Emerging Topics in Life Sciences (2021) **5** 729–745
https://doi.org/10.1042/ETLS20210246

Review Article

Artificial intelligence, machine learning, and deep learning for clinical outcome prediction

Rowland W. Pettit[1], Robert Fullem[2], Chao Cheng[1,3] and Christopher I. Amos[1,3,4]

[1]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, U.S.A.; [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, U.S.A.; [3]Section of Epidemiology and Population Sciences, Department of Medicine, Baylor College of Medicine, Houston, TX, U.S.A.; [4]Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, U.S.A.

Correspondence: Christopher I. Amos (Chris.Amos@bcm.edu)

AI is a broad concept, grouping initiatives that use a computer to perform tasks that would usually require a human to complete. AI methods are well suited to predict clinical outcomes. In practice, AI methods can be thought of as functions that learn the outcomes accompanying standardized input data to produce accurate outcome predictions when trialed with new data. Current methods for cleaning, creating, accessing, extracting, augmenting, and representing data for training AI clinical prediction models are well defined. The use of AI to predict clinical outcomes is a dynamic and rapidly evolving arena, with new methods and applications emerging. Extraction or accession of electronic health care records and combining these with patient genetic data is an area of present attention, with tremendous potential for future growth. Machine learning approaches, including decision tree methods of Random Forest and XGBoost, and deep learning techniques including deep multi-layer and recurrent neural networks, afford unique capabilities to accurately create predictions from high dimensional, multimodal data. Furthermore, AI methods are increasing our ability to accurately predict clinical outcomes that previously were difficult to model, including time-dependent and multi-class outcomes. Barriers to robust AI-based clinical outcome model deployment include changing AI product development interfaces, the specificity of regulation requirements, and limitations in ensuring model interpretability, generalizability, and adaptability over time.

Introduction

In the modern era, the volume and variability of data available to understand and predict clinical outcomes are beyond the scope of singular human comprehension. For this reason, artificial intelligence (AI) methods are well-positioned to meaningfully assist in the clinical practice of medicine. AI is a broad concept, grouping together many initiatives that use a computer to perform tasks that would usually require a human to complete [1]. Examples of computational solutions which fall under the category of AI include perceiving visual stimuli, understanding speech, making decisions based on input data, and language translation [2]. ML is a sub-concept of AI, which focuses on having a machine perform an otherwise intelligent task by learning based on its errors to improve its capabilities with experience [3]. ML adapts and learns iteratively, without human feedback, by applying statistical models that identify patterns in data and draw useful inferences [4]. Finally, deep learning (DL) is a specific category within ML that uses various artificial neural network architectures to extract and process features within data. This hierarchy, narrowing in from broad to specific, can be appreciated in Figure 1, adapted from Min et al. [5]. The tangible products in the field of AI have evolved greatly over the last 30 years. We point the reader here for historical [6,7,8,9], technical [10], medically focused [11], and failure highlighting [12], reviews on the evolution of AI methods. This review will outline the progress, uses, and barriers to comprehensively integrating these emerging statistical and machine learning (ML) tools into clinical practice.

Received: 1 October 2021
Revised: 3 December 2021
Accepted: 7 December 2021

Version of Record published: 20 December 2021

Emerging Topics in Life Sciences (2021) **5** 729–745
https://doi.org/10.1042/ETLS20210246

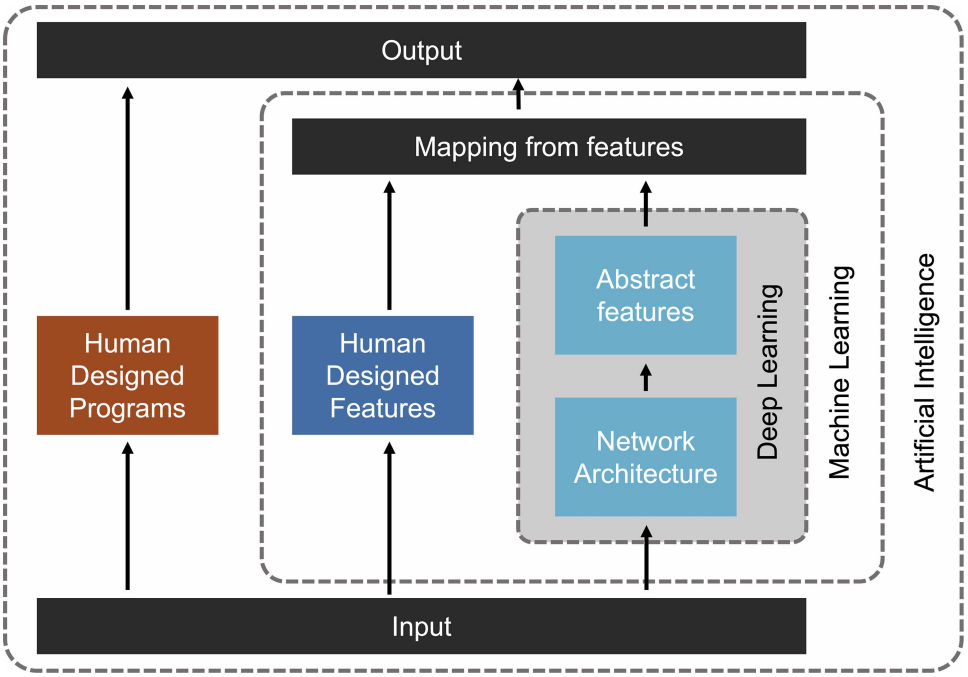

Figure 1. Representation of concepts: artificial intelligence, machine learning, and deep learning.

Clinical outcome predictions

Conceptual framework

Clinical outcome predictive models can be generalized down to the following concept. A model, represented as a function '*f*' is applied to input data X to represent a known outcome variable y [13]

$$f(X) \rightarrow y$$

The model *f* is 'fit,' or trained on the input data X so that when it encounters new data it can predict what y will be, or $\hat{y}$ [13]. The goal of creating a clinical outcome predictor is for your *f* to work well enough so that predicted values of $\hat{y}$ on new data would be correct if the actual status or value of y was known. Input data X can take many forms. However, usually input information is cleaned and processed into matrix form [14]. In this data object, each row, or 'instance,' represents a single entity or observation of the data (i.e. an individual patient), and each column, or 'feature,' represents a property of the data (i.e. the patient's age, or blood type). Models where only linear operations (*f*) act on input features (X) to predict an outcome (*y*) are referred to as generalized linear models, or GLMs [15]. Various non-linear, or other operations, however, may be implemented on these data (X). In general, a weight, *w,* is given to each feature as the features are combined. Using matrix notation, our general equation can be conceptualized as follows [13]:

$$f(w \cdot X) \rightarrow y$$

While the exact weighting and feature manipulation methods (*f*) are unique per ML method, generally, models are built through iterative training [13]. The ML method *f* is initiated with random parameter weights w_0, the model is fit to the input $f(w_0 \cdot X)$, and an initial outcome prediction $\hat{y}_0$ is generated. Then a loss function is created, which can take various forms (i.e. root mean squared error [16] or cross-entropy loss [17]) to compare each $\hat{y}$ prediction to the known outcome *y*. Loss functions (*L*) operate such that the closer the $\hat{y}$ prediction comes to accurately getting the real value of y, the smaller the function output is [18]. It is then relatively straightforward to optimize an algorithm by minimizing your defined loss function. During iterative training, the gradient of the loss function dictates how the weights of features (w) should be changed so that the loss

function is minimized. The goal of loss optimization is for the models' predictions ($\hat{y}$) to get closer and closer to their actual outcome value (y), which is observed as loss function converges to 0 or is minimized:

$$L(y, \hat{y})|w \rightarrow 0, \quad \min$$

What is an outcome, and how will you measure it?

With a general framework established, it is next necessary to consider the outcome, y, and decide how it should appropriately be measured. This step is essential as different ML methods lend themselves most appropriately to modeling differing outcome types. For example, the most straightforward clinical outcome that can be observed is that of a binary outcome. A binary outcome variable can only take two values and often represents a 'yes'/'no' event [19]. Clinically, these could include an outcome describing treatment failure versus success or patient mortality within a defined time period from an intervention. Almost all ML methods can be used to perform binary classification [13] However, perhaps you are interested in an outcome with more than two classes and need to create a multi-class classifier [20,21]. This could be the case when trying to differentiate among multiple types of dermatological lesions, including benign, melanoma, basal cell, and squamous cell carcinoma lesions. The next step up from a multi-class classifier is creating an ML model that can predict a continuous [22] clinical outcome. Perhaps you wish to predict the expected level for a biomarker or hope to predict the length of stay for patients who receive a procedure. An important consideration for a classifier is understanding if an outcome occurs over a time horizon. Is there a time component to incorporate into the model? For example, when looking at cancer recurrence after chemotherapy, recurrence rates are only an interesting if you know the period of remission, after chemotherapy but before recurrence. ML models can perform 'survival analysis' [23] tasks. Utilizing a time-dependent outcome variable comes with constraints, however. Instances where patients are in your data but have not yet experienced an outcome must be 'censored' appropriately. Censoring is accomplished in a nonparametric manor in classical statistics through the Kaplan–Meier [24] method. This compensatory method is imperfect, however, and fails to appropriately account for the informative censoring of competing risks [25], where patient dropout may be non-random and censored individuals have risk factors influencing the survival outcome of interest. Other survival analysis specific analyses requirements include assessing the informative missingness [26] of covariates [27], the impact of confounders [28], and latent heterogeneity of patient cohorts [29]. These considerations can be handled to a degree with advanced statistical methods, but not yet by ML methods. Further detailed insights into survival analysis are available in the literature [30,31].

Emerging methods and emerging applications

ML methods are rapidly being applied in novel avenues to predict clinical outcomes. To visualize, we have generated Figure 2 charting AI related PubMed searches by year. In Supplementary Table S1, we have complied recent review articles detailing emerging examples of how statistical and ML methods are being utilized for clinical outcome prediction in major medical specialities. Applications are found in the fields of Anesthesiology [32,33,34], Dermatology [35,36,37], Emergency Medicine [38,39], Family Medicine [40,40], Internal Medicine [41,42,43], Interventional Radiology [44,45], Medical Genetics [46], Neurological Surgery [47], Neurology [48,49,50], Obstetrics and Gynecology [51,52], Ophthalmology [53,54,55], Orthopaedic Surgery [56], Otorhinolaryngology [57,58], Pathology [59,60,61], Pediatrics [62], Physical Medicine and Rehabilitation [63,64], Plastic and Reconstructive Surgery [65,66], Psychiatry [67,68], Radiation Oncology [69,70], Radiology [71,72], General Surgery [73,74], Cardiothoracic Surgery [75,76], Urology [77,78], Vascular Surgery [79,80]. These papers introduce terms describing ML models as 'supervised' or 'unsupervised'. Supervised ML methods are trained to predict a specified outcome, while unsupervised models are given no outcome target and seek to identify patterns in data unguided [81]. Almost all clinical outcome statistical and ML models are supervised models [82] where a target is set beforehand. Let's investigate these methods mentioned individually.

