ANNUAL REVIEWS

Annual Review of Public Health

Machine Learning in Epidemiology and Health Outcomes Research

Timothy L. Wiemken[1] and Robert R. Kelley[2]

[1]Center for Health Outcomes Research, Saint Louis University, Saint Louis, Missouri 63104, USA; email: timothy.wiemken@health.slu.edu

[2]Department of Computer Science, Bellarmine University, Louisville, Kentucky 40205, USA; email: rkelley@bellarmine.edu

Annu. Rev. Public Health 2020. 41:21–36

First published as a Review in Advance on October 2, 2019

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-040119-094437

Keywords

predictive modeling, artificial intelligence, deep learning, treatment effects, walkthrough, biostatistics

Abstract

Machine learning approaches to modeling of epidemiologic data are becoming increasingly more prevalent in the literature. These methods have the potential to improve our understanding of health and opportunities for intervention, far beyond our past capabilities. This article provides a walkthrough for creating supervised machine learning models with current examples from the literature. From identifying an appropriate sample and selecting features through training, testing, and assessing performance, the end-to-end approach to machine learning can be a daunting task. We take the reader through each step in the process and discuss novel concepts in the area of machine learning, including identifying treatment effects and explaining the output from machine learning models.

INTRODUCTION TO EPIDEMIOLOGY

Epidemiology is defined as the study of the distribution and determinants of disease (25). Improvements in population health and increased survival rates in humans can be traced to interventions developed from evidence obtained through epidemiologic study (41). Data analytics are one of the most critical underlying aspects of epidemiology; increasing computational power over the past decade has vastly expanded our modeling capabilities and approaches (23). Due to the variety of areas of study in epidemiology and the unique needs of each, novel computational modeling strategies are highly prevalent in the scientific literature.

Testing: the process of passing an independent data set (typically the remaining data not used in the training of the model) to the trained model, producing various performance metrics

Machine learning: an umbrella term encompassing a multitude of algorithms used for prediction and estimation of treatment effects

Supervised machine learning: a machine learning approach focused on predicting an outcome of interest

HISTORICAL RATIONALE FOR STATISTICAL MODELING

Frequentist statistical methodologies are the most commonly used approaches to analytics in epidemiologic studies to date (27). These methods are often confusing to nonstatisticians and are steeped in the development of hypotheses and the calculation of probabilities that offer support for or against rejection of these hypotheses. Basic statistical tests and multivariable regression modeling are commonly used for testing hypotheses to define associations or treatment effects between predictor and outcome variables under study. These traditional statistical approaches are used in what is coined as the "data culture" (12).

Traditional regression-type modeling of health outcomes in epidemiology can be categorized by the purpose of the model, whether it is necessary to predict a dependent variable given multiple independent variables (e.g., predictive models) or to produce a measure of treatment effect or magnitude and statistical association of individual independent variables on the dependent variable (e.g., explanatory models) (57). Each modeling strategy provides useful information for investigators and practitioners. Most traditional modeling approaches are data focused and make various assumptions about the data used within the model (12). Assumptions such as linearity, lack of multicollinearity, and proportional risk/odds/hazards over time are well understood by epidemiologists. As more data become available for analytics, Richard Bellman's "curse of dimensionality" becomes apparent (7). In this state, research questions become more advanced, traditional modeling assumptions become more difficult to meet, relationships are highly nonlinear, and new methods must be utilized. Novel approaches in machine learning have become a focus in medicine, with more limited use in population health (22) over the past several years. The purpose of this review is to document the uses, strategies, and approaches, as well as the advantages and disadvantages of machine learning models in the field of epidemiology and health outcomes research, with a main focus on supervised machine learning methods.

INTRODUCTION TO MACHINE LEARNING

"Machine learning" is an umbrella term used to describe a wide variety of models and strategies that focus on algorithmic modeling (45). In contrast, the term regression in epidemiology typically refers to a wide variety of frequentist regression models such as logistic, linear, and Cox proportional hazards often used in epidemiology and biostatistics (28).

The concept of machine learning has existed from the early 1950s to address the possibility of having computers approximate the human thought process through pattern matching, recognition, and decision making (64). This work continued through the research of Arthur Samuel, who wrote a program to learn to play the board game checkers (53), and that of Frank Rosenblatt (50), who designed the first artificial neural network, which used the principles of neural biology to perform computation. Since that time, numerous machine learning algorithms have been developed

Table 1 Linking terms and phrases in epidemiology and machine learning

Term in epidemiology and biostatistics	Term in machine/statistical learning
Dependent variable; outcome variable; response variable	Label/class
Independent variable; predictor variable; explanatory variable	Feature
Contingency table; 2 × 2 table	Confusion matrix
Sensitivity	Recall
Positive predictive value	Precision
Deep learning	Artificial neural network with more than 1 hidden layer
Outcome group with the highest frequency	Majority class
Outcome group with the lowest frequency	Minority class
Proportion of cases in each category of the outcome variable (when outcome is categorical)	Class balance

to solve many learning problems. These algorithms are generally grouped into supervised or unsupervised models. Supervised models are typically used to predict an outcome (known as a label in machine learning), similar to predictive modeling using regression. Unsupervised models are typically used to discover unknown patterns in data, without respect to a particular label. In this review, we focus primarily on supervised models in epidemiology. Although these techniques have been around for many years, machine learning was not accepted as a Medline Medical Subject Heading (MeSH) term until 2016 (**https://www.ncbi.nlm.nih.gov/mesh/2010029**).

