ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

ISPOR Report

Machine Learning Methods in Health Economics and Outcomes Research—The PALISADE Checklist: A Good Practices Report of an ISPOR Task Force

William V. Padula, PhD, Noemi Kreif, PhD, David J. Vanness, PhD, Blythe Adamson, PhD, Juan-David Rueda, MD, PhD, Federico Felizzi, PhD, MBA, Pall Jonsson, PhD, Maarten J. IJzerman, PhD, Atul Butte, MD, PhD, William Crown, PhD

ABSTRACT

Advances in machine learning (ML) and artificial intelligence offer tremendous potential benefits to patients. Predictive analytics using ML are already widely used in healthcare operations and care delivery, but how can ML be used for health economics and outcomes research (HEOR)? To answer this question, ISPOR established an emerging good practices task force for the application of ML in HEOR.

The task force identified 5 methodological areas where ML could enhance HEOR: (1) cohort selection, identifying samples with greater specificity with respect to inclusion criteria; (2) identification of independent predictors and covariates of health outcomes; (3) predictive analytics of health outcomes, including those that are high cost or life threatening; (4) causal inference through methods, such as targeted maximum likelihood estimation or double-debiased estimation—helping to produce reliable evidence more quickly; and (5) application of ML to the development of economic models to reduce structural, parameter, and sampling uncertainty in cost-effectiveness analysis.

Overall, ML facilitates HEOR through the meaningful and efficient analysis of big data. Nevertheless, a lack of transparency on how ML methods deliver solutions to feature selection and predictive analytics, especially in unsupervised circumstances, increases risk to providers and other decision makers in using ML results.

To examine whether ML offers a useful and transparent solution to healthcare analytics, the task force developed the PALISADE Checklist. It is a guide for balancing the many potential applications of ML with the need for transparency in methods development and findings.

Keywords: artificial intelligence, machine learning.

VALUE HEALTH. 2022; 25(7):1063–1080

Introduction

The term “machine learning” (ML) refers to a family of statistical methods that generally focus on classification, ranking, and prediction.[1] Modern healthcare data are characterized by high dimensionality, massive volume, rapid turnover, and complexity of structure. These characteristics require efficient methods to generate evidence where traditional approaches are costly or limited. There is a growing recognition of the applicability of ML approaches to address healthcare problems.[2] In this article, we focus upon the potential applications of ML in health economics and outcomes research (HEOR).[1]

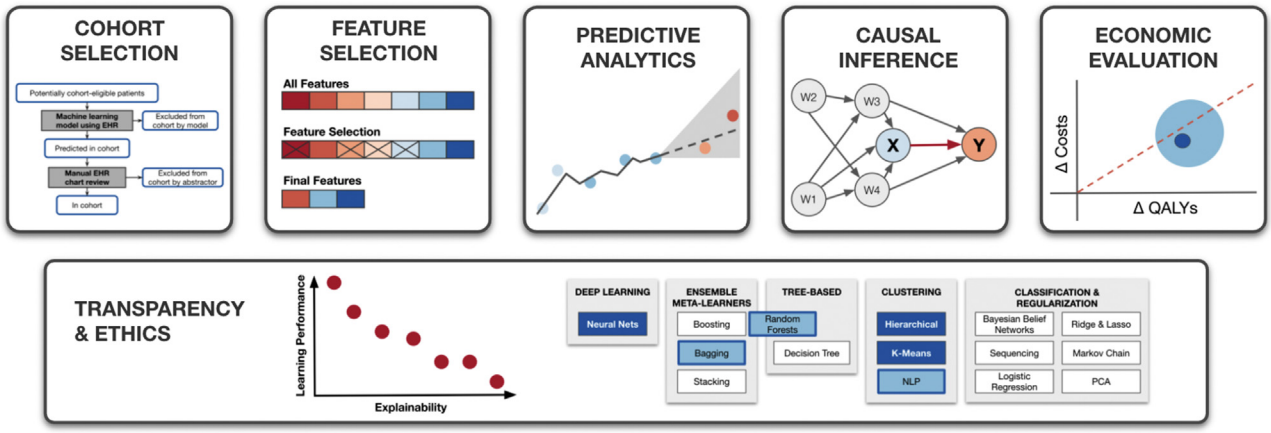
There are 2 broad categories of ML methods—supervised and unsupervised.[3] Supervised methods require specification of an outcome variable to perform classification, ranking, or prediction.[4] Unsupervised methods are focused mainly on dimension reduction and identifying the underlying structure of the data without specifying outcomes.

Traditional ML models rely upon the development of features—or variables—that are defined based on researcher domain knowledge. In contrast, representation learning methods, including so-called deep learning models, extract features directly from the data itself, thereby enhancing our understanding of the structure or relationship between causes and effects in the data that may have been previously unknown.[5,6]

ML is a potentially valuable addition to the HEOR toolkit. ML can facilitate the search for complex relationships in high-dimensional data sets, such as those generated by electronic health records (EHRs) or mobile health devices. These relationships can be used to improve detection and classification of disease, to identify cohorts of patients sharing characteristics that might not be obvious when considering only a small set of variables using traditional methods, and to forecast trajectories of health outcomes under alternative personalized treatment options.[7]

Using supervised or unsupervised ML can enhance the value of healthcare delivery by potentially triggering interventions before

Figure 1. Conceptual diagram of machine learning applications in HEOR.



HEOR indicates health economics and outcomes research; NLP, natural language processing; PCA, principal component analysis; QALY, quality-adjusted life-year.

adverse outcomes occur, but ML is subject to all of the usual challenges encountered in other forms of observational data analysis. In particular, the fact that ML methods operate on big data does not necessarily protect against bias. Increasing sample size—for example, obtaining additional administrative healthcare data—does not correct the problem of bias in treatment effect estimates if the data set is lacking in key clinical severity measures, such as cancer stage in a model of breast cancer outcomes, or contains confounders that we know little about with respect to a causal pathway.[8,9]

This criticism especially applies to situations where the objective is to estimate causal effects in observational studies using routinely collected healthcare data. For prediction, the cross-validation machinery will do its best to minimize prediction error, but this may not minimize bias because of a lack of comparability between treatment and control groups—particularly when the groups differ in features not measured in the data.

The ISPOR ML Methods Emerging Good Practices Task Force developed guidance for HEOR and decision makers in the use of ML methods. This report considers 5 applications of ML methods that are important to HEOR: (1) ML-assisted cohort selection, (2) feature selection, (3) predictive analytics, (4) causal inference, (5) health economic evaluation, and reflection on transparency and explainability (Fig. 1). We present these considerations in an order reflecting a standard approach to performing HEOR: identifying a study population, classifying exposures that can alter outcomes,

Table 1. ML applications for HEOR.

