Tutorial

Machine Learning and Health Science Research: Tutorial

Hunyong Cho[1*], PhD; Jane She[1*], BA; Daniel De Marchi[1], BSc; Helal El-Zaatari[1], BSc; Edward L Barnes[2,3], MPH, MD; Anna R Kahkoska[4,5,6], MD, PhD; Michael R Kosorok[1], MM, PhD; Arti V Virkud[7], PhD

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[2]Division of Gastroenterology and Hepatology, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[3]Center for Gastrointestinal Biology and Diseases, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[4]Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[5]Division of Endocrinology and Metabolism, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[6]Center for Aging and Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[7]Kidney Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States
[*]these authors contributed equally

Corresponding Author:
Jane She, BA
Department of Biostatistics
University of North Carolina at Chapel Hill
3101 McGavran-Greenberg Hall
CB #7420
Chapel Hill, NC, 27599-7420
United States
Phone: 1 (919) 966 7250
Email: jane.she@unc.edu

Abstract

Machine learning (ML) has seen impressive growth in health science research due to its capacity for handling complex data to perform a range of tasks, including unsupervised learning, supervised learning, and reinforcement learning. To aid health science researchers in understanding the strengths and limitations of ML and to facilitate its integration into their studies, we present here a guideline for integrating ML into an analysis through a structured framework, covering steps from framing a research question to study design and analysis techniques for specialized data types.

(J Med Internet Res 2024;26:e50890) doi: 10.2196/50890

KEYWORDS

health science researcher; machine learning pipeline; machine learning; medical machine learning; precision medicine; reproducibility; unsupervised learning

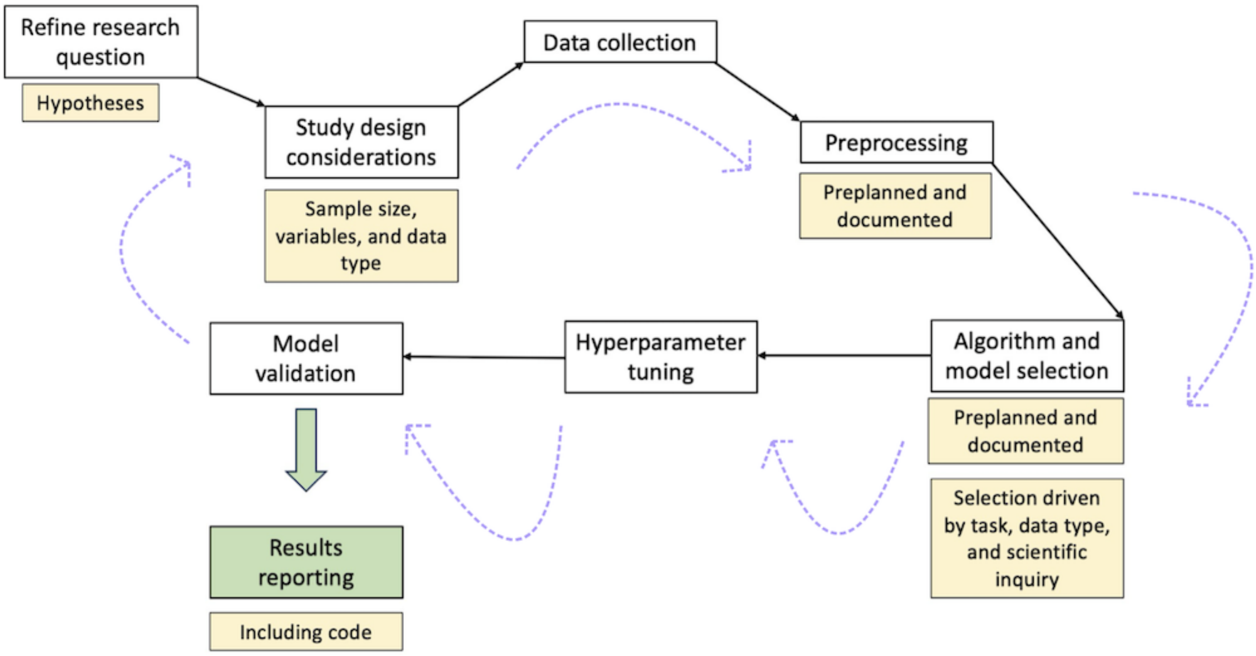
Introduction

As a brief overview, machine learning (ML) is generally characterized by model complexity and capacity for processing high-dimensional or complicated data forms and is often mentioned as an antonym to traditional statistical learning algorithms. However, this division is not clear, and ML algorithms range from traditional statistical analysis tools such as simple linear regression to cutting-edge deep neural network algorithms. While often used interchangeably with artificial intelligence (AI), ML is a subset of AI and seeks to use data-driven methods to identify patterns and make decisions. This can then be used in the field of AI to allow problem-solving and decision-making.

ML is becoming increasingly popular in the research community due to the proliferation of complex or unstructured data sets and the increased capacity and access to computing power needed to run these models. ML models can often discover sophisticated and surprising patterns in these data sets that would be difficult to discover using classical methods [1,2]. The health science research domain has been no exception to this paradigm, as the health science fields have an abundance of data well suited for these models, such as genomics sequencing data and electronic health records (EHR) data [3-6]. Applications of ML to the health field can lead to targeted interventions to provide support for health care professionals [7]. ML has also become almost indispensable to the fast-growing field of PM, which uses rich patient information to precisely target interventions [8].

This paper provides a sequential framework for health scientists intending to use ML in a research proposal and discusses types of analyses that can be done and factors to consider. It will also include a special introduction to the field of PM, which has become a popular research area with the development of new ML methodologies. Finally, we discuss some unique data types and analysis techniques specific to those application areas. In general, throughout the study design process, documentation and preplanning are highly recommended for the sake of reproducibility of the work carried out. For a visual illustration of the research pipeline flowchart, see Figure 1. There are also existing pipelines such as MLOps and CRISP used in business and industry settings that may be adapted to health science research fields; however, this paper will follow a framework more commonly seen in health science research. We relegate some technical topics, such as general sample size calculations, model training, and model tuning and validation to Multimedia Appendix 1 [9-21]. Readers are also encouraged to reference other ML primers, such as one for epidemiologists [22] and 1 for biologists [23].

Figure 1. Machine learning workflow for a health science research question, from research question refinement to results reporting, with additional considerations. The cyclic nature of the process is reflected in the arrows, as several different iterations may be considered before narrowing down to a decisive pipeline, leading to result reporting.


Refine research question
Hypotheses
Study design considerations
Sample size, variables, and data type
Data collection
Preprocessing
Preplanned and documented
Algorithm and model selection
Preplanned and documented
Selection driven by task, data type, and scientific inquiry
Hyperparameter tuning
Model validation
Results reporting
Including code

Experimentation

This section introduces step-by-step core considerations for designing an ML-involved research project.

Refining Research Questions: What Can Machine Learning Do?

ML methods can be used to answer questions for studies that may fall within the following categories: prediction, estimation, understanding causal associations, and decision support. ML can also help support main analyses as an auxiliary tool through missing data imputation, inverse propensity score weighting, dimensionality reduction, and variable selection. This last point will be covered further in the Data Collection and Preprocessing section.

