Building artificial intelligence and machine learning models : a primer for emergency physicians

RAMLAKHAN, Shammi, SAATCHI, Reza <http://orcid.org/0000-0002-2266-0187>, SABIR, Lisa, VENTOUR, Dale, HUGHES, Ruby, SHOBAYO, Olamilekan and SINGH, Yardesh

Citation:

RAMLAKHAN, Shammi, SAATCHI, Reza, SABIR, Lisa, VENTOUR, Dale, HUGHES, Ruby, SHOBAYO, Olamilekan and SINGH, Yardesh (2022). Building artificial intelligence and machine learning models : a primer for emergency physicians. Emergency Medical Journal. [Article]

Building artificial intelligence and machine learning models : a primer for emergency physicians

Shammi L Ramlakhan[1], Reza Saatchi[2], Lisa Sabir[1], Dale Ventour[3], Olamilekan Shobayo[2], Ruby Hughes[4], Yardesh Singh[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

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

ABSTRACT

There has been a rise in the number of studies relating to the role of artificial intelligence (AI) in healthcare. Its potential in Emergency Medicine (EM) has been explored in recent years with operational, predictive, diagnostic and prognostic emergency department (ED) implementations being developed. For EM researchers building models de novo, collaborative working with data scientists is invaluable throughout the process. Synergism and understanding between domain (EM) and data experts increases the likelihood of realising a successful real-world model. Our linked manuscript provided a conceptual framework (including a glossary of AI terms) to support clinicians in interpreting AI research. The aim of this paper is to supplement that framework by exploring the key issues for clinicians and researchers to consider in the process of developing an AI model.

INTRODUCTION

There has been a rise in the number of studies relating to the role of artificial intelligence (AI) in healthcare. Its potential in Emergency Medicine (EM) has been explored in recent years with operational, predictive, diagnostic and prognostic emergency department (ED) implementations being developed.(1) For EM researchers building models de novo, collaborative working with data scientists is invaluable throughout the process. Synergism and understanding between domain (EM) and data experts increases the likelihood of realising a successful real-world model. Our linked manuscript provided a conceptual framework (including a glossary of AI terms) to support clinicians in interpreting AI research.(1) The aim of this paper is to supplement that framework by exploring the key issues for clinicians and researchers to consider in the process of developing an AI model.

STAGES IN DEVELOPING AN AI (MACHINE LEARNING) MODEL

EM research in AI is almost entirely focussed on Machine Learning (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'. (2) 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 1. Traditional reporting guidelines such as TRIPOD (and eventually TRIPOD-AI) for prognostic models or STARD for diagnostic models(3) can supplement these steps and serve as checkpoints while building a model.

