Foundations of Machine Learning-Based Clinical Prediction Modeling: Part I—Introduction and General Principles

2

Julius M. Kernbach and Victor E. Staartjes

2.1 Introduction

Although there are many applications of machine learning (ML) in clinical neuroscience, including but not limited to applications in neuroimaging and natural language processing, classical predictive analytics still form the majority of the body of evidence that has been published on the topic.

When reviewing or working on research involving ML-based predictive analytics—which is becoming increasingly common—it is important to do so with a strong methodological basis. Especially considering the "democratization" of ML methods through libraries and the increasing computing power, as well as the exponentially increasing influx of ML publications in the clinical neurosciences, methodological rigor has become a major issue. This chapter and in fact the entire five-part series (cite Chaps. 3–6) is intended to convey that basic conceptual and programming knowledge to tackle ML tasks with some basic prerequisite R knowledge, with a particular focus on predictive analytics.

At this point, it is important to stress that the concepts and methods presented herein are intended as an entry-level guide to ML for clinical outcome prediction, presenting one of many valid approaches to clinical prediction modeling, and thus does not encompass all the details and intricacies of the field. Further reading is recommended, including but not limited to Max Kuhn's "Applied Predictive Modeling" [1] and Ewout W. Steyerberg's "Clinical Prediction Models" [2].

This first part focuses on defining the terms ML and AI in the context of predictive analytics, and clearly describing their applications in clinical medicine. In addition, some of the basic concepts of machine intelligence are discussed and explained. Part II goes into detail about common problems when developing clinical prediction models: What overfitting is and how to avoid it to arrive at generalizable models, how to select which input features are to be included in the final model (feature selection) or how to simplify highly dimensional data (feature reduction). We also discuss how data splits and resampling methods like cross-validation and the bootstrap can be applied to validate models before clinical use. Part III touches on several topics including how to prepare your data correctly (standardization, one-hot encoding) and evaluate models in terms of discrimination and calibration, and points out some recalibration methods. Some other points of significance and caveats that the reader may encounter while developing a clinical prediction model are discussed: sample size, class imbalance, missing data and how to impute it, extrapolation, as well as how to choose a cutoff for binary classification. Parts IV and V present a practical approach to classification and regression problems, respectively. They contain detailed instructions along with a downloadable code for the R statistical programming language, as well as a simulated database of Glioblastoma patients that allows the reader to code in parallel to the explanations. This section is intended as a scaffold upon which readers can build their own clinical prediction models, and that can easily be modified. Furthermore, we will not in detail explain the workings of specific ML algorithms such as generalized linear models, support vector machines, neural networks, or stochastic gradient boosting. While it is certainly important to have a basic understanding of the specific

J. M. Kernbach and V. E. Staartjes have contributed equally to this series, and share first authorship.

J. M. Kernbach
Neurosurgical Artificial Intelligence Laboratory Aachen (NAILA), Department of Neurosurgery, RWTH Aachen University Hospital, Aachen, Germany

V. E. Staartjes (✉)
Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Zurich, Switzerland
e-mail: victoregon.staartjes@usz.ch; https://micnlab.com/

V. E. Staartjes et al. (eds.), *Machine Learning in Clinical Neuroscience*, Acta Neurochirurgica Supplement 134,
https://doi.org/10.1007/978-3-030-85292-4_2

algorithms one applies, these details can be looked up online [3] and detailed explanations of these algorithms would go beyond the scope of this guide. The goal is instead to convey the basic concepts of ML-based predictive modeling, and how to practically implement these.

2.2 Machine Learning: Definitions

As a field of study, ML in medicine is positioned between statistical learning and advanced computer science, and typically evolves around *learning problems*, which can be conceptually defined as optimizing a performance measure on a given task by learning through training experience on prior data. A ML algorithm inductively learns to automatically extract patterns from data to generate insights [4, 5] without being explicitly programmed. This makes ML an attractive option to predict even complex phenomena without prespecifying an a priori theoretical model. ML can be used to leverage the full granularity of the data richness enclosed in the *Big Data* trend. Both the complexity and dimensionality of modern medical data sets are constantly increasing and nowadays comprise many variables per observation, much so that we speak of "wide data" with generally more variables (in ML lingo called *features*) than observations (samples) [6, 7]. This has given rise to the so-called *omics* sciences including radiomics and genomics [8–10]. The sheer complexity and volume of data ranging from hundreds to thousands of variables at times exceeds human comprehension, but combined with increased computational power enables the full potential of ML [3, 11].