Although the term machine learning is often used in today's environment, "statistical learning" is also commonly used in the literature. This variation in terminology is due to several novel strategies that combine traditional frequentist biostatistical approaches, such as hypothesis testing, with algorithmic approaches typical of machine learning models (7). This practice further blurs the lines between traditional biostatistics and machine learning, resulting in the combined phrasing: statistical learning. Regardless, the machine learning literature utilizes different terms for similar concepts used in epidemiology. For the purposes of this review, we use machine learning terminology. For clarity, **Table 1** displays common terms used in epidemiology with their corollaries in machine learning. Most notably, the term features refers to what epidemiologists would consider independent variables, whereas the term label refers to the dependent variable.

Unsupervised (machine learning): a machine learning approach with no outcome, used for clustering and data reduction

Label: the outcome variable in a supervised machine learning model

Feature: the variable(s) used to assist the model in predicting an outcome or those used in a cluster or data reduction algorithm

THE MACHINE LEARNING APPROACH

Setting up a machine learning model such that the predictions are valid and accurate can be a daunting task, not substantially different from developing a biostatistical regression model. Below, we describe the end-to-end process of machine learning (**Figure 1**), with examples in epidemiology and public health.

Sample Size

The literature includes numerous approaches for identifying appropriate sample sizes for machine learning models (6, 21, 47). However, sample size estimates are difficult to compute because machine learning models are largely algorithmic based. Most do not utilize frequentist statistical measures such as *p*-values, nor do they focus on effect sizes, two concepts central to the traditional calculation of sample sizes.

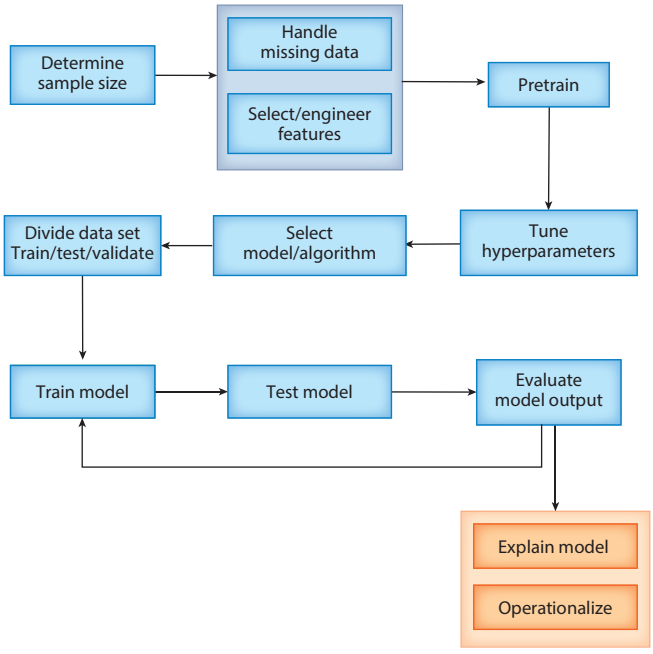

Figure 1

The end-to-end supervised machine learning workflow.

For unsupervised models, sample sizes should be based on the research question and inherent variability of the data under consideration. Because unsupervised models can be used for hypothesis generation or data reduction, sample size calculations may be unnecessary. For example, very small data sets (13 unique individuals) have been used to detect clusters of patients with similar inflammatory markers among hospitalized patients with pneumonia (70).

In short, many loose recommendations and heuristics are available for machine learning sample size estimation. Publications in machine learning sample size estimations are often specific to a discipline. For example, genetic epidemiology may require smaller data sets (e.g., <100 rows) (12) as compared with a moderate size (several hundred cases) necessary for a behavioral/cognition outcome evaluating functional magnetic resonance imaging (fMRI) (11). However, it is important to note that the optimal sample size is dependent on the data available and is based on the number of rows and the number and quality of features. If features included are redundant or not predictive of the label, the model may be inaccurate regardless of the volume of features. Furthermore, if there are many features but few instances of the label, then models may have difficulty matching patterns to the label for the full feature space and the model will be unlikely to function appropriately in production (29). Much like any modeling scheme, results can be generated regardless of the sample size. In machine learning, for the results to be accurate and generalizable, the overall sample size needed should be carefully considered a priori and may be much larger than anticipated (65).

Feature Selection

Parsimony is a central tenet of regression model building in epidemiology to prevent overfitting. Selecting the relevant predictors with the appropriate level of explanation is critical to the model's

success. Like regression modeling, overfitting is a major concern in machine learning modeling; parsimony is accomplished through feature selection in which the features in data sets are thoughtfully chosen for the model. This approach is especially important for machine learning models because they are applied to data sets that were collected for reasons other than a specific hypothesis, as is the case with electronic medical record data (32) or genetic epidemiology data (40). These types of data sets typically contain a large number of features compared to regression models; many of them are irrelevant to the model being constructed. One example outlines a technique to detect and limit the set of variables to be used for modeling (known as the feature set) in antigen discovery for vaccinology (18), which could be applied to any 'omics data set.