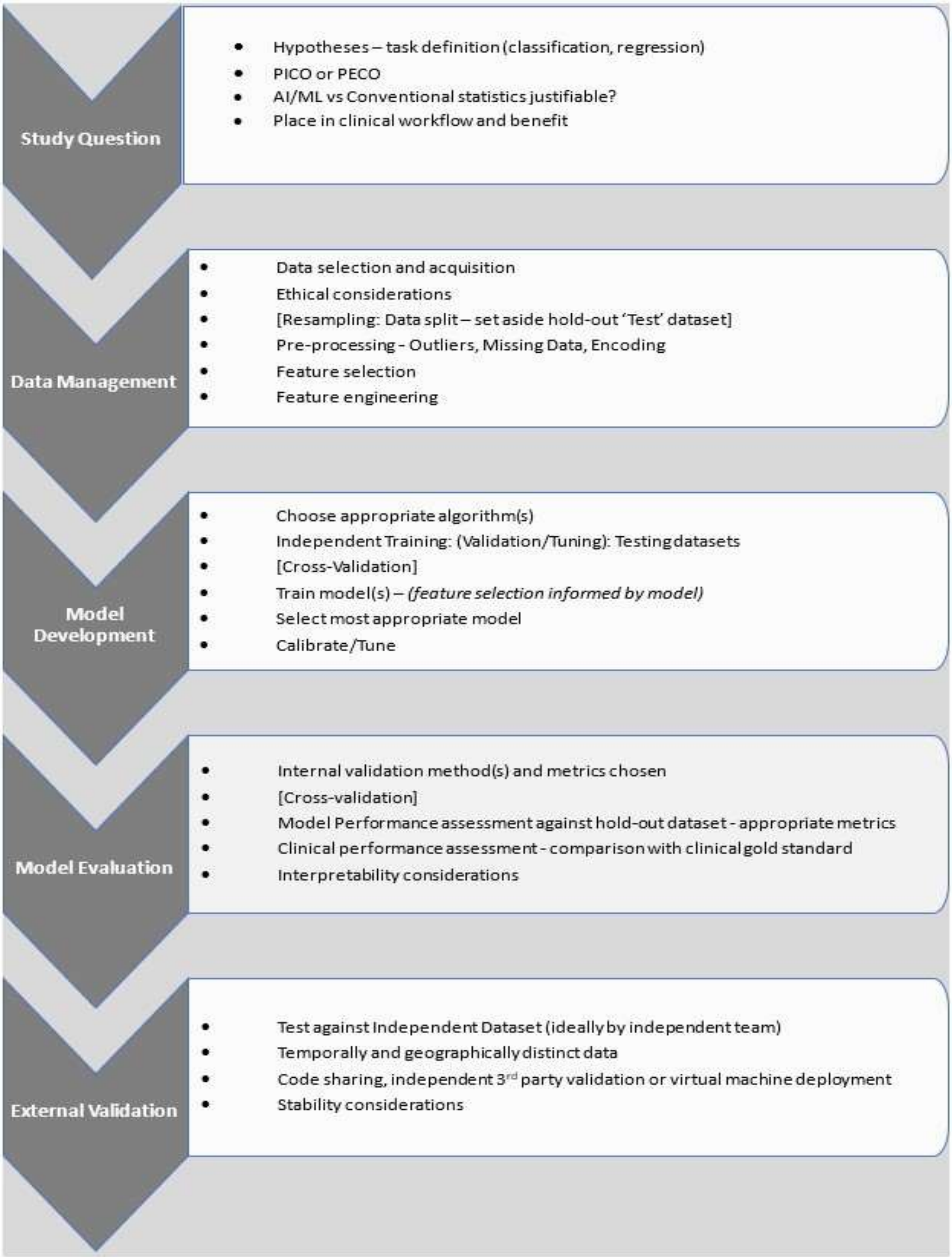

Study Question
Hypotheses – task definition (classification, regression)
PICO or PECO
AI/ML vs Conventional statistics justifiable?
Place in clinical workflow and benefit
Data Management
Data selection and acquisition
Ethical considerations
[Resampling: Data split – set aside hold-out 'Test' dataset]
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 1 Steps in 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. AI, artificial intelligence; ML, machine learning; PECO, Population-Exposure- Comparison- Outcome; PICO, Population-Intervention-Comparison-Outcome.

Study Question

Although this may seem simplistic, it is critical that the problem or question be clearly defined as well as the proposed clinical benefit. Unless one is undertaking data mining or exploratory analysis of a novel dataset (although arguably, an idea of what is being asked is still important in these activities), this will act as an anchor further in the model development pipeline to mitigate against bias or misinterpretation. Domain knowledge, i.e. real-world clinical EM experience of the clinical workflow or problem is key to developing a relevant, interpretable and generalisable model.

In addition, as more open access ML libraries (containing ML algorithms and other related code), such as scikit-learn.org and CRAN-R packages (cran.r-project.org) and even automated ML (AutoML) become freely available, the temptation is to apply ML to a problem when traditional solutions work well or are more suitable. (4–6) The current clinical status quo should be examined, and any proposed improvement in its performance considered in terms of potential clinically significant patient or system benefit. A priori determination of an ideally patient centred (or shared) minimal acceptable difference in outcome should be made. (7) The performance measures by which this outcome is evaluated should also be defined a priori.

Consideration of these issues at conception is important – even traditionally derived diagnostic and prognostic models often make no difference in clinical outcomes or are never used in a clinical setting. (8,9) Many published ML models are likely to suffer the same fate, with evidence suggesting that ML does not consistently outperform clinicians, (6,10,11) with inappropriate usage and methods likely to result in unsuccessful deployment. Even accurate ML models may not necessarily confer a clinical benefit – the so-called AI chasm. (12)

Finally, as with almost all types of EM relevant research, patient and public engagement should be sought where appropriate. This not only pertains to outcomes, but also regarding the use of routine personal patient data for AI purposes, and the acceptability of a ML model informing their clinical management and profiling. (13,14)

Data Management

Unsurprisingly, data preparation is the most time-consuming aspect of ML model development. ML is dependent on data, therefore the quality and relevance of data used in developing a model is of paramount importance - the programming cliché of *'garbage in, garbage out'* is particularly true in ML. Most EM datasets will be retrospectively curated, with only a few prospective collected. The challenge is to ensure that the data available is homogenous and representative of the population and setting in which the model will be deployed. Any potential for bias should be addressed, as failure to do so may lead to significant disadvantages to certain groups when the model is implemented. (15) In addition, data collection methods at the point of acquisition should be as objective as possible to minimise heterogeneity due to human interpretation. As data is often historical, some sense checking should take place to identify clinical practice evolutions which may alter outcomes; hence negating or amplifying the impact of features used in a model developed from temporally distinct data.

Recently, *Federated Learning* has been developed, whereby disparate data sources from multiple sites are used to train models, without the data being transferred outside their original source location. This allows a range of data to be used, and in turn may produce more generalisable models. (16)

Sample size

Although a core consideration in classical statistical methods, sample size calculations are uncommon in ML research. (17) Rule of thumb estimates from traditional regression methods are often used as a default, and varies from 10-50 events per feature (18), the square of the number of features or fitting a weighted learning curve. (19) The constant however, is that the quality of the data is more important, and although in general, more data is better, it depends on the specific task and model. With inaccurate (high bias) models, more data is unlikely to help with underfitting, whereas in high variance (overfitted) models, more data will probably help. (20)

Once a suitable dataset is identified, it must be cleaned and presented in a format which is interpretable to the selected ML algorithm, and which allows reproducibility and ideally reversibility. The following preparation/pre-processing activities should be considered.