1. *Linear Models.* Logistic and linear regression (LR) are the most straightforward predictive models. These are classical statistical techniques and are most accurately regarded as statistical learning methods [13]. LR methods combine input features in a linear combination to predict an outcome. When input features are independently correlated with the outcome, linear models perform very well, on par or even better than new ML methods [83,84]. LR methods do not capture non-linear relationships between variables, and,

Emerging Topics in Life Sciences (2021) **5** 729–745
https://doi.org/10.1042/ETLS20210246

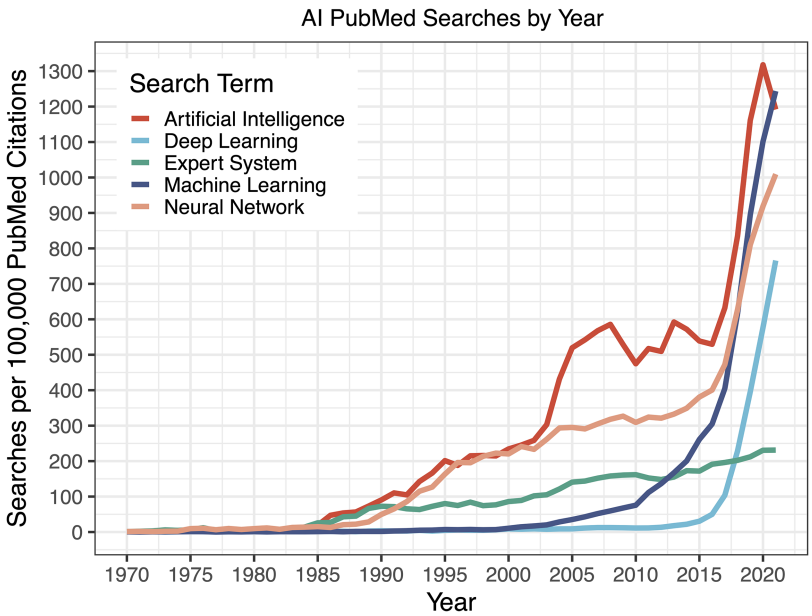

Figure 2. AI PubMed searches per 100 000 citations by Year.

without specific feature construction, treat all features independently [85]. LR models will continue to provide excellent insight into clinical outcome predictions as LR models are both computationally efficient and highly interpretable [15]. LR feature weights can be tested for individual significance and be understood as feature multipliers in relation to an outcome metric [86]. It is standard practice to benchmark performance of ML models to that of statistical learning LR methods to critically evaluate the need for a more complex, often less interpretable ML model [87]. As example, LR methods are recently being used to predict clinical outcomes of COVID-19 mortality [88], the development of chronic diseases such as HTN and DM [84], stroke risk [89], and predicting acute myeloid leukemia outcomes from patient gene signatures [90]. When outcomes evolve over time, linear cox-proportional hazards statistical models are used to estimate baseline and feature specific hazard ratios of an outcome continuously [91]. Cox models are statistically entrenched due to interpretability, simplicity, and enduring widespread incorporation [92]. Although not a regression technique, the Naïve Bayes (NB) ML method also appreciates data features independently toward an outcome of interest [93]. For binary outcome prediction, NB calculates the posterior probability of the positive outcome class for each numerical feature and each sub-category within categorical features totaling probabilities [94]. NB has been recently used to predict responses to chemotherapy [95] and to predict the development of Alzheimer's disease from genomic data [96].

2. *Decision Tree Methods.* An individual decision tree is a top-down flowchart-like structure in which nodes represent a decision point, determined by a single input feature, and branches from nodes continue to diverge reaching more terminal nodes (Figure 3). Node decision points are created from features through information theory in which features are split based on entropy or variance regarding the known outcome of interest. After training the outcomes into a decision tree, then a new data instance of information can be fed into the tree, and the node decisions can be followed to predict the likely clinical outcome.

The method may be used to create both Classification and Regression Trees, which produces the acronym CART [97] for brevity. Many tree-based methods exist. A random forest trains multiple decision trees on input data, with each subtree having only a subset of the total column feature variables to consider. After training, all trees of the forest are run in parallel on new data entries, and the majority prediction opinion of the forest determines the model's final prediction. This method has the advantage of making decision nodes to be created at minor features, forcing their appreciation, and avoiding a few strong predictors driving prediction in all scenarios. This ability has led to excellent clinical outcome predictions, including recently to predict stroke outcomes [98], drug response from clinical and serological markers [99], or mortality after traumatic brain injury

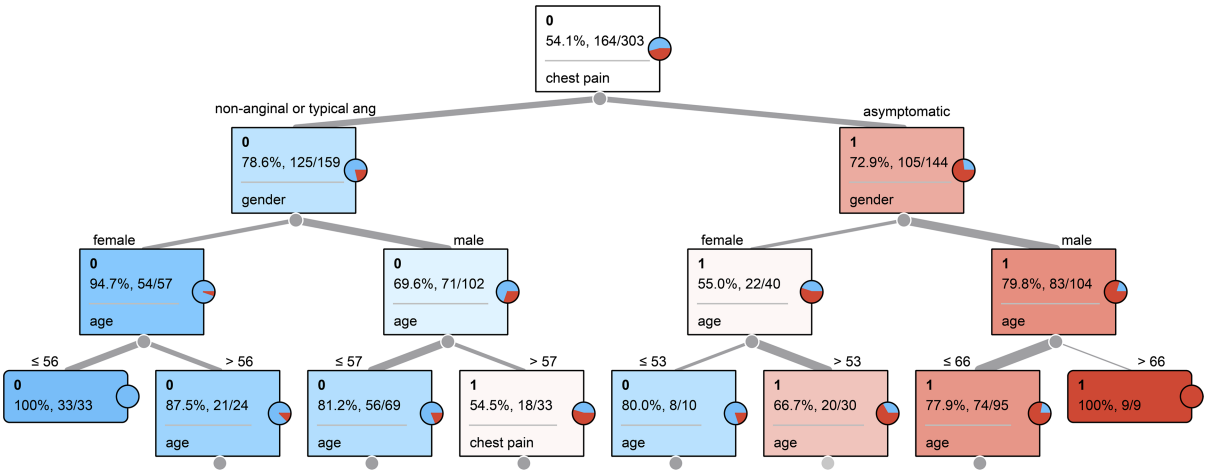

Figure 3. Example of a decision tree to predict coronary artery narrowing (1, red) vs no narrowing (0, blue) using input features of age, gender, and type of chest pain.

[100]. On small datasets with only a few highly correlated features, a random forest model may not perform better than simpler methods [101]. Two key concepts are introduced with the random forest. Combining several individual models to create one is known as ensemble modeling. Training multiple base models in parallel is known as 'bootstrap aggregation,' or 'bagging' [102]. Bagging is used in various statistical applications and does not require decision trees to exclusively serve as the underlying base models [103]. Boosting ensemble methods [61], by contrast, take a different strategy and train multiple models in series. By training sequentially, boosting affords later models the opportunity to learn from the previous models, or 'learners,' weaknesses. Popular boosting methods used in clinical modeling include XGBoost [104,105] and AdaBoost [106,107], which are often multi-decision tree ensemble methods [108,109]. Finally, when a time horizon outcome variable is used, novel random survival forest [110] methods can be used for time-dependent clinical predictions [111]. In general, tree-based methods are interpretable, can appropriately model non-linear relationships, and feature rankings of relative importance can be readily retrieved [112]. Tree-based methods are limited in that they require manual feature construction to appreciate multiple variables concurrently [113].

3. *Clustering, Kernel and Non-deterministic* methods. Clustering methods, in general, are unsupervised.

ML methods, however, they can be used for clinical outcome predictions. In the k-Nearest Neighbor (kNN) approach, clusters are found within data through 'k' number of random centroid placements, iterative Euclidian distance calculation between all data and centroids, recentering centroids to be in the center of 'nearest points', and reassignment of data cluster labels. KNNs have been utilized to cluster large-scale microRNA expression profiles into correctly classified human cancers [114,115,116]. Support vector machines [117] (SVM) are a kernel-based ML method that attempts to represent data into a higher dimensional feature space and find a hyperplane to separate samples by their outcome status [13]. SVM has limited utility when your input data has a large number of dimensions, as projecting all features into a higher dimensional space is computationally intensive, especially when using a non-linear kernels. They are used to predict outcomes when datasets are manageable [118]. Non-Deterministic methods are machine learning methods where a model is not constrained to create predictions in the context of the known outcome. For example, a non-deterministic classifier [119] trained to predict a binary outcome may be allowed to predict three or more states. The advantage of this is that the model has more 'options' to bucket borderline negative instances into when it is unsure of the appropriate class designation. This principle can ultimately lead to more correct classifications of the positive class. Such methods have been applied to clinical outcome prediction [120], where they demonstrated utility in predicting clinical cancer type [119]. More information on non-deterministic algorithms may be found here [121].

4. *Deep Learning*. Deep learning is a sub-category within ML, defined by the use of neural network architectures [5,122]. The most basic neural network architecture is a fully connected, or 'dense' [123], feed-forward network, or a 'multi-layer perceptron' [125]. In Figure 5 we can see how a deep neural network [124]

Emerging Topics in Life Sciences (2021) **5** 729–745
https://doi.org/10.1042/ETLS20210246

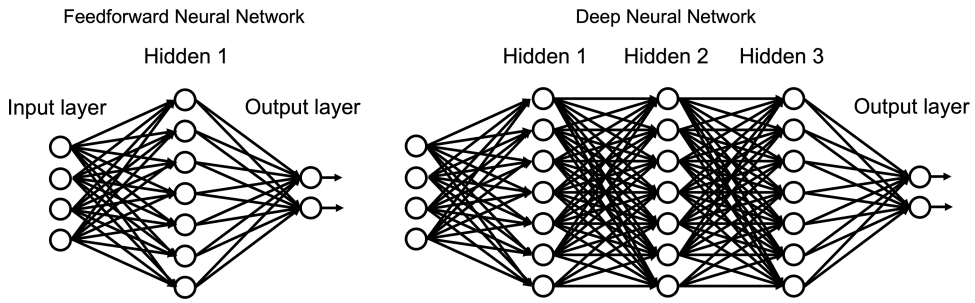

Figure 4. Fully connected (Dense) neural network versus deep neural network.

simply refers to having more than one hidden layer of interconnected nodes. In a general neural network architecture, the value of circles, or nodes, is the weighted sum of the outputs of nodes connected to it [125]. Line connections each have a weight, which is an individual parameter that is tuned during model training, modulated by optimizing your determined loss function [126]. To introduce non-linearity into the network, an activation function (ReLu [127], Sigmoid [128]) acts on threshold weighted inputs into a node. Feedforward neural networks are useful for clinical outcome prediction [129]. Deep neural networks (Figure 4) increase the number of hidden layers [130] between input and output and add advantages of more abstract feature representations [5]. Deep learning methods are being used extensively to predict clinical outcomes [131,132,133] When limited training instances are available, transfer learning [134] is appropriate. In transfer learning, a deep neural network model is pre-trained on a large adjacent type dataset, such as the ImageNet [135] database of 3.2 million images. This pre-trained model is then transferred and refitted with your smaller dataset. During this second step, the early hidden node layers of the network are 'frozen,' and only deep layer parameter weights can be iteratively modified. Freezing the weights and values of nodes in the first few layers protects fundamental information learned on the large dataset and only allows for 'fine tuning' of later nodes so that your desired outcome can be predicted. Transfer learning is widely utilized for clinical outcome prediction [136,137,138]. To facilitate transfer learning, initiatives exist to train large general base models on broad datasets to be utilized for future downstream tasks [139]. An example of such a foundational model includes Med-BERT [140] which is deep neural network model with pre-trained contextual embeddings specific for predicting disease from electronic health records. While experimental, and seemingly poised for powerful clinical modeling, caution prior to implementation is rightfully being taken to understand limitations of the foundational model which would be inherited by all downstream functions [139]. Dropout [141], or randomly removing nodes temporarily during training iterations, can prevent overfitting and improve model performance [142]. Survival neural networks [143,144] exist and are used for predicting time-dependent and censored clinical outcomes [145].