With the exponential demand of AI and ML in modern medicine, a lot of confusion was introduced regarding the separation of these two terms. AI and ML are frequently used interchangeably. We define ML as subset of AI—to quote Tom Mitchell—ML "is the study of computer algorithms that allow computer programs to automatically improve through experience" [12], involving the concept of "learning" discussed earlier. In contrast, AI is philosophically much vaster, and can be defined as an ambition to enable computer programs to behave in a human-like nature. That is, showing a certain human-like intelligence. In ML, we learn and optimize an algorithm from data for maximum performance on a certain learning task. In AI, we try to emulate natural intelligence, to not only learn but also apply the gained knowledge to make elaborate decisions and solve complex problems. In a way, ML can thus be considered a technique towards realizing (narrow) AI. Ethical considerations on the "AI doctor" are far-reaching [13, 14], while the concept of a clinician aided by ML-based tools is well accepted.

The most widely used ML methods are either supervised or unsupervised learning methods, with the exceptions of semi-supervised methods and reinforcement learning [6, 15]. In supervised learning, a set of input variables are used as training set, e.g. different meaningful variables such as age, gender, tumor grading, or functional neurological status to predict a known target variable ("label"), e.g. overall survival. The ML method can then learn the pattern linking input features to target variable, and based on that enable the prediction of new data points—hence, *generalize* patterns beyond the present data. We can train a ML model for survival prediction based on a retrospective cohort of brain tumor patients, since we know the individual length of survival for each patient of the cohort. Therefore, the target variable is *labeled*, and the machine learning-paradigm *supervised*. Again, the actually chosen methods can vary: Common models include support vector machines (SVMs), as example of a *parametric* approach, or *the k*-nearest neighbor (KNN) algorithm as a *non-parametric* method [16]. On the other hand, in *unsupervised* learning, we generally deal with *unlabeled* data with the assumption of the structural coherence. This can be leveraged in clustering, which is a subset of unsupervised learning encompassing many different methods, e.g. hierarchical clustering or *k*-means clustering [4, 17]. The observed data is partitioned into clusters based on a measure of similarity regarding the structural architecture of the data. Similarly, dimensionality reduction methods—including principal component analysis (PCA) or autoencoders—can be applied to derive a low-dimensional representation explicitly from the present data [4, 18].

A multitude of diverse ML algorithms exist, and sometimes choosing the "right" algorithm for a given application can be quite confusing. Moreover, based on the so-called *no free lunch theorem* [19] no single statistical algorithm or model can generally be considered superior for all circumstances. Nevertheless, ML algorithms can vary greatly based on the (a) representation of the candidate algorithm, (b) the selected performance metric, and (c) the applied optimization strategy [4, 5, 20]. Representation refers to the learner's hypothesis space of how they formally deal with the problem at hand. This includes but is not limited to instance-based learners, such as KNN, which instead of performing explicit generalization compares new observations with similar instances observed during training [21]. Other representation spaces include hyperplane-based models, such as logistic regression or naïve Bayes, as well as rule-based learners, decision trees or complex neural networks, all of which are frequently leveraged in various ML problems across the neurosurgical literature [22, 23]. The evaluated performance metrics can vary greatly, too. Performance evaluation and reporting play a pivotal role in predictive analytics (c.f. cite Chap. 4). Lastly, the applied ML algorithm is *optimized* by a so-called objective function such as greedy search or unconstrained continuous optimization options, including different choices of gradient descent [24, 25]. Gradient descent represents the most common optimization strategy for neural networks and can take different forms, e.g. batch- (“vanilla”), stochastic- or mini-batch gradient descent [25]. We delve deeper into optimization to illustrate how it is used in learning.

