target patient group and assess the potential data sets for bias. If the data exclude certain patient groups, the developed model will be limited in relevance and may bias treatment against those groups, especially if they are at higher risk. For example, a heart failure prediction model developed primarily on non-Māori may underpredict events in Māori, perpetuating inequities.

(3) To establish sufficient data rules of thumb of 10 to 20 (minimum) outcome events per variable have been popularized, but Richard Riley's recent work provides tools to make *a priori* calculations which inform the maximum number of variables in a model.[21–23] If some variable data are missing, imputation is accepted as better than discarding data incomplete patients because it reduces the risk of bias and reduces uncertainties in final estimates. However, there is a need to assess first if the data are missing at random.[24–28]

(4) Stepwise methods of variable selection, which use arbitrary *P* value thresholds to reduce the number of variables, are outdated and can result in biased and poorly transferable models.[5,6] The starting point for variable selection is to consult experts to determine the commonly collected predictors in clinical practice.[29] If the model relies on variables that are only accessible retrospectively, there needs to be a strong reason for inclusion, as it would also require their prospective collection for the model's future use.

In the past, perhaps because of the need for simple algorithms able to be used at the bedside, continuous variables like age and systolic blood pressure were dichotomized. This process discards information and is unnecessary with today's technology.[30]

(5) Model outcomes must be relevant to the model users. For example, if the aim is to develop an in-hospital prediction model for cardiologists, is this of relevance to the cardiologist and what level of improvement over current predictions would be of use?

(6) Changing clinical practice is difficult, yet it is the implied goal of all ML. Before undertaking ML research, establish the need for the model and the barriers to implementation. These may include the likes of the "black-box" nature of some ML, management priorities, and IT resources. Codesign may help overcome such barriers.

(7) An external validation data set is not strictly required and there are good internal validation techniques that should always be applied. Best practice will avoid data-splitting for smaller data sets.[31,32] However, generalizability is still best assessed with external data sets. Identify these prior to model development and choose the largest data set to be the development data set.

(8) Finally, if the exercise is to be more than an academic one, it must be compared with standard practice. While not doing so is not fatal, it certainly means a model is much less likely to be adopted by others.

Choosing algorithms

While diagnosis involves classification, ML models do not have to output a classification. A model that produces a probability output may be more informative. Logistic regression can be thought of as a baseline algorithm that is easy to implement and interpret. Ensemble models such as random forests or gradient boosting can improve on simple logistic regression. This is the familiar wisdom of the crowd. There are several forms of ensemble models by combining multiple models or using bootstrapping. A recent survey of methods can be found in Sagi and Rokach.[33]

Reporting

It is important to report how the "best" model for validation was chosen and the methodology for this should be prespecified. It need not be based on the AUC. As figure 1 illustrates, there are other considerations depending on the intended purpose of the model. This includes calibration. Metrics that should be considered include log-likelihood, Brier skill (the relative change in Brier score from a baseline model), Nagelkerk's R^2, and the integrated discrimination improvement reported separately for those with and without events. If the intended use does not include presenting predicted probabilities, then the best model may be the one that has the highest specificity for a prespecified minimum sensitivity (for identifying low-risk patients) or vice versa (for identifying high-risk patients).

The primary outcome measures are both the discrimination (predefined) and calibration in the validation cohort of the best model. These must be reported in the abstract along with a confidence interval.

PART 3: EXAMPLES OF ML IN CARDIOLOGY

Myocardial infarction decision support in the emergency department

Troponin is the strongest predictor of a myocardial infarction (not surprisingly as it is also part of the definition) on presentation to the emergency department. Troponin concentrations are also associated with age and sex, both of which are also associated with myocardial infarction. The kinetics of high-sensitivity troponin, and in particular the rate of change, are also related to the outcome.[34]

Decision support tools for risk stratification in the emergency department have been constructed from one person's expert opinion (HEART), have used scores developed for other purposes (ADAPT), have developed a fit-for-purpose score using logistic regression (EDACS), and have used simple decision trees based on a single variable, namely the troponin result.[35–39] Application of ML that can provide support via providing predicted probabilities is new. Body et al.[40] T-MACS combined high-sensitivity troponin with electrocardiogram, vomiting, sweating and the nature of the pain in a logistic regression in 2017, which outputs a myocardial infarction probability. Discrimination was high (0.90 in the external validation data set) and a threshold chosen to aid decision-making meant that the model could be applied clinically with a high degree of safety (very high sensitivity). Performance at the threshold was good and the model has been implemented in the Greater Manchester region of the United Kingdom.[41] No calibration was provided, and the algorithm was not well calibrated in a further external validation.[42] This indicates the need for recalibration prior to application in new settings.