Inquiries in traditional studies are often limited to the discovery and measurement of the size of certain effects or to establishing the causal relationship between variables. These are called estimation and causal inference, respectively. The recent realm of research has also expanded to prediction [24], where algorithms can predict an outcome for a patient when given a set of input variables. Functionalities such as estimation, establishing causal relationships, and prediction are meant to, in an indirect manner, support clinical decision-making. However, there are decision-support frameworks that explicitly provide recommendations through reinforcement learning (RL; defined in the section Advanced Concepts). One of the most important applications in the medical domain is PM, where the goal is to provide an optimal treatment for individual patients each with unique characteristics [8].

Prediction

As a concrete example, suppose a researcher is interested in investigating the health effects of electronic screen time use over several months [25]. This is an example of a study where the research question hinges on accurate prediction of the use of a screen. Since it is impractical and unethical to monitor an individual for several months and self-reported measures can be unreliable, the use of ML algorithms for predicting screen time is useful [26]. Prediction has also been used in the identification of early cancer diagnoses using image data analysis [27]. Classification of patients for disease screening, a prediction task, can be performed with high accuracy using ML.

Although the accuracy of a predictive algorithm is considered one of the most important virtues, interpretability [28] is another important aspect to consider, especially in health science research. Interpretability often comes at the price of reduced accuracy, which is sometimes framed as the "interpretability-accuracy trade-off." More complex models, which may improve prediction accuracy, maybe less interpretable, as it can be difficult to trace why the model arrived at such a decision, how the predictors relate to the outcome, and how to interpret the results. Interpretability generally is used to mean being able to understand the inner workings of a model, but as evidenced by the previous sentence, it can encompass several different aspects. These can range from overall model structure, ability to explain individual predictors, transparency of decision-making processes, and more. Measuring interpretability is a challenge, as it can be context-dependent for the problem you are working on; more information on interpretability can be found at [29]. In the screen time prediction example above, interpretability is not of concern but rather maximizing the prediction accuracy, as it may not be of interest how the algorithm predicted the values, but rather the predicted values themselves.

Estimation

ML algorithms can also be used to estimate associations between exposures and health outcomes [30]. Examples include calculating the odds ratio of obesity while comparing 2 socioeconomic statuses, measuring the association between physical activity and mortality [31], and estimating the association between sociodemographic traits and diabetes prevalence [32]. However, the estimation procedures of ML algorithms are often limited to point estimation and usually lack inferential abilities such as *P* values and Bayes factors. Meaning, estimation procedures can usually find an approximate value of a parameter (like an average) through point estimation but are not usually able to output other quantities such as CIs or hypothesis tests which provide information on the population as a whole. This is because models that are nonparametric or complex may not make certain distributional assumptions, which makes quantifying CIs for a point estimate not easily doable. For an investigator wanting to confirm the positive effects of a medical treatment on patient health outcomes, ML often cannot discern whether the estimated effect size is statistically significant or not. Rather, this can be done through classical statistical tests, which possess inferential capabilities. However, this limitation is not the same as generating CIs for model performance, which is a separate procedure and generally more straightforward as model evaluation may involve data splitting or repeated sampling.

That being said, certain ML algorithms still have the potential for inferential capacity. Recently, a random forest-based framework for judging the statistical significance of heterogeneous treatment effects for individuals with specific covariate values has been developed [33]. Additionally, many other algorithms, such as support vector machine (SVM) and *k*-nearest neighbors (*k*-NN), can output CIs and *P* values for estimated effects [34,35]. However, these approaches are, in general, much less efficient than classical statistical tests and thus should be used after carefully considering the trade-off between flexibility (model specification) and efficiency (power of the test).

Causal Inference

Understanding causal associations is the activity of investigating the cause of an outcome, such as the occurrence of disease. In the statistical literature, it is known as causal inference, which provides a foundation for establishing causality [36,37]. Research questions related to understanding causal associations include estimation of average or individual treatment effects (ATE and ITE, respectively) and identification of important risk factors or subgroups for a health outcome. A rich literature for causal inference methods has been developed in statistics. For example, when estimating average treatment effects [38] from observational data, propensity score matching [39] is frequently used, which is often done using flexible models such as random forests [40]. However, its use should be carefully considered due to potential small sample size issues and covariate imbalance.

Study Design Considerations

Quality of data is a key design consideration for the successful use of ML. Given the complexity of ML, which often involves managing a vast range of input variables coming in various formats, it is crucial to plan the identification, collection, and management of these variables. Data from multiple sources—for example, clinical information, genomics data, and medical images have different dimensions—eventually needs to be aligned for downstream analyses using techniques such as feature concatenation, feature extraction, and tree and metric-based learning, so planning the process ahead of time is essential to consider any feasibility issues [41,42].

In ML studies, missing data are one of the most frequently observed issues that can harm the quality of data and can lead to bias. Thus, planning the data collection process to minimize missing data and setting up quality control checks on data entry errors is essential. Approaches to data missingness will be described in the Data Collection and Preprocessing section.

Sample Size and Strategies for Sample Size Determination

In general, ML models with tunable parameters require much larger sample sizes than traditional statistical models to achieve the same level of estimation or prediction accuracy. Since ML models usually have much weaker model assumptions than traditional parametric models, when the dimension of a parameter is much larger, more data are needed for the estimator to determine the model structure on top of estimating the mean outcome or the parameters of interest. The phenomenon when the required sample size grows exponentially with the dimension of the parameter is called the "curse of dimensionality," which is attributable to the nonparametric nature of ML models.

This relatively large sample size requirement is not the only issue, but precisely calibrating the required size is another challenge. Unlike traditional clinical trials, where the sample size of a study is planned to achieve a certain amount of power to detect a certain effect size [43,44], the sample size determination for ML has a different meaning, and there is no generic framework for it [45]. In ML, where the model performance is often measured in terms of prediction accuracy; measures, such as mean squared error and classification error rate, are meant to be controlled under a predefined level, and the sample size that meets such prediction accuracy is to be derived.

Popular choices of evaluation metrics include mean squared error and R^2 for continuous outcomes, Brier scores for survival outcomes, classification error rate (accuracy rate) for categorical outcomes, and area under the receiver operating characteristics curve (AUC) for binary outcomes. However, these evaluation metrics should be chosen after consideration of the cost of wrong predictions and the benefits of correct predictions. For example, a model for predicting cancer may have to impose a higher cost for false negative than for false positive. Thus, a true negative rate (TNR) or a partial AUC could be considered for its evaluation measure after considering threshold selection and other possible reporting metrics. There are no “best” evaluation metrics, as this is highly dependent on the problem itself beyond the characterization of a classification or regression framework; differences in metrics can emerge when there are outliers in the data set, model comparison, and differential penalties for errors.

Although there is no deterministic sample size formula for predictive models, one can fit a learning curve on the training data for a given ML algorithm based on some evaluation measures such as the prediction error rate and AUC, which quantifies the overall accuracy of a binary classification model [45]. Essentially, the researcher is required to run the ML algorithm for the pilot data using training data and project the evaluation measure based on the fitted learning curve through the evaluation of the testing data. This evaluation measure is then used to inform the sample size or amount of data needed for the specific accuracy or statistical power desired [46].