The difficulty in the selection of a parsimonious feature set is more complex than just the feature set's impact on a particular study outcome (9). There are many other reasons to reduce the number of features in one's training data set prior to using a machine learning model. First, the model will train faster, which is particularly attractive with complex modeling schemes under local computation as opposed to cluster computing. Second, reducing the number of redundant features or features that do not affect the outcome may decrease the likelihood of overfitting the model.

Feature selection can be done in numerous ways including selecting clinically meaningful features, simple correlations between features, and feature importance scores. Other machine learning models such as least absolute shrinkage and selection operator (LASSO) regression may also be useful for feature selection (62, 69). Genetic algorithms for feature selection have become popular and have been used for various purposes (63), including to understand the impact of uncontrolled comorbidities on clinical outcomes in hospitalized patients with pneumonia (3). Regardless of the method used for feature selection, investigators have suggested that the accuracy and stability of the model should be considered when using feature selection algorithms (21). Otherwise, these models risk overfitting.

In the era of 'omics data, the feature set provided to epidemiologists for analytics has expanded substantially (22). Many approaches for selecting features have been proposed (24, 25). One example is ranked guided iterative feature elimination (RGIFE), which shows promise for identifying enhanced clinically relevant biomarkers (26) (see the sidebar titled Ranked Guided Iterative Feature Elimination for Feature Selection). Regardless, much like in explanatory regression model building, strict automation of feature selection is likely not an appropriate solution on its own. In nearly all areas, domain experts should be enlisted to assist in feature selection for meaningful models to be developed.

Feature selection: the process of selecting a subset of features to include in a machine learning model

Training: the supervised machine learning process of creating a predictive model through using a subset of data, often around 70%

RANKED GUIDED ITERATIVE FEATURE ELIMINATION FOR FEATURE SELECTION

Issue: Data sets with a large number of features are difficult to use because relevant features are difficult to identify.

Solution: Ranked guided iterative feature elimination (RGIFE), an algorithm that uses cross-validation to identify relevant features in classification scenarios, is proposed. RGIFE first estimates the performance of a model with the original feature set with *k*-fold cross-validation. The model then ranks the importance of the features to the classification task. From there the model removes attributes from the end of the feature rank (lowest ranking features) and runs the model again. Reduced feature sets that perform within a tolerable level are accepted until the performance of the model fails below a specified threshold. The performance of this feature selection model was compared with several commonly used feature selection methods.

Conclusion: RGIFE provided similar prediction performance with few features for several cancer-related data sets.

Feature Engineering

Feature engineering: the process of creating new features from existing features using mathematical and various combinatorial approaches

Performance: the various statistical values gathered from machine learning models, used to assess how well the model achieves its intended purpose

Creating or engineering new features from available data to capture latent effects is another important facet of machine learning. Historically, feature engineering has been a manual and laborious process, limited by many factors including the mathematical expertise, time available for analysis, and domain knowledge of the study team.

Simple feature engineering, such as taking the logarithm of a continuous variable to change its distribution or aggregating two variables to account for multicollinearity, is sometimes necessary in traditional regression modeling. For example, the performance accuracy of machine learning models to predict early sepsis was improved by multiplying a shock index by age to derive a new feature for regression (19).

Feature engineering is becoming more complex, with the potential to uncover latent effects that would not be accounted for otherwise. One novel automated feature engineering approach is deep feature synthesis, which combines multiple feature transformations and aggregations, of any type or complexity, to create new features (33). Individual variables are precursors to these new deep features, created through primitives, mathematical formulas used to transform or aggregate. Primitives can range from simple to complex, including means, sums, principal components, or even predicted probabilities or error terms from traditional regression models. Aggregations are very useful for longitudinal features, whereas transformations are typically used for time-invariant features.

In our experience, approaches to feature engineering often merge with feature selection, especially for longitudinal data sets. For instance, in clinical epidemiology, when selecting features, investigators often must limit which laboratory values to include in a model because the number of time points and frequency at which laboratory data are collected during a hospital stay vary.

External data sources are becoming important components of accurate machine learning models. Affixing data collected outside the primary data set provides machine learning algorithms additional features from which to learn. In fact, researchers have suggested that different machine learning algorithms are unlikely to provide a substantial improvement in model performance if the same feature set is used for each (19, 25).

External data sets can be combined with primary data sets, including, among others, geographic location, weather data, and aggregated population statistics. For example, area deprivation indices have been used to predict health care outcomes. This deprivation score is an aggregate score developed from US Census data (58), which can be linked to individual-level data to provide some estimates of cluster effects. It has been successfully used to assist in the prediction of hospital readmission (35) and outcomes in hospitalized patients with community-acquired pneumonia (68). Investigators have used machine learning to aggregate Web search and location data, linked with restaurant data, to identify potentially unsafe restaurants (51). Aggregate data such as these have the capability to revolutionize the performance of model predictions in epidemiology.