Than et al.[43] evaluated an ML model in a large international data set developed by Abbott diagnostics that used a very simple feature set: age, sex, 2 troponin measurements, and the time between the 2 measurements. The model had been developed using gradient boosting, had high-discrimination, and was well calibrated in the validation data set. It was not compared against other models. In a further external validation, discrimination remained high and performance metrics at thresholds remained good.[44] However, the model underpredicted myocardial infarction at prediction values $< 50\%$. Again, this points to the importance of checking calibration for each cohort in which a model is likely to be applied.

Heart failure diagnosis

ML models have been developed for heart failure diagnosis and have been summarized in 2 systematic reviews.[45,46] These reviews highlight the diverse range of both situations, in which ML is thought to be of use, and the techniques employed. For instance, convoluted neural networks are used to enhance pathologists' reading of whole-slide biopsy images,[47] various methods predict readmission but with poor discrimination,[48] and deep neural networks used demographic and electrocardiogram features to identify heart failure show good discrimination, but no comparison to current methods.[49]

Two notable ML cardiology articles display the features of quality model development and assessment. The first, by Dana Sax et al.,[50] aimed to improve 30-day adverse event prediction in patients presenting to emergency departments with acute heart failure. They compared an existing tool, STRATIFY, with their ML model made with 13 variables from STRATIFY and 58 possible additional variables. Data were divided into test (20%) and development (80%) data sets with simple imputation for missing variables and 10-fold cross validation for hyperparameter tuning. Model assessment was, overall, by the AUC, ROC and calibration curves. Precision/recall curves were also presented and, at prespecified risk thresholds, sensitivity, specificity, negative and positive likelihood ratios, PPV, NPV and F1 score. Across all thresholds net reclassification was used to compare a logistic regression model with an XGBoost model. The authors thoroughly discussed the limitations, including possible bias from its retrospective nature, and future plans around implementation. They demonstrated improvement with an ML model over STRATIFY (AUC, 0.76 *cf* 0.68). While using only the AUC for overall comparison is limiting, and an AUC of 0.76 may not be sufficient to precipitate a change of practice, the detailed assessment of performance at clinically relevant thresholds mitigated this limitation. The other limitation is there is no assessment of performance in key demographics.

The second, by Kuan Lee et al.,[51] used research data from 14 studies and 13 countries with adjudicated outcomes to develop and validate a model to aid decision support for the diagnosis of heart failure. The data sets were identified using a systematic review methodology including assessment for risk of bias. Four ML models were developed using multiply imputed data sets. Validation involved treating each of the 14 data sets as an external data set (without imputation) and using the remaining data sets for model development. Several criteria including calibration, Brier score, AUC, and proportions of patients above and below specified probability criteria were used to judge model performance. Additionally, decision curves (published in the supplement) were produced. Diagnostic performance was assessed for a wide range of demographic subgroups. The limitations of the study were well addressed, particularly acknowledging the possibility of selection bias as 16 of the 30 studies eligible for inclusion did not participate. This was another excellent study, with few weaknesses. One, the use of NPV and PPV to compare patient subgroups, rather than sensitivity and specificity, is to be avoided because the differences in these metrics may be driven by difference in the prevalence of heart failure in each subgroup rather than true differences in model performance.