Duplicate removal

Because most ML algorithms consider each individual sample of input data in relation to the rest of the dataset, having duplicates will bias the model towards the features in that sample. If duplicates are used in training and test sets, it will also contribute to data leakage.

Missing Data

Some advanced algorithms can handle missing data. However, it is worth considering whether the missing data is random, or represents bias in data collection or recording. The impact of the method of dealing with missing values needs to be considered in the context of the effect on the model - some methods will reduce the available data, while others may introduce new categories, and fail to consider covariance, for example. Common approaches are: removal of samples/subjects with missing data; imputation – where an estimate (mean, median, mode) based on the remainder of data is input; where interpolated data (for example between two valid times in a time-series variable) or last-observed-carry-forward (for example in longitudinal physiological data) are used; unsurprisingly, using a DNN to impute missing data into the dataset; or finally, simply encoding as null/missing and letting the algorithm deal with it.

Outliers

Nonsense values will be obvious to an Emergency Physician (EP) looking at a dataset, although not necessarily to a data analyst, reinforcing the importance of combining domain knowledge and expertise. True outliers can affect model performance and metrics, making identification important before training. However, some outliers can be hard to define and may represent important data relationships. Caution must therefore be exercised before removing or changing values, especially if the sample size is small. (21) Similarly, extraneous or irrelevant data (noise) can affect model performance, and consideration of how to handle it is necessary.(22) Identification of irrelevant or impractical data or variables will be enabled by sound EM domain expertise.

Feature (Variable) Selection

As should be apparent, basic ML algorithms work best with data that is relevant to the problem. Although databases can have high numbers of variables (high dimensionality), not all will be useful, and may simply add to the complexity and computational requirements of the model pipeline. In addition, there may be a tendency to overfit as with outliers or noise. Good ML features should be unambiguous, represent the EM operational/clinical meaning and context of the data and their inter-relationships. Confounders and covariates should be addressed, as they are unlikely to add to the model if there are other mutually uncorrelated features. (23)

In some classification tasks, there may be only a few instances of an uncommon value for a categorical variable. By chance, these may have the same outcome - being 'perfect' predictors - and hence result in overfitting. This may be addressed by using a modelling method with feature selection, and removing rare categories which appear highly important to the model. (24) An alternative method removes dummy variables (with less than say 10 observations) as if they were an outlier.

In most EM applications, feature selection by clinical domain experts (EPs) will be sufficient based on clinical experience, known predictive or prognostic variables, and some common sense. The idea of scientific plausibility is of particular relevance. For example, an imaging Deep Neural Network (DNN) determined that images taken by a *portable* x-ray machine predicted significant pathology rather than using the images themselves. This would have been apparent given that sicker patients in ED would have portable films, but this knowledge would not be discernible to the algorithm. (25) There are other, non-human methods of selecting features, such as removal of those with low statistical variance, using an algorithm which shows the importance of features in the dataset (e.g. random forest or linear regularisation models such as LASSO regression) or using a grid search where multiple models with a selection of features are compared to identify the best feature set. (26)

Feature Engineering

There are several ways to represent variables, and some types of models handle different types of data differently to others. Feature engineering is the process of transforming data in order to improve the effectiveness of the model being used for a particular task. (27) EM databases may include data from various sources, and with a range of values. Algorithms may be biased towards variables with larger values, in which case the data can be standardised by rescaling i.e., changing the range of values (usually between 0 and 1), or normalisation i.e., changing to a normal distribution.

Another EM phenomenon is of imbalanced datasets, where patients with an event/outcome are relatively uncommon compared with the rest of the ED population. In this case, an algorithm may be biased to learning characteristics of the majority (without the outcome) rather than the group we are generally interested in – those with the outcome. The effect of imbalanced datasets can be offset by stratified re-sampling, over- and under- sampling minority and majority classes respectively, increasing the weight of samples in the minority class during training, or by adding artificially generated data which approximates the minority class. (28) Care must be taken with interpreting performance metrics, such as AUROC, in imbalanced datasets, particularly when artificial data is introduced.

It may be desirable to aggregate some features, for example grouping components of a recognised score together rather than each separate component being considered individually. Provided that this is done in a contextually appropriate manner, it can reduce the dimensionality of the dataset. Conversely, it may be desirable to decompose a complex feature into constituent parts which are useful for the problem and algorithm, for example separating the Glasgow Coma Scale into motor, verbal and eye opening.

Encoding

Machines do not generally understand non-numerical input, and therefore even image and text data have to be converted into a format which can be used by an algorithm. Most data used in EM applications will be numerical, however care should be taken that encoding preserves interval/ratio relationships. For nominal categorical data, one-hot encoding is commonly used. This assigns a binary style of coding, where the position of a ‘1’ will be interpreted by the machine as the particular category. For example, 3 labels for ethnicity will be encoded as 100 (White), 010 (Black), 001 (Asian), and a dummy variable - 000. However, consideration should be given to the effect of the selected encoding. For example, if rather than 3 ethnicity codes, we used 15, one-hot encoding of this higher cardinality, categorical variable will substantially increase dimensionality, and thus the number of features and variables the model needs to consider - just for ethnicity. This makes overfitting much more likely as it can easily fit relatively unimportant variables to the model. Similarly, care should be taken to encode ordinal data so that its clinical meaning is preserved, for example with the AVPU scale or A-E triage.