Recurrent neural networks allow for information stored in a node at a previous time point to connect with nodes at later time points. This historical feedback is the hallmark of a recurrent neural network, which allows for sequence or time-series data to be captured. RNNs are used to time-dependent outcomes such as epileptic siezures [146] and cancer treatment response [147]. Two common types of RNN are long short term memory (LSTM [148]) and gated recurrent units (GRU [149]) RNNs which allow for information to be carried and accessed for longer periods without information loss. Convolutional neural networks uniquely capture spatial information within data, and adjacent inputs must be related for CNN to be useful. CNNs have been utilized to predict malignancies from pulmonary nodules [150]. Overall, deep neural networks demonstrate superior performance on nearly all multimodal and image-based classification tasks, but are on par with other methods in regard to purely tabular inputs [151]. A limitation is their interpretability, as no singular features or direct feature weights are carried forward.

Data extraction and preprocessing

Each of these emerging methods requires access to reliable, standardized data input (X) that is appropriately captured to model an outcome of interest (y) [152]. To obtain and maintain X, extraction pipelines and preprocessing steps much be carefully attended. Often, this is the most time-consuming step in developing an ML

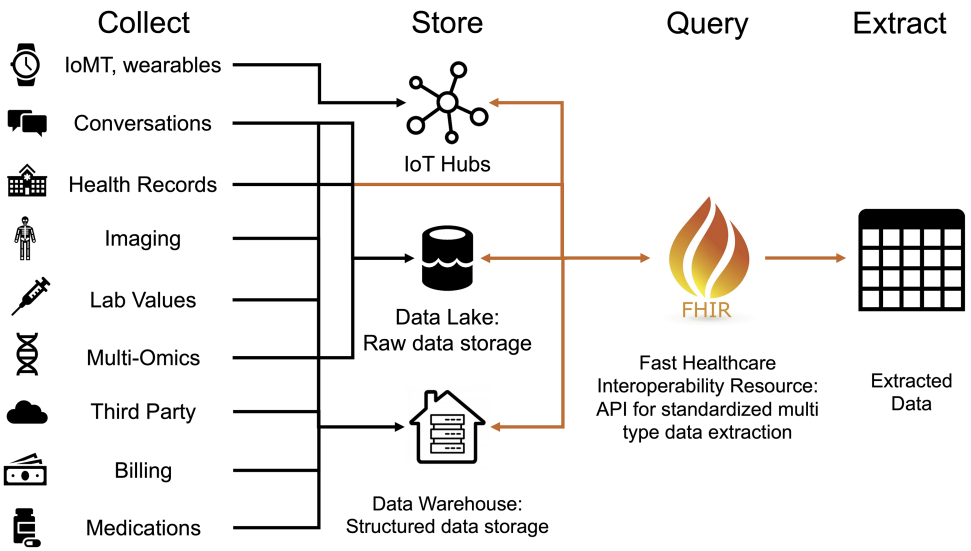

Figure 5. Healthcare data extraction standard pipeline.

model [153]. To predict our outcome ŷ accurately on new seen data, we will need to ensure that our training data X is generalizable [154] and representative of the population for which we aim to perform clinical outcome predictions.

1. *Data Accession and Extraction.* A convenient method of data storage is a clinical repository. Here data is stored in a data frame or table such that X is already formatted with patients listed as rows, and relevant feature variables for each patient are listed as columns. However, the necessary data will often not be available in this format and must be collated and transformed into the proper input format. The 2010 passage of the Affordable Care Act [155] included a mandate for health care providers to adopt electronic health record (EHR) systems. EHR records require a large amount of storage space, and due to their nature, cannot be recorded as one data table. Instead, data is often decentralized and made available through encrypted linking of data lakes [156] (raw unstructured) or data warehouses [157] (semi-structured or structured) data storage. Figure 5 shows how information is stored by hospital systems and can be collated on request or query. To perform clinical predictions, interfacing with these raw outputs, data lakes, and warehouses is required, and currently, several modalities exist to do so. One popular standard is FHIR [158] or the fast healthcare interoperability resource. This API Python coding tool provides a standard format in which, as a researcher, you can submit a query to FHIR, which provides the correct back-end commands to retrieve a properly formatted output table [158]. These tools become particularly useful when trying to concatenate clinical data with genetic sampling data and other individual lab or other biomarker values that might exist in various datasets. In general, such processes outside of FHIR can be accomplished on multiple platforms through merging datasets [159] with overlapping patient instances or concatenating data instances to an already existing data set. FHIR [160] directed EHR extraction to clinical outcome prediction pipelines [161] are incipient, and examples include predicting opioid use after spine surgery [162], outcomes and superiority of chronic disease treatment methods [163], and others [164]. Data extraction, or accession pipelines [162] far more complex than these, are also being explored and implemented to conduct clinical outcome predictions. To circumnavigate inter-institutional competition, privacy, permission, and remote storage issues, the use of blockchain technology for data accession rather than extraction is an emerging method being pursued [165]. Specifically, 'swarm learning' allows for decentralization and confidentiality of data to be maintained, which may increase intra-institution EHR center participation, and the overall sample sizes available for clinical outcome predictions [165].
2. *Data Preprocessing.* Now that the raw data is extracted (or accessed) and combined into a meaningful data set representation, the often tedious and challenging work of data cleaning and preprocessing can take place. Inattention to these steps that can lead to inaccurately performing ML models. We will draw on both well-established statistical data preparation methods, and more recent preprocessing approaches to prepare

Emerging Topics in Life Sciences (2021) **5** 729–745
https://doi.org/10.1042/ETLS20210246

these data. The first consideration should be the appropriateness of features included [166]. Given the natural variance or 'noise' of real-world data, features may spuriously have predictive properties in your testing data that are not reflected in real clinical settings. Therefore, ask yourself if the inclusion of a variable makes sense for your outcome of interest and discard features not expected to provide predictive value. Also, ask yourself, 'how was this variable recorded?' ML models can draw inferences from variables that may not be expected [167]. For example, if a relatively innocuous feature such as functional status in the clinical data set is only recorded for very sick individuals, then a model predicting death may be weighting that variable for its presence or absence independent of the score itself. After this first crucial step, a second step would involve feature construction [168]. Are there already established metrics or outputs that need to be constructed from the features at hand? Feature construction can be very beneficial to overall model performance [169], as it forces the model to consider combinations of features during training explicitly. Feature transformation [170] is an additionally critical. It is often advantageous to convert data from one form to another, such as continuous to discrete. For example, 'age' is a popular continuous variable that can be helpful to bin into discrete intervals. Outlier removal [171] is often warranted when using real-world data, which may include erroneous results or other unhelpful extreme values. Any method to remove outliers, however, should be standardized [171]. Several methods exist, including removing an outlier with a one-class support vector machine [172], a covariance estimator [173], or an isolation forest [174]. Finally, it may be appropriate to perform feature scaling [175], which could involve applying log transformations or other scales to create features with better value distributions and ranges.

3. *Optimizing row instances.* Generally, increasing your training data strengthens final model performance in predicting outcomes [176]. If you have few training instances, it may be appropriate to increase your data set size through resampling methods or synthetic data generation. Resampling methods [177] include different cross-validation procedures which involve more complex partitioning and reuse of training and testing samples. To increase the sample size, you could also create synthetic data. Many statistical methods exist for generating synthetic data [178], such as SMOTE [179], and all of which serve to produce new synthetic samples that are similar but plausibly different from the original 'true' samples. Synthetic data allows the model to improve performance by giving additional samples to iterate over while optimizing feature weights. Adding synthetic data can improve the number of minority class samples in the dataset. It is a common challenge of ML modeling for a model to underweight rare minority classes [180]. To force the model to more reliably predict the minority class, you can up-sample that class through synthetic data generation. Alternatively, if your sample size is sufficient, you may down sample the majority class [181] to better balance your input data, although care should be taken not to exclude relevant subgroups. An important distinction is that while synthetic or resampled can be applied to model training data, it is generally not acceptable to include synthetic or resampled data in testing datasets.

Similar to synthetic data generation, the statistical imputation [182] of missing values is a powerful tool during data preprocessing. You may have nearly complete data, where a variable may not be populated for a few samples. Depending on the data type, you can impute the missing data. Imputing is the idea of using the context clues from the surrounding features and what has been observed elsewhere in the dataset to estimate what the missing value or parameter should be. Imputation is common in biological contexts [182].

4. *Optimizing column features.* Metrics such as information gain [183] either through the GINI index, or the information gain ratio [184] are statistical metrics that can be used to determine how much information from a potential input feature is given to the outcome. Features giving relatively no information gain may be candidates for removal. Several methods can capture the information stored in multiple features but convey them in fewer features. This concept is known as dimensionality reduction [185], and it can be useful when consistent standardized data are inaccessible and features are abundant. Statistical methods, such as principal component analysis [186], and unsupervised ML methods, including t-SNE [187], and UMAP [188], can serve this purpose. Also useful for clustering analysis, these methods can reduce high dimensional data into a lower-dimensional representation, to accomplish the dimension reduction goal. These methods can be described as 'representation learning' or identifying a lower-dimensional feature to represent higher dimensional data

A common challenge arises when dealing with categorical variables, such as blood type that lack ordinal relationships. Unaltered input of this feature as integer representations (1, 2, 3, etc.) would falsely convey a natural

		Model Prediction		Metrics
Ground Truth	Total Population = TP + TN + FP + FN	Test +	Test -	Accuracy = (TP + TN) / Total
	Known +	True Positive (TP)	False Negative (FN)	Sensitivity, Recall, True positive rate = TP / (TP + FN)
	Known -	False Positive (FP)	True Negative (TN)	Specificity, True negative rate = TN / (FP + TN)
	Metrics	Positive predictive value, Precision = TP / (TP + FP)	Negative predictive value = TN / (FN + TN)	F1 score = 2 × (precision × recall) / (precision + recall)

Figure 6. Confusion matrix for model evaluation and formulas for calculating summary statistics.

ordering to these data that is not present biologically. To address this, the statistical method of one hot encoding [189] exists, which converts categorical variable features into individual columns for each of its subcomponents. One hot encoding is useful, but it can lead to expansive datasets. As a consequence of training a neural network on a categorical variable, the network 'embedding' representation of the categorical variable holds representational value. These categorical embeddings [190] can be extracted from the trained neural network and used in place of a categorical variable to represent the feature, uniquely capturing unobvious inter-feature relationships accurately [191], improving the end trained ML model's performance [192].