To accurately estimate this curve, at least 2 or 3 points are required [47]. This means that the researcher is required to take at least 2 subsets of the available data and calculate 2 respective error rates. However, the pilot data might not capture all the biases present in the larger data set, as the sample may not be fully representative of the population or phenomena of interest. The researcher must be wary of generalizations using this pilot data. It is therefore recommended that a statistician trained in ML be present to assist with these technical sample size estimation procedures.

More details are included in section A of Multimedia Appendix 1 [9-21], which also includes information on how to mitigate the large sample requirement in neural networks through augmentation techniques and transfer learning (defined in the section Advanced Concepts).

Data Collection and Data Preprocessing

As previously mentioned, EHR, administrative claims, clinical trials, and longitudinal cohort data are major data types in the ML world. However, there are also “specialized data types,” which require their own distinct methods of analysis due to their unique qualities. These include textual or language data, imaging, and genomics, and will be discussed in the Applications section. Due to the highly complex nature of the data being used, ML analyses often involve heavy data preprocessing. This step often requires more time than the main analysis itself and not only includes screening for erratic values, detecting and understanding outliers, and handling of missing values, but also transformation of the data into a software-friendly format, feature scaling, feature selection, dimensionality reduction, and sample splitting for validation, among others [48].

These procedures, while seemingly not important, may bring significant changes to the conclusion. For example, data preprocessing is an essential step for categorical features when using certain gradient boosting algorithms such as XGBoost, as the algorithm requires the categorical variables to be coded through mean coding or one hot coding before use in the model. Additionally, feature scaling would change the results of any methods involving Euclidean distance metrics such as principal component analysis (PCA), *k*-means, and *k*-NN.

As mentioned in the section Refining Research Questions: What Can Machine Learning Do? ML can be used as an auxiliary tool for missing data imputation [49], dimensionality reduction [50] before regression analysis, and variable selection, all of which can make an analysis more manageable.

Missing data, which typically arises in survival analysis, longitudinal studies, among other scientific studies, has great potential to create statistical bias if not accounted for in an auxiliary analysis [51]. Simply discarding observations with missing data may lead to selection bias and reduced sample size, resulting in incorrect estimation of relationships. Instead, the mechanism behind the missing data can be accounted for through an auxiliary analysis to mitigate the effects of the bias using tools such as imputation and maximum likelihood estimation. See, for example, “missForest” for imputation based on random forests [52]. As an example, [49] provides a real-use case of how ML methods can be used to impute missing data in a breast cancer problem.

Algorithm and Model Selection

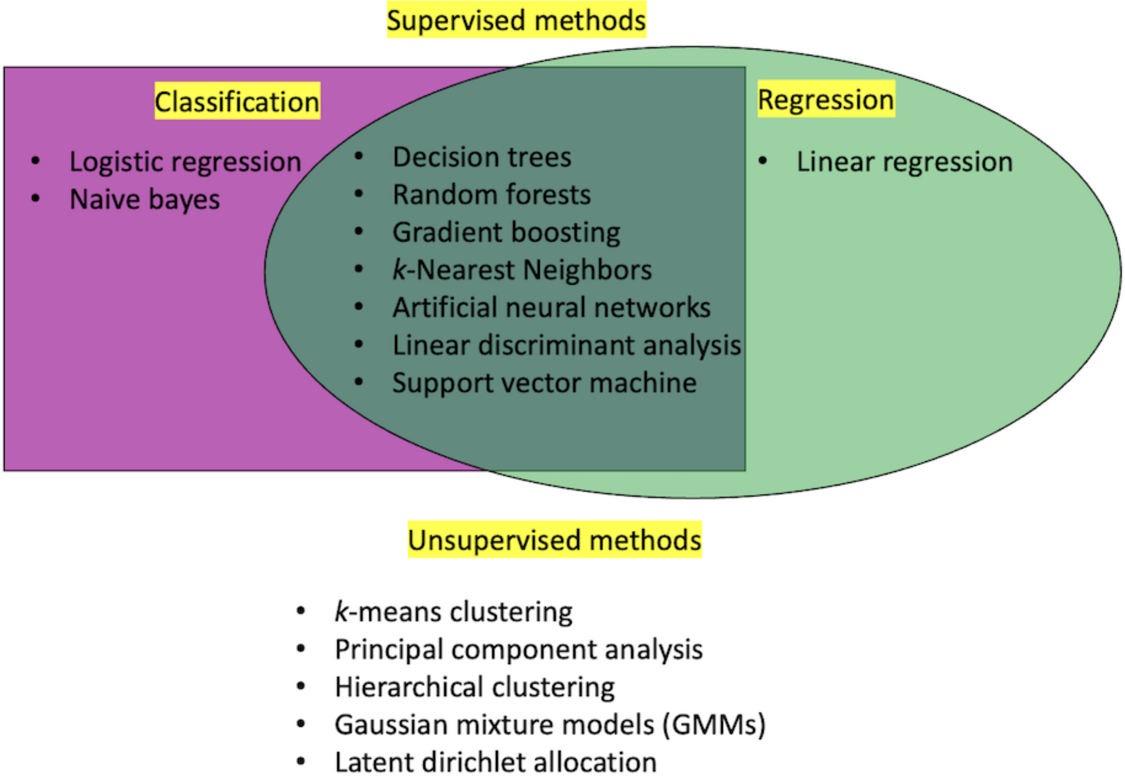
The choice of an ML method largely depends on the type of task and data type. For example, linear discriminant analysis (LDA) and *k*-means clustering can only be used with continuous predictors; SVM and support vector regression can be used for classification and regression problems, respectively; and random forests and neural networks are capable of both classification and regression. Table 1 lists commonly available algorithms in each category and summarizes their benefits and drawbacks.

Once the candidate algorithms are identified, the choice of the algorithm may be driven by the scientific inquiry, as discussed in the section Refining Research Questions: What Can Machine Learning Do? Additional factors for algorithm choice may include computing resources, data limitations, and data assumptions. Figure 2 gives a list of common ML algorithms and the purposes they may be used for. The nature of the scientific study will determine the importance of interpretability in the prediction of particular phenomena. A “black box” predictive model may not clearly explain why such predictions were made, only what the predictions are [53]. For clinicians who want to attribute a specific cause of an output, these methods may be less suited for their research question, and it is suggested to use a more interpretable model. As an extreme example, consider using an ML algorithm to support the decision of no amputation, minor amputation, or major amputation for a patient with diabetic foot ulcer. One can imagine that an interpretable ML algorithm must be preferred, as was proposed in [54], as the decision-making process needs to be clear before an amputation is carried out.

Table 1. Benefits and drawbacks of common machine learning methods in supervised and unsupervised settings. This list is not exhaustive and includes popular machine-learning algorithms in each category.