Coronary artery disease probability

Forrest et al.[52] developed an ML from electronic health records to use as an in-silico marker of coronary artery disease producing a probability of coronary artery disease. Training and validation were conducted in a United States cohort, and external testing in a United Kingdom cohort. The AUC was the principal diagnostic metric. Sensitivity, specificity, accuracy, PPV and NPV were also reported, but there is no indication of the threshold used to determine these later metrics. This is an example of the difficulty with computer science language not being familiar to clinical readers, as it is commonplace within computer science to use a probability of 0.5 as a classification threshold. However, if this was the case in this article, it should have been reported. Also missing was a calibration plot. Brier scores were referred to, but these are inadequate and noninformative on their own. Not referred to in the article was a precision-recall curve and calibration plot shown in the appendix. The latter showed the algorithm overpredicted the probability of coronary artery disease for all but the highest predictions. A very positive feature of the study was the demonstration of the association of the probabilities with coronary artery stenosis and all-cause death. There was no comparison with other prediction models.

SUMMARY

While the age of ML is upon us, it has not yet been translated into practice. This is partly because there is a learning curve among decision makers to understand what constitutes a well executed ML study. Here I have attempted to provide some translation between the language of computer science and the more familiar language of the medical statistician and have then highlighted some useful metrics and graphical tools to evaluate ML models. Finally, I identified some studies demonstrating different elements of ML in cardiology.

FUNDING

The author is supported by charitable donations to the Christchurch Heart Institute and the Emergency Care Foundation, both of Christchurch New Zealand.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