Evaluation

With our outcome identified, method selected, and input data preprocessed, it is time to train our model. We need to separate the data set X into training and testing data. An 80%: 20% training to testing split ratio of the total data is common [193]. In sectioning data, it is useful to ensure that all output classes are represented and any sub-type demographics of interest in both the testing and training sections. ML development occurs on the 80% training set. Model training should never involve the test set [13]. We want the test set to be an objective example of what new data would look like if given to the model. Cross validation [194] and random sampling with replacement [195] are useful repeated sampling metrics to estimate average model performance, in the case your initial train-test split happened to be unusually favorable or unfavorable toward model generation.

During training, the ML method will iteratively attempt to minimize the loss function. Training will stop after a preset number of training iterations, or when a certain loss function threshold is reached. At this point, your model can be trialed against the test data, and the difference between predictions generated on the test data $\hat{y}$ and the ground truth y can be compared. For classification problems, ML models will output a class probability score.

Many performance evaluation metrics are available for understanding how $\hat{y}$ compares to y. This class probability score can be visualized as an area under the receiver operator curve (AUC –ROC [196,197]) or an area under the precision-recall curve (AUC-PR [198]). Thresholding the probability score will allow for a class prediction to be made. A confusion matrix can next be generated, indicating how many of the test set instances were correctly classified as positive (true positive) or negative (true negative) and incorrectly classified as false positive and false negative outcomes. Figure 6 shows how to calculate relevant outcome summary evaluation metrics from these findings. F1, which is the harmonic mean of precision and recall, is commonly used to compare ML methods, including in cases of class imbalance [199].

Conclusion and future directions

AI methods are well suited to predict clinical outcomes. A great amount of methodological and application development has occurred and now serves as a precedent, inspiring even further method trialing and advancement. Currently, the methods for cleaning, creating, accessing, extracting, augmenting, and representing data are well defined. Appropriate procedures are available to model different outcome types, and methods are ever

Emerging Topics in Life Sciences (2021) **5** 729–745
https://doi.org/10.1042/ETLS20210246

evolving to predict clinical outcomes more accurately. We still find that barriers to robust ML implementation into clinical practice, however, remain. The interfaces used for implementing AI methods are undergoing rapid competitive selection. In addition to a coding background and familiarity with statistics, to perform ML well, one needs to have access to professional statistical software (STATA [200], SAS [201], Matlab [202]) or know how to code in R [203] or Python [204]. We view the continued development of user-friendly ML software, including Excel plug ins [203], Orange [205], and KNIME [206] development for visual coding, will increase accessibility, understanding, and use. An additional barrier to robust implementation is regulation. Currently, diagnostic predictors or clinician assist tools are viewed as 'software-as-medical' devices [207]. To get regulatory approval, a model must demonstrate superiority in the clinical predictions it creates over a physician or group of physicians. As the FDA navigates these uncharted regulatory waters, perhaps a perspective shift to non-inferiority will occur, allowing for more ML model large-scale adoption in clinical settings. To hire a radiologist, does the new individual need to be better than all the radiologists that came before them? Should this be the same logic used in approving clinician assist and support tools? Finally, ML methods will need to increase in their proof of interpretability, generalizability, and adaptability. As deep learning gets increasingly complex, how can we verify predictions are not being biased? How will models be updated to avoid becoming stagnant, no longer representing shifted populations and parameters? These key questions will need to be addressed for widescale deployment and acceptance of ML clinical outcome predictors going forward. In addition to overcoming these barriers, we suggest the future direction of AI clinical outcome predictions to increase focus on personalized medicine and create strong models that can be used for person-specific goals.

Summary

- AI methods are well suited to predict clinical outcomes.
- Current methods for cleaning, creating, accessing, extracting, augmenting, and representing data for the use of AI clinical prediction are well defined and ready for implementation.
- The use of AI to predict clinical outcomes is a dynamic and rapidly evolving arena, with new methods and applications emerging.
- Barriers to robust AI clinical outcome prediction include changing AI development interfaces, regulation requirements, and limitations in model interpretability, generalizability, and adaptability over time.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

Cancer Prevention Research Interest of Texas (CPRIT) award: RR170048 (CIA); National Institutes of Health (NIH) for INTEGRAL consortium: U19CA203654 (CIA); National Institutes of Health (NIH): R01CA139020 (MLB); NIH T32ES027801 (RWP).

Author Contributions

All authors contributed to the writing and editing of the manuscript.

Acknowledgements

The authors would like to thank the Baylor College of Medicine Medical Scientist M.D./Ph.D. Training Program for their support (RWP). The authors would like to thank BRASS: Baylor Research Advocates for Student Scientists for their support (RWP).

Abbreviations

AI, Artificial Intelligence; AUC-PR, Area under precision-recall curve; AUC-ROC, Area under receiver operator curve; CNN, Convolutional Neural Network; DL, Deep Learning; EHR, Electronic Health record; FDA, The U.S. Food and Drug Administration; FHIR, Fast Healthcare Interoperability Resources; FN, False Negative; FP, False Positive; GRU, Gated recurrent units recurrent neural network; IoMT, Internet of Medical Things; IoT, Internet of Things; kNN, k-Nearest neighbor; LR, Linear regression; LSTM, Long short term memory recurrent neural network; ML, Machine Learning; NB, Naïve Bayes; PCA, principal component analysis; RNN, Recurrent neural network; SVM, Support vector machines; TN, True Negative; TP, True Positive; t-SNE, t-stochastic neighbor embedding; UMAP, Uniform Manifold Approximation and Projection.

References

1 Dobrev D. A Definition of Artificial Intelligence. Published online October 3, 2012. Accessed September 26, 2021. https://arxiv.org/abs/1210.1568v1
2 McCracken, J. (2003) Oxford dictionary of English. In:123. https://doi.org/10.3115/1067737.1067764
3 Lv, H. and Tang, H. (2011) Machine learning methods and their application research. *Proceedings - 2011 International Symposium on Intelligence Information Processing and Trusted Computing*, IPTC. Published online 2011:108–110. https://doi.org/10.1109/IPTC.2011.34
4 Wang, H., Ma, C. and Zhou, L. (2009) A brief review of machine learning and its application. Proceedings - 2009 International Conference on Information Engineering and Computer Science, ICIECS 2009. Published online 2009. https://doi.org/10.1109/ICIECS.2009.5362936
5 Min, S., Lee, B. and Yoon, S. (2017) Deep learning in bioinformatics. *Brief. Bioinformatics* **18**, 851–869 https://doi.org/10.1093/BIB/BBW068
6 Haenlein, M. and Kaplan, A. (2019) A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence; 61(4):5–14. https://doi.org/10.1177/0008125619864925
7 Benko, A. and Sik Lányi, C. (2009) History of Artificial Intelligence. *Encyclopedia of Information Science and Technology*, Second Edition. 1759–1762. https://doi.org/10.4018/978-1-60566-026-4.CH276
8 Nilsson, N. (2009) *The Quest for Artificial Intelligence*. United States, Cambridge University Press. Accessed December 2, 2021. https://www.google.com/books/edition/The_Quest_for_Artificial_Intelligence/nUJdAAAAQBAJ?hl=en&gbpv=0
9 Dick, S. (2019) Artificial intelligence. *Harvard Data Sci. Rev.* **1**, 1 https://doi.org/10.1162/99608F92.92FE150C
10 Coolen, ACC, Keuhn, R and Sollich, P. (2005) *Theory of Neural Information Processing Systems*. OUP Oxford Accessed December 2, 2021. https://www.google.com/books/edition/Theory_of_Neural_Information_Processing/bVpnKFLM4RcC?hl=en&gbpv=0
11 Kaul, V., Enslin, S. and Gross, S.A. (2020) History of artificial intelligence in medicine. *Gastrointest. Endosc.* **92**, 807–812 https://doi.org/10.1016/J.GIE.2020.06.040
12 Smith, B. (2019) *The Promise of Artificial Intelligence Reckoning and Judgment*. Cambridge, MA, United States, MIT Press
13 Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning. Published online 2009. https://doi.org/10.1007/978-0-387-84858-7
14 El Naqa, I. and Murphy, M.J. (2015) What Is Machine Learning? *Machine Learning in Radiation Oncology*. Published online:3–11. https://doi.org/10.1007/978-3-319-18305-3_1
15 Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *J. R. Stat. Soc. A (Gen)* **135**, 370–384 https://doi.org/10.2307/2344614
16 Wallach, D. and Goffinet, B. (1989) Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecol. Model.* **44**, 299–306 https://doi.org/10.1016/0304-3800(89)90035-5
17 Li, L., Doroslovacki, M. and Loew, M.H. (2020) Approximating the gradient of cross-entropy loss function. *IEEE Access* **8**, 111626–111635 https://doi.org/10.1109/ACCESS.2020.3001531
18 Bottou, L. (2010) Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT 2010 - 19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers*. Published online 177–186. https://doi.org/10.1007/978-3-7908-2604-3_16
19 Koyejo, O., Natarajan, N., Ravikumar, P. and Dhillon, I.S. Consistent Binary Classification with Generalized Performance Metrics
20 Aly, M. (2005) undefined. Survey on multiclass classification methods. Citeseer. Published online 2005. Accessed December 11, 2021.
21 Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. Published online August 13, 2020. Accessed December 11, 2021. https://arxiv.org/abs/2008.05756v1
22 Wang, Y. and Witten, I.H. (1997) Induction of model trees for predicting continuous classes. Proceedings of the 9th European Conference on Machine Learning Poster Papers. Published online 1997:128–137. Accessed September 30, 2021. https://researchcommons.waikato.ac.nz/handle/10289/1183
23 Ohno-Machado, L. (2001) Modeling medical prognosis: survival analysis techniques. *J. Biomed. Inform.* **34**, 428–439 https://doi.org/10.1006/JBIN.2002.1038
24 Bland, J.M. and Altman, D.G. (1998) Survival probabilities (the kaplan-Meier method). *BMJ* **317**, 1572–1580 https://doi.org/10.1136/BMJ.317.7172.1572
25 Satagopan, J.M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D. and Auerbach, A.D. (2004) A note on competing risks in survival data analysis. *Br. J. Cancer* **91**, 1229–1235 https://doi.org/10.1038/sj.bjc.6602102
26 Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y. and Ranganath, R. (2020) A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits on Translational Science Proceedings*. 2020;2020:191. Accessed December 1, 2021. /pmc/articles/PMC7233077/
27 Katki H, LyX SMUS with, 2008 undefined. Survival analysis for cohorts with missing covariate information. 19221812911. Accessed December 1, 2021. https://scholar.google.com/ftp://192.218.129.11/pub/CRAN/doc/Rnews/Rnews_2008-1-1.pdf#page=14
28 Nieto, F.J. and Coresh, J. (1996) Adjusting survival curves for confounders: a review and a new method. *Am. J. Epidemiol.* **143**, 1059–1068 https://doi.org/10.1093/OXFORDJOURNALS.AJE.A008670
29 Aalen, O.O. (1988) Heterogeneity in survival analysis. *Stat. Med.* **7**, 1121–1137 https://doi.org/10.1002/SIM.4780071105
30 Flynn, R. (2012) Survival analysis. *J. Clin. Nurs.* **21**, 2789–2797 https://doi.org/10.1111/j.1365-2702.2011.04023.x