2.3 Optimization: The Central Dogma of Learning Techniques

At the heart of nearly all ML and statistical modeling techniques used in data science lies the concept of *optimization*. Even though optimization is the backbone of algorithms ranging from linear and logistic regression to neural networks, it is not often stressed in the non-academic data science space. Optimization describes the process of iteratively adjusting parameters to improve performance. Every optimization problem can be decomposed into three basic elements: First, every algorithm has *parameters* (sometimes called *weights*) that govern how the values of the input variables lead to a prediction. In linear and logistic regression, for example, these parameters include the coefficients that are multiplied with the input variable values, as well as the intercept. Second, there may be realistic *constraints* within which the parameters, or their combinations, must fall. While simple models such as linear and logistic regression often do not have such constraints, other ML algorithms such as support vector machines or *k*-means clustering do. Lastly and importantly, the optimization process is steered by evaluating a so-called *objective function* that assesses how well the current iteration of the algorithm is performing. Commonly, these objective functions are *error* (also called *loss*) functions, describing the deviation of the predicted values from the true values that are to be predicted. Thus, these error functions must be *minimized*. Sometimes, you may choose to use indicators of performance, such as accuracy, which conversely need to be *maximized* throughout the optimization process.

The optimization process starts by randomly *initializing* all model parameters—that is, assigning some initial value for each parameter. Then, predictions are made on the training data, and the error is calculated. Subsequently, the parameters are adjusted in a certain direction, and the error function is evaluated again. If the error increases, it is likely that the direction of adjustment of the parameters was awry and thus led to a higher error on the training data. In that case, the parameter values are adjusted in different directions, and the error function is evaluated again. Should the error decrease, the parameter values will be further modified in these specific directions, until a *minimum* of the error function is reached. The goal of the optimization process is to reach the *global minimum* of the error function, that is, the lowest error that can be achieved through the combination of parameter values within their constraints. However, the optimization algorithm must avoid getting stuck at *local minima* of the error function (see Fig. 2.1).

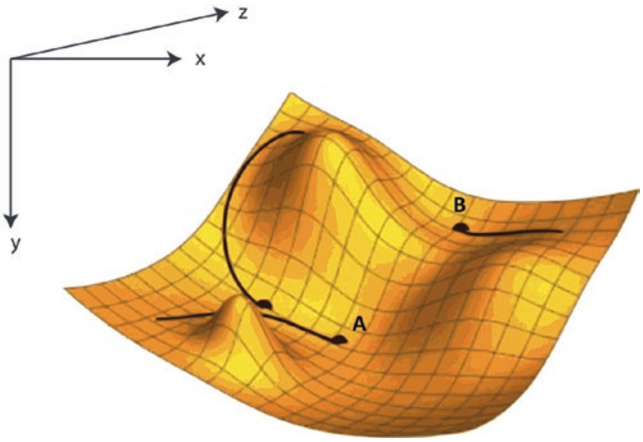

Fig. 2.1 Illustration of an optimization problem. In the *x* and *z* dimension, two parameters can take different values. In the *y* dimension, the error is displayed for different values of these two parameters. The goal of the optimization algorithm is to reach the *global minimum* (**A**) of the error through adjusting the parameter values, without getting stuck at a *local minimum* (**B**). In this example, three models are initialized with different parameter values. Two of the models converge at the global minimum (**A**), while one model gets stuck at a local minimum (**B**). Illustration by Jacopo Bertolotti. (This illustration has been made available under the Creative Commons CC0 1.0 Universal Public Domain Dedication)

The way in which the parameters are adjusted after each iteration is governed by an *optimization algorithm*, and approaches can differ greatly. For example, linear regression usually uses the ordinary least square (OLS) optimization method. In OLS, the parameters are estimated by solving an equation for the minimum of the sum of the square errors. On the other hand, *stochastic gradient descent*—which is a common optimization method for many ML algorithms—iteratively adjusts parameters as described above and as illustrated in Fig. 2.1. In stochastic gradient descent, the amount by which the parameters are changed after each iteration (also called *epoch*) is controlled by the calculated derivative (i.e. the slope or *gradient*) for each parameter with respect to the error function, and the *learning rate*. In many models, the learning rate is an important hyperparameter to set, as it controls how much parameters change in each iteration.