Concept	Definition
Big data	Referring to data with high dimensionality based on complex combinations of large numbers of observations and features—often with less structure than typical observational data. ML methods can support high throughput analytics of big data more efficiently than traditional statistical and econometric methods in HEOR.
Cohort selection	The process of using a set of inclusion and exclusion criteria to select a set of patients upon which to perform a retrospective study. ML methods, sometimes combined with the use of natural language processing on unstructured data, empower researchers to identify individuals with specific characteristics from large volumes of observations in a time-efficient manner.
Feature selection	The process of selecting a subset of relevant features (key independent variables, predictors, and covariates) for use in model construction. ML can be used to identify variables that have associations with the main outcome measure from a large volume of potential predictors.
Predictive analytics	Systematic examination of the associative structure of observed data for the purpose of generating estimates of outcomes not yet observed. It comprises a set of statistical models and/or algorithms used to represent the associative structure of observed data along with a set of rules to generate estimated values from those models.
Causal inference	The process of drawing a conclusion about a causal effect of an intervention conditional upon controlling for the effects of confounding influences. Statistical approaches for causal inference that incorporate ML can have lower biases than traditional parametric approaches, because of the flexibility of ML algorithms used.
Economic evaluation	ML methods can reduce uncertainty in economic models with respect to structural, parameter, and sampling uncertainty given that data are more fit for purpose of the exact model needs. Furthermore, ML offers analysis of vast amounts of data that would be required to facilitate dynamic simulation modeling that represents more real-world aspects of health systems.

HEOR indicates health economics and outcomes research; ML, machine learning.

Table 2. ML methods that could become common in support of HEOR practices

Method	ML classification	Definition	Example applications
Classification and regression			
Bayesian belief networks	Supervised	Bayesian belief networks express proposed relationships (edges) among variables (nodes) as a network. When expressed as a DAG, the causal relationship among variables can be expressed as a set of conditional distributions that can be estimated from the data and combined with beliefs (priors) about potential relationships.	Economic evaluation predictive analytics
Hidden Markov chains	Supervised	Models used to explore dependency between adjacent time points to establish temporal order of exposures leading to a series of common outcomes. Hidden traits allow the machine to identify transitions between potentially unobserved health states.	Economic evaluation • Transition probability extraction • Health state designations
Ridge and LASSO regression, elastic net	Supervised	These are penalized regression methods. Elastic nets are regularized regression models that penalize the least squares criterion by adding a function of the magnitude of the parameter estimates, with the goal of reducing the influence of weak predictor variables, thus reducing variance of predictions. When the function is the sum of the absolute value of parameters, LASSO is obtained; when using the sum of the squared value of parameters, ridge regression is obtained. Although prediction accuracy may be improved, individual parameter estimates may be biased.	Feature selection, predictive analytics, causal inference (propensity score, outcome regression, "double variable selection" to select confounders)
Tree-based methods			
Decision tree	Supervised	Decision tree, or classification tree, is used to predict a qualitative response for an observation belonging to the most commonly occurring class of training observations in the region to which it belongs.	Economic evaluation • Determining clinical pathways • Structuring a decision model • Predictive analytics
Random forests	Unsupervised or supervised	Random forest is an ensemble method comprising multiple tree-based (recursive partitioning) models. Each tree is generated following prespecified rules, but with a different sample of observations (see bagging) or using a different subset of variables (see feature selection).	Predictive analytics, feature selection, causal inference (propensity score, outcome regression, causal forests for treatment effect heterogeneity)
Ensemble meta-learners			
Boosting	Supervised	Boosting is an ensemble method in which sequences of models are generated, each subsequent model being built based on prediction errors from the previous stage. Adaptive boosting upweights poorly predicted observations in the subsequent stage, whereas gradient boosting estimates the prediction error from the previous stage.	Predictive analytics, causal inference
Bagging	Unsupervised or supervised	Bagging is a process for generating ensembles of models using repeated Bootstrap resamples applying ML model to each resample and aggregating them into a consensus prediction.	Feature selection, predictive analytics
Stacking	Supervised	Stacking is a meta-learning algorithm to learn how to best combine the predictions (eg, using weights) from 2 or more base ML algorithms. It can harness the capabilities of several well-performing models and make predictions that have better performance than any single model in the ensemble. "Super learning" is an example of stacking.	Predictive analytics, causal inference (propensity score, outcome regression)
Clustering			
Hierarchical clustering	Unsupervised	Bottom-up agglomeration or top-down division of observations into groups based on strength of association to common features. Number of clusters are learned ex poste rather than prespecified.	Cohort selection, feature selection

continued on next page

Table 2. Continued

Method	ML classification	Definition	Example applications
K-means clustering/ partitioning around medoids	Unsupervised	Partitioning observations into prespecified numbers of clusters based on common degrees of association to features of interest. It requires that the programmer specify the number of clusters a priori.	Cohort selection, feature selection
PCA	Unsupervised	A dimensionality reduction method that identifies a set of projections that represent the majority of variability in the data	Feature selection
Deep learning			
Neural networks	Supervised or unsupervised	A series of algorithms that endeavor to recognize underlying relationships between exposures and outcomes based on layers of interactions and intermediate outcomes. Methods include artificial neural networks, convolution neural networks, and recurrent neural networks.	Feature selection, predictive analytics, causal inference
Data-specific approaches used with ML			
Text: NLP	Supervised or unsupervised	To read, decipher, and understand language encoded in medical records to extract data for patient cohorts of interest based on common keywords or phrases	Cohort selection
Imaging: Image recognition/ computer vision	Supervised or unsupervised	Ability of software to identify objects, change, and disease in image data such as video sequences, views from multiple cameras, multidimensional data from a 3D scanner, or medical scanning devices, such as CT and MRI. Computer vision systems are used to process, correct, and analyze the images, for example, of surgical tools and the patient's body.	Predictive analytics, economic evaluation • Transition probability extraction • Health state designations
Audio: DSP	Supervised or unsupervised	Processing and detection of signals in sound files of voice and other audio, with health applications in hearing aids	Predictive analytics, causal inference, economic evaluation • Health state designations

CT indicates computed tomography; DAG, directed acyclic graph; DSP, digital signal processing; HEOR, health economics and outcomes research; LASSO, least absolute shrinkage and selection operator; ML, machine learning; MRI, magnetic resonance imaging; NLP, natural language processing; PCA, principal component analysis.

predicting the association between exposures and outcomes, assessing causal effects of interventions, and understanding whether or not interventions or healthcare policy decisions add value (Table 1). The intent is to introduce these concepts at a high level and refer readers to sources where they can learn more about theory and techniques that can support and advance the HEOR field.

Cohort Selection

Researchers are challenged when data elements necessary to apply study inclusion and exclusion criteria are trapped within the unstructured portions of the patient's EHR.[10] Traditionally, human review to abstract these unstructured elements is expensive and time consuming.[11] This limits the size of cohorts that can be studied, especially for rare criteria in which a large number of charts need to be abstracted to find just a few cohort-eligible patients.

One of many ML techniques and applications described in Table 2 is to use natural language processing (NLP) to help in defining cohort selection criteria. In this process, abstractors first label the patients (ie, is cancer metastatic?), then train a model on the labeled data, and apply the trained model as an extra prefilter to new data. When the probability of eligibility is above a specified threshold, a patient is queued for further human abstraction. This process increases the efficiency by reducing the total number of patients needing abstraction by human reviewers, allowing more precise study eligibility criteria at scale.

When labeled data are available, ML methods can also be used to identify features that are correlated with the probability of observing the outcome of interest without previous specification of these features. Patients who have similar probabilities of the outcome but do not have the treatment of interest can serve as comparators for treatment effectiveness estimation. The use of ML for identification of comparison groups has been demonstrated to reduce bias in treatment effect estimates relative to traditional approaches.[12,13]

If unstructured data are required to determine cohort eligibility, NLP techniques may sometimes be used to convert unstructured data to structured data. Relevant methods include term-frequency inverse-document frequency, named entity recognition, or deep learning methods such as word embeddings, document embeddings, or language models such as Bidirectional Encoder Representations from Transformers.[14] If labeled data are incomplete, unsupervised techniques such as clustering may also be used to assign labels.