31 Kleinbaum, D. and Klein, M. (2010) Survival Analysis. Accessed December 1, 2021. https://link.springer.com/content/pdf/10.1007/978-1-4419-6646-9.pdf
32 Chae, D. (2020) Data science and machine learning in anesthesiology. *Korean J. Anesthesiol.* **73**, 285 https://doi.org/10.4097/KJA.20124
33 Alexander, J.C., Romito, B.T. and Çobanoğlu, M.C. (2020) The present and future role of artificial intelligence and machine learning in anesthesiology. *Int. Anesthesiol. Clin.* **58**, 7–16 https://doi.org/10.1097/AIA.0000000000000294
34 Hashimoto, D.A., Witkowski, E., Gao, L., Meireles, O. and Rosman, G. (2020) Artificial intelligence in anesthesiology current techniques, clinical applications, and limitations. *Anesthesiology* **132**, 379–394 https://doi.org/10.1097/ALN.0000000000002960
35 Du, A.X., Emam, S. and Gniadecki, R. (2020) Review of machine learning in predicting dermatological outcomes. *Front. Med.* **7**, 266 https://doi.org/10.3389/FMED.2020.00266/BIBTEX
36 Thomsen, K., Iversen, L., Titlestad, T.L. and Winther, O. (2019) Systematic review of machine learning for diagnosis and prognosis in dermatology. *J. Dermatolog. Treat.* **31**, 496–510 https://doi.org/10.1080/09546634.2019.1682500
37 Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., Brownstone, N. and Liao, W. (2020) Machine learning in dermatology: current applications, opportunities, and limitations. *Dermatol. Ther.* **10**, 365–386 https://doi.org/10.1007/S13555-020-00372-0/FIGURES/3
38 Stewart, J., Lu, J., Goudie, A., Bennamoun, M., Sprivulis, P., Sanfillipo, F. et al. (2021) Applications of machine learning to undifferentiated chest pain in the emergency department: a systematic review. *PLoS ONE* **16**, e0252612 https://doi.org/10.1371/JOURNAL.PONE.0252612
39 Tang, K.J.W., Ang, C.K.E., Constantinides, T., Rajinikanth, V., Acharya, U.R. and Cheong, K.H. (2021) Artificial intelligence and machine learning in emergency medicine. *Biocybernet. Biomed. Eng* **41**, 156–172 https://doi.org/10.1016/J.BBE.2020.12.002
40 Kueper, J.K., Terry, A.L., Zwarenstein, M. and Lizotte, D.J. (2020) Artificial intelligence and primary care research: a scoping review. *Ann. Fam. Med.* **18**, 250–258 https://doi.org/10.1370/AFM.2518
41 Ben-Israel, D., Jacobs, W.B., Casha, S., Lang, S., Ryu, W.H.A., de Lotbiniere-Bassett, M. et al. (2020) The impact of machine learning on patient care: a systematic review. *Artif. Intell. Med.* **103**, 101785 https://doi.org/10.1016/J.ARTMED.2019.101785
42 Pandit, A. and Radstake, T.R.D.J. (2020) Machine learning in rheumatology approaches the clinic. *Nat. Rev. Rheumatol.* **16**, 69–70 https://doi.org/10.1038/s41584-019-0361-0
43 Stafford, I.S., Kellermann, M., Mossotto, E., Beattie, R.M., MacArthur, B.D. and Ennis, S. (2020) A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit. Med.* **3**, 1–11 https://doi.org/10.1038/s41746-020-0229-3
44 Mazaheri, S., Loya, M.F., Newsome, J., Lungren, M. and Gichoya, J.W. (2021) Challenges of implementing artificial intelligence in interventional radiology. *Semin. Interv. Radiol.* **38**, 554–559 https://doi.org/10.1055/S-0041-1736659
45 Desai, S.B., Pareek, A. and Lungren, M.P. (2021) Current and emerging artificial intelligence applications for pediatric interventional radiology. *Pediatr. Radiol.* **2021**, 1–5 https://doi.org/10.1007/S00247-021-05013-Y
46 Rauschert, S., Raubenheimer, K., Melton, P.E. and Huang, R.C. (2020) Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. *Clin. Epigenet.* **12**, 1–11 https://doi.org/10.1186/S13148-020-00842-4/TABLES/2
47 Buchlak, Q.D., Esmaili, N., Leveque, J.C., Farrokhi, F., Bennett, C., Piccardi, M. et al. (2019) Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurg. Rev.* **43**, 1235–1253 https://doi.org/10.1007/S10143-019-01163-8
48 Myszczynska, M.A., Ojamies, P.N., Lacoste, A.M.B., Neil, D., Saffari, A., Mead, R. et al. (2020) Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat. Rev. Neurol.* **16**, 440–456 https://doi.org/10.1038/s41582-020-0377-8
49 Sirsat, M.S., Fermé, E. and Câmara, J. (2020) Machine learning for brain stroke: a review. *J. Stroke Cerebrovasc. Dis.* **29**, 105162 https://doi.org/10.1016/J.JSTROKECEREBROVASDIS.2020.105162
50 Yuan J, Ran X, Liu K, et al. Machine Learning Applications on Neuroimaging for Diagnosis and Prognosis of Epilepsy: A Review. Published online February 5, 2021. Accessed December 1, 2021. https://arxiv.org/abs/2102.03336v3
51 Iftikhar, P.M., Kuijpers, M.V., Khayyat, A., Iftikhar, A. and DeGouvia De Sa, M. (2020) Artificial intelligence: a new paradigm in obstetrics and gynecology research and clinical practice. *Cureus* **12**, e7124 https://doi.org/10.7759/CUREUS.7124
52 Sone, K., Toyohara, Y., Taguchi, A., Miyamoto, Y., Tanikawa, M., Uchino-Mori, M. et al. (2021) Application of artificial intelligence in gynecologic malignancies: a review. *J. Obstet. Gynaecol. Res.* **47**, 2577–2585 https://doi.org/10.1111/JOG.14818
53 Sengupta, S., Singh, A., Leopold, H.A., Gulati, T. and Lakshminarayanan, V. (2020) Ophthalmic diagnosis using deep learning with fundus images–a critical review. *Artif. Intell. Med.* **102**, 101758 https://doi.org/10.1016/J.ARTMED.2019.101758
54 Sarhan, M.H., Nasseri, M.A., Zapp, D., Maier, M., Lohmann, C.P., Navab, N. et al. (2020) Machine learning techniques for ophthalmic data processing: a review. *IEEE J. Biomed. Health Inform.* **24**, 3338–3350 https://doi.org/10.1109/JBHI.2020.3012134
55 Armstrong, G.W. and Lorch, A.C. (2020) A(eye): a review of current applications of artificial intelligence and machine learning in ophthalmology. *Int. Ophthalmol. Clin.* **60**, 57–71 https://doi.org/10.1097/IIO.0000000000000298
56 Ogink, P.T., Groot, O.Q., Karhade, A.V., Bongers, M.E.R., Oner, F.C., Verlaan, J.J. et al. (2021) Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review. *Acta Orthop.* **92**, 526–531 https://doi.org/10.1080/17453674.2021.1932928
57 Standiford, T.C., Farlow, J.L., Brenner, M.J., Conte, M.L. and Terrell, J.E. (2021) Clinical decision support systems in otolaryngology–head and neck surgery: a state of the art review. *Otolaryngol. Head Neck Surg.* **165**,1–227 https://doi.org/10.1177/01945998211004529
58 Crowson, M.G., Ranisau, J., Eskander, A., Babier, A., Xu, B., Kahmke, R.R. et al. (2020) A contemporary review of machine learning in otolaryngology–head and neck surgery. *Laryngoscope* **130**, 45–51 https://doi.org/10.1002/LARY.27850
59 Thakur, N., Yoon, H. and Chong, Y. (2020) Current trends of artificial intelligence for colorectal cancer pathology image analysis: a systematic review. *Cancers* **12**, 1884 https://doi.org/10.3390/CANCERS12071884
60 Sultan, A.S., Elgharib, M.A., Tavares, T., Jessri, M. and Basile, J.R. (2020) The use of artificial intelligence, machine learning and deep learning in oncologic histopathology. *J. Oral Pathol. Med.* **49**, 849–856 https://doi.org/10.1111/JOP.13042
61 McAlpine, E.D., Michelow, P. and Celik, T. (2021) The utility of unsupervised machine learning in anatomic pathology. *Am. J. Clin. Pathol.* **156**, 1–166 https://doi.org/10.1093/AJCP/AQAB085
62 Hoodbhoy, Z., Jeelani, S.M., Aziz, A., Habib, M.I., Iqbal, B., Akmal, W. et al. (2021) Machine learning for child and adolescent health: a systematic review. *Pediatrics* **147**, e2020011833 https://doi.org/10.1542/PEDS.2020-011833/33441