On the one hand, small learning rates can take many iterations to converge and make getting stuck at a local minimum more likely—on the other hand, a large learning rate can overshoot the global minimum. As a detailed discussion of the mathematical nature behind different algorithms remains beyond the scope of this introductory series, we refer to popular standard literature such as “Elements of Statistical Learning” by Hastie and Tibshirani [4], “Deep Learning” by Goodfellow et al. [26], and “Optimization for Machine Learning” by Sra et al. [27].

2.4 Explanatory Modeling Versus Predictive Modeling

The “booming” of applied ML has generated a methodological shift from *classical statistics* (experimental setting, hypothesis testing, group comparison, inference) to data-driven *statistical learning* (empirical setting, algorithmic modeling comprising ML, AI, pattern recognition) [28]. Unfortunately, the two statistical cultures have developed separately over the past decades [29] leading to incongruent evolved terminology and misunderstandings in the absence of an agreed-upon technical theorem (Table 2.1). This already becomes evident in the basic terminology describing model inputs and outputs: *predictors* or *independent variables* refer to model inputs in classical statistics, while *features* are the commonly used term in ML; outputs, known as *dependent variable* or *response*, are often labeled *target variable* or *label* in ML instead [30]. The duality of language has led to misconceptions regarding the fundamental difference between inference and prediction, as the term *prediction* has frequently been used incompatibly as in-sample correlation instead of out-of-sample generalization [31, 32]. The variation of one variable with a subsequent correlated variable later in time, such as the outcome, in the same group (in-sample correlation) does not imply prediction, and failure to account for this distinction can lead to false clinical decision-making [33, 34]. Strong associations between variables and outcome in a clinical study remain averaged estimates of the evaluated patient cohort, which does not necessarily enable predictions in unseen new patients. To shield clinicians from making wrong interpretations, we clarify the difference between explanatory modeling and predictive modeling, and highlight the potential of ML for strong predictive models.

Table 2.1 A comparison of central concepts in classical/inferential statistics versus in statistical/machine learning

Classical/inferential statistics	Statistical/machine learning
Explanatory modeling	**Predictive modeling**
An a priori chosen theoretical model is applied to data in order to test for causal hypotheses.	The process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations.
Focus on in-sample estimates	**Focus on out-of-sample estimates**
Goal: to confirm the existence of an effect in the entire data sample. Often using significance testing.	Goal: Use the best performing model to make new prediction for single new observations. Often using resampling techniques.
Focus on model interpretability	**Focus on model performance**
The model is chosen a priori, while models with intrinsic means of interpretability are preferred, e.g. a GLM, often parametric with a few fixed parameters.	Different models are applied and the best performing one is selected. Models tend to be more flexible and expressive, often non-parametric with many parameters adapting to the present data.
Experimental data	**Empirical data**
Long data (n samples > p variables)	**Wide data (n samples $\ll p$ variables)**
Independent variables	**Features**
Dependent variable	**Target variable**
Learn deductively by model testing	**Learn a model from data inductively**

Knowledge generation in clinical research has nearly exclusively been dominated by classical statistics with the focus on *explanatory modeling* (EM) [32]. In carefully designed experiments or clinical studies, a constructed theoretical model, e.g. a regression model, is applied to data in order to test for causal hypotheses. Based on theory, a model is chosen a priori, combining a fixed number of experimental variables, which are under the control of the investigator. Explicit model assumptions such as the Gaussian distribution assumption are made, and the model, which is believed to represent the true *data generating process*, is evaluated for the entire present data sample based on hypothesis and significance testing (“inference”). In such association-based modeling, a set of independent variables (X) are assumed to behave according to a certain mechanism (“theory”) and ultimately cause an effect measured by the dependent variable (Y). Indeed, the role of *theory* in explanatory modeling is strong and is always reflected in the applied model, with the aim to obtain the most accurate representation of the underlying theory (technically speaking, classical statistics seeks to minimize *bias*). Whether *theory* holds true and the effect actually exists is then confirmed in the data, hence the overall analytical goal is *inference*.