ML-assisted cohort selection has been used in numerous HEOR studies (Table 3).[5,15-38] For example, the model-assisted cohort selection application at Flatiron Health is used to source and select patients for oncology studies from deidentified EHR-derived data.[30] Other ML case studies have demonstrated the ability to identify larger volumes of patients for studies.[39] Sohn et al[29] built a system to extract physician-asserted side effects from EHR

Table 3. Case studies of ML methods applied in HEOR.

Publication	Title	Study design	Health outcome	Data type	ML approach	ML method used	How ML was applied.	Key findings	Source
Ting et al, *JAMA* 2017	Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images from Multiethnic Populations With Diabetes	Cohort study	Diabetic retinopathy	EHR	Predictive analytics	Deep learning	Deep learning improved diagnostic accuracy for diabetic retinopathy by training the algorithm on 494 661 retina images.	Diabetic retinopathy can be diagnosed with greater specificity and sensitivity using deep learning methods than traditional n-of-1 diagnosis.	[15]
Nori et al, *Alzheimers & Dementia* 2019	ML models to predict onset of dementia: A label learning approach	Case-control	Dementia	Claims and EHR	Predictive analytics	Gradient boosting	Combined claims and EHR data to test whether label learning methods could improve prediction of Alzheimer's disease.	Label learning methods did not significantly improve predictive models of Alzheimer's disease.	[16]
Cole et al, *Pediatr Rheumatol* 2013	Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for EHR-based research	Cohort study	Uveitis	EHR	Feature selection; predictive analytics	Logistic regression; unsupervised hierarchical clustering	ML improved specificity of medication delivery to patients of varying characteristics with chronic uveitis.	Text analytics can care of future patients based on n-of-1 samples of previous rare-disease cases.	[17]
Hong et al, *PLoS One* 2018	Predicting hospital admission at emergency department triage using ML	Cohort study	High-risk hospital admission	EHR	Predictive analytics	Logistic regression; boosting; neural network	ML algorithms improved predictive validity of efforts to anticipate a hospital admission.	ML can accurately predict hospital admission based on patient history in the EHR.	[18]
Futoma et al, *J Biomed Inform* 2015	A comparison of models for predicting early hospital readmissions	Cohort study	30-day hospital readmission	EHR	Feature selection; predictive analytics	Deep learning; random forest	Random forests selected features that appropriately differentiate readmission risk between cohorts, and deep learning improves readmission prediction accuracy.	Deep learning can improve prediction of 30-day readmission.	[19]
Rajkomar et al, *NPJ Digit Med* 2018	Scalable and accurate deep learning with electronic health records	Cohort study	In-hospital mortality	EHR	Predictive analytics	Deep learning	Deep learning increased sensitivity and specificity of predicting in-hospital mortality and 30-day readmission.	Deep learning outperforms statistical methods to improve prediction of measures of hospital performance.	[5]

continued on next page

Table 3. Continued

Publication	Title	Study design	Health outcome	Data type	ML approach	ML method used	How ML was applied.	Key findings	Source
Xu et al, *J Biomed Inform* 2011	Applying semantic-based probabilistic context-free grammar to medical language processing–a preliminary study on parsing medication sentences	Cohort study	Colorectal cancer	EHR	Cohort selection	NLP	Algorithm combined ML and NLP to detect patients with colorectal cancer.	A 2-step method extracted disease concepts from clinical notes followed by confirmation of cases using aggregated information from narratives and billing data.	20
Jiao et al, *Nature Commun* 2020	A deep learning system accurately classifies primary and metastatic cancers using passenger mutation	Cohort study	Cancer	Genomic data	Predictive analytics	Deep learning	Predicted cancer type based on patterns of somatic passenger mutations detected in whole genome sequencing of 2606 tumor archives.	Passenger mutations can inform detection of circulating tumor DNA.	21
Liu et al, *AMIA Annu Symp Proc* 2012	A study of transportability of an existing smoking status detection module across institutions	Cohort study	Smoking	EHR	Cohort selection	NLP	Detected smoking status in patient charts.	A customized module achieved significantly higher F-measures than direct applications.	22
Padula et al, *BMJ Qual Saf* 2019	Value of Hospital resources for effective pressure injury prevention: a cost-effectiveness analysis	Markov modeling	Hospitalized patients at risk of pressure injuries	EHR	Economic evaluation	Hidden Markov chain	Hidden Markov chain provided structure for the economic model by identifying transition probabilities for unobserved health states in data.	Predicting high-risk patients using ML can reduce risk of patients and conserve costly labor/ nursing time.	23
Kreif et al, *Am J Epidemiol* 2017	Estimating the Comparative Effectiveness of Feeding Interventions in the Pediatric Intensive Care Unit: A Demonstration of Longitudinal Targeted Maximum Likelihood Estimation	Longitudinal analysis of clinical trial data	Pediatric ICU	EHR	Causal inference	TMLE using super learning	TMLE using super learning adjusted for time-dependent confounding.	The study estimates the probability of a child's being discharged alive from the pediatric ICU by a given day, under a range of longitudinal feeding regimes. It demonstrates the benefit of the flexible TMLE approach.	24

continued on next page

Table 3. Continued

Publication	Title	Study design	Health outcome	Data type	ML approach	ML method used	How ML was applied.	Key findings	Source
Padula et al, *JAMIA* 2017	Using clinical data to predict high-cost performance coding issues associated with pressure ulcers: a multilevel cohort model	Cohort study	Hospitalized patients at risk of pressure injuries	EHR	Feature selection; predictive analytics	Random forests; multilevel logistic regression	Random forests identified key variables for the predictive model from an EHR of > 10 000 potential predictors. Logistic regression was used to derive a statistical model that derived a decision rule for differentiating low-risk and high-risk patients.	Accurately predicting patients who are at high risk of pressure injury requires an understanding of clinical characteristics not previously hypothesized in studies driven by clinical judgment.	[25]
Arora et al, *Value Health* 2019	Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine	Systematic review	Population health	EHR	Predictive analytics	Bayesian belief network	Predictive analytics for precision medicine.	Meaningful for predictive analytics—risk prediction.	[26]
Liu et al, *JAMIA* 2012	Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs.	Cohort study	Adverse drug reactions	EHR	Predictive analytics; causal inference	Support vector machine; cluster analysis	Filtered notes, added new annotated data for training the ML classifier (SVM with a radial basis function kernel), and added rules to the rule-based classifier.	Smoking detection module in cTAKES, developed at Mayo Clinic to yield a significantly better classifier in terms of the F-measure on a data set obtained from Vanderbilt University.	[27]
Xu et al, *AMIA Annu Symp Proc* 2011	Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases	Cohort study	Colorectal cancer	EHR	Cohort selection	NLP	Algorithm combined ML and NLP to detect patients with colorectal cancer.	Their 2-step method extracted CRC concepts from clinical notes followed by determination of CRC cases using aggregated information from narratives and billing data.	[28]
Sohn et al, *JAMIA* 2011	Drug side effect extraction from clinical narratives of psychiatry and psychology patients.	Cohort study	Mental health disorders	EHR	Cohort selection	NLP	System extracted physician-asserted side effects from EHR clinical narratives of psychiatry and psychology patients.	Their system leverages NLP using cTAKES along with decision trees (C4.5) using side effect keyword features and pattern-matching rules.	[29]