63 Khera, P. and Kumar, N. (2020) Role of machine learning in gait analysis: a review. *J. Med. Eng. Technol.* **44**, 441–467 https://doi.org/10.1080/03091902.2020.1822940
64 Amorim, P., Paulo, J.R., Silva, P.A., Peixoto, P., Castelo-Branco, M. and Martins, H. (2021) Machine learning applied to low back pain rehabilitation–a systematic review. *Int. J. Digit. Health* **1**, 10 https://doi.org/10.29337/IJDH.34
65 Mantelakis, A., Assael, Y., Sorooshian, P. and Khajuria, A. (2021) Machine learning demonstrates high accuracy for disease diagnosis and prognosis in plastic surgery. *Plastic Reconstr. Surg. Glob. Open.* **9**, e3638 https://doi.org/10.1097/GOX.0000000000003638
66 Huang, S., Dang, J., Sheckter, C.C., Yenikomshian, H.A. and Gillenwater, J. (2021) A systematic review of machine learning and automation in burn wound evaluation: a promising but developing frontier. *Burns* **47**, 1691–1704 https://doi.org/10.1016/J.BURNS.2021.07.007
67 Le Glaz, A., Haralambous, Y., Kim-Dufor, D.H., Lenca, P., Billot, R., Ryan, T.C. et al. (2021) Machine learning and natural language processing in mental health: systematic review. *J. Med. Internet. Res.* **23**, e15708 https://doi.org/10.2196/15708
68 Bracher-Smith, M., Crawford, K. and Escott-Price, V. (2020) Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Mol. Psychiatry* **26**, 70–79 https://doi.org/10.1038/s41380-020-0825-2
69 Field, M., Hardcastle, N., Jameson, M., Aherne, N. and Holloway, L. (2021) Machine learning applications in radiation oncology. *Phys. Imaging Radiat. Oncol.* **19**, 13–24 https://doi.org/10.1016/J.PHRO.2021.05.007
70 el Naqa, I. (2021) Prospective clinical deployment of machine learning in radiation oncology. *Nat. Rev. Clin. Oncol.* **18**, 605–606 https://doi.org/10.1038/s41571-021-00541-w
71 Rajkumar, D. Applications of Machine Learning in Radiology-A review. *Journal For Innovative Development in Pharmaceutical and Technical Science (JIDPTS*). Published online 2020:8. Accessed December 1, 2021. www.jidps.com
72 Wichmann, J.L., Willemink, M.J. and de Cecco, C.N. (2020) Artificial intelligence and machine learning in radiology: current state and considerations for routine clinical implementation. *Invest. Radiol.* **55**, 619–627 https://doi.org/10.1097/RLI.0000000000000673
73 Elfanagely, O., Toyoda, Y., Othman, S., Bormann, B., Mellia, J.A., Broach, R.B. et al. (2021) Machine learning and surgical outcomes prediction: a systematic review. *J. Surg. Res.* **264**, 346–361 https://doi.org/10.1016/J.JSS.2021.02.045
74 Henn, J., Buness, A., Schmid, M., Kalff, J.C. and Matthaei, H. (2021) Machine learning to guide clinical decision-making in abdominal surgery—a systematic literature review. *Langenbeck's Arch. Surg.* **1**, 1–11 https://doi.org/10.1007/S00423-021-02348-W
75 Kilic, A. (2020) Artificial intelligence and machine learning in cardiovascular health care. *Ann. Thorac. Surg.* **109**, 1323–1329 https://doi.org/10.1016/J.ATHORACSUR.2019.09.042
76 Dias, R.D., Shah, J.A. and Zenati, M.A. (2020) Artificial intelligence in cardiothoracic surgery. *Miner. Cardioangiol.* **68**, 532 https://doi.org/10.23736/S0026-4725.20.05235-4
77 Suarez-Ibarrola, R., Hein, S., Reis, G., Gratzke, C. and Miernik, A. (2019) Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. *World J. Urol.* **38**, 2329–2347 https://doi.org/10.1007/S00345-019-03000-5
78 Salem, H., Soria, D., Lund, J.N. and Awwad, A. (2021) A systematic review of the applications of expert systems (ES) and machine learning (ML) in clinical urology. *BMC Med. Inform. Decis. Mak.* **21**, 1–36 https://doi.org/10.1186/S12911-021-01585-9
79 Zarkowsky, D.S. and Stonko, D.P. Artificial intelligence's role in vascular surgery decision-making. *Seminars in Vascular Surgery*. Published online October 27, 2021. https://doi.org/10.1053/J.SEMVASCSURG.2021.10.005
80 Boyd, C., Brown, G., Kleinig, T., Dawson, J., McDonnell, M.D., Jenkinson, M. et al. (2021) Machine learning quantitation of cardiovascular and cerebrovascular disease: a systematic review of clinical applications. *Diagnostics* **11**, 551 https://doi.org/10.3390/DIAGNOSTICS11030551
81 Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. and Aljaaf, A.J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Published online 2020:3–21. https://doi.org/10.1007/978-3-030-22475-2_1
82 Lo Vercio, L., Amador, K., Bannister, J.J., Crites, S., Gutierrez, A., MacDonald, M.E. et al. (2020) Supervised machine learning tools: a tutorial for clinicians. *J. Neural Eng.* **17**, 062001 https://doi.org/10.1088/1741-2552/ABBFF2
83 Matloff N. Statistical regression and classification: From linear models to machine learning. *Statistical Regression and Classification: From Linear Models to Machine Learning*. Published online January 1, 2017:1–493. https://doi.org/10.1201/9781315119588
84 Nusinovici, S., Tham, Y.C., Chak Yan, M.Y., Wei Ting, D.S., Li, J., Sabanayagam, C. et al. (2020) Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **122**, 56–69 https://doi.org/10.1016/J.JCLINEPI.2020.03.002
85 Pregibon, D. (1981) Logistic Regression Diagnostics. **9**, 705–724 https://doi.org/101214/aos/1176345513
86 Searle, S.R. and Gruber, M.H.J. (2016) *Linear Models*, Wiley, New Jersey, United States
87 Mehryar, M., Rostamizadeh, A. and Talwalkar, A. (2018) *Foundations of Machine Learning*, MIT Press, Cambridge, Massachusetts, USA Accessed December 1, 2021. https://www.google.com/books/edition/Foundations_of_Machine_Learning_second_e/dWB9DwAAQBAJ?hl=en&gbpv=0
88 Ghosal, S., Sengupta, S., Majumder, M. and Sinha, B. (2020) Linear regression analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). *Diabetes Metab. Syndr.* **14**, 311–315 https://doi.org/10.1016/J.DSX.2020.03.017
89 Lip, G.Y.H., Genaidy, A., Tran, G., Marroquin, P., Estes, C. and Sloop, S. (2021) Improving stroke risk prediction in the general population: a comparative assessment of common clinical rules, a new multimorbid index, and machine-learning-based algorithms. *Thromb. Haemost.* **19**, 603–883 https://doi.org/10.1055/A-1467-2993
90 Sha, K., Lu, Y., Zhang, P., Pei, R., Shi, X., Fan, Z. et al. (2020) Identifying a novel 5-gene signature predicting clinical outcomes in acute myeloid leukemia. *Clin. Transl. Oncol.* **23**, 648–656 https://doi.org/10.1007/S12094-020-02460-1
91 Lin, D.Y. and Wei, L.J. (1989) The robust inference for the cox proportional hazards model. *J. Am. Stat. Assoc.* **84**, 1074–1078 https://doi.org/10.1080/01621459.1989.10478874
92 Zhao, X. and Zhou, X. (2006) Proportional hazards models for survival data with long-term survivors. *Stat. Probab. Lett.* **76**, 1685–1693 https://doi.org/10.1016/J.SPL.2006.04.018
93 Wolfson, J., Bandyopadhyay, S., Elidrisi, M., Vazquez-Benitez, G., Vock, D.M., Musgrove, D. et al. (2015) A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. *Stat. Med.* **34**, 2941–2957 https://doi.org/10.1002/SIM.6526
94 Murphy KP. Naive Bayes classifiers

95 Yang, L., Fu, B., Li, Y., Liu, Y., Huang, W., Feng, S. et al. (2020) Prediction model of the response to neoadjuvant chemotherapy in breast cancers by a Naive Bayes algorithm. *Comput. Methods Prog. Biomed.* **192**, 105458 https://doi.org/10.1016/J.CMPB.2020.105458
96 Wei, W., Visweswaran, S. and Cooper, G.F. (2011) The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *J. Am. Med. Inform. Assoc.* **18**, 370–375 https://doi.org/10.1136/AMIAJNL-2011-000101
97 Lewis R. An introduction to classification and regression tree (CART) analysis. 2000 Annual Meeting of the Society for Academic Emergency Medicine. Published online 2000. Accessed November 30, 2021. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf
98 Fernandez-Lozano, C., Hervella, P., Mato-Abad, V., Rodríguez-Yáñez, M., Suárez-Garaboa, S., López-Dequidt, I. et al. (2021) Random forest-based prediction of stroke outcome. *Sci. Rep.* **11**, 1–12 https://doi.org/10.1038/s41598-021-89434-7
99 Li, Y., Pan, J., Zhou, N., Fu, D., Lian, G., Yi, J. et al. (2021) A random forest model predicts responses to infliximab in Crohn's disease based on clinical and serological parameters. *Scand. J. Gastroenterol.* **56**, 1030–1039 https://doi.org/10.1080/00365521.2021.1939411
100 Hanko, M., Grendár, M., Snopko, P., Opšenák, R., Šutovský, J., Benčo, M. et al. (2021) Random forest–based prediction of outcome and mortality in patients with traumatic brain injury undergoing primary decompressive craniectomy. *World Neurosurg.* **148**, e450–e458 https://doi.org/10.1016/J.WNEU.2021.01.002
101 Couronné, R., Probst, P. and Boulesteix, A.L. (2018) Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* **19**, 1–14 https://doi.org/10.1186/S12859-018-2264-5
102 Breiman, L. (1996) Bagging predictors. *Mach. Learn.* **24**, 123–140 https://doi.org/10.1007/BF00058655
103 Bühlmann, P. and Yu, B. (2002) Analyzing bagging. *Ann. Statist.* **30**, 927–961 https://doi.org/10.1214/AOS/1031689014
104 Chen, T. and Guestrin, C. (2016) XGBoost: a scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 13–17-August-2016:785–794. https://doi.org/10.1145/2939672.2939785
105 Jiang, Y.Q., Cao, S.E., Cao, S., Chen, J.N., Wang, G.Y., Shi, W.Q. et al. (2021) Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning. *J. Cancer Res. Clin. Oncol.* **147**, 821–833 https://doi.org/10.1007/S00432-020-03366-9/FIGURES/5
106 Schapire RE. Explaining AdaBoost. *Empirical Inference: Festschrift in Honor of Vladimir N Vapnik*. Published online January 1, 2013:37–52. https://doi.org/10.1007/978-3-642-41136-6_5
107 Huang, Q., Chen, Y., Liu, L., Tao, D. and Li, X. (2020) On combining biclustering mining and adaboost for breast tumor classification. *IEEE Trans. Knowl. Data Eng.* **32**, 728–738 https://doi.org/10.1109/TKDE.2019.2891622
108 Hyoungju, C.J. and Park, J.H. (2021) Prediction of critical care outcome for adult patients presenting to emergency department using initial triage information: an XGBoost algorithm analysis. *JMIR Med. Inform.* **9**, e30770 https://doi.org/10.2196/30770
109 Varghese, B.A., Shin, H., Desai, B., Hwang, D.H., Aron, M., Siddiqui, I. et al. (2021) Predicting clinical outcomes in COVID-19 using radiomics on chest radiographs. *Br. J. Radiol.* **94**, 20210221 https://doi.org/10.1259/BJR.20210221
110 Ishwaran, H. (2008) UKEBML. random survival forests. *Ann. Appl. Stat.* **2**, 841–860 https://doi.org/10.1214/08-aoas169
111 Akai, H., Yasaka, K., Kunimatsu, A., Nojima, M., Kokudo, T., Kokudo, N. et al. (2018) Predicting prognosis of resected hepatocellular carcinoma by radiomics analysis with random survival forest. *Diagn. Interv. Imaging* **99**, 643–651 https://doi.org/10.1016/J.DIII.2018.05.008
112 Deng, H. (2018) Interpreting tree ensembles with in trees. *Int. J. Data Sci. Anal.* **7**, 277–287 https://doi.org/10.1007/S41060-018-0144-8
113 Breiman, L. (2001) Random forests. *Mach. Learn.* **45**, 5–32 https://doi.org/10.1023/a:1010933404324
114 Parry, R.M., Jones, W., Stokes, T.H., Phan, J.H., Moffitt, R.A., Fang, H. et al. (2010) k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics J.* **10**, 292–309 https://doi.org/10.1038/tpj.2010.56
115 Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M. et al. (2008) MicroRNAs accurately identify cancer tissue origin. *Nat. Biotechnol.* **26**, 462–469 https://doi.org/10.1038/nbt1392
116 Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A. et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. U.S.A.* **102**, 13550–13555 https://doi.org/10.1073/PNAS.0506230102
117 Pisner, D.A and Schnyer, D.M. Support vector machine. *Machine Learning: Methods and Applications to Brain Disorders*. Published online January 1, 2020:101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7
118 Cui, S., Wang, D., Wang, Y., Yu, P.W. and Jin, Y. (2018) An improved support vector machine-based diabetic readmission prediction. *Comput. Methods Prog. Biomed.* **166**, 123–135 https://doi.org/10.1016/J.CMPB.2018.10.012
119 José Del Coz, J., Díez, J. and Bahamonde, A. (2009) Learning nondeterministic classifiers. *J. Mach. Learn. Res.* **10**, 2273–2293 https://www.jmlr.org/papers/volume10/delcoz09a/delcoz09a.pdf
120 Žohar, P., Kovačič, M., Brezocnik, M. and Podbregar, M. (2005) Prediction of maintenance of sinus rhythm after electrical cardioversion of atrial fibrillation by non-deterministic modelling. *Europace* **7**, 500–507 https://doi.org/10.1016/J.EUPC.2005.04.007
121 Floyd, R.W. (1967) Nondeterministic algorithms. *J. ACM (JACM)* **14**, 636–644 https://doi.org/10.1145/321420.321422
122 Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S. (2016) Deep learning for visual understanding: a review. *Neurocomputing* **187**, 27–48 https://doi.org/10.1016/J.NEUCOM.2015.09.116
123 Wang, S., Tang, C., Sun, J. and Zhang, Y. (2019) Cerebral micro-bleeding detection based on densely connected neural network. *Front. Neurosci.* **13**, 422 https://doi.org/10.3389/FNINS.2019.00422
124 Sharma, P. and Singh, A. Era of deep neural networks: A review. *8th International Conference on Computing, Communications and Networking Technologies*, ICCCNT 2017. Published online December 13, 2017. https://doi.org/10.1109/ICCCNT.2017.8203938
125 Voulodimos, A., Doulamis, N., Doulamis, A. and Protopapadakis, E. (2018) Deep learning for computer vision: a brief review. *Comput. Intell. Neurosci.* **2018**, 7068349 https://doi.org/10.1155/2018/7068349
126 Janocha, K. and Czarnecki, W.M. (2016) On loss functions for deep neural networks in classification. *Schedae Informaticae* **25**, 49–59 https://doi.org/10.4467/20838476SI.16.004.6185
127 Agarap AF. Deep Learning using Rectified Linear Units (ReLU). <arXiv. Published online March 2018
128 Narayan, S. (1997) The generalized sigmoid activation function: competitive supervised learning. *Inform. Sci.* **99**, 69–82 https://doi.org/10.1016/S0020-0255(96)00200-9