Machine learning-based *predictive modeling* (PM) is defined as the process of applying a statistical model or data mining algorithm to data for the purpose of predicting future observations. In a heuristic approach, ML or PM is applied to *empirical data* as opposed to experimentally controlled data.

As the name implies, the primary focus lays on optimizing the prediction of a target variable (Y) for new observations given their set of features (X). As opposed to explanatory modeling, PM is *forward looking* [32] with the intention of predicting new observations, and hence *generalization beyond the present data* is the fundamental goal of the analysis. In contrast to EM, PM seeks to minimize both *variance* and *bias* [35, 36], occasionally sacrificing the theoretical interpretability for enhanced predictive power. Any underlying method can constitute a predictive model ranging from parametric and rigid models to highly flexible non-parametric and complex models. With a minimum of a priori specifications, a model is then heuristically derived from the data [37, 38]. The true data generating process lays in the data, and is inductively learned and approximated by ML models.

2.5 Workflow for Predictive Modeling

In clinical predictive analytics, *generalization* is our ultimate goal. To answer different research objectives, we develop, test, and evaluate different models for the purpose of clinical application (for an overview see https://topepo.github.io/caret/available-models.html). Many research objectives in PM can be framed either as the prediction of a continuous endpoint (regression) such as progression-free survival measured in months or alternatively as the prediction of a binary endpoint (classification), e.g. survival after 12 months as a dichotomized binary. Most continuous variables can easily be reduced and dichotomized into binary variables, but as a result data granularity is lost. Both regression and classification share a common analytical workflow with difference in regard to model evaluation and reporting (c.f. *cite* Chap. 5 *Classification problems* and *cite* Chap. 6 *Regression problems* for a detailed discussion). An adaptable pipeline for both regression and classification problems is demonstrated in Parts IV and V. Both sections contain detailed instructions along with a simulated dataset of 10,000 patients with glioblastoma and the code based on the statistical programming language R, which is available as open-source software.

For a general overview, a four-step approach to PM is proposed (Fig. 2.2): First and most important (1) all data needs to be pre-processed. ML is often thought of as *letting data do the heavy lifting*, which in part is correct, however, the raw data is often not suited to learning well in its current form. A lot of work needs to be allocated to preparing the input data including data cleaning and pre-processing (imputation, scaling, normalization, encoding) as well as *feature engineering* and *selection*. This is followed by using (2) resampling techniques such as *k*-fold cross-validation (c.f. cite Chap. 3 *generalization and overfitting*) to train different models and perform hyperparameter tuning. In a third step (3), the different models are compared and evaluated for generalizability based on a chosen out-of-sample performance measure in an independent testing set. The best performing model is ultimately selected, the model's out-of-sample calibration assessed (c.f. cite Chap. 4 *Evaluation and points of significance*), and, in a fourth step (4) the model is externally validated—or at least prospectively internally validated—to ensure clinical usage is safe and generalizable across locations, different populations and end users (c.f. *cite* Chap. 3 *Generalization and overfitting*). The European Union (EU) and the Food and Drug Administration (FDA) have both set standards for classifying machine learning and other software for use in healthcare, upon which the extensiveness of validation that is required before approved introduction into clinical practice is based. For example, to receive the CE mark for a clinical decision support (CDS) algorithm—depending on classification—the EU requires compliance with ISO 13485 standards, as well as a clinical evaluation report (CER) that includes a literature review and clinical testing (validation) [39].