Missing Data

Missing data, regardless of the mechanism creating the missingness, is an issue across all analytics. Many traditional regression models will drop cases with missing data and run the model. The majority of machine learning models will not run with missing data; therefore, care to ensure data are complete is critical. One solution to this problem is data imputation, a technique to generate reasonable synthetic values at random when data are missing completely. These approaches have the opportunity to reduce error in missingness by accounting for nonlinear relationships in the imputator (42). Examples of machine learning missing data imputers are ripe in the literature,

largely basing the models on random forest approaches (11, 55, 60). In epidemiology, variations on this theme have been used to impute missing data in some studies to better define the role of age-mixing patterns in HIV transmission dynamics (5), defining burnout and stress relationships among health care workers (13), health care utilization in patients with spinal cord injuries (49), and treatment completion prediction in patients with rape-onset post-traumatic stress disorder (34).

Hyperparameter: option(s) required in many machine learning algorithms to fine-tune or optimize the training of the model

Classification or Regression?

Another decision to make in machine learning model building is to determine the type of outcome the investigator is interested in predicting. In machine learning, classification models are considered in the context of categorical labels, whereas regression models are used for continuous labels; each model has different ramifications for model building. Nearly all supervised machine learning models can handle both classification and regression problems. In-depth review of common methods used in health research is beyond the scope of this review but can be found elsewhere (71).

Pretraining/Hyperparameter Optimization

All supervised machine learning algorithms have various hyperparameters that should be adjusted in order to provide a valid and accurate prediction. Some examples include the learning rate in neural networks, C and sigma in a support vector machine, or *k* in the *k*-nearest neighbor algorithm. The process of adjusting these hyperparameters is called tuning. Although most machine learning models have default values for each hyperparameter, it is worth the effort to optimize these parameters. To tune hyperparameters, a subset of the data is needed. There are many heuristics to determine how much data should be used for tuning these parameters, but there is no consensus. We recommend that ~50% of the available data be randomly selected for hyperparameter tuning through cross-validation. This is only a generic heuristic and should be modified on the basis of the variation present in the features and outcome.

The rationale for utilizing a large portion of the data for hyperparameter tuning is that the optimal parameters cannot be known before running a model. An invalid model may result if investigators do not provide appropriate values (59). Several approaches to tuning have been described (8, 31). Grid search approaches are also easy to implement and allow for prespecification of a multitude of possible values for many or all the necessary hyperparameters required by the model. The limitation of grid searching for hyperparameters is the computationally intensive computation required. Since the investigator specifies a set of values for several hyperparameters in a grid search, models must be built for all combinations of values. Model tuning is critical and continues to be discussed as a salient concept in epidemiology. Tessmer and colleagues (61) showcase this with respect to improving R_0 calculations in infectious disease epidemiology and dynamics.

Another consideration is specifically for classification models. In this context, the pretraining data set may need to be balanced with respect to the class label frequency. Here, the class label with the lowest frequency of cases is termed the minority class, whereas the higher frequency is termed the majority class. For many classification algorithms, having a relatively balanced outcome is critical (what this means is debated, though as close to 50/50 as possible is ideal). Imbalance in the outcome of a model is an issue when evaluating model performance statistics. If one class label has a much higher prevalence than another, predictive accuracy may look good while the model is predicting only the majority class. In this context, downsampling, upsampling, and synthetic minority oversampling technique (SMOTE) (14) sampling of the data are common

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE FOR HANDLING CLASS IMBALANCE

Issue: Class imbalance in the outcome for clinical and epidemiological data sets prevents machine learning algorithms from learning accurately.

Solution: Synthetic minority oversampling technique (SMOTE) is a technique in which the minority class (e.g., the group with the lowest frequency) in a classification problem is oversampled by creating synthetic samples that are similar to actual samples. With this approach, the machine learning algorithm has more examples of the minority class from which to learn. The algorithm can further be combined with undersampling of the majority class to create more balanced outcomes in the data set.

Conclusion: The combination of SMOTE and undersampling performs better than undersampling alone because it focuses learning on the minority class.

approaches (see the sidebar titled Synthetic Minority Oversampling Technique for Handling Class Imbalance). Class balancing should be done only after splitting the data and should be independent of both the training and the testing data sets. Balancing methods have been utilized in many areas of epidemiology, including in cancer survivorship prediction (24, 67), groundwater contamination (38), and mesothelioma patients (17). Alghamdi and colleagues (1) used data from the Henry Ford Health system to predict incident diabetes with cardiorespiratory fitness data using SMOTE to balance the outcome.

Furthermore, continuous data typically should be normalized prior to training to standardize the scale of multiple continuous features and improve computational performance. It is important to normalize the data after splitting the data sets into pretraining, training, and testing. It is not advised to normalize and then split, as data leakage may occur, resulting in aberrant model performance statistics. The goal of the test set is to make it as independent of the other data as possible. If the cases ending up in the test set have features that have been scaled in consideration of some of the data in the training set, leakage will become an issue. Standardization and normalization of continuous data are necessary for many machine learning models. Seligman and colleagues (54) used data standardization approaches to understand social determinants of health in the Health and Retirement Study.

Training

After identification of optimal hyperparameters, the next step is to split the remaining data into training and testing data sets. When defining the proportion of cases to use for a training data set, researchers face many considerations, but there is no proportion that should be deemed always acceptable. When selecting the proportion of cases for a training set, major considerations include (*a*) number of cases, (*b*) number of features, and (*c*) amount of variation in the features. The importance lies with how well one's training data set describes all the possible patterns of data and their potential prediction of the label. In the literature reviewed, 80% of the cases are most often used for training, although this is simply a heuristic and is not evidence based.