129 Abbasi, S.F., Ahmad, J., Tahir, A., Awais, M., Chen, C., Irfan, M. et al. (2020) EEG-based neonatal sleep-wake classification using multilayer perceptron neural network. *IEEE Access* **8**, 183025–183034 https://doi.org/10.1109/ACCESS.2020.3028182
130 Waldrop, M.M. (2019) News feature: what are the limits of deep learning? *Proc. Natl Acad. Sci. U.S.A.* **116**, 1074–1077 https://doi.org/10.1073/PNAS.1821594116
131 Diamant, A., Chatterjee, A., Vallières, M., Shenouda, G. and Seuntjens, J. (2019) Deep learning in head & neck cancer outcome prediction. *Sci. Rep.* **9**, 1–10 https://doi.org/10.1038/s41598-019-39206-1
132 Tjepkema-Cloostermans, M.C., da Silva Lourenço, C., Ruijter, B.J., Tromp, S.C., Drost, G., Kornips, F.H.M. et al. (2019) Outcome prediction in postanoxic coma with deep learning. *Crit. Care Med.* **47**, 1424–1432 https://doi.org/10.1097/CCM.0000000000003854
133 Che, Z., Purushotham, S., Khemani, R. and Liu, Y. (2016) Interpretable Deep Models for ICU Outcome Prediction. *AMIA Annual Symposium Proceedings*; 2016:371. Accessed September 30, 2021. /pmc/articles/PMC5333206/
134 Morid, M.A., Borjali, A. and del Fiol, G. (2021) A scoping review of transfer learning research on medical image analysis using imageNet. *Comput. Biol. Med.* **128**, 104115 https://doi.org/10.1016/J.COMPBIOMED.2020.104115
135 Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Accessed September 25, 2021. http://code.google.com/p/cuda-convnet/
136 López-García, G., Jerez, J.M., Franco, L. and Veredas, F.J. (2020) Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PLoS ONE* **15**, e0230536 https://doi.org/10.1371/journal.pone.0230536
137 He, L., Li, H., Wang, J., Chen, M., Gozdas, E., Dillman, J.R. et al. (2020) A multi-task, multi-stage deep transfer learning model for early prediction of neurodevelopment in very preterm infants. *Sci. Rep.* **10**, 1–13 https://doi.org/10.1038/s41598-020-71914-x
138 Han, W., Qin, L., Bay, C., Chen, X., Yu, K.H., Miskin, N. et al. (2020) Deep transfer learning and radiomics feature prediction of survival of patients with high-grade gliomas. *Am. J. Neuroradiol.* **41**, 40–48 https://doi.org/10.3174/ajnr.A6365
139 Bommasani R. On the opportunities and risks of foundation models. arxiv.org. Published online 2021. Accessed December 1, 2021. https://arxiv.org/abs/2108.07258
140 Rasmy, L., Xiang, Y., Xie, Z., Tao, C. and Zhi, D. (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 1–13 https://doi.org/10.1038/s41746-021-00455-y
141 Wager, S., Wang, S. and Liang, P. (2013) Dropout training as adaptive regularization. In: *Advances in Neural Information Processing Systems*. Accessed September 30, 2021. https://papers.nips.cc/paper/4882-dropout-training-as-adaptive-regularization
142 Wei, C., Kakade, S. and Ma, T. (2020) The implicit and explicit regularization effects of dropout. In: *37th International Conference on Machine Learning*, ICML 2020. Vol PartF16814: 10112–10123. Accessed September 30, 2021. http://proceedings.mlr.press/v119/wei20d.html
143 Zhu, X., Yao, J. and Huang, J. (2016) Deep convolutional neural network for survival analysis with pathological images. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine*, BIBM Published online January 17, 2017:544–547. https://doi.org/10.1109/BIBM.2016.7822579
144 de Laurentiis, M. and Ravdin, P.M. (1994) A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Lett.* **77**, 127–138 https://doi.org/10.1016/0304-3835(94)90095-7
145 Ryu, S.M., Seo, S.W. and Lee, S.H. (2020) Novel prognostication of patients with spinal and pelvic chondrosarcoma using deep survival neural networks. *BMC Med. Inform. Decis. Mak.* **20**, 1–10 https://doi.org/10.1186/S12911-019-1008-4
146 Buteneers, P., Verstraeten, D., van Mierlo, P., Wyckhuys, T., Stroobandt, D., Raedt, R. et al. (2011) Automatic detection of epileptic seizures on the intra-cranial electroencephalogram of rats using reservoir computing. *Artif. Intell. Med.* **53**, 215–223 https://doi.org/10.1016/J.ARTMED.2011.08.006
147 Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I. et al. (2019) Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin. Cancer Res.* **25**, 3266–3275 https://doi.org/10.1158/1078-0432.CCR-18-2495
148 Reddy, B.K. and Delen, D. (2018) Predicting hospital readmission for lupus patients: an RNN-LSTM-based deep-learning methodology. *Comput. Biol. Med.* **101**, 199–209 https://doi.org/10.1016/J.COMPBIOMED.2018.08.029
149 Wang, J.M., Liu, W., Chen, X., McRae, M.P., McDevitt, J.T. and Fenyö, D. Predictive modeling of morbidity and mortality in COVID-19 hospitalized patients and its clinical implications. medRxiv. Published online March 29, 2021. https://doi.org/10.1101/2020.12.02.20235879
150 Baldwin, D.R., Gustafson, J., Pickup, L., Arteta, C., Novotny, P., Declerck, J. et al. (2020) External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax* **75**, 306–312 https://doi.org/10.1136/THORAXJNL-2019-214104
151 Sahoo, A.K., Pradhan, C. and Das, H. (2020) Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. In: *Studies in Computational Intelligence*. Vol SCI 871. Springer, Cham; 201–212. https://doi.org/10.1007/978-3-030-33820-6_8
152 García, S., Luengo, J. and Herrera, F. (2015) Data Preprocessing in Data Mining. Accessed September 30, 2021. https://link.springer.com/content/pdf/10.1007/978-3-319-10247-4.pdf
153 Kotsiantis S, Kanellopoulos D, of PPI journal, 2006 undefined. Data preprocessing for supervised leaning. *Citeseer*. Published online 2006. Accessed September 30, 2021. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.8413&rep=rep1&type=pdf
154 Therrien, R. and Doyle, S. (2018) Role of training data variability on classifier performance and generalizability. In: Https://Doi.Org/10.1117/12.2293919. Vol 10581. SPIE; 5. https://doi.org/10.1117/12.2293919
155 Kocher, R., Emanuel, E.J. and DeParle, N.A.M. (2010) The affordable care act and the future of clinical medicine: The opportunities and challenges. *Ann. Intern. Med.* **153**, 536–539 https://doi.org/10.7326/0003-4819-153-8-201010190-00274
156 Mathis, C. (2017) Data lakes. *Datenbank-Spektrum* **17**, 289–293 https://doi.org/10.1007/s13222-017-0272-7
157 Matthias, J., Lenzerini, M., Vassiliou, Y., Vassiliadis, P. and Hoffner, V. (2003) Fundamentals of data warehouses. *SIGMOD Record* **32**, 55–56 https://doi.org/10.1145/776985.776995
158 Bender D, Sartipi K. HL7 FHIR: An agile and RESTful approach to healthcare information exchange. In: *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems*; 2013:326–331. https://doi.org/10.1109/CBMS.2013.6627810
159 Liu, R., Simon, E., Amann, B. and Gançarski, S. (2020) Discovering and merging related analytic datasets. *Inform. Syst.* **91**, 101495 https://doi.org/10.1016/j.is.2020.101495