In ML, the model being developed is meant to be used for making predictions on unseen data when clinically or operationally deployed. Before the model can be finalised, an estimate of how it will perform on new data is necessary. Resampling is the process of splitting the dataset, so that this estimate can be made on a subset of data which will be new to the model, but where the outputs are in fact labelled, but known only to the analyst.

The most robust methodology involves splitting data early on so that data leakage and the inevitable overfitting is avoided. Data leakage occurs when *any* data from outside the training set manages to leak into the model building (training) process. The model therefore has already ‘seen’ the test dataset, hence inappropriately influencing the model and its validation. A list of pre-processing techniques should be estimated and developed in the presence of the training data *only*, and the same list then applied to future data. (27) This is even more important where transformations which use group parameters are used, such as imputation or use of measures of spread. In order to mitigate against data leakage, there is a compelling case for splitting off the test dataset and setting it aside before *anything* is done with *any* of the data.

The terminology used in splitting data can be confusing, with different terms being used to describe the same procedures. For example, when evaluating a final model, a test dataset in ML is analogous to an internal validation dataset in conventional statistical modelling, even if temporally split. (29) During model development, the training dataset is often split to provide data to facilitate feature selection, model refinement (tuning of hyperparameters), between-model performance comparison and give an idea as to how a model will perform in practice. This split is commonly termed a validation set, or sometimes a tuning or calibration set and is used to assess the model which was fit on the training set.

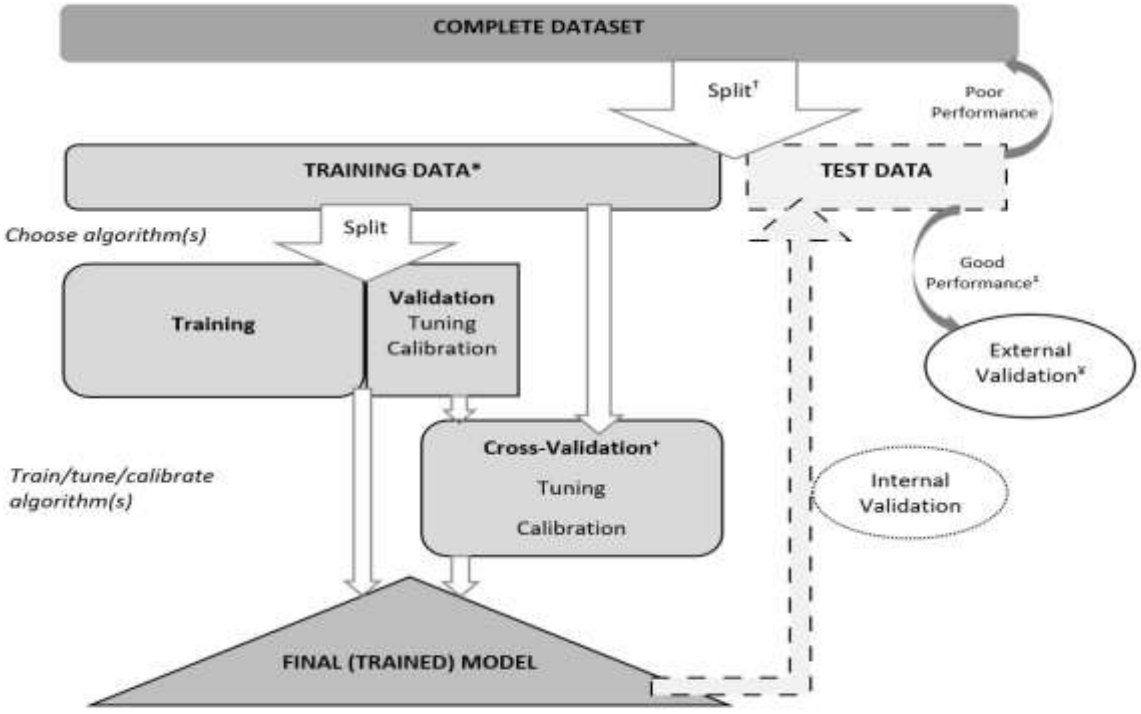


Figure 2 Overview of data flow in Resampling (Train:Test) †No data transformation should occur before the test set is set aside. *Training set can also be used to determine the final model, without splitting into Train:Validation or Cross-Validation (CV). +CV splits Training set, hence a manual Train:Validation split may be unnecessary. ‡The final (good) model can be trained on the entire dataset before deployment/external validation. ¥External validation dataset - ideally temporally and geographically distinct.