Similar to the hyperparameter tuning set described above, training data must be balanced with respect to the outcome for many models in the context of classification. As above, they should be balanced after splitting and be independent of all data in pretraining and testing data sets. Again as described above for hyperparameter tuning, normalization or standardization of continuous variables should be conducted after splitting.

Testing

Once again, testing the performance of the tuned and trained machine learning model in a separate data set (the test set) necessitates a proportion of the total data. The goal is to have a useful representation of real life in the testing data set. One must be careful to ensure that there is no spillover from training data sets. Here, no balancing of the outcome minority class should occur; however, if features are normalized for training, they should also be normalized in the testing set. Here, the normalization factors should be applied from the training set. For example, if a column in the training set is normalized such that the individual value is subtracted from the column mean and divided by the column standard deviation in the training set (a common method of normalization), the mean and standard deviation from the training set should be applied to the values in the testing data set. The rationale for this approach arises when a model is in production. In this scenario, a single row of data (i.e., an individual's data) is supplied to the model for prediction. Here, there would be no other data from which to standardize this individual, other than the training data mean and standard deviation.

Estimating Treatment Effects

Machine learning has traditionally been focused strictly on predictive modeling, at the expense of determining treatment effects. However, causal inference and treatment effect estimation are central considerations for epidemiologists. In machine learning, treatment effects have historically been of little interest because models are created to produce predictions of the future as opposed to direct interpretation of predictor–outcome relationships (e.g., average treatment effects). This approach does not translate to a model that is effective for causal inference of model parameters. However, investigators have developed several methods for estimating treatment effects from machine learning models (20).

Heterogeneous Treatment Effects

Investigators utilizing machine learning approaches have begun to explore heterogeneous treatment effects as opposed to the overall average treatment effects (2, 10, 16, 26, 30, 44). Heterogeneous treatment effects are those that are systematically different within different groups of study subjects, often called conditional average treatment effects. One can think of identifying heterogeneous treatment effects as identifying effect modification; however, exploration of heterogeneous treatment effects can be much more rigorous, comprising multiple features as opposed to just a one-way or two-way interaction term in a regression model. Investigators have developed machine learning models to detect these very specific clusters of individuals who showcase different treatment effects within their cluster of similar individuals. The most prominent example of cluster detection for calculating heterogeneous treatment effects is within causal forest models, a form of random forests that allows for the detection of subgroups of similar individuals who display different predictor–outcome effects (66). These models have been sparsely used in epidemiology; a 2017 example from Baum and colleagues (4) evaluates heterogeneous treatment effects in the Look AHEAD trial, an evaluation of weight loss interventions for reducing cardiovascular complications of type 2 diabetes (see the sidebar titled Heterogeneous Treatment Effects in the Look AHEAD Trial). Other methods are available for identifying these effects in machine learning, including through the use of Bayesian additive regression trees (BART) and artificial neural networks (ANNs). In 2019, Künzel and colleagues (36) presented X-learner, a unified method to calculate heterogeneous treatment effects that allows for computation in the presence of complex distributions of treatment effects.

HETEROGENEOUS TREATMENT EFFECTS IN THE LOOK AHEAD TRIAL

Issue: The Look AHEAD trial found no significant reduction in cardiovascular events in type 2 diabetic patients when undergoing weight loss interventions. Therefore, the average treatment effect of the weight loss intervention was not significantly associated with cardiovascular events in type 2 diabetics.

Solution: Using a causal forest, a type of machine learning algorithm, investigators were able to identify a subset of 75% of patients enrolled, in whom the intervention was significantly associated with reductions in cardiovascular events.

Conclusion: Randomized trials, although providing the highest level of evidence, focus largely on average treatment effects. Given the varied patients enrolled, investigators may be able to identify an ineffective treatment on average. In this setting, there may be subpopulations in whom the treatments are effective or are harmful.

Defining heterogeneous treatment effects can be particularly useful in the case of a negative study, when the results are inconclusive or suggest that the intervention is not effective. In these negative studies, there may be subpopulations or clusters of individuals who will have a different and sometimes clinically meaningful treatment effect (10). Although one can use traditional methods to identify subpopulations through stratified regression modeling or the inclusion of interaction terms, the epidemiologist will run into issues with multiple testing bias, a common pitfall in frequentist statistics. In methods driven by hypothesis testing, each additional test run on the data increases the amount of statistical error present. Therefore, if an epidemiologist wishes to evaluate 10 different variables as providing different effects among the study sample, the level of statistical error increases substantially. These novel machine learning methods do not suffer from this issue because they are algorithmic approaches to defining treatment effects and not focused on hypothesis testing. Therefore, one can evaluate as many variables as desired for defining heterogeneous treatment effects without increasing statistical error. Furthermore, because these methods are not bound by the same assumptions as are frequentist statistical approaches, there is no concern for the increase in statistical error when performing repeated hypothesis tests on the same data (i.e., multiple comparisons bias).

These approaches are very timely as we move toward personalized medicine and personalized health (46). To this end, epidemiologists can have a much better analytical handle on individual-level variations in treatment effects.