Methods	Strengths	Limitations
Benefits and drawbacks of some supervised methods		
Logistic regression and linear regression	• Interpretable • Easy implementation	• Overfitting with highly correlated data (use variable selection or shrinkage methods) • Poor performance for nonlinearly separable data
Naive Bayes	• Performs well even without conditional independence • Easy implementation	• Simple; outperformed by well-tuned, more complex models
k-nearest neighbors	• Nonparametric (no model assumptions needed) • High level of flexibility; performs well for nonlinear boundaries	• Low interpretability • Poor performance for high-dimensional data • Difficulty dealing with missing values
Support vector machine	• Performs well with high-dimensional data and nonlinear boundaries	• Low interpretability • Poor performance for imbalanced data • Usually outperformed by newer methods
Decision trees	• Interpretable for trivial data sets • Nonparametric (no model assumptions needed) • Works with nonlinear relationships • Classification for more than 2 classes	• Prone to overfitting • Difficult to interpret for nontrivial data sets
Random forest	• Handles high-dimensional data well • Reduces overfitting from decision trees • Reduces variance	• Low interpretability • Poor performance for sparse data
Gradient boosted trees	• Increased accuracy over random forests	• More difficult to implement due to tuning parameter selection
Artificial neural networks	• Works well with many data types (images, text, audio, etc) • Adaptable architecture	• Low interpretability • Overfitting if trained too long • Requires a great deal of data
Benefits and drawbacks of some unsupervised methods		
k-means clustering	• Fast, easy implementation	• Only quantitative data • No clear best way to choose k • Poor performance for noncircular cluster
Hierarchical cluster	• Reproducible • Visually interpretable by dendrograms • Cluster shape not assumed to be globular	• Poor performance on high-dimensional data • Hierarchy level must be selected
Gaussian mixture models	• Flexibility since clusters can have irregular shapes • No assumption of cluster number or level • Accommodates mixed cluster membership	• Poor performance on high-dimensional data
Linear discriminant analysis	• Interpretable • Can lower model variance over logistic regression if model assumptions are met	• Can only be used with continuous predictors • Poor performance for nonlinearly separable data (try quadratic discriminant analysis)

Figure 2. Commonly used algorithms in the supervised setting by algorithm type distinguished between classification and regression problems, as well as methods used in unsupervised learning.



Being too open-ended about model possibilities may lead to nonreproducibility or phenomena analogous to *P*-hacking, where researchers may choose the model that leads to the highest accuracy for the data at hand after trials of multiple approaches. This highlights the importance of having a held-out test set, which is used only at the end of model development to report model performance results, as well as having an appropriate justification for model selection. More details about a held-out test set can be found in the Hyperparameter and Model Validation section. It is important to be specific enough about the goals of the analysis to justify the use of different algorithms. At the same time, being too specific can put too many undesired constraints on research, unduly limiting the use of adequate algorithms and models [55].

Hyperparameter Tuning and Model Validation

Assuming an appropriate model has been chosen, hyperparameter tuning and model validation become the next steps for an ML practitioner. In section B of Multimedia Appendix 1 [9-21], we provide guidelines for tuning hyperparameters for 2 popular ML methods—tree-based methods and neural networks. The relatively high performance of these models is achieved by adequately tuning the hyperparameters.

Model performance assessment metrics are used to determine how well a trained model performs on new, unseen data. Popular model assessment metrics include in regression: R^2 values, mean absolute error, mean square error, and in classification: recall, F_1-score, and AUC. Beyond these metrics, aspects of model performance can also include ease of use and deployment feasibility. In many health care cases, understanding how a model reached the conclusion as well as interpreting the results of the conclusion may be preferred over blackbox models, as medical decisions are made based on the results of the model. The model deployment aspect focuses on its practical use; a complicated model may not be used in resource-constrained settings, so its use may not be feasible.

Typically, the preprocessed data are split into separate training and testing sets. The term “validation set” is often used interchangeably with the term “test set” and usually refers to a portion of the data that are not used in training the model. The model is evaluated on the test set to give an unbiased estimate of model performance on unseen data. This test set cannot be touched before model fitting and is not used for training the model or tuning model parameters.

In some literature, the training data themselves may also be split into 2 separate data sets: one, dubbed the “training set,” is used to train the models to get parameter estimates, and the other, the “validation set,” is used to help tune parameters. Therefore, rather than a split into a training and testing set as previously mentioned, we have a training, validation, and testing split, where the test set is held out until model performance evaluation. The interchangeable use of “test set” and “validation set” in this case may be confusing, as they do not refer to the same thing when data are portioned this way—one must be careful in reading to understand which scenario is occurring.

When partitioning data, it is important for the test set to be representative of the data rather than having different characteristics than the training set. There are various factors to consider when forming a test set, which can depend on the use case. For example, one may want the training and test sets to contain records from different individuals for diagnosis purposes, or for the training and test sets to contain observations from different time points on the same individuals for prognosis. Saeb et al [56] discuss examples of using different types of splitting in cross-validation and how results can differ based on the partitioning.

Section B of Multimedia Appendix 1 [9-21] continues to discuss how to validate these models using *k*-fold cross-validation, which is a validation method that uses a given sample, assuming collecting additional data is difficult. For high-dimensional data where *k*-fold cross-validation is infeasible to implement due to computational costs, an alternative approach is introduced. It should be noted that cross-validation methods still require a held-out test set to evaluate model performance at the very end; *k*-fold cross-validation is used on the training data set to tune the parameters, but it does not replace the need for a separate testing set. An introduction to how ML models are trained is also discussed in section C of Multimedia Appendix 1 [9-21] for additional information and completion in the understanding of model training.

Results and Reproducibility

The final step of any project is to report the results. Luo and colleagues [57] set up reporting standards of ML predictive model-based research for biomedical researchers, which include a list of reporting items to be included. Reporting of such items is essential to promoting reproducibility in research. Among the items are details including the nature of the study along with a background, objectives, clinical rationale, data sources, type of modeling, inclusion and exclusion criteria, time span, model validation strategies, handling of missing values, cleaning and transformation, candidate modeling techniques with justification, model selection criteria, clinical implications, and model limitations [57].

Several of these reporting items have been discussed in the paper and fall under the outlined categories of research question refinement, study design considerations, data collection and preprocessing, algorithm and model selection, hyperparameter tuning, and model validation. The goal of the list of reporting standards for predictive modeling is to encourage transparency and reproducibility to ensure credibility in the scientific community and the methodological soundness of research.

The need for transparent reporting is even more apparent when considering the nuances of ML. While studies involving ML have experimental design steps that overlap with general study considerations such as refining a research question, study design, and data collection, the use of ML in health science studies requires ML-specific considerations in terms of quality and size of data, adequacy of methods, and reproducibility. These issues are inherent in ML-involved research owing to both the complexity of data and ML models and a wide spectrum of ML methods. Readers are encouraged to read the literature [58-60] for guidelines on general study considerations. Therefore, there is a necessity for conscientious approaches to reporting. Within each ML method, there are usually one or more hyperparameters, such as the depth and node size in tree methods and the penalty term in kernel regression [61]. Cherry-picking a hyperparameter after looking at the data multiple times may result in irreproducibility. This is why the held-out test set can only be used for result reporting and should not be used for further model development. As previously mentioned, for the sake of reproducibility, keeping documentation of the research pipeline from start to finish as outlined is also necessary.