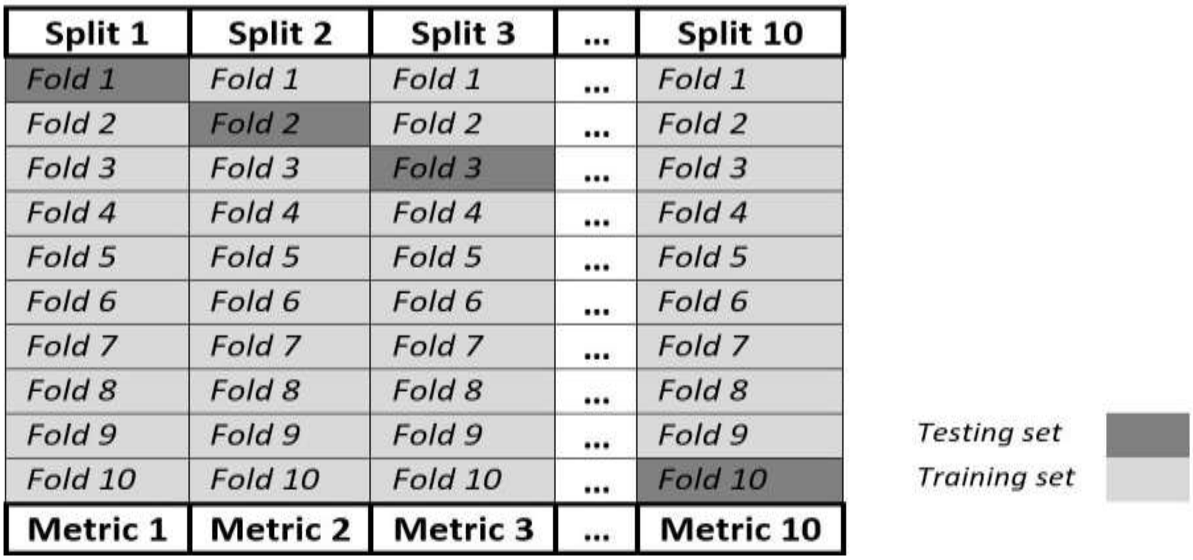
The ratio for splitting the complete dataset into *train:test* or the training set into *train:validate* can be set based on the size of the dataset, the purpose of the model and the algorithm being used. Most EM applications will use complex models which requires more training and will need a larger proportion of data in the training set. This is commonly set at 70:30 (*train:test*) or 70:10:20 (*train:validate:test*), however if the model has a large number of hyperparameters to tune, then a larger validation proportion is prudent. The probability distributions of features in each set should also be the same i.e., stratified if possible.

Cross Validation

There are several concerns with the traditional *train:validate* approach to resampling. Particularly in smaller datasets, or those with uncommon outcomes, data is wasted insofar as it is not available to train the model. The model is also dependent on the random selection of the subsets, and this may bias it towards a particular subset. Increasingly, a cross-validation approach is being preferred, whereby the training set is split as before, but the process is repeated a number of times with different subsets of data. Multiple performance measures of the model can then be made, which gives a better estimate of generalisation. In addition, this method allows the optimal hyperparameters to be assessed and chosen for the final model. The common methods for cross validation are bootstrapping and *k*-fold cross validation. In bootstrapping, random samples from the training dataset are selected (with replacement - the size of the whole set is maintained) for training, and the non-selected samples are used for testing. This tends to reduce variance, but increase bias due to the stochastic nature of the sampling. (21)

k-fold cross-validation is becoming increasingly popular as an alternative resampling method as it gives a good estimate of the generalisation properties of the model as well as low bias.

The training set is divided into *k* equal folds or subsets (*k* is usually set at 5-10) and *k*-1 folds are used to train, while the remaining fold is held as the test set. (Figure 3) The process is repeated with each successive fold being held back as the test set, while the remaining *k*-1 are used to train. The performance over the *k* cycles is averaged to give an estimate of the performance of the model, and also an estimate of the expected variance in new, out-of-sample data (generalisation). (21,23) For small classes or rare categorical features, stratified *k*-fold cross validation should be used. (24) *K*-fold cross validation is particularly useful when evaluating several models in order to choose the best performing.



Split 1	Split 2	Split 3	...	Split 10
Fold 1	*Fold 1*	*Fold 1*	...	*Fold 1*
Fold 2	*Fold 2*	*Fold 2*	...	*Fold 2*
Fold 3	*Fold 3*	*Fold 3*	...	*Fold 3*
Fold 4	*Fold 4*	*Fold 4*	...	*Fold 4*
Fold 5	*Fold 5*	*Fold 5*	...	*Fold 5*
Fold 6	*Fold 6*	*Fold 6*	...	*Fold 6*
Fold 7	*Fold 7*	*Fold 7*	...	*Fold 7*
Fold 8	*Fold 8*	*Fold 8*	...	*Fold 8*
Fold 9	*Fold 9*	*Fold 9*	...	*Fold 9*
Fold 10	*Fold 10*	*Fold 10*	...	*Fold 10*
Metric 1	**Metric 2**	**Metric 3**	...	**Metric 10**

Testing set ■
Training set □

Figure 3 k-fold cross validation. Training set divided into k folds (k=10 in this example). k-1 folds are used to train; the remaining fold is used as the test set. The process is repeated with each fold being held back as the test set, and the remaining 9 used to train. Performance metrics are averaged over k splits to give an estimate of the out-of-sample performance of the model.

When a model is evaluated, all of the steps leading to the final model are in fact being evaluated. This includes data preparation/transformation, the algorithm(s), training and tuning/calibration.

It is clear that much of the process of model development is iterative, and although the steps are described in order, some anticipation of later steps is necessary at the outset.

Conversely, adjustments in earlier steps may be undertaken based on experience and information gained later in the model development pipeline.