Defining Model Performance

Many methods are available to define whether a trained model performs adequately to predict the outcome with little error in the testing data set. Little novelty has been observed in the area of model performance definition; most epidemiologists focus on mean squared error, root mean squared error, accuracy, precision–recall, area under the receiver operating characteristic curves (AUC ROC), and F1 statistics, depending on whether the epidemiologist is modeling continuous or categorical outcomes.

All methods provide some evidence of overall model performance, though none should be considered ideal in any particular circumstance. For example, the AUC in an ROC curve is often used to define model performance. This value can be deceiving if modeling a categorical outcome that is imbalanced. In this case, an acceptable AUC can be obtained even when the model predicts only the majority class and predicts next to zero of the minority class. In this context, the early retrieval rate from the ROC curve or precision–recall curves can be used (52). The early retrieval rate can be obtained from the left-most area (generally 1-specificity of greater than 80%) of the

ROC curve. Here, if the model predicts only the majority class, one will see a low AUC because the model has a low sensitivity where the false positive rate is low. This result suggests that the model does not predict positive cases (typically the minority class). Care must be taken, when evaluating any classification model, to ensure that if the labels are not balanced then appropriate measures of model performance are used.

Explaining the Model

Machine learning algorithms are often called black boxes because most of these predictive models are difficult to explain for a single individual. Here, the practitioner is forced to believe the model, without any understanding of the potential for false positives or negatives for an individual prediction. For example, if a classification model that has an 85% balanced accuracy is being used in practice (e.g., accounting for any imbalance), it will still be wrong 15% of the time. If a new data point is supplied to the model and provides a prediction, the practitioner cannot understand whether this particular prediction is more or less likely to be in error. To correct this deficiency, novel approaches such as the local interpretable model-agnostic explanations (LIME) tests have been developed (48) and successfully used in epidemiologic studies. For example, Pereira and colleagues (43) used LIME to unbox a random forest classifier to enhance local interpretation of features in brain lesion research.

FUTURE DIRECTIONS

Deep Learning

ANNs have been used in machine learning for many years. The phrase deep learning has become popular to describe ANNs with many hidden layers. These models, while very complex, are extremely flexible and allow the epidemiologist to include an almost unlimited number of features for classification or regression tasks, and they are very accurate at modeling highly nonlinear relationships. The limitation of these approaches has largely been in hyperparameter selection and training, as large ANNs may be too computationally intensive to run on local machines, reducing their wide adoption. In 2018, Mocanu and colleagues (39) provided an alternative to traditional ANN training, using a sparse evolutionary approach that mimics natural evolution: building models and adding features as long as the model performance improves. Taking a cue from network analytics and graph theory, they provide a solution to the training time limitations without a decrease in model performance.

One novel area of deep learning in clinical epidemiology and medicine is image recognition and computer vision. This has been a broad area of research in the biomedical sciences, with reviews published on the topic (15) and many international competitions devoted to image analytics, such as the Medical Image Computing and Computer Assisted Intervention conference and the ImageCLEF evaluation campaigns. Areas of work range from image segmentation to object tracking and image detection. Evaluations of the work in this area found that many concerns exist with respect to the generalizability of research and competition findings (37). Regardless of these concerns, this area will continue to be important, with applications in epidemiology for many years.

Interventional Machine Learning

The black box issue of many machine learning models imposes some difficulty for the initiation of interventions to reduce the risk of poor health outcomes. We offer two concerns: First, because models do not explicitly provide estimates of the impact of each individual feature on the label

of the model, targeting interventions on modifiable risk factors is difficult; and second, if interventions do take place to reduce the risk of a poor health outcome, model performance statistics will be poor without constant retraining. For example, if a model identifies an individual with a high probability of death, but an intervention occurs to reduce the probability and the individual survives, the model's prediction was actually incorrect. The model predicted death, but the individual survived. A workaround could be to influence the model with an indicator for intervention; however, on initial training, this indicator would be zero for all individuals and the model would fail to learn the patterns from it at a reasonable pace.

Novel neural networks have been developed to assist in reducing these issues. A 2019 example from Shickel and colleagues (56) used gated recurrent unit neural networks to identify individuals at risk of in-hospital mortality. This model not only allowed practitioners to model the probability of death longitudinally, but also provided the ability to document the magnitude of how various time points contributed to the prediction. Therefore, the investigators could provide targeted intervention on the basis of the model predictions as well as model changes in the probability over time. Models such as these have a bright future in epidemiology, finally allowing researchers to model many nonlinear relationships in a longitudinal manner while reducing our reliance on blindly obeying what the model predicts.

PUTTING IT ALL TOGETHER AND LEARNING MORE

Machine learning continues to be a burgeoning area in data analytics. Computing software is making it increasingly easier to learn to create, build, tune, and implement machine learning models. Open-source software such as H2O (**https://www.h2o.ai**) and Keras (**https://keras.io**) as well as the multitude of pay-for-model platforms are widely available and extremely powerful, many of which limit or eliminate the need for any knowledge of computer programming. Learning from scratch can be a daunting process, though many online and in-person courses are available, as well as programs resulting in Bachelor, Master, and Doctoral degrees in Computer Science, Machine Learning, Data Analytics, and Data Science.