Application

This section includes specialized data types in ML.

Natural Language Processing

Medical notes of physicians may contain important information beyond quantitative clinical records. They, however, are not readily analyzable without processing such as transcription and topic extraction. Natural language processing (NLP) does what was previously considered impossible by processing such nonstandard form of massive data into a readily analyzable format, opening huge opportunities for health science research.

NLP as a field has undergone a revolution since 2018. The seminal paper, "BERT" [62], delivered unprecedented performance on almost every major language task. NLP models using transformers (defined in the section Advanced Concepts), such as BERT and GPT [63,64], can be used for various language tasks, including classification, summarization, imputation, and prediction. The most common and useful tasks in medical NLP usually deal with hospital documents and patient interactions. For instance, NLP models can be used to automatically transcribe patient conversations, predict disease from medical notes, or impute missing values in medical forms [65]. There are many high-quality models trained on massive text corpora that can, out of the box, deliver state-of-the-art performance on almost any task.

Therefore, the first step in any NLP project is to select a pretrained model closest to the language domain being used and perform transfer learning. Transfer learning is where a model pretrained on 1 task is then trained on a related but different task. For example, for medical tasks, "Med-BERT: pretrained contextualized embeddings on large-scale structured EHR for disease prediction" and "SciBERT: a pretrained language model for scientific text" may be of use, as they are trained on similar language as is used in medical contexts [60,61,66]. The original BERT model will also work well for most purposes. The common structure of language and the size of most of these training data sets (terabytes for some models) mean that a general model will have almost certainly been exposed to any sort of text problem a researcher may be interested in due to the sheer breadth of data.

The next step is to preprocess the data so that text is converted into simple numeric tokens that can be used as inputs and process those tokens into small sets for the model to interpret. Once this is done, the model can be fine-tuned on the new data. This is done by taking the previously selected pretrained model and carefully training it with a low learning rate on the new data. Once this is done, the model should be ready for use.

Extreme care needs to be taken to derive an optimal train and test split.

Imaging

Imaging research has long been the most high-profile ML task. High-quality benchmark data sets such as the ImageNet challenge have provided robust methods for model assessment with useful pretrained networks for transfer learning [67,68]. The uses of imaging in biomedical applications are myriad. Diagnostics, such as automatic reading and classification of radiological scans or tissue biopsies, are an active area of research. Computer-assisted decision support, where ML algorithms mark anomalous areas for clinicians to investigate, is also relatively well developed. Each use case often requires highly specific knowledge and training data; we leave the specifics to clinical experts.

Convolutional neural networks are a type of deep learning model heavily used in image analyses, such as medical imaging. Convolutional neural networks can extract patterns from image pixels and are thus widely used in abnormality detection, segmentation, and classification [69].

For imaging, as with language, it is strongly encouraged to take advantage of transfer learning. High-quality models are available on many tasks. In addition, vision has demonstrated similar properties as language, and seemingly unconnected tasks often turn out to be very similar, such as categorizing pastries and segmenting tumors. Even if the images in question are very different sizes, it can still be effective to simply resize them to fit the network in question. As in all things, the best strategy is simply to experiment [70].

Genomics

With its high-dimensional nature and the growing availability of large-scale data, genomics has become one of the largest research areas where ML is used [71]. The capacity of traditional statistics is often limited without the support of ML, especially in “multi-omics,” where multiple modes of genomics data, such as DNA-seq, RNA-seq, ChIP-seq, proteomics, and metagenomics, are analyzed together.

ML is used in genomics in multiple ways. For example, ML can be used to predict a certain gene’s expression level given the corresponding DNA-seq information. Genome-wide association studies (GWAS) that aim to identify genetic variants associated with a medical condition of interest, frequently involve ML algorithms such as neural networks and random forests [72]. Zou et al [73] provide more examples of deep learning applications to genomics.

The usefulness of ML as an auxiliary tool should not be underestimated. The overwhelming number of genes is often screened using the variable importance of random forests before downstream analyses. High-dimensional features can be reduced using autoencoders [74,75] to lower-resolution data, which can then be analyzed using traditional statistical analysis tools. Graphical illustrations of the data can also be created through 2D or 3D summaries using algorithms such as tSNE and PCoA [76,77].

Despite its broad capacity in genomics research and ability to handle high-dimensional data, ML has limitations. In genomics data, the number of features outnumbering the sample size, or high dimensionality, is a commonly seen attribute; even with the use of ML, the relatively small sample size can cause reliability and reproducibility issues.

Precision Medicine

As previously mentioned, the goal of the research question of interest may be to indirectly support clinical decision-making. RL is a subset of ML that explicitly provides recommendations for decision-making at sequential time points. PM is a field where such algorithms can be applied to make treatment recommendations for individuals according to their unique characteristics.

PM starts from patient heterogeneity, where reactions to treatment vary from patient to patient [78]. For many illnesses, no panacea exists. PM seeks to recommend different treatment options for unique individuals based on their characteristics; this is formally called individualized treatment rules (ITRs) [79,80]. ITR forms the basis of PM by providing the best treatment recommendations tailored for each patient, as treatment effects can be heterogeneous among individuals. These rules are best identified with rich information about patients such as sociodemographic, clinical, and genomic data. Recently, a wealth of ITR methods have been developed [79-82].

Health care professionals and clinicians are often faced with treating a patient multiple times based on changes in response. For example, researchers can plan adaptive intervention programs for weight loss where later interventions are adjusted depending on responses to the previous treatment [83]. Such dynamic strategies are called dynamic treatment regimes (DTRs) [84,85]. DTRs aim to provide tailored decisions over more than 1 time point based on subject characteristics and their evolving contexts so that a long-term outcome of interest is optimized. The literature on this subtopic is fast-growing [86-88].

For example, patients with cancer may be given frontline chemotherapy followed by a salvage treatment if the response to the initial treatment is not successful [89]. A DTR can then be used to account for potential changes in a patient so that optimal recommendations can be made for each patient for unique stages in their disease to optimize a long-term outcome of interest, such as patient survival. An estimation of such DTR may require a large sample size. To address this issue, investigators can design a multistage randomized trial in an adaptive way through a sequential multiple assignment randomized trial (SMART) [83,90]. For instance, if a patient responded well to the first treatment, increasing the dose may not be particularly effective, and assigning a continued or decreased dose would be worth exploring. ITR and DTR are decision-support tools that provide the best treatment recommendations for patients.

SMARTs are an adaptive study design approach to finding a DTR. While SMARTs are used only for a fixed number of time points, the number of decision time points could be arbitrarily large for some problems. This is formally called an infinite-horizon setting. For example, artificial pancreas programs decide the amount of insulin infused every minute, so that numerous actions are taken even during a day [91]. A class of DTRs that provide essentially continuous recommendations is called just-in-time adaptive intervention [92]. A Markov decision process (MDP) is often used for these problems. MDP is a class of dynamic decision rules that base their decision only on the current state information, not necessarily depending on the history of the change. V-learning [93] is an example of such an infinite-horizon DTR that uses MDP structure.