Model Development

At the initial stages of identifying the question being asked of the proposed ML model, the type of solution (or task) should also be defined. The majority of ML algorithms used in EM will be for supervised classification (discrete binary or multi-class prediction) or regression (predicting a continuous value) tasks. Even complex tasks such as image recognition or text analysis can usually be broken down into either of these. Less commonly in EM, unsupervised learning models can find relationships with unlabelled data by using clustering, association or dimensionality-reduction algorithms, for example. It is therefore important to determine which of these tasks will best answer the question posed, given the characteristics of the available data. Once this is decided, the most appropriate algorithms can then be chosen for developing the model.

Unfortunately, there is no single algorithm that works best for a given task or clinical question. Interdependencies such as the type, quantity and quality of data as well as the proposed place in the ED workflow for the deployed models are key considerations. The aim is to choose an algorithm or group of algorithms which suits the task, performs best and generalises appropriately. It is generally preferable to use simple algorithms, such as decision trees, and then progress to more complex ones, such as ensembles or DNNs if necessary, as simpler methods do not necessarily equate to poorer performance compared with more complex algorithms. (6) The process often involves a degree of trial and error, with the caveat of more complex models generally being more difficult to interpret or visualise and requiring more computational power.

Ensemble learning combines diverse simpler algorithms with the aim of improving their performance. It is based on considering several algorithms' output in a democratic manner and using this to find the best combination. (30) The simplest ensembling methods use mean, weighted mean or mode to combine model outputs. More advanced ensembles include *stacking/blending* (sequentially building new models on each other using the performance of the previous model); *b*ootstrap *agg*regation (*bagging* – multiple independent models running in parallel on subsets of data and giving a combined output); and *boosting* (models created in sequence with each successive model correcting the errors of the previous model thus combining the strengths of weaker models into a strong model). (31)

Ensemble learning should create models with relatively fixed (or similar) bias and reduce variance by combining outputs. It should be noted that the ensemble may not necessarily perform better than any individual contributing algorithm/model, and in some cases, can perform worse than a single strong contributing model. (30) These more complex learning models are being used more frequently, but benchmarking them against their simpler constituent algorithms is essential.

Model Tuning & Calibration

Whilst model tuning can occur after validation against the hold-out test set, it is preferable that this be done during model development. It should be apparent that training, validation and tuning/re-tuning occurs iteratively and encompasses model evaluation using defined metrics and feedback. Tuning of model hyperparameters can improve performance by improving the skill of the model at making predictions, allowing recalibration and minimising overfitting by regularisation and optimisation. Examples of tuneable hyperparameters include the maximum number of features, number of trees, depth of forest, samples at each leaf node (Random Forest models), number of nodes and depth of layers (DNNs).

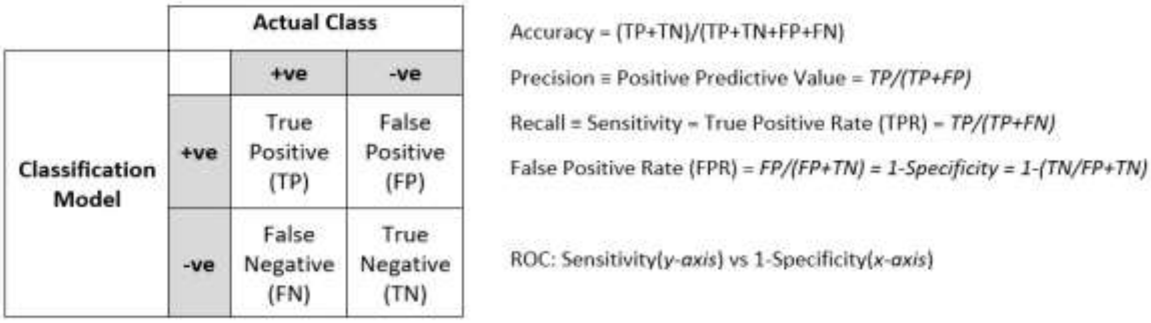
Calibration

In EM classification studies, it may be more practical to determine the probability of an input falling into a particular output class. This calibration allows an assessment of uncertainty of the model's prediction which is useful in clinical decision making and in communicating risk. This should also be the aim in regression where calibration is the adjustment in the model's predictions to more closely match actual outcomes. (21) Some algorithms will calibrate their predictions (such as logistic regression algorithms) automatically, whereas others such as DNNs or tree-based algorithms do not. There are several ways of assessing model calibration with varying robustness, however notably, the commonly used Hosmer-Lemeshow method is not recommended as it artificially risk stratifies patients and is uninformative regarding the type of miscalibration. (32)

Model Evaluation

Models are repeatedly evaluated - the process is a fundamental part of model development, internal validation and estimates of generalisability. Predictions from initial model configurations are compared with actual/target outputs, and readjusted iteratively to attain optimal performance or for between-model comparison by comparing performance metrics. Activities such as feature selection, cross-validation, tuning or comparison of several candidate algorithms are all grounded in robust model evaluation. Even post-deployment evaluation for stability and external generalisability are premised on defined evaluation metrics. Clearly, an objective measurement of the model's performance is central to model evaluation at all stages.