Partnering with new team members is another pathway to the creation of machine learning models for various needs. Various academic areas such as data science, biostatistics, data analytics, business, computer science, and computational biology are helpful places to begin a search for experts. Outside of academia, many businesses and various industries are employing experts in the fields of machine learning. Networking through colleagues and at local, regional, national, and international congresses can build a team of experts quickly.

CONCLUSIONS

In conclusion, machine learning is becoming increasingly popular not only for developing predictive modeling, but also for defining treatment effects in epidemiology. Improving these approaches, explaining risk factors, and producing full-scale production algorithms for rapid prediction and improvements in population health all serve as ripe areas for continued research.

SUMMARY POINTS

1. Machine learning is a rapidly advancing area of data analytics.
2. Traditional regression models can be used for machine learning needs, though more algorithmic methods are often considered machine learning.

3. Development of a machine learning model includes feature selection, feature engineering, dealing with missing data, training the model, tuning the hyperparameters, testing the model, evaluating its performance, and explaining or operationalizing the final trained model in production.
4. The complexities of machine learning applications are being greatly reduced with the introduction of open-source machine learning platforms, many of which have a point-and-click interface as opposed to tools that necessitate in-depth knowledge of computer or statistical programming.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLOS ONE* 12(7):e0179805
2. Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113(27):7353–60
3. Bandhary S, Contreras-Mora BY, Gupta R, Fernandez P, Jimenez P, et al. 2017. Clinical outcomes of community-acquired pneumonia in patients with diabetes mellitus. *J. Respir. Infect.* 1(1):23–28
4. Baum A, Scarpa J, Bruzelius E, Tamler R, Basu S, Faghmous J. 2017. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial. *Lancet Diabetes Endocrinol.* 5(10):808–15
5. Beauclair R, Hens N, Delva W. 2018. The role of age-mixing patterns in HIV transmission dynamics: novel hypotheses from a field study in Cape Town, South Africa. *Epidemics* 25:61–71
6. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. 2013. Sample size planning for classification models. *Anal. Chim. Acta* 760:25–33
7. Bellman R. 2015. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univ. Press
8. Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13:281–305
9. Blum AL, Langley P. 1997. Selection of relevant features and examples in machine learning. *Artif. Intell.* 97(1):245–71
10. Bonetti M, Gelber RD. 2004. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 5(3):465–81
11. Breiman L. 2001. Random forests. *Mach. Learn.* 45:5–32
12. Breiman L. 2001. Statistical modeling: the two cultures. *Stat. Sci.* 16(3):199–215
13. Büssing A, Falkenberg Z, Schoppe C, Recchia DR, Poier D. 2017. Work stress associated cool down reactions among nurses and hospital physicians and their relation to burnout symptoms. *BMC Health Serv. Res.* 17(1):551
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* 16:321–57
15. Chen W, Li W, Dong X, Pei J. 2018. A review of biological image analysis. *Curr. Bioinform.* 13:337–43
16. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, et al. 2016. Double/debiased machine learning for treatment and causal parameters. arXiv:160800060 [Econ. Stat.]
17. Chicco D, Rovelli C. 2019. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PLOS ONE* 14(1):e0208737

18. De La Fuente J, Villar M, Estrada-Peña A, Olivas JA. 2018. High throughput discovery and characterization of tick and pathogen vaccine protective antigens using vaccinomics with intelligent Big Data analytic techniques. *Expert Rev. Vaccines* 17(7):569–76
19. Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. 2019. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Ann. Emerg. Med.* 73:334–44
20. Fang G, Annis IE, Elson-Lafata J, Cykert S. 2019. Applying machine learning to predict real-world individual treatment effects: insights from a virtual patient cohort. *J. Am. Med. Inf. Assoc.* 26(10):977–88
21. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. 2012. Predicting sample size required for classification performance. *BMC Med. Inform. Decis. Mak.* 12(1):8
22. Flaxman AD, Vos T. 2018. Machine learning in population health: opportunities and threats. *PLOS Med.* 15(11):e1002702
23. Forbes. 2018. The rise in computing power: why ubiquitous artificial intelligence is now a reality. *Forbes*, July 17. **https://www.forbes.com/sites/intelai/2018/07/17/the-rise-in-computing-power-why-ubiquitous-artificial-intelligence-is-now-a-reality/#22a73011d3f3**
24. Fotouhi S, Asadi S, Kattan MW. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* 90:103089
25. Frérot M, Lefebvre A, Aho S, Callier P, Astruc K, Aho Glélé LS. 2018. What is epidemiology? Changing definitions of epidemiology 1978–2017. *PLOS ONE* 13(12):e0208442
26. Green DP, Kern HL. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* 76(3):491–511
27. Greenland S, Poole C. 2013. Living with *p* values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* 24(1):62–68
28. Hastie T, Tibshirani R, Friedman JH. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. 2nd ed.
29. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. 2005. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* 21(8):1509–15
30. Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7(1):443–70
31. Jaderberg M, Dalibard V, Osindero S, Czarnecki WM, Donahue J, et al. 2017. Population based training of neural networks. arXiv:1711.09846 [Cs]
32. Jensen PB, Jensen LJ, Brunak S. 2012. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* 13(6):395–405
33. Kanter JM, Veeramachaneni K. 2015. Deep feature synthesis: towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris*, pp. 1–10. New York: IEEE
34. Keefe JR, Wiltsey Stirman S, Cohen ZD, DeRubeis RJ, Smith BN, Resick PA. 2018. In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depress. Anxiety* 35(4):330–38
35. Kind AJH, Jencks S, Brock J, Yu M, Bartels C, et al. 2014. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. *Ann. Intern. Med.* 161(11):765–74
36. Künzel SR, Sekhon JS, Bickel PJ, Yu B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *PNAS* 116:4156–65
37. Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, et al. 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* 9(1):5217
38. Messier KP, Wheeler DC, Flory AR, Jones RR, Patel D, et al. 2019. Modeling groundwater nitrate exposure in private wells of North Carolina for the Agricultural Health Study. *Sci. Total Environ.* 655:512–19
39. Mocanu DC, Mocanu E, Stone P, Nguyen PH, Gibescu M, Liotta A. 2018. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nat. Commun.* 9(1):2383
40. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD. 2008. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet. Epidemiol.* 32(4):325–40
41. *Nat. Commun.* Editors. 2018. Epidemiology is a science of high importance. *Nat. Commun.* 9(1):1703

42. Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, et al. 2014. Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods Ecol. Evol.* 5(9):961–70
43. Pereira S, Meier R, McKinley R, Wiest R, Alves V, et al. 2018. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-random forest system on brain lesion segmentation. *Med. Image Anal.* 44:228–44
44. Powers S, Qian J, Jung K, Schuler A, Shah NH, et al. 2018. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat. Med.* 37(11):1767–87
45. R. Soc. (G. B.). 2017. *Machine learning: the power and promise of computers that learn by example*. Rep. DES4702, R. Soc. G. B., London. **https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf?**
46. Ramaswami R, Bayer R, Galea S. 2018. Precision medicine from a public health perspective. *Annu. Rev. Public Health* 39:153–68
47. Raudys SJ, Jain AK. 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.* 13(3):252–64
48. Ribeiro MT, Singh S, Guestrin C. 2016. "Why should I trust you?": explaining the predictions of any classifier. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–44. New York: ACM
49. Ronca E, Scheel-Sailer A, Koch HG, Gemperli A, Group SwiSCI Study, et al. 2017. Health care utilization in persons with spinal cord injury: part 2—determinants, geographic variation and comparison with the general population. *Spinal Cord* 55(9):828–33
50. Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65(6):386–408
51. Sadilek A, Caty S, DiPrete L, Mansour R, Schenk T Jr., et al. 2018. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *npj Digit. Med.* 1(1):36
52. Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* 10(3):0118432
53. Samuel AL. 1959. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* 3(3):210–29
54. Seligman B, Tuljapurkar S, Rehkopf D. 2018. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Popul. Health* 4:95–99
55. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am. J. Epidemiol.* 179(6):764–74
56. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. 2019. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci. Rep.* 9(1):1879
57. Shmueli G. 2010. To explain or to predict? *Stat. Sci.* 25(3):289–310
58. Singh GK. 2003. Area deprivation and widening inequalities in US mortality, 1969–1998. *Am. J. Public Health* 93(7):1137–43
59. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15:1929–58
60. Stekhoven DJ, Buhlmann P. 2012. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28(1):112–18
61. Tessmer HL, Ito K, Omori R. 2018. Can machines learn respiratory virus epidemiology?: A comparative study of likelihood-free methods for the estimation of epidemiological dynamics. *Front. Microbiol.* 9:343
62. Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58(1):267–88
63. Tsai C-F, Eberle W, Chu C-Y. 2013. Genetic algorithms in feature and instance selection. *Knowl.-Based Syst.* 39:240–47
64. Turing AM. 1950. Computing machinery and intelligence. *Mind* 59(236):433–60
65. van der Ploeg T, Austin PC, Steyerberg EW. 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* 14(1):137
66. Wager S, Athey S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113:1228–42

67. Wang Y, Wang D, Ye X, Wang Y, Yin Y, Jin Y. 2019. A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction. *Inf. Sci.* 474:106–24
68. Wiemken TL, Carrico RM, Furmanek SP, Guinn BE, Mattingly WA, et al. The impact of socioeconomic position on the incidence, severity, and clinical outcomes of hospitalized patients with community-acquired pneumonia. *Public Health Rep*. In press
69. Wiemken TL, Furmanek SP, Mattingly WA, Guinn BE, Cavallazzi R, et al. 2017. Predicting 30-day mortality in hospitalized patients with community-acquired pneumonia using statistical and machine learning approaches. *J. Respir. Infect.* 1(3):50–56
70. Wiemken TL, Kelley RR, Fernandez-Botran R, Mattingly WA, Arnold FW, et al. 2017. Using cluster analysis of cytokines to identify patterns of inflammation in hospitalized patients with community-acquired pneumonia: a pilot study. *J. Respir. Infect.* 1(1):3–11
71. Wiemken TL, Kelley RR, Mattingly WA, Ramirez JA. 2019. Clinical research in pneumonia: role of artificial intelligence. *J. Respir. Infect.* 3(1):1–4