PM has a strong connection with ML. Treatment effects are often dictated by thousands of patient characteristics, such as sociodemographic, genetic, clinical, and behavioral factors. Genetic factors alone are high-dimensional and can contain millions of traits. The goal of PM is to recommend the best treatment for a patient given their unique characteristics by providing them with an ITR. For example, a clinician may be interested in delineating an optimal ITR for each patient that best achieves cancer remission [94].

Limitations and Optimizations

ML models trained on data that inaccurately represent the population cause fundamental issues such as biased prediction and suboptimality of decisions. Data, where the healthy population is poorly represented, may not be used to make a conclusion for the general population without adjustment. As ML can be used to support decision-making processes, it is also crucial that these decisions arising from the data are not discriminatory toward certain populations [95]. In the ML world, this term is called fairness. When the underlying data are biased, the ML algorithms that are trained on such data may produce biased results, which can lead to inaccurate predictions or withholding of resources. Bias in data may result from measurement bias, representation bias, and sampling bias, among others [96,97]. As an example, unbalanced gender data in the medical imaging field has led to algorithmic underperformance [98]. This discussion should be considered in the data quality assessment when planning data collection.

Advanced Concepts

Transformers

Transformers in the context of NLP are a type of deep learning model architecture that was first introduced by Vaswani in 2017 [63] and have outperformed other model types such as neural networks in both language generation and language interpretation [99]. Transformer models have a unique self-attention mechanism where the model can weigh the importance of different pieces of input data.

Transfer Learning

Transfer learning is the process of taking knowledge from one task and applying it to a different task. This can be useful in several scenarios, such as when training data for models can be difficult to collect or it is computationally expensive to train a new model. In such cases, a pretrained model used in one task can then be trained using data from a different but related task, and the information learned from the previous task can be useful in the new task. This can have reduced data requirements and improved performance as opposed to training a brand-new model, especially when using large, pretrained, publicly available models [100].

Reinforcement Learning

RL is a type of ML focused on training an algorithm to make sequential decisions in potentially changing environments to maximize a cumulative reward [101]. An agent, or decision maker, will receive some quantification of the current environment, also known as the state; the agent will then take an action that will change the state of the environment. The value associated with taking the action and transitioning to the next state is quantified by a “reward”; the agent should choose actions to maximize long-term reward. The goal may be to find the optimal sequence of actions to take to maximize the long-term reward. An application of this is decision support, which has previously been discussed. Other applications of RL range from PM to the development of self-driving cars to financial trading to the creation of ChatGPT. Those interested in the applications of RL in health care should see reference [102] for examples of use and [103] for guidelines of use.

Outlook

The field of ML has a vast trove of tools and resources for use. Its potential, though impressive and exciting, can also be a drawback, as the inherent flexibility in the analysis process gives room for the researcher to arbitrarily or questionably choose methods that result in overfitting and false discovery. This paper has provided a framework for steps involved in using ML in research, discusses analyses for specialized data formats, reviews decision support and bias in ML, and introduces PM, a popular field in the health research domain. Consulting ML experts throughout the process will not only streamline the analysis but also play a large role in legitimizing justifications for choice selections.

Furthermore, the paper has highlighted the importance of preplanned documentation to ensure transparency and foster credibility within the scientific community. The incorporation of concrete examples within the health care domain, in addition to the provision of techniques involved in specialized data types, illustrates the vast applications of ML methodologies and their potential impact in the health science field.

Overall, the specifics of different data types and the wide variety of research goals make it difficult to make more specific recommendation guidelines. However, major ML considerations and how to approach them are discussed, with specific examples. While this paper provides a general recommended research framework and major considerations for the use of ML, it is not comprehensive, as the aim was to provide a general overview of potential methods and considerations.

While ML has strong performance potential in a variety of situations, its use needs to be carefully planned through the aforementioned steps and justified to obtain the best results, as ML cannot overcome poor study design or data quality despite all its virtues. By acknowledging these limitations, the research community can better strive for high-quality data and reproducible results to continue driving innovation in society.

Acknowledgments

JS, DDM, HE, and MRK were supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health (grant UL1TR002489).

ARK is supported by the National Center for Advancing Translational Sciences, National Institutes of Health (grant KL2TR002490). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. ARK also reports receiving research grants from the Diabetes Research Connection and the American Diabetes Association and a prize from the National Academy of Medicine outside the submitted work.

AVV is supported by the National Institutes of Health T32 grant for the UNC Kidney Center.

ELB is supported by the National Institutes of Health (grant K23DK127157).

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

HC, JS, DDM, HE, and MRK were responsible for paper concept and design. HC, JS, DDM, and HE contributed to the drafting of the manuscript. AVV and MRK are joint senior authors. All authors were involved in the critical revision of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary information on general sample size calculations, model training, and model tuning and validation.
[DOCX File , 34 KB-Multimedia Appendix 1]

References

1. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
2. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-1219. [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
3. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
4. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
5. Greely HT. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annu Rev Genomics Hum Genet. 2007;8:343-364. [doi: 10.1146/annurev.genom.7.080505.115721] [Medline: 17550341]
6. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014;42(Database issue):D975-d979. [FREE Full text] [doi: 10.1093/nar/gkt1211] [Medline: 24297256]
7. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149-153. [FREE Full text] [doi: 10.1093/cid/cix731] [Medline: 29020316]
8. Kosorok MR, Laber EB. Precision medicine. Annu Rev Stat Appl. 2019;6(1):263-286. [FREE Full text] [doi: 10.1146/annurev-statistics-030718-105251]
9. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv. Preprint posted online on November 19, 2015. [doi: 10.48550/arXiv.1511.06348]
10. Incorporating nesterov momentum into adam. International Conference of Learning Representations. URL: https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ [accessed 2024-01-16]
11. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv. Preprint posted online on June 12, 2017. [doi: 10.48550/arXiv.1706.03825]
12. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. 2019. Presented at: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining; August 4-8, 2019; Anchorage, AK. [doi: 10.1145/3292500.3330701]
13. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):60. [doi: 10.1186/s40537-019-0197-0]