Reported ML performance metrics should mirror conventional statistical measures, and be standardised to allow interpretability in a clinical context. For both classification and regression tasks, the metric should reflect the intended clinical deployment. AUROC is a measure of discrimination (i.e., the ability to predict higher risk in patients with the outcome than those without), however, it is often used for a blanket comparison in classification tasks, but without consideration of the impact of clinical usage or class imbalance. For example, a good AUROC for a sepsis score does not necessarily translate to clinical utility for excluding a poor outcome. (33) Imbalanced classes is a recognised problem in AI, and metrics should be interpretable based on distribution and prevalence. (34,35) For example, Precision- Recall curves may be preferable in imbalanced classes where missing a true positive is undesirable. A confusion matrix (akin to a *n* x *n* contingency table, Figure 4) is usually reported for classification models (or regression models where a class threshold is set) and this represents good practice. It facilitates assessment of individual components of model performance, and allows determination of performance at the intended place in the ED clinical workflow. The standard statistical terms and metrics should be used rather than the ML terms as far as possible, for example sensitivity rather than recall.



		Actual Class	
		+ve	**-ve**
Classification Model	**+ve**	True Positive (TP)	False Positive (FP)
	-ve	False Negative (FN)	True Negative (TN)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision ≡ Positive Predictive Value = *TP/(TP+FP)*

Recall = Sensitivity = True Positive Rate (TPR) = *TP/(TP+FN)*

False Positive Rate (FPR) = *FP/(FP+TN) = 1-Specificity = 1-(TN/FP+TN)*

ROC: Sensitivity(*y-axis*) vs 1-Specificity(*x-axis*)

Figure 4 Confusion matrix with ML and standard statistical terms. ROC, receiver operator curve.

In regression tasks, the most commonly used metrics are related to the error (i.e. the difference between predicted and actual values) of the model. This can simply be the Mean Absolute Error (MAE) which weighs differences equally and is robust to outliers. However, it may not be as amenable to tuning and calibration as other metrics. Mean Squared Error (MSE), Root MSE and Coefficient of Determination (R^2) are other metrics which are preferable when large errors are undesirable or to compare with a constant baseline performance. In most cases, it is helpful if the error is interpretable in a clinical context, and so may require re-conversion into an absolute value. Similarly, interpretability of performance is facilitated by Bland-Altman plots when predictions are tied to a clinical measurement, and measures such as interobserver reliability may also be appropriate. (36)

The final model(s) will be assessed against the hold-out (*test*) dataset, where ideally, acceptable performance will be demonstrated using pre-determined metrics. An estimate of performance would also have been available from cross-validation. Confidence limits (or non-parametric estimates from cross-validation) are helpful in assessing the expected generalisability of performance metrics. Sensitivity analysis is suggested as a standard for model interpretation, with a range of performance reported (features of instances where it was most/least confident in classification tasks or had largest/smallest errors in regression tasks). (37)

External Validation and Generalisability

The ultimate aim in ML is to create generalisable models – those which can perform predictably well on new, unseen (clinical or non-clinical) information. The majority of published EM models do not consider this in detail, and are commonly framed on the data on which they were trained or tested in internal validation. In most cases, the testing data is derived from the same repository or hospital(s) as the training data, and therefore is too homogenous to confidently extrapolate to other temporally or geographically distinct settings or datasets.

There are measures which may reduce overfitting to training/internal validation data and the resulting non-generalisability concerns. Some of these have been discussed earlier, but in general features should be chosen which are reasonable clinically or causal predictors of the outcome of interest. This human plausibility or *explainability* is fundamental to the generalisability as well as clinical trustworthiness of the model.

Models vary in their complexity and often in vitro performance is inversely related to this complexity. Interpretability is often difficult with more complex models, meaning that the decision-making steps are not open to interrogation to determine why a model made a particular prediction. A balance between performance and interpretability/ explain-ability is important for AI models to be considered for adoption and for patients, clinicians and regulators to trust the outputs when deployed for actual patient decisions. (14,38) This is being facilitated by more transparent reporting of model pipelines and increasingly, visualisations of model decision-making are being harnessed in order to translate the mechanics of model predictions into a digestible and interpretable format for clinicians. (39)

The best immediate assessment of generalisability is the use of an *external validation* test dataset. Whereas a temporally separate validation sample is often acceptable in traditionally derived prediction models, this is not necessarily the case in ML models. (40) At least geographically distinct (and ideally temporally, by independent investigators if possible) is preferable to allow testing of performance on a dataset which is completely new to the model. This will approximate real-world deployment, and combined with estimates of spread/variance from internal validation, should provide confidence in model predictions when used clinically. The latter is of importance, as due to the stochastic nature of most algorithms, some variability in predictions is expected.

It is recognised that a truly independent dataset may not be available to model developers. Good practice entails provision of the code and processes used in model development so that, like with traditional research, enough detail is available for reproducibility by independent researchers. In some instances, where intellectual property concerns may be an issue, independent third-party evaluation may be utilised. Alternatively, a virtual machine can be made available, which allows the input of data from a clinician's own setting to test performance. (36,37) Not only can this approach to reproducibility provide an assessment of generalisability, but it also facilitates adaptation of the model to fit a particular setting, hence avoiding duplication of effort and research wastage. Furthermore, it will also allow consideration of model stability across variably resourced settings and as clinical care and knowledge evolves over time.