continued on next page

Table 3. Continued

Publication	Title	Study design	Health outcome	Data type	ML approach	ML method used	How ML was applied.	Key findings	Source
Birnbaum et al, *arXiv e-Print* 2019	MACS with bias analysis for generating large-scale cohorts from the EHR for oncology research	NLP	Oncology	EHR	Cohort selection	NLP; logistic regression	Flatiron Health MACS trained a model on 17 263 patients using term-frequency inverse-document-frequency and logistic regression. To refresh models and continually test them at scale, Flatiron Health developed model integration, monitoring, and serving architecture (Mimosa) to continually monitor the performance and potential bias of MACS models in production. It trains and evaluates new models to adapt to changes in treatment patterns, documentation patterns, and Flatiron's network of oncology clinics.	The algorithm had an AUC of 0.976, a sensitivity of 96.0%, and an abstraction efficiency gain of 77.9%.	[30]
Hansen et al *Circulation: Cardiovascular Quality & Outcomes* 2016	Identifying drug-drug interactions by data mining: A pilot study of warfarin-associated drug interactions	Random forests	Patients prescribed with warfarin	EHR	Feature selection	Random forests; logistic regression	Random forest was set up to predict altered INR levels after novel prescriptions. The most important drug groups from the analysis were further investigated using logistic regression in a new data set.	Identified known warfarin-drug interactions without a previous hypothesis using clinical registries. Additionally, the study discovered a few potentially novel interactions.	[31]

continued on next page

Table 3. Continued

Publication	Title	Study design	Health outcome	Data type	ML approach	ML method used	How ML was applied.	Key findings	Source
Churpek et al, *Critical Care Med* 2016	Multicenter Comparison of ML Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards	Random forests; logistic regression	Decompensation among hospitalized patients	EHR	Feature selection; predictive analytics	Random forests; logistic regression	Compared with MEWS, authors hypothesized that random forests or logistic regression with improved feature selection could improve sensitivity and specificity for intervening with hospitalized patients at risk of additional decompensation.	In the validation data set, the random forest model was the most accurate model (AUC 0.80). The logistic regression model with spline predictors was more accurate than the model using linear predictors (AUC 0.77 vs 0.74; $P < .01$), and all models were more accurate than the MEWS (AUC 0.70).	32
Henry et al, *Science Translational Medicine* 2015	A targeted real time early warning score (TREWScore) for septic shock	Cox proportional hazard model	Hospitalized patients at risk of septic shock	EHR	Feature selection; predictive analytics	Imputation of missing data; LASSO regression	EHR data provided rich information to predict septic shock risk. Researchers imputed missing data to then inform a LASSO model on the features that best predicted septic shock. These features where then fit to a Cox proportional hazard model to compute risk in real time.	Increased accuracy of patients at risk of septic shock compared with existing models (eg, MEWS approach).	33
Ali et al, *Biophys Rev* 2019	ML and feature selection for drug response prediction in precision oncology applications	Cluster analysis				Bayesian efficient multiple kernel learning method	Multiple methods used to explore efficiency.	Compare ML methods for using genomic, epigenomic, and proteomic data to predict individual response to cancer drugs, finding superior performance of the Bayesian efficient multiple kernel learning model.	34

continued on next page

Table 3. Continued

Publication	Title	Study design	Health outcome	Data type	ML approach	ML method used	How ML was applied.	Key findings	Source
Miotto et al, *Sci Rep* 2016	Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records	Deep learning	Diabetes; schizophrenia; cancer	EHR	Feature selection; predictive analytics	Unsupervised deep feature selection	Unsupervised deep feature learning created representative archetypes of patients from high-dimensional (> 40 000 variables) EHR data, finding that it improved the ability to predict occurrence of future disease and used that information to predict development of severe diabetes, schizophrenia, and several cancers.	Created representative archetypes of patients from high-dimensional (> 40 000 variables) EHR data.	35
Neugebauer et al, *Stat Med* 2014	Targeted learning in real-world comparative effectiveness research with time-varying interventions	Cohort study	Diabetes	EHR	Causal inference	TMLE with super learner marginal structural model	TMLE with super learning used in marginal structural models to adjust for confounding and selection bias in causal inference pathway.	The causal parameter of interest is the effect of dynamic treatment strategies, intensifying medication when blood A1c level reached threshold. In particular, the authors estimate the cumulative risk of a failure event, albuminuria, under different hypothetical interventions. The main estimation method used is targeted learning using the super learner, and an increased precision of this approach is demonstrated compared with traditional methods.	36

continued on next page

Table 3. Continued

Publication	Title	Study design	Health outcome	Data type	ML approach	ML method used	How ML was applied.	Key findings	Source
Kempa-Liehr et al, *Int J Med Inform* 2020	Healthcare pathway discovery and probabilistic ML	Cohort study	Appendicitis	EHR	Predictive analytics; economic evaluation	Pathway discovery; probabilistic regression modeling	ML was used to define clinical care pathways taken by patients with appendicitis. Probabilistic regression predicted potential outcomes based on variability in patient pathway and time to successful treatment.	Pathway discovery can reduce variability in care pathways that increase likelihood of desirable outcomes, improve clinical scheduling, and improve patient recovery time.	[37]
Liu et al, *Nature* 2021	Evaluating eligibility criteria of oncology trials using real-world data and AI	Cohort study	Oncology	EHR	Cohort selection; feature selection; predictive analytics; causal inference	MACS; Shapley values	To evaluate the influence of an individual criterion, the Shapley value was used to assess the average expected marginal contribution of adding one criterion to the hazard ratio after all possible combinations of criteria have been considered.	With a data-driven approach to broaden restrictive criteria, the pool of eligible patients more than doubled on average and the hazard ratio of the overall survival decreased by an average of 0.05. This suggests that many patients who were not eligible under the original trial criteria could potentially benefit from the treatments.	[38]

AI indicates artificial intelligence; AUC, area under the curve; CRC, colorectal cancer; cTAKES, clinical Text Analysis and Knowledge Extraction System; EHR, electronic health record; HEOR, health economics and outcomes research; INR, interventional normalized ratio; LASSO, least absolute shrinkage and selection operator; MACS, model-assisted cohort selection; MEWS, modified early warning score; ML, machine learning; NLP, natural language processing; TMLE, targeted maximum likelihood estimate.

clinical narratives of psychiatry and psychology patients, leveraging NLP along with decision trees using side effect keyword features and pattern-matching rules.

Cohort selection with ML still faces risk of selection bias. When abstractors review patients above a threshold, specificity is maximized (ie, there are few false-positives). Nevertheless, if the algorithm produces a high number of false-negatives (ie, has low sensitivity), then patients could be erroneously excluded from a cohort. If these false-negatives are not randomly sampled, then the resulting research cohort could be biased. The impact of bias can be assessed by validating results from ML-generated cohorts against results using manual abstraction-generated cohorts.

Cohort selection bias may also occur by race or ethnicity traits.[40] This is, at its core, an issue of utilization of healthcare services by different populations that may be related to discrimination or sociodemographic strata facing structural barriers to healthcare access and use. Although there are no standard solutions to the problem, stratification and weighting methods may be helpful in overcoming such bias.