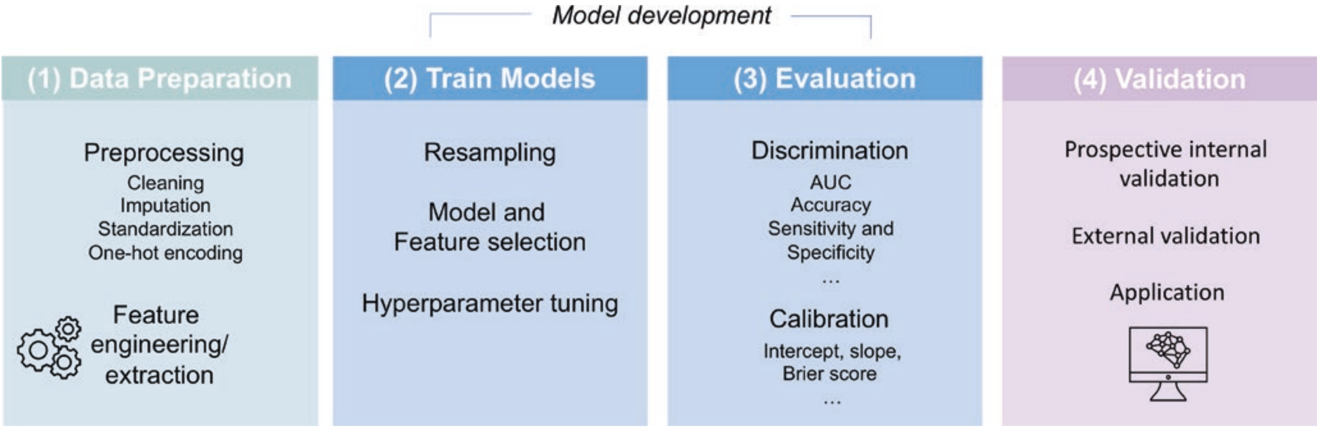

Fig. 2.2 A four-step predictive modeling workflow. (1) Data preparation includes cleaning and featurization of the given raw data. Data preprocessing combines cleaning and outlier detection, missing data imputation, the use of standardization methods, and correct feature encoding. The pre-processed data is further formed into features—manually in a process called *feature engineering* or automatically deduced by a process called *feature extraction*. In the training process (2) resampling techniques such as *k*-fold cross-validation are used to train and tune different models. Most predictive features are identified in a *feature selection* process. (3) Models are compared and evaluated for generalizability in an independent testing set. The best performing model is selected, and out-of-sample discrimination and calibration are assessed. (4) The generalizing model is prospectively internally and externally validated to ensure safe clinical usage across locations and users

2.6 Conclusion

We appear to be at the beginning of an accelerated trend towards data-driven decision-making in biomedicine enabled by a transformative technology—machine learning [5]. Given the ever-growing and highly complex “big data” biomedical datasets and increases in computational power, machine learning approaches prove to be highly successful analytical strategies towards a patient-tailored approach regarding diagnosis, treatment choice, and outcome prediction. Going forward, we expect that training neuroscientists and clinicians in the concepts of machine learning will undoubtably be a corner stone for the advancement of individualized medicine in the realm of precision medicine. With the series “*Machine learning-based clinical prediction modeling*,” we aim to provide both a conceptual and practical guideline for predictive analytics in the clinical routine to strengthen every clinician’s competence in modern machine learning techniques.

Disclosures

Funding No funding was received for this research.

Conflict of Interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers’ bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent No human or animal participants were included in this study.