Conclusion

Interpretation of AI research in EM requires clear reporting and studies should meet research reporting standards by describing methodology transparently. The preceding overview of AI model development supplements the conceptual framework for interpretation and appraisal of AI studies covered in the companion paper. It should support clinicians in interpreting and undertaking AI studies (collaboratively with data scientists) and in critically appraising models which are proposed for use in their setting. This will facilitate a better understanding of the methods used in model pipelines and how these methods can be contextually adapted for various EM and AI tasks.

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

References

1. Ramlakhan S, Saatchi R, Sabir L, Singh Y, Hughes R, Shabayo O, et al. Understanding and Interpreting Artificial Intelligence, Machine Learning and Deep Learning in Emergency Medicine. Emerg Med J.

2. 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/

3. The EQUATOR Network. Diagnostic and prognostic studies [Internet]. [cited 2022 Feb 11]. Available from: https://www.equator-network.org/?post_type=eq_guidelines&eq_guidelines_study_design=diagnostic-prognostic-studies&eq_guidelines_clinical_specialty=0&eq_guidelines_report_section=0&s=

4. Marincowitz C, Paton L, Lecky F, Tiffin P. Predicting need for hospital admission in patients with traumatic brain injury or skull fractures identified on CT imaging: a machine learning approach. Emerg Med J [Internet]. 2021 Apr 7 [cited 2021 Oct 1]; Available from: https://emj.bmj.com/content/early/2021/04/07/emermed-2020-210776

5. 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

6. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019 Jun 1;110:12–22.

7. Hess EP, Hollander JE, Schaffer JT, Kline JA, Torres CA, Diercks DB, et al. Shared decision making in patients with low risk chest pain: prospective randomized pragmatic trial. BMJ. 2016 Dec 5;355:i6165.

8. Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, Ioannidis JPA. Diagnostic tests often fail to lead to changes in patient outcomes. Journal of Clinical Epidemiology. 2014 Jun 1;67(6):612–21.

9. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.

10. 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. Academic Emergency Medicine [Internet]. [cited 2021 Feb 7];n/a(n/a). Available from: https://www.onlinelibrary.wiley.com/doi/abs/10.1111/acem.14190

11. 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. The Lancet Digital Health. 2019 Oct 1;1(6):e271–97.

12. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. npj Digital Medicine. 2018 Aug 28;1(1):1–3.

13. 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

14. 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

15. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447–53.

16. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med. 2021 Sep 15;1–9.

17. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, et al. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. Can Assoc Radiol J. 2019 Nov 1;70(4):344–53.

18. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine. 2000;19(8):1059–79.

19. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making. 2012 Feb 15;12(1):8.

20. Ng A. Neural Networks and Deep Learning [Internet]. Coursera. [cited 2021 Jan 21]. Available from: https://www.coursera.org/learn/neural-networks-deep-learning

21. Kuhn M, Johnson K. Applied Predictive Modeling. 1st ed. New York: Springer; 2013. 613 p.

22. Gupta S, Gupta A. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. Procedia Computer Science. 2019 Jan 1;161:466–74.

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

24. 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. Journal of Medical Internet Research. 2016;18(12):e323.

25. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine. 2018 Nov 6;15(11):e1002683.

26. Kravchenko A. Feature Engineering and Feature Selection [Internet]. kaggle.com. [cited 2021 Jan 22]. Available from: https://kaggle.com/kashnitsky/topic-6-feature-engineering-and-feature-selection

27. Kuhn M, Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Models. 1st edition. Chapman and Hall/CRC; 2019. 310 p.

28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002 Jun 1;16:321–57.

29. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1–73.

30. Zhang C, Ma Y, editors. Ensemble Machine Learning: Methods and Applications. 1st ed. New York: Springer; 2012. 340 p.

31. Singh A. Ensemble Learning | Ensemble Techniques [Internet]. Analytics Vidhya. 2018 [cited 2020 Dec 27]. Available from:

https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

32. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019 Dec 16;17(1):230.

33. Sabir L, Ramlakhan S, Goodacre S. Comparison of qSOFA and Hospital Early Warning Scores for prognosis in suspected sepsis in emergency department patients: a systematic review. Emerg Med J [Internet]. 2021 Aug 17 [cited 2021 Oct 1]; Available from: https://emj.bmj.com/content/early/2021/08/16/emermed-2020-210416

34. Wainer J, Franceschinell RA. An empirical evaluation of imbalanced data strategies from a practitioner's point of view. arXiv:181007168 [cs, stat] [Internet]. 2018 Oct 16 [cited 2021 Feb 1]; Available from: http://arxiv.org/abs/1810.07168

35. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019 Mar 19;6(1):27.

36. 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: Cardiovascular Imaging. 2020 Sep 1;13(9):2017–35.

37. 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. Nature Medicine. 2020 Sep;26(9):1320–4.

38. Council on Artificial Intelligence. OECD Legal Instruments [Internet]. OECD. [cited 2021 Jan 30]. Available from: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

39. Sacha D, Sedlmair M, Zhang L, Lee JA, Peltonen J, Weiskopf D, et al. What you see is what you can change: Human-centered machine learning by interactive visualization. Neurocomputing. 2017 Dec 13;268:164–75.

40. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns. 2020 Nov;1(8):100129.