1. Gartner Hype Cycle. Available at: https://www.gartner.com/en/research/methodologies/gartner-hype-cycle. Accessed 30 Aug 2022.
2. Faes L, Sim DA, van Smeden M, Held U, Bossuyt PM, Bachmann LM. Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins? *Front Digit Health.* 2022;4:833912.
3. Pearson KX. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag.* 1900;50:157–175.
4. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. *Am Stat.* 2016;70:129–133.
5. Smith G. Step away from stepwise. *J Big Data.* 2018;5:32.
6. Steyerberg EW, Uno H, Ioannidis JPA, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol.* 2018;98:133–143.
7. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc.* 2022;29:1525–1534.
8. Google Developers. Machine Learning Glossary. Available at: https://developers.google.com/machine-learning/glossary?hl=en. Accessed 30 Nov 2022.
9. Tseng AS, Shelly-Cohen M, Attia IZ, et al. Spectrum bias in algorithms derived by artificial intelligence: a case study in detecting aortic stenosis using electrocardiograms. *Eur Heart J Digit Health.* 2021;2:561–567.
10. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med.* 2020;383:874–882.
11. Carrington AM, Fieguth PW, Mayr F, et al. The ROC Diagonal is not Layperson's Chance: a New Baseline Shows the Useful Area. In: Holzinger A, Kieseberg P, Tjoa AM, Weipp E, eds. In: *Machine Learning and Knowledge Extraction. Lecture Notes in Computer Science.* Springer; 2022:100–113.
12. Than MP, Herbert M, Flaws D, et al. What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: a clinical survey. *Int J Cardiol.* 2013;166:752–754.
13. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests. *P values confidence intervals and power: a guide to misinterpretations Eur J Epidemiol.* 2016;31:337–350.
14. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; on behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230.
15. van Smeden M, Heinze G, Calster BV, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. *Eur Heart J.* 2022;43:2921–2930.
16. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ.* 2020;368:l6927.
17. Liu Y, Chen PHC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA.* 2019;322:1806.
18. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med.* 2019;170:W1.
19. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med.* 2019;170:51.
20. The White House Office of Science and Technology. Blueprint for AI Bill of Rights. Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/. Accessed 14 Dec 2022.
21. Riley RD. Correction to: Minimum sample size for developing a multivariable prediction model: Part II-binary and time-to-event outcomes by Riley RD, Snell KI, Ensor J, et al. *Stat Med.* 2019;38:5672.
22. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med.* 2019;38:1276–1296.
23. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ.* 2020;368:m441.
24. Altman DG, Bland JM. Missing data. *BMJ.* 2007;334:424.
25. Newgard CD, Lewis RJ. Missing Data: How to Best Account for What Is Not Known. *JAMA.* 2015;314:940.
26. He Y. Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter. *Circ Cardiovasc Qual Outcomes.* 2010;3:98–105.
27. Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA.* 2015;314:1966.
28. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35:1925–1931.
29. Kaufman S, Rosset S, Perlich C. *Leakage in data mining: formulation, detection, and avoidance. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'11*. New York: ACM Press; 2011:556–563.
30. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ.* 2006;332:1080.
31. Steyerberg EW. Validation in prediction research: the waste by data splitting. *J Clin Epidemiol.* 2018;103:131–133.
32. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245–247.
33. Sagi O, Rokach L. Ensemble learning: A survey. *WIREs Data Mining Knowl Discov.* 2018;8:e1249.
34. Pickering JW, Young JM, George PM, et al. Early kinetic profiles of troponin I and T measured by high-sensitivity assays in patients with myocardial infarction. *Clin Chim Acta.* 2020;505:15–25.
35. Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. *Neth Heart J.* 2008;16:191–196.
36. Than MP, Cullen LA, Aldous S, et al. 2-Hour accelerated diagnostic protocol to assess patients with chest pain symptoms using contemporary troponins as the only biomarker: the ADAPT trial. *J Am Coll Cardiol.* 2012;59:2091–2098.
37. Than MP, Pickering JW, Aldous SJ, et al. Effectiveness of EDACS versus ADAPT accelerated diagnostic pathways for chest pain: A pragmatic randomized controlled trial embedded within practice. *Ann Emerg Med.* 2016;68:93–102.e1.
38. Than MP, Flaws D, Sanders S, et al. Development and validation of the Emergency Department Assessment of Chest pain Score and 2h accelerated diagnostic protocol. *Emerg Med Australas.* 2014;26:34–44.
39. Reichlin T, Schindler C, Drexler B, et al. One-hour rule-out and rule-in of acute myocardial infarction using high-sensitivity cardiac troponin T. *Arch Intern Med.* 2012;172:1211–1218.
40. Body R, Carlton E, Sperrin M, et al. Troponin-only Manchester Acute Coronary Syndromes (T-MACS) decision aid: single biomarker re-derivation and external validation in three cohorts. *Emerg Med J.* 2017;34:349–356.
41. Health Innovation Manchester. T-MACS: Troponin only Manchester Acute Coronary Syndromes. Available at: https://healthinnovationmanchester.com/our-work/t-macs/. Accessed 18 Nov 2022.
42. Greenslade JH, Nayer R, Parsonage WA, et al. Validating the Manchester Acute Coronary Syndromes (MACS) and Troponin-only Manchester Acute Coronary Syndromes (T-MACS) rules for the prediction of acute myocardial infarction in patients presenting to the emergency department with chest pain. *Emerg Med J.* 2017;34:517–523.
43. Than MP, Pickering JW, Sandoval Y, et al. Machine Learning to Predict the Likelihood of Acute Myocardial Infarction. *Circulation.* 2019;140:899–909.
44. Doudesis D, Lee KK, Yang J, et al. Validation of the myocardial-ischaemic-injury-index machine learning algorithm to guide the diagnosis of myocardial infarction in a heterogenous population: a prespecified exploratory analysis. *Lancet Digit Health.* 2022;4:e300–e308.
45. Olsen CR, Mentz RJ, Anstrom KJ, Page D, Patel PA. Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. *Am Heart J.* 2020;229:1–17.
46. Bazoukis G, Stavrakis S, Zhou J, et al. Machine learning versus conventional clinical methods in guiding management of heart failure patients—a systematic review. *Heart Fail Rev.* 2021;26:23–34.

47. Nirschl JJ, Janowczyk A, Peyster EG, et al. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue, Marsden A, ed. *PLoS One.* 2018;13:e0192726.

48. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC Heart Fail.* 2019;6:428–435.

49. Kwon J, myoung. Kim KH, Jeon KH, et al. Development and Validation of Deep-Learning Algorithm for Electrocardiography-Based Heart Failure Identification. *Korean Circ J.* 2019;49:629.

50. Sax DR, Mark DG, Huang J, et al. Use of Machine Learning to Develop a Risk-Stratification Tool for Emergency Department Patients With Acute Heart Failure. *Ann Emerg Med.* 2021;77:237–248.

51. Lee KK, Doudesis D, Anwar M, et al. Development and validation of a decision support tool for the diagnosis of acute heart failure: systematic review, meta-analysis, and modelling study. *BMJ.* 2022;377:e068424.

52. Forrest IS, Petrazzini BO, Duffy Áaue, et al. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. *Lancet.* 2023;401:215–225.