Feature Selection

The number of observations in many healthcare data sets, combined with high dimensionality, creates challenges for researchers by exceeding the threshold for analytics with classical HEOR methods. Imbalances in these healthcare data sets such as possessing far more predictors than observations—commonly known as the *large p, small n* problem—further complicate analytic tasks by increasing the risk of overfitting. Feature selection methods can reduce the risk of overfitting and better estimate some causal parameters such as average treatment effects (ATEs).

Classical statistical methods often fail in high-dimensional data. For example, r-square in ordinary least squares methods

Table 4. Approaches to conducting feature selection.

Approach	Description	Considerations and examples
Filter	Generate and evaluate the subset of variables without the involvement of a model. normally used as a preprocessing step. Does not leverage ML methods.	The advantage of using filter methods is that it is fast to set up and run and has a low impact on computational memory. In contrast, these methods are crude compared with wrappers because they are limited to application on relatively smaller data sets, the univariate techniques ignore dependencies, and there is potential for the selection of redundant variables.[48]
Wrapper	"Brute force" feature selection techniques. In wrappers, the subset of variables is measured by the performance of the model.	The advantage of wrappers relies on the fact that the subset interacts with the classifier. Therefore, they have high prediction performance because they are designed to maximize model performance.[47] In contrast, wrappers tend to overfit, are computationally intensive, and need arbitrary stopping criteria. Examples of wrappers include recursive feature elimination, sequential forward selection, and genetic algorithms.
Embedded	Propose and evaluate a subset of variables during the construction of the model.	Historically common approaches in HEOR include ridge and LASSO regression methods.
Hybrid	Combine filter and wrapper approaches.	Hybrid methods first apply filtering and then follow this with the application of a wrapper method.

HEOR indicates health economics and outcomes research; LASSO, least absolute shrinkage and selection operator.

approaches one as the number of features approaches the sample size, but such models are unlikely to provide robust classifications for new data for 2 reasons. First, when the counts of observations and predictors are similar, ordinary least squares models will have excessive variability and poor predictions. This case is apparent in predicting outcomes associated with multiple gene expressions.[41]

Second, when encountering a *large p, small n* situation, the resulting variance can be infinite. This commonly occurs when evaluating n-of-1 trials, but also for data with high dimensionality.[42] The curse of dimensionality refers to the fact that when the dimensionality of the data increases, the available data become sparse, resulting in high levels of missing data.[43]

To address these problems, feature selection identifies a subset of the predictors for the analysis based on a balanced combination of improvements in correlation and reductions in error. The objective is to obtain an optimal classifier by using a minimal number of features and excluding the redundant ones.[44] In contrast, feature extraction creates new features as a function of other features and common methods for feature creation including linear discriminant analysis, principal component analysis, autoencoders, and neural networks (see Table 2).[45]

Despite having less discriminatory power, feature selection has the advantage over feature extraction in preserving the original data.[46] Other advantages include finding a more parsimonious model, improving model generalization by limiting the risk of overfitting and enhanced performance, avoidance of collinearity, reduction in time and computational resources for model fitting, allowance for a deeper insight into the underlying processes that generated the data, increased probability of finding predictors in real-world data, and safety surveillance for anomaly detection. Most of the methods for feature selection rely on supervised data with some techniques for unsupervised data, which will not be covered here.[47]

Feature selection methods can be classified into 4 approaches: filter, wrapper, embedded, and hybrid (Table 4[47,48]).

There is a lack of consensus on the preferred methods concerning performance, although there is a propensity for preference of wrapper methods over filtering.[47,49,50] Wrappers often provide the best subset of variables compared with filtering methods but are prone to overfitting. We recommend trying multiple feature selection algorithms and evaluating what works best for the specific problem of interest.

The Challenge of Missing Data

An underdeveloped area within feature selection is the handling of missing data.[51] There are 2 possible methods for conducting feature selection in the presence of missing data: feature selection and then imputation of missing values or imputation of missing values before feature selection. The choice of approach depends on the data being used for the feature selection process; some techniques can handle missing data (tree-based methods), but others are intolerant to missing data (Support Vector Machine [SVM], Neural Network [NN], Generalized Linear Modeling [GLMnet], etc).

The latter approach in imputing missing values before feature selection is prone to imputation bias. In imputation bias, the predictor becomes important after the imputation process (false-positive).[52] Therefore, the selection of the approach depends on the purpose of your model (eg, inference or prediction). Our recommendation for most HEOR scenarios would be to perform feature selection after imputation given that the field of biostatistics has introduced methods to impute or censor observations with missing data. Nevertheless, the correct approach will depend on the number of predictors and the computational capabilities, type of problem, amount of missingness, and type of missingness.

Predictive Analytics

A core value of the ML predictive analytics approach is the use of procedures for model construction to increase the accuracy of out-of-sample prediction—often at the expense of explaining variation in sample data.[53,54] Predictive analytics is rooted in statistical decision theory, where losses incurred by a decision maker depend on how far predictions are from actual outcomes when the algorithm is applied to new input data.[55] In the context of classification, we wish to "predict" whether or not an individual belongs to an explicitly defined class (eg, diabetic) based on the observation of related predictor variables (eg, age, body mass index, historical measures of blood glucose); losses are incurred for

Table 5. The PALISADE Checklist—key considerations for evaluating the transparency of ML to stakeholders and decision makers.

Element	Definition
Purpose	Is the purpose of the algorithm clearly stated at the outset? Is the implementation of the algorithm in a healthcare setting fair and ethical?
Appropriateness	Is there a clear justification that the algorithm is acceptable in the context within which it is being applied?
Limitations	Have the strengths and limitations, in the context of the purpose, been identified? This should cover both the algorithm and any data used.
Implementation	Consideration of access, implementation, and resource issues when implemented in healthcare settings.
Sensitivity and specificity	For classification algorithms, has the model performance and accuracy (specificity and sensitivity) been appropriately evaluated?
Algorithm characteristics	Has the ML mechanism been clearly characterized and described? Is there sufficient transparency for the results to be reproducible?
Data characteristics	Is the selection of data sets justified and are the key characteristics known? This should extend to training sets, test sets and validation sets.
Explainability	Are the outputs of the algorithm clearly understandable by both the healthcare professional and the patient?

ML indicates machine learning.

false-positives (patient is predicted to be diabetic when not) and false-negatives (patient is predicted not to be diabetic when they are).

The relative consequences of false-positives and false-negatives can be explicitly set using a loss matrix, steering the final algorithm more toward sensitivity or specificity. In the context of regression, we wish to predict the occurrence (binary, multinomial) or magnitude (continuous) of an outcome: for example, "will this patient respond or not respond to treatment given a profile of SNPs and biomarkers?" or "how much will this patient's treatment episode cost given multiple demographic characteristics, measures of disease severity, and functional performance?"

Losses are incurred when more or fewer outcomes are predicted than actually occur, as in the log-loss or Brier score metrics, or when the magnitude of a continuous outcome is more or less than its prediction, as in absolute loss (L1) or squared loss (L2) metrics. A core feature of predictive analytics then is to systematically explore large model spaces (including many potential predictor variables, functional forms, interactions, etc) and select one or more models from that space to generate predictions that minimize loss because of inaccuracy.