14. Xiao D, Huang Y, Qin C, Liu Z, Li Y, Liu C. Transfer learning with convolutional neural networks for small sample size problem in machinery fault diagnosis. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 2019;233(14):5131-5143. [doi: 10.1177/0954406219840381]
15. Tsinganos P, Cornelis B, Cornelis J, Jansen B, Skodras A. Data augmentation of surface electromyography for hand gesture recognition. Sensors (Basel). 2020;20(17). [FREE Full text] [doi: 10.3390/s20174892] [Medline: 32872508]
16. Stone CJ. Optimal rates of convergence for nonparametric estimators. Ann Stat. 1980:1348-1360. [FREE Full text]
17. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. [FREE Full text] [doi: 10.1186/1471-2105-7-91] [Medline: 16504092]
18. Qi Y. Random forest for bioinformatics. Ensemble Mach Learn. 2012:307-323. [FREE Full text]
19. Zeiler M. ADADELTA: An adaptive learning rate method. arXiv. [FREE Full text]
20. Bergstra DY, James, Cox D. HYPEROPT: A python library for optimizing the hyperparameters of machine learning algorithms. 2013. Presented at: Proceedings of the 12th Python in Science Conference; July 29-August 2, 2013; Austin, TX. [doi: 10.25080/Majora-8b375195-003]
21. Kingma D, Ba J. Adam: a method for stochastic optimization. arXiv. [FREE Full text]
22. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 2019;188(12):2222-2239. [FREE Full text] [doi: 10.1093/aje/kwz189] [Medline: 31509183]
23. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40-55. [doi: 10.1038/s41580-021-00407-0] [Medline: 34518686]
24. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. N Engl J Med. 2017;376(26):2507-2509. [FREE Full text] [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
25. Stiglic N, Viner RM. Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews. BMJ Open. 2019;9(1):e023191. [FREE Full text] [doi: 10.1136/bmjopen-2018-023191] [Medline: 30606703]
26. Fletcher RR, Chamberlain D, Richman D, Oreskovic N, Taveras E. Wearable sensor and algorithm for automated measurement of screen time. 2016. Presented at: 2016 IEEE Wireless Health (WH); October 25-27, 2016:1-8; Bethesda, MD. [doi: 10.1109/WH.2016.7764564]
27. Debelee TG, Schwenker F, Ibenthal A, Yohannes D. Survey of deep learning in breast cancer image analysis. Evol Syst. 2019;11(1):143-163. [doi: 10.1007/s12530-019-09297-2]
28. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell. 2019;267:1-38. [FREE Full text] [doi: 10.1016/j.artint.2018.07.007]
29. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. [FREE Full text] [doi: 10.48550/arXiv.1702.08608]
30. Levin KA. Study design v. case-control studies. Evid Based Dent. 2006;7(3):83-84. [FREE Full text] [doi: 10.1038/sj.ebd.6400436] [Medline: 17003803]
31. Rockhill B, Willett WC, Manson JE, Leitzmann MF, Stampfer MJ, Hunter DJ, et al. Physical activity and mortality: a prospective study among women. Am J Public Health. 2001;91(4):578-583. [FREE Full text] [doi: 10.2105/ajph.91.4.578] [Medline: 11291369]
32. Cowie CC, Eberhardt MS. Sociodemographic characteristics of persons with diabetes. In: Diabetes in America. 2nd Edition. Bethesda, MD. National Institute of Diabetes and Digestive and Kidney Diseases; 1995:85-116.
33. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113(523):1228-1242. [doi: 10.1080/01621459.2017.1319839]
34. Jiang B, Zhang X, Cai T. Estimating the confidence interval for prediction errors of support vector machine classifiers. J Mach Learn Res. 2008;9:521-540. [FREE Full text]
35. Arbach L, Reinhardt J, Bennett D, Fallouh G. Mammographic masses classification: comparison between Backpropagation Neural Network (BNN), K Nearest Neighbors (KNN), and human readers. 2003. Presented at: CCECE 2003—Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436); May 4-7, 2003:1441-1444; Montreal, QC. [doi: 10.1109/ccece.2003.1226174]
36. Hernan MA, Robins JM. Causal inference. GRASS (Grup de Recerca en Anàlisi eStadística de la Supervivència). 2010. URL: https://grass.upc.edu/en/seminar/presentation-files/causal-inference/chapters-1-i-2/@@download/file/BookHernanRobinsCap1_2.pdf [accessed 2024-01-10]
37. Imbens GW, Rubin DB. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York, NY. Cambridge University Press; 2015.
38. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A. 2019;116(10):4156-4165. [FREE Full text] [doi: 10.1073/pnas.1804597116] [Medline: 30770453]
39. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010;29(3):337-346. [FREE Full text] [doi: 10.1002/sim.3782] [Medline: 19960510]
40. Zhao P, Su X, Ge T, Fan J. Propensity score and proximity matching using random forest. Contemp Clin Trials. 2016;47:85-92. [FREE Full text] [doi: 10.1016/j.cct.2015.12.012] [Medline: 26706666]

41. Li Y, Ngom A. Data integration in machine learning. 2015. Presented at: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 9-12, 2015:1665-1671; Washington, DC. [doi: 10.1109/bibm.2015.7359925]
42. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform. 2018;19(2):325-340. [FREE Full text] [doi: 10.1093/bib/bbw113] [Medline: 28011753]
43. Zhong B. How to calculate sample size in randomized controlled trial? J Thorac Dis. 2009;1(1):51-54. [FREE Full text] [Medline: 22263004]
44. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003;20(5):453-458. [FREE Full text] [doi: 10.1136/emj.20.5.453] [Medline: 12954688]
45. Vapnik VN. The Nature of Statistical Learning Theory. New York, NY. Springer; 1999.
46. Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: a learning-curve analysis. J Mach Learn Res. 2003;4:211-255. [FREE Full text]
47. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Anal Chim Acta. 2013;760:25-33. [doi: 10.1016/j.aca.2012.11.007] [Medline: 23265730]
48. Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Riche NH, et al. Research directions in data wrangling: visualizations and transformations for usable and credible data. Inf Vis. 2011;10(4):271-288. [doi: 10.1177/1473871611415994]
49. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010;50(2):105-115. [doi: 10.1016/j.artmed.2010.05.002] [Medline: 20638252]
50. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241. [FREE Full text] [doi: 10.1186/s13059-015-0805-z] [Medline: 26527291]
51. Guo CY, Yang YC, Chen YH. The optimal machine learning-based missing data imputation for the cox proportional hazard model. Front Public Health. 2021;9:680054. [FREE Full text] [doi: 10.3389/fpubh.2021.680054] [Medline: 34291028]
52. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-118. [FREE Full text] [doi: 10.1093/bioinformatics/btr597] [Medline: 22039212]
53. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. 2018. Presented at: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); October 01-03, 2018:80-89; Turin, Italy. [doi: 10.1109/dsaa.2018.00018]
54. Xie P, Li Y, Deng B, Du C, Rui S, Deng W, et al. An explainable machine learning model for predicting in-hospital amputation rate of patients with diabetic foot ulcer. Int Wound J. 2022;19(4):910-918. [FREE Full text] [doi: 10.1111/iwj.13691] [Medline: 34520110]
55. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv. [FREE Full text]
56. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. Gigascience. 2017;6(5):1-9. [FREE Full text] [doi: 10.1093/gigascience/gix019] [Medline: 28327985]
57. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323. [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
58. Peat JK, Mellis C, Williams K, Xuan W. Health Science Research: A Handbook of Quantitative Methods. London, UK. Routledge; 2020.
59. Woodward M. Epidemiology: Study Design and Data Analysis, 3rd Edition. Boca Raton, FL. CRC press; 2013.
60. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16(1):29. [FREE Full text] [doi: 10.1186/s12916-018-1017-7] [Medline: 29490655]
61. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing. 2020;415:295-316. [doi: 10.1016/j.neucom.2020.07.061]
62. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. [FREE Full text]
63. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017). New York, NY. Curran Associates; 2017.
64. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Papers With Code. 2018. URL: https://paperswithcode.com/paper/improving-language-understanding-by [accessed 2024-01-19]
65. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):86. [FREE Full text] [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
66. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. Association for Computational Linguistics; 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3–7, 2019:3615-3620; Hong Kong, China. URL: https://aclanthology.org/D19-1371.pdf [doi: 10.18653/v1/d19-1371]

67. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;6(6):84-90. [doi: 10.1145/3065386]
68. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009. Presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009:248-255; Miami, FL. [doi: 10.1109/cvpr.2009.5206848]
69. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst. 2018;42(11):226. [doi: 10.1007/s10916-018-1088-1] [Medline: 30298337]
70. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016;3:9. [FREE Full text] [doi: 10.1186/s40537-016-0043-6]
71. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321-332. [FREE Full text] [doi: 10.1038/nrg3920] [Medline: 25948244]
72. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front Genet. 2020;11:350. [FREE Full text] [doi: 10.3389/fgene.2020.00350] [Medline: 32351543]
73. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12-18. [doi: 10.1038/s41588-018-0295-5] [Medline: 30478442]
74. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389-403. [doi: 10.1038/s41576-019-0122-6] [Medline: 30971806]
75. Tan J, Ung M, Cheng C, Greene CS. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pac Symp Biocomput. 2015;20:132-143. [FREE Full text] [Medline: 25592575]
76. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579-2605. [FREE Full text]
77. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17(4):628-641. [FREE Full text] [doi: 10.1093/bib/bbv108] [Medline: 26969681]
78. Grutters JPC, Sculpher M, Briggs AH, Severens JL, Candel MJ, Stahl JE, et al. Acknowledging patient heterogeneity in economic evaluation: a systematic literature review. Pharmacoeconomics. 2013;31(2):111-123. [doi: 10.1007/s40273-012-0015-4] [Medline: 23329430]
79. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc. 2012;107(449):1106-1118. [FREE Full text] [doi: 10.1080/01621459.2012.695674] [Medline: 23630406]
80. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012;68(4):1010-1018. [FREE Full text] [doi: 10.1111/j.1541-0420.2012.01763.x] [Medline: 22550953]
81. Zhang Y, Laber EB, Tsiatis A, Davidian M. Using decision lists to construct interpretable and parsimonious treatment regimes. Biometrics. 2015;71(4):895-904. [FREE Full text] [doi: 10.1111/biom.12354] [Medline: 26193819]
82. Athey S, Wager S. Efficient policy learning. Stanford Institute for Economic Policy Research (SIEPR). 2017. URL: https://siepr.stanford.edu/publications/working-paper/efficient-policy-learning [accessed 2024-01-10]
83. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. Transl Behav Med. 2014;4(3):260-274. [FREE Full text] [doi: 10.1007/s13142-014-0265-0] [Medline: 25264466]
84. Kosorok MR, Moodie EE. Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine. Philadelphia, PA. Society for Industrial and Applied Mathematics; 2015.
85. Chakraborty B, Murphy SA. Dynamic treatment regimes. Annu Rev Stat Appl. 2014;1:447-464. [FREE Full text] [doi: 10.1146/annurev-statistics-022513-115553] [Medline: 25401119]
86. Zhao YQ, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. J Am Stat Assoc. 2015;110(510):583-598. [FREE Full text] [doi: 10.1080/01621459.2014.937488] [Medline: 26236062]
87. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika. 2013;100(3):681-684. [FREE Full text] [doi: 10.1093/biomet/ast014] [Medline: 24302771]
88. Zhang Y, Laber EB, Davidian M, Tsiatis AA. Interpretable dynamic treatment regimes. J Am Stat Assoc. 2018;113(524):1541-1549. [FREE Full text] [doi: 10.1080/01621459.2017.1345743] [Medline: 30774169]
89. Song Z, Yu X, Lou G, Shi X, Zhang Y. Salvage treatment with apatinib for advanced non-small-cell lung cancer. Onco Targets Ther. 2017;10:1821-1825. [FREE Full text] [doi: 10.2147/OTT.S113435] [Medline: 28367065]
90. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA. A "SMART" design for building individualized treatment sequences. Annu Rev Clin Psychol. 2012;8:21-48. [FREE Full text] [doi: 10.1146/annurev-clinpsy-032511-143152] [Medline: 22224838]
91. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: a systematic review. Artif Intell Med. 2020;104:101836. [doi: 10.1016/j.artmed.2020.101836] [Medline: 32499004]
92. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med. 2018;52(6):446-462. [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]

93. Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR. Estimating dynamic treatment regimes in mobile health using V-learning. J Am Stat Assoc. 2020;115(530):692-706. [FREE Full text] [doi: 10.1080/01621459.2018.1537919] [Medline: 32952236]
94. Friedman AA, Letai A, Fisher DE, Flaherty KT. Precision medicine for cancer with next-generation functional diagnostics. Nat Rev Cancer. 2015;15(12):747-756. [FREE Full text] [doi: 10.1038/nrc4015] [Medline: 26536825]
95. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. Nature. 2018;559(7714):324-326. [doi: 10.1038/d41586-018-05707-8] [Medline: 30018439]
96. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):1-35. [doi: 10.1145/3457607]
97. Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv. [FREE Full text] [doi: 10.48550/arXiv.1808.00023]
98. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020;117(23):12592-12594. [FREE Full text] [doi: 10.1073/pnas.1919012117] [Medline: 32457147]
99. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. Association for Computational Linguistics; 2020. Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020:38-45; Virtual. URL: https://aclanthology.org/2020.emnlp-demos.6/ [doi: 10.18653/v1/2020.emnlp-demos.6]
100. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. Proc IEEE. 2021;109(1):43-76. [doi: 10.1109/jproc.2020.3004555]
101. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res. 1996;4:237-285. [FREE Full text] [doi: 10.1613/jair.301]
102. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv. 2021;55(1):1-36. [doi: 10.1145/3477600]
103. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25(1):16-18. [FREE Full text] [doi: 10.1038/s41591-018-0310-5] [Medline: 30617332]

Abbreviations

AI: artificial intelligence
ATE: average treatment effects
AUC: area under the receiver operating characteristics curve
DTR: dynamic treatment regime
EHR: electronic health record
GWAS: genome-wide association studies
ITE: individual treatment effects
ITR: individualized treatment rule
k-NN: k-nearest neighbors
LDA: linear discriminant analysis
MDP: Markov decision process
ML: machine learning
NLP: natural language processing
PCA: principal component analysis
PM: precision medicine
RL: reinforcement learning
SMART: sequential multiple assignment randomized trial
SVM: support vector machine
TNR: true negative rate

Edited by A Mavragani; submitted 15.07.23; peer-reviewed by C Sáez, Z Ren, M Aczon; comments to author 11.10.23; revised version received 30.11.23; accepted 21.12.23; published 30.01.24

Please cite as:
Cho H, She J, De Marchi D, El-Zaatari H, Barnes EL, Kahkoska AR, Kosorok MR, Virkud AV
Machine Learning and Health Science Research: Tutorial
J Med Internet Res 2024;26:e50890
URL: https://www.jmir.org/2024/1/e50890
doi: 10.2196/50890
PMID: 38289657