References

1. Kuhn M, Johnson K. Applied predictive modeling. New York, NY: Springer Science & Business Media; 2013.
2. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY: Springer Science & Business Media; 2008.
3. Senders JT, Zaki MM, Karhade AV, Chang B, Gormley WB, Broekman ML, Smith TR, Arnaout O. An introduction and overview of machine learning in neurosurgical care. Acta Neurochir. 2018;160:29. https://doi.org/10.1007/s00701-017-3385-8.
4. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York, NY: Springer Science & Business Media; 2013.
5. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349:255. https://doi.org/10.1126/science.aaa8415.
6. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. New York, NY: Chapman and Hall; 2015. https://doi.org/10.1201/b18401.
7. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58:267. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
8. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006. https://doi.org/10.1038/ncomms5006.
9. Li H, Zhu Y, Burnside ES, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ Breast Cancer. 2016;2:16012. https://doi.org/10.1038/npjbcancer.2016.12.
10. Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, Madabhushi A. Radiomics and radiogenomics in lung cancer: a review for the clinician. Lung Cancer. 2018;115:34. https://doi.org/10.1016/j.lungcan.2017.10.015.
11. Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. PLoS One. 2019;14(3):e0214365.
12. Mitchell TM. The discipline of machine learning. Mach Learn. 2006;17:1. https://doi.org/10.1080/026404199365326.
13. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci. 2019;64:277. https://doi.org/10.1016/j.jocn.2019.03.001.
14. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320:2199. https://doi.org/10.1001/jama.2018.17163.
15. Grigsby J, Kramer RE, Schneiders JL, Gates JR, Smith WB. Predicting outcome of anterior temporal lobectomy using simulated neural networks. Epilepsia. 1998;39:61. https://doi.org/10.1111/j.1528-1157.1998.tb01275.x.
16. Bzdok D, Krzywinski M, Altman N. Points of significance: machine learning: supervised methods. Nat Methods. 2018;15:5. https://doi.org/10.1038/nmeth.4551.
17. Altman N, Krzywinski M. Points of significance: clustering. Nat Methods. 2017;14:545. https://doi.org/10.1038/nmeth.4299.
18. Murphy KP. Machine learning: a probabilistic perspective. Cambridge, MA: MIT Press; 2012.
19. Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural Comput. 1996;8:1341. https://doi.org/10.1162/neco.1996.8.7.1341.
20. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012;55(10):78.
21. Armañanzas R, Alonso-Nanclares L, DeFelipe-Oroquieta J, Kastanauskaite A, de Sola RG, DeFelipe J, Bielza C, Larrañaga P. Machine learning approach for the outcome prediction of temporal lobe epilepsy surgery. PLoS One. 2013;8:e62819. https://doi.org/10.1371/journal.pone.0062819.
22. Bydon M, Schirmer CM, Oermann EK, Kitagawa RS, Pouratian N, Davies J, Sharan A, Chambless LB. Big data defined: a practical review for neurosurgeons. World Neurosurg. 2020;133:e842. https://doi.org/10.1016/j.wneu.2019.09.092.
23. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, Smith TR, Arnaout O. Machine learning and neurosurgical outcome prediction: a systematic review.

World Neurosurg. 2018;109:476. https://doi.org/10.1016/j.wneu.2017.09.149.

24. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Proc COMPSTAT2010; 2010. https://doi.org/10.1007/978-3-7908-2604-3_16.
25. Ruder S. An overview of gradient descent optimization algorithms. ArXiv. 2017:160904747. Cs.
26. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
27. Sra S, Nowozin S, Wright SJ. Optimization for machine learning. Cambridge, MA: MIT Press; 2012.
28. Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, Steyerberg EW, CENTER-TBI Collaborators. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J Clin Epidemiol. 2020;122:95–107.
29. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199–231.
30. Bzdok D. Classical statistics and statistical learning in imaging neuroscience. Front Neurosci. 2017;11:543. https://doi.org/10.3389/fnins.2017.00543.
31. Gabrieli JDE, Ghosh SS, Whitfield-Gabrieli S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron. 2015;85:11. https://doi.org/10.1016/j.neuron.2014.10.047.
32. Shmueli G. To explain or to predict? Stat Sci. 2011;25(3):289–310.
33. Whelan R, Garavan H. When optimism hurts: inflated predictions in psychiatric neuroimaging. Biol Psychiatry. 2014;75:746. https://doi.org/10.1016/j.biopsych.2013.05.014.
34. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci. 2017;12:1100. https://doi.org/10.1177/1745691617693393.
35. Domingos P. A unified bias-variance decomposition and its applications. In: Proc 17th Int. Conf Mach. Learn. San Francisco, CA: Morgan Kaufmann; 2000. p. 231–8.
36. James G, Hastie T. Generalizations of the bias/variance decomposition for prediction error. Stanford, CA: Department of Statistics, Stanford University; 1997.
37. Abu-Mostafa YS, Malik M-I, Lin HT. Learning from data: a short course. Chicago, IL: AMLBook; 2012. https://doi.org/10.1108/17538271011063889.
38. Van der Laan M, Hubbard AE, Jewell N. Learning FROM DATA. Epidemiology. 2010;21:479. https://doi.org/10.1097/ede.0b013e3181e13328.
39. Harvey H. How to get clinical AI tech approved by regulators. Medium; 2019. https://towardsdatascience.com/how-to-get-clinical-ai-tech-approved-by-regulators-fa16dfa1983b. Accessed 3 May 2020.