In predictive analytics, the algorithm is "trained" on an available sample of data but its performance is assessed on data not used in the training. The most common approach is n-fold cross-validation, which involves partitioning a subset of all of the available data into a training subset fraction (N-1)/N and a validation fraction (1/N) and repeating the process N times so that each observation is in the training subset N-1 times and in the validation subset once. An algorithm is then run to fit the observed outcomes or labels in the training subset. Each algorithm may have embedded within it a method for selecting predictors, functional forms, and interactions (see feature selection section) and is typically controlled by a set of parameters governing variable selection, model construction, and overall complexity.

Predictions are then generated by applying predictors for observations in the validation subset to the algorithm estimated on the training subset. In the single-model approach, the prediction is the result of running the predictor variables for observations in the validation set through the single best-fitting model equation. In the ensemble approach, a set or sequence of models is generated, often using bootstrapped resamples of the training data (sometimes upweighting observations that were poorly predicted in previous iterations) and randomly selected subsets of predictors.[56]

A consensus prediction is generated by running the validation subset of observations through the ensemble of models to obtain multiple predictions and typically selecting the mean or median or last prediction. Performance of the algorithm is assessed by applying the loss metric to the generated predictions and the corresponding true values. Generally speaking, performance varies depending on the parameters used in the algorithm to explore the model space and select one or more models. These parameters can be "tuned" (rerunning the cross-validation procedure with different parameter values) to yield the best performance in the validation subset (or average of the multiple validation subsets in the case of cross-validation and repeated cross-validation).

Final performance of the optimally tuned algorithm can then be assessed by applying the loss metric to a holdout test subset that was not used for algorithm generation or tuning. For true external validation, the tuned algorithm should be applied to a new set of data from a different source.[57] If practiced carefully and transparently, predictive analytics can offer a rigorous and reproducible approach to modeling, particularly with "big data" where the number of potential predictors can be enormous.

Why is the HEOR field investing so much effort in developing and applying new methods to make predictions? First, accurate predictions have high stakes: involving life, death, health-related quality of life, and trillions of dollars of expenditures. To maximize the value of healthcare, a system must deliver the right interventions to the right individuals at the right time and in the right setting. Better predictions are necessary for better decision making.[58]

Second, the "big data" revolution is itself a major driver of innovation in prediction methods in healthcare.[43] The amount of data generated on a continual basis in the course of healthcare delivery and in daily life (eg, through mHealth applications and geo-tracking) go far beyond what we have traditionally seen in the EHR. As the cost of processing and storing biometric, imaging, and genomic (and other -omics) data has fallen, their potential for practical use in predicting health outcomes and improving the value of healthcare has grown. Numerous efforts have bootstrapped themselves seemingly out of the desire to create value by mining these new—and often voluminous and messy—data resources.[59,60]

Third, HEOR practitioners are increasingly aware that traditional predefined, single-model approaches designed for explanation may be less than optimal for prediction.[61] Prioritizing unbiasedness of parameter estimates over predictive accuracy is important for research based on hypothesis testing. Predictive analytics attempts to reconcile the realities of the "bias-variance tradeoff" to find the optimal degree of model complexity that minimizes losses caused by inaccurate predictions.[1] As a field that routinely practices decision analysis for economic evaluation, HEOR practitioners should find the grounding of the predictive analytic approach in statistical decision theory appealing; the more accurately the field is able to predict health outcomes into the future, the more informed are the economics of decision making in the long run for investment purposes.

The algorithms typically used in predictive analytics often mimic processes of biological learning—the most common being supervised learning where feedback for models is obtained and acted upon based on the deviation between predicted and observed outcomes (or classifications/labels).[62] Recently, an increasing number of applications are using the process of self-organized (unsupervised) learning—searching for patterns of association in data without an explicit outcome or classification.[35]

Causal Inference

In contrast to the objectives of predictive analytics, causal inference asks questions about the effects of interventions or policies.[63,64] The "fundamental problem of causal inference" is that for each individual we only observe the outcome corresponding to the treatment actually received, but not the one corresponding to another potential treatment of interest (eg, if the individual had not been treated). Hence, estimating causal parameters always relies on crucial, untestable assumptions.[65] A commonly made assumption is that all the covariates that influence treatment assignment and are prognostic of the outcome have been observed.[66]

ML can play a role in selecting potential confounders, for example, by prescreening of covariates based on variable importance in the outcome model (see Feature selection section) or by using variable selection approaches designed specifically for causal inference.[67,68] Nevertheless, these approaches cannot replace formal causal reasoning (using, eg, directed acyclic graphs) and subject matter knowledge.[69] Sensitivity analysis approaches are increasingly recommended to assess how sensitive estimates of causal effects are to the presence of potential unobserved confounders.[70]

Causal inference in HEOR is often performed by fitting propensity score (PS) models to create matched samples and estimate ATEs by comparing these samples.[13,71] Although logistic regression is often used to fit PS models, it has been shown that "off the shelf" ML methods for prediction and classification (eg, random forests, classification and regression trees, least absolute shrinkage and selection operator) are more flexible and can lead to lower bias in the treatment effect estimates.[12] Nevertheless, these approaches in themselves are imperfect given that they are tailored to minimize root mean square error as opposed to targeting the causal parameter.

Some extensions of these methods have addressed specific challenges of using the PS for confounding adjustment, by customizing the loss function of the ML algorithms (eg, instead of minimizing classification error, to maximize balance in the matched samples).[72] Nevertheless, the issue remains that giving equal importance to many covariates when creating balance may not actually minimize bias (eg, if many of the covariates are only weak confounders). It is recommended that balance on variables that are thought to be the most prognostic to the outcome should be prioritized; nevertheless, this ultimately requires subjective judgment.[73]

An approach specifically designed to overcome this challenge is targeted maximum likelihood estimation (TMLE), also called the targeted minimum loss–based method.[74,75] TMLE was originally developed as a doubly robust approach—a method that gives unbiased estimates of treatment effects if at least one of the PS or outcome regression models is correctly specified—without the consideration of ML. Nevertheless, because of the complexity of trying to specify the exposure and outcome mechanisms, it was seen as optimal to use ML when implementing TMLE.

The super learner is an ensembling ML approach that is recommended to be used with TMLE to help overcome bias because of model misspecification.[63,76,77] Super learning can draw upon the full repertoire of ML and traditional econometric/epidemiological methods and produce estimates that are asymptotically as good as the best performing model—eliminating the need to make strong assumptions about functional form and estimation method up front.[77] TMLE can be used to estimate ATEs, but also more complex causal parameters such as the impact of longitudinal interventions such as dynamic treatment regimens (see Kreif et al[24] and Neugebauer et al[36] in Table 3[5,15-38]).