160 Huber L, Honeder T, dHealth WH. 2020 undefined. FHIR Analytics-Pragmatic Review of Recent Studies. books.google.com. 2020;271:110–112. https://doi.org/10.3233/SHTI200083
161 Sun H, Depraetere K, Meesseman L, … JDRJ of B, 2021 undefined. A scalable approach for developing clinical risk prediction applications in different hospitals. *Elsevier*. Accessed September 30, 2021. https://www.sciencedirect.com/science/article/pii/S153204642100112X
162 Franz, L., Shrestha, Y.R and Paudel, B. A Deep Learning Pipeline for Patient Diagnosis Prediction Using Electronic Health Records. 2020;10. Accessed September 30, 2021. http://arxiv.org/abs/2006.16926
163 Tarumi, S., Takeuchi, W., Chalkidis, G., Rodriguez-Loya, S., Kuwata, J., Flynn, M. et al. (2021) Leveraging artificial intelligence to improve chronic disease care: methods and application to pharmacotherapy decision support for type-2 diabetes mellitus. *Methods Inf. Med.* **60**, E32–E43 https://doi.org/10.1055/s-0041-1728757
164 Kawaler, E., Cobian, A., Peissig, P., Cross, D., Yale, S. and Craven, M. (2012) Learning to Predict Post-Hospitalization VTE Risk from EHR Data. *AMIA Annual Symposium Proceedings*. 2012; 436. Accessed September 30, 2021. /pmc/articles/PMC3540493/
165 Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V. et al. (2021) Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 https://doi.org/10.1038/s41586-021-03583-3
166 Cai, J., Luo, J., Wang, S. and Yang, S. (2018) Feature selection in machine learning: a new perspective. *Neurocomputing* **300**, 70–79 https://doi.org/10.1016/j.neucom.2017.11.077
167 Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H. and Ferrante, E. (2020) Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl Acad. Sci.* **117**, 12592–12594 https://doi.org/10.1073/PNAS.1919012117
168 Sondhi, P. (2010) Feature construction methods: a survey. sifaka cs uiuc edu. 69:70–71
169 Piramuthu, S., Ragavan, H. and Shaw, M.J. (1998) Using feature construction to improve the performance of neural networks. *Manag. Sci.* **44**, 416–430 https://doi.org/10.1287/MNSC.44.3.416
170 Liu H, Appl HMIIST, 1998 undefined. Feature transformation and subset selection. *Citeseer*. Accessed September 30, 2021. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.102&rep=rep1&type=pdf
171 Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T. and Fränti, P. (2005) Improving K-means by outlier removal. *Lect. Notes Comput. Sci.* **3540**, 978–987 https://doi.org/10.1007/11499145_99
172 Choi, Y.S. (2009) Least squares one-class support vector machine. *Pattern Recognit. Lett.* **30**, 1236–1240 https://doi.org/10.1016/J.PATREC.2009.05.007
173 Hardin, J. and Rocke, D.M. (2004) Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Comput. Stat. Data Anal.* **44**, 625–638 https://doi.org/10.1016/S0167-9473(02)00280-3
174 Cheng, Z., Zou, C. and Dong, J. (2019) Outlier detection using isolation forest and local outlier. In: *Proceedings of the 2019 Research in Adaptive and Convergent Systems, RACS 2019*. Association for Computing Machinery, Inc; 161–168. https://doi.org/10.1145/3338840.3355641
175 Juszczak, P., Tax, D. asci RDProc, 2002 undefined. Feature scaling in support vector data description. *Citeseer*. Accessed September 30, 2021. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.6071&rep=rep1&type=pdf
176 Figueroa, R.L., Zeng-Treitler, Q., Kandula, S. and Ngo, L.H. (2012) Predicting sample size required for classification performance. *BMC Med. Inform. Decis. Mak.* **12**, 1–10 https://doi.org/10.1186/1472-6947-12-8
177 Batuwita, R. and Palade, V. (2010) Efficient resampling methods for training support vector machines with imbalanced datasets. In: *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2010.5596787
178 Nikolenko, S.I. (2021) Synthetic data for deep learning. In: *Springer Optimization and Its Applications*. Vol 174. Springer Optimization and Its Applications. Springer International Publishing; 1–54. https://doi.org/10.1007/978-3-030-75178-4_1
179 Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 https://doi.org/10.1613/JAIR.953
180 Weng, C.G. and Poon, J. A New Evaluation Measure for Imbalanced Datasets
181 Lin, H.I. and Nguyen, M.C. (2020) Boosting minority class prediction on imbalanced point cloud data. *Appl. Sci. (Switzerland)* **10**, 973 https://doi.org/10.3390/app10030973
182 Jiang, Y., Chen, S., McGuire, D., Chen, F., Liu, M., Iacono, W.G. et al. (2018) Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. *PLoS Genet.* **14**, e1007452 https://doi.org/10.1371/JOURNAL.PGEN.1007452
183 Kent, J.T. (1983) Information gain and a general measure of correlation. *Biometrika* **70**, 163–173 https://doi.org/10.1093/biomet/70.1.163
184 Raileanu, L.E. and Stoffel, K. (2004) Theoretical comparison between the gini index and information gain criteria. *Ann. Math. Artif. Intell.* **41**, 77–93 https://doi.org/10.1023/B:AMAI.0000018580.96245.C6
185 Pratihar, D.K. (2011) Non-Linear Dimensionality Reduction Techniques. In: Encyclopedia of Data Warehousing and Mining, *Second Edition*. Springer; 308. https://doi.org/10.4018/9781605660103.ch219
186 Wold, S. and Esbensen, K. Laboratory PGC and intelligent, 1987 U. Principal component analysis. *Elsevier*. Accessed September 30, 2021. https://www.sciencedirect.com/science/article/pii/0169743987800849
187 Zhu, W., Webb, Z.T., Mao, K. and Romagnoli, J. (2019) A deep learning approach for process data visualization using t-distributed stochastic neighbor embedding. *Ind. Eng. Chem. Res.* **58**, 9564–9575 https://doi.org/10.1021/ACS.IECR.9B00975
188 McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018) UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 https://doi.org/10.21105/joss.00861
189 Seger, C. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. DEGREE PROJECT TECHNOLOGY. Published online 2018. Accessed September 30, 2021. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426
190 Guo, C. and Berkhahn, F. Entity Embeddings of Categorical Variables. Published online April 22, 2016. Accessed September 30, 2021. https://arxiv.org/abs/1604.06737v1
191 Hill, F., Cho, K.H., Jean, S., Devin, C. and Bengio, Y. (2015) Embedding word similarity with neural machine translation. In: *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR. Accessed September 30, 2021. https://arxiv.org/abs/1412.6448v4

192 Jian, S., Cao, L., Pang, G., Lu, K. and Gao, H. (2017) Embedding-based representation of categorical data by hierarchical value coupling learning. *IJCAI International Joint Conference on Artificial Intelligence*;0:1851-1857. Accessed September 30, 2021. https://opus.lib.uts.edu.au/handle/10453/126349
193 Larsen, J. and Goutte, C. (1999) On optimal data split for generalization estimation and model selection. In: *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*. IEEE; 225–234. https://doi.org/10.1109/nnsp.1999.788141
194 Schaffer, C. (1993) Selecting a classification method by cross-validation. *Mach. Learn.* **13**:135–143. https://doi.org/10.1007/BF00993106
195 Pathak, P.K. (1962) On simple random sampling with replacement. *Sankhyā: The Indian Journal of Statistics*; Series A:287–302. Accessed September 30, 2021. https://www.jstor.org/stable/25049220
196 Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 https://doi.org/10.1016/j.patrec.2005.10.010
197 Muschelli, J. (2019) ROC and AUC with a binary predictor: a potentially misleading metric. *J. Classif.* **37**, 696–708 https://doi.org/10.1007/S00357-019-09345-1
198 Boyd, K., Eng, K.H. and Page, C.D. (2013) Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 8190 LNAI(PART 3):451–466. https://doi.org/10.1007/978-3-642-40994-3_29
199 Yacouby, R. and Axman, D. (2020) Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. In: Association for Computational Linguistics (ACL); 79–91. https://doi.org/10.18653/v1/2020.eval4nlp-1.9
200 Cerulli, G. Machine Learning using Stata/Python. Published online March 3, 2021. Accessed September 30, 2021. https://arxiv.org/abs/2103.03122v1
201 Kolosova, T. and Berestizhevsky, S. (2016) *Supervised Machine Learning: Optimization Framework and Applications with SAS and R*. Vol 4. Accessed September 30, 2021. https://books.google.com/books?hl=en&lr=&id=3sb2DwAAQBAJ&oi=fnd&pg=PP1&dq=SAS+machine+learning&ots=_gWMyWHFV-&sig=LeCrYXb2k5RGdA6xKFGXFx1jL9A
202 Paluszek, M. and Thomas, S. (2020) MATLAB Machine Learning Toolboxes. In: *Practical MATLAB Deep Learning* 25–41. https://doi.org/10.1007/978-1-4842-5124-9_2
203 Lantz, B. (2019) *Machine Learning with R: Expert Techniques for Predictive Modeling*. Accessed September 30, 2021. https://books.google.com/books?hl=en&lr=&id=iNuSDwAAQBAJ&oi=fnd&pg=PP1&dq=r+machine+learning&ots=O84Sra7zP-&sig=KzZNoNgNl4kpAjOh8xBRd3zYxN4
204 Oliphant, T.E. (2007) Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20 https://doi.org/10.1109/MCSE.2007.58
205 Demšar, J., Zupan, B., Leban, G. and Curk, T. (2004) Orange: From experimental machine learning to interactive data mining. *Lect. Notes Comput. Sci.* **3202**, 537–539 https://doi.org/10.1007/978-3-540-30116-5_58
206 Berthold, M., Cebron, N., Dill, F., Gabriel, R., Kötter, T., Meini, T. et al. 2009. KNIME-the Konstanz information miner: version 2.0 and beyond. *dl.acm.org*. Published online 2006:58–61. https://doi.org/10.1145/1656274.1656280
207 Gerke, S., Babic, B., Evgeniou, T. and Cohen, I.G. (2020) The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit. Med.* **3**, 1–4 https://doi.org/10.1038/s41746-020-0262-2

Supplementary Table 1. Reviews per medical specialty of the current use of machine learning methods in clinical outcome prediction.

Medical Specialty	Review Articles
Anesthesiology[32–34]	Chae et al. 2020, Alexander et al. 2020, Hashimoto et al. 2020
Dermatology[35–37]	Du et al. 2020, Thomsen et al. 2019, Chan et al. 2020
Emergency Medicine[38,39]	Stewart et al. 2021, Tang et al. 2021
Family Medicine[40,41]	Kueper et al. 2020, Ben-Israel et al. 2020
Internal Medicine[41–43]	Ben-Israel et al. 2020, Pandit et al. 2020, Stafford et al. 2020
Interventional Radiology[44,45]	Mazaheri et al 2021, Desai et al. 2021
Medical Genetics[46]	Rauschert et al 2020
Neurological Surgery[47]	Buchlak et al. 2019
Neurology[48-50]	Myszczynska et al. 2020, Sirsat et al. 2020, Yuan et al. 2021
Obstetrics and Gynecology[51,52]	Iftikhar et al. 2020, Sone et al. 2021
Ophthalmology[53-55]	Sengupta et al. 2020, Sarhan et al. 2020, Armstrong et al. 2020
Orthopaedic Surgery[56]	Orink et al. 2021
Otorhinolaryngology[57,58]	Standiford et al. 2021, Crowson et al. 2020
Pathology[59–61]	Thakur et al. 2020, Sultan et al. 2020, McAlpine et al. 2021
Pediatrics[62]	Hoodbhoy et al. 2021
Physical Medicine and Rehabilitation[63,64]	Khera et al. 2020, Amorim et al. 2021
Plastic and Reconstructive Surgery[65,66]	Mantelakis et al. 2021, Huang et al. 2021
Psychiatry[67,68]	Le Glaz et al. 2021, Bracher-Smith et al. 2020
Radiation Oncology[69,70]	Field et al. 2021, El Naqa et al. 2021
Radiology[71,72]	Rajkumar et al. 2020, Wichmann et al. 2020
General Surgery[73,74]	Elfanagely et al. 2021, Henn et al. 2021
Cardiothoracic Surgery[75,76]	Kilic et al. 2020, Dias et al. 2020
Urology[77,78]	Suarez-Ibarrola et al. 2019, Salem et al. 20201
Vascular Surgery[79,80]	Zarkowsky et al. 2021, Boyd et al. 2021