In an approach similar to TMLE, double/debiased ML estimators use ML to obtain estimates of the PS and outcome regression and combine these using an augmented probability weighted score equation, using sample splitting—estimating the nuisance parameters on different parts of the sample than the treatment effect—to guarantee good asymptotic performance.[78,79]

To meet the needs of HEOR researchers interested in personalized medicine, the causal ML methods introduced earlier can also estimate heterogeneous treatment effects.[80-83] These approaches are particularly promising for HEOR because they can potentially inform individualized treatment decisions ("who to treat") by learning granular treatment effects down to the individual level.[84] A related set of approaches uses ML to directly learn the optimal treatment allocation rule.[84,85]

Economic Evaluation

Health policy and regulatory agencies have vocalized concerns about the use of economic evaluation (eg, cost-effectiveness analysis) for resource allocation of new healthcare services and biomedical technologies in real-world settings given the vast amount of uncertainty and assumptions economic models entail.[86] Although the minimization of bias through trial-based evidence generation remains a gold standard to gain understanding about clinical benefit and comparative effectiveness, randomized-controlled trials often fall short of expectations in HEOR by their inability to represent real-world outcomes to generate the necessary parameters to fulfill the needs of economic models.[87]

Health economic models need to better address the complexity of care delivery by accounting for health service delivery constraints that affect timely delivery such as in cases of reduced capacity.[88] Economic modeling faces an array of uncertainties, including parameter uncertainty, structural uncertainty, and sampling uncertainty.[89]

Parameter Uncertainty

Economic modeling often depends on a patchwork of data parameters from multiple sources that reduce internal model validity. Using a single database, such as EHR and administrative

data to inform probabilities of clinical outcomes and endpoints and the real-world costs of healthcare, could improve upon existing methods of parameter abstraction. ML methods could help to appropriately mine and select features that meet model needs in terms of relevancy and accuracy. Nevertheless, the generalizability of the results would also depend critically on the patient populations represented in the data.

Structural Uncertainty

Studies such as the Dartmouth Atlas have demonstrated that routine clinical practice is characterized by substantial variability rather than guideline-based care.[90] Combining big data with concepts in predictive analytics could facilitate enhanced structure of economic models that encompass more practice variability in healthcare. This would enable models to control for common inefficiencies that would not be accounted for if they only portrayed guideline-based care. Sequencing and unsupervised clustering methods offer a starting point to inform model structure based on the common order of events that patients experience in a clinical care pathway, either through observation or imputation.[91] Furthermore, ML could enhance the ability for models to accurately reflect outcomes over longer terms horizons if the data are calibrated to prespecified time horizons.

Sampling Uncertainty

Sampling uncertainty stems from the lack of available or representative observations from existing data sources. Cohort selection methods empowered with NLP could support economic modeling approaches by mining data on a wider range of individuals who experience real-world outcomes in healthcare that are pertinent to the economic research question. When health outcomes classify as rare events, feature selection methods can aid in expanding on n-of-1 samples to reconfigure economic models to reflect the predictors and endpoints most reflective of health outcomes on a case-by-case basis. In turn, these methodologies offer the field a pathway to more efficient subgroup analysis by drawing diverse observations from which outcomes can either be classified or imputed based on common population-based trends within a contiguous database.

Transparency and Explainability

Transparency is not just desirable but a necessary requirement of decision making in healthcare.[92] Key decision makers—regulators, health technology assessment bodies, and payers—are increasingly embedding elements of transparency in their operating models, such as providing access to documents and data on which decisions are based.[93-95]

Transparency is especially important in the context of ML. The "black box" nature of some ML algorithms (such as deep learning), combined with potentially flawed training sets, may sometimes introduce unintended biases.[96-98] In many cases, these biases reflect patterns in the data stemming from healthcare disparities in access or treatment and wasteful or potentially harmful clinical care.[40]

The application of ML can straddle the whole spectrum from exploratory to hypothesis testing work. The ethical and transparency requirements will be influenced by this, and applications toward hypothesis testing and clinical decision making will require greater transparency and scrutiny.[99] Ultimately, explainability (ie, the extent to which methods and results can be understood by humans) is important to healthcare decision makers and patients. Explainability varies with different methods (Fig. 1); nevertheless, explainability cannot be achieved unless developers are as transparent as possible about methods and execution of ML.

ML is a rapidly evolving field. Several new reporting guidelines and standards have recently appeared, calling for increased transparency around methods, model parameters, and more specifics on training, test, and validation cohorts and standards in the testing of these models in clinical scenarios.[100-103] One area that has not received adequate attention is that of transparency in the application of ML. The TRIPOD-AI and PROBAST-AI reporting standards provide an important structure for reporting that developers of ML should adhere to.[104]

Nevertheless, the stakeholder communities, such as HEOR practitioners, healthcare decision makers, and patients, who are often not experts in ML may have broader concerns about the use of ML in healthcare. As a consequence, this task force has identified key elements in the "PALISADE" Checklist. The checklist provides prompts that developers of ML can use to structure their thinking about how the appropriateness of ML methods can be communicated to stakeholders and healthcare decision makers (Table 5).

Methods in ML That Could Add or Detract From Transparency

Regulators are increasingly asking that decisions made by automated means can be explained and justified. As an example, the 2018 European Union General Data Protection Regulation stipulates the "right to explanation" of decisions from algorithms.[105] The ease with which ML outputs can be explained varies between classes of methods according to their complexity (and performance) and explainability (Fig. 1). Simpler approaches including tree-based methods or gradient boosting machines enable an external auditor to see exactly how the decision problem is being assessed. Conversely, more complex algorithms including neural nets involve so many layers of interactions among model parameters that any intuitive interpretation is hopeless.[106]

The Healthcare Decision Makers' Viewpoint of ML Technologies

There are emerging signs that decision makers including regulatory, health technology assessors, payers, and providers are considering ML and developing assessment frameworks.[107-116] There is an ongoing debate about who is responsible for recognizing and assessing elements relating to safety, quality, efficacy, effectiveness, and appropriate system implementation. Therefore, a multiagency/multistakeholder approach may be useful.

Unlike the single approval process for medicines, a more iterative process of monitoring may also be required. This is due to the uncertainty of performance over time, potential ethical issues that may arise, and the sometimes-evolving nature of both an algorithm and its implementation in practice. Given the impact and sensitivities of ML, the decision maker space should be extended beyond the regulators to include a dialog with patients and society.

Discussion

Potential applications of ML methods in research are numerous and can augment the scientific methods of HEOR. All HEOR objectives start with a research question. Whatever the question may be, ML cohort selection methods can help researchers to identify study samples from big data that are pertinent to the investigation.

Feature selection methods can support HEOR by refining the list of exposure variables to those that have strong associations with clinical or economic outcomes of interest. Feature selection cannot replace the generation of hypotheses that HEOR studies might establish with subject matter experts, but it can complement expected exposures and identify more for further investigation.

The use of ML has the potential to more efficiently curate complex data to be used in a variety of different types of HEOR models. This could certainly benefit economic modelers, whose efforts are often questioned based on the wide-ranging uncertainty and assumptions of such models. ML models complement other emerging HEOR methods such as dynamic simulation models and constrained optimization models.[88,89,117] For example, there is an increasing interest to adopt patient-level dynamic simulation models in precision medicine, to model optimal treatment pathways for individual patients.[118,119]

A challenge that the field of ML will need to overcome is educating clinicians so that they can better understand what ML methods produce and why they should be trusted.[120] There will remain an ongoing balance between using the most rigorous method fit for purpose of the data and research question, combined with consideration for what clinicians and patients are comfortable adopting.

A related issue is the danger that ML algorithms can reinforce bias in healthcare delivery. Underserved populations, by definition, have less representation in healthcare databases. ML algorithms inherently reinforce these patterns in their algorithms. This is an extremely difficult problem to overcome. At a minimum, decision makers using ML coupled with clinical data need to be aware of the potential for bias and proactively seek to overcome it.

Consistent with the criteria for developing ISPOR Good Practices in HEOR, this task force established important general considerations to evaluate the transparency of ML models to support HEOR, which can be found in the PALISADE Checklist.[121] More collaboration between communities of HEOR scientists and computer scientists with ML expertise is encouraged to more rapidly enable learning from each other. Given that the use of ML continues to expand, trust in ML methods will depend on transparency in reporting and advancement in validation techniques to ensure that methods are reproducible.

Article and Author Information

Accepted for Publication: March 25, 2022

Published Online: May 6, 2022

doi: https://doi.org/10.1016/j.jval.2022.03.022

Author Affiliations: Department of Pharmaceutical and Health Economics, School of Pharmacy, University of Southern California, Los Angeles, CA, USA (Padula); The Leonard D. Schaeffer Center for Health Policy & Economics, University of Southern California, Los Angeles, CA, USA (Padula); Centre for Health Economics, University of York, York, England, UK (Kreif); Department of Health Policy and Administration, College of Health and Human Development, Pennsylvania State University, Hershey, PA, USA (Vanness); Flatiron Health, New York, NY, USA (Adamson); AstraZeneca, Cambridge, England, UK (Rueda); Novartis, Basel, Switzerland (Felizzi); National Institute for Health and Care Excellence, Manchester, England, UK (Jonsson); Centre for Health Policy, School of Population and Global Health, University of Melbourne, Melbourne, Australia (IJzerman); School of Medicine, University of California, San Francisco, San Francisco, CA, USA (Butte); The Heller School for Social Policy and Management, Brandeis University, Waltham, MA, USA (Crown).

Correspondence: William Crown, PhD, The Heller School for Social Policy and Management, Brandeis University, 415 South St MS 035, Waltham, MA 02453, USA. Email: wcrown@brandeis.edu; William V. Padula, PhD, USC Schaeffer Center, 635 Downey Way (VPD), Los Angeles, CA 90089, USA. Email: padula@usc.edu

Author Contributions: *Concept and design:* Padula, Kreif, Vanness, Adamson, Rueda, IJzerman, Crown
Acquisition of data: Padula, Adamson
Analysis and interpretation of data: Padula, Adamson, IJzerman
Drafting of the manuscript: Padula, Kreif, Vanness, Adamson, Rueda, Felizzi, Jonsson, IJzerman, Butte, Crown
Critical revision of the paper for important intellectual content: Padula, Kreif, Vanness, Adamson, Rueda, Felizzi, Jonsson, IJzerman, Butte, Crown
Statistical analysis: Padula
Administrative, technical, or logistic support: Padula, Felizzi, Jonsson, Butte, Crown
Supervision: Padula
Other (Figure development): Felizzi

Conflict of Interest Disclosures: Dr Padula reported receiving grants from the National Institutes of Health/Office of Extramural Research during the conduct of the study and reported receiving personal fees from Monument Analytics outside the submitted work. Dr Adamson reported receiving personal fees from Flatiron Health and Infectious Economics and reported stock ownership in Roche. Dr Rueda is employed by AstraZeneca. Dr Felizzi reported personal fees from Novartis and Roche and reported stock ownership in Roche. Dr IJzerman reported receiving grants and speaker fees from Illumina outside the submitted work. Dr Butte reported receiving personal fees from Samsung, Mango Tree Corporation, 10x Genomics, Helix, Pathway Genomics, Verinata (Illumina), Personalis NuMedii, Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck, Roche, Johnson and Johnson, Pfizer, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, and Westat; reported stock ownership in Personalis, NuMedii, Apple, Facebook, Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, CVS, Nuna Health, Assay Depot, Vet24seven, Regeneron, Sanofi, Royalty Pharma, AstraZeneca, Moderna, Biogen, Parexel, and Sutro; and reported royalties and stock from Stanford University for several patents and other disclosures licensed to NuMedii and Personalis outside the submitted work. Drs Padula, Vanness, and IJzerman are editors for *Value in Health* and had no role in the peer-review process of this article. No other disclosures were reported.

Funding/Support: Dr Padula declares support by an unrestricted grant during the conduct of this task force from the US National Institutes of Health (KL2 TR001854). There are no other funding sources to declare.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgment: Collaborator: Suzanne Belinson, PhD (Tempus Inc, Chicago, IL). The task force extends a special thanks to Elizabeth Molsen and Hanke Zheng for their administrative support of this effort. The task force wishes to acknowledge the efforts of many referees who participated in the peer-review process of this manuscript: Adrian Jinich, Luke Benson, Sang Kyu Cho, Catharina Groothuis-Oudshoorn, Tadesse Abegaz, Sandipan Bhattacharjee, Richard Birnie, James Murray, Doug Faries, Alan Brnabic, Jenny Lo-Ciganic, Juliana Setyawan, Dawn Lee, Gurmit Sandhu, Ilya Lipkovich, Alexandre Batisse, Bill Marder, and Karl Patterson.

REFERENCES

1. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Germany: Springer Science and Business Media; 2009.
2. Wiens J, Saria S, Sendak M, et al. Do no harm: a road map for responsible machine learning for health care [published correction appears in *Nat Med*. 2019;25(10):1627]. *Nat Med*. 2019;25(9):1337–1340.
3. Crown WH. Real-world evidence, causal inference, and machine learning. *Value Health*. 2019;22(5):587–592.
4. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak*. 2010;10:16.

5. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
7. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine Learning in Health Care: a critical appraisal of challenges and opportunities. *EGEMS (Wash DC)*. 2019;7(1):1.
8. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health*. 2015;18(2):137–140.
9. Titiunik R. Can big data solve the fundamental problem of causal inference? *PS Pol Sci Pol*. 2015;48(1):75–79.
10. Berger ML, Curtis MD, Smith G, Harnett J, Abernethy AP. Opportunities and challenges in leveraging electronic health record data in oncology. *Future Oncol*. 2016;12(10):1261–1274.
11. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof*. 2014;10:12.
12. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546–555.
13. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010;63(8):826–833.
14. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint. Posted online October 11, 2018. arXiv:181004805. https://doi.org/10.48550/arXiv.1810.04805.
15. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211–2223.
16. Nori VS, Hane CA, Crown WH, et al. Machine learning models to predict onset of dementia: a label learning approach. *Alzheimers Dement (N Y)*. 2019;5:918–925.
17. Cole TS, Frankovich J, Iyer S, Lependu P, Bauer-Mehren A, Shah NH. Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for EHR-based research. *Pediatr Rheumatol Online J*. 2013;11(1):45.
18. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One*. 2018;13(7):e0201016.
19. Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform*. 2015;56:229–238.
20. Xu H, AbdelRahman S, Lu Y, Denny JC, Doan S. Applying semantic-based probabilistic context-free grammar to medical language processing–a preliminary study on parsing medication sentences. *J Biomed Inform*. 2011;44(6):1068–1075.
21. Jiao W, Atwal G, Polak P, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat Commun*. 2020;11(1):728.
22. Liu M, Shah A, Jiang M, et al. A study of transportability of an existing smoking status detection module across institutions. *AMIA Annu Symp Proc*. 2012;2012:577–586.
23. Padula WV, Pronovost PJ, Makic MBF, et al. Value of hospital resources for effective pressure injury prevention: a cost-effectiveness analysis. *BMJ Qual Saf*. 2019;28(2):132–141.
24. Kreif N, Tran L, Grieve R, De Stavola B, Tasker RC, Petersen M. Estimating the comparative effectiveness of feeding interventions in the pediatric intensive care unit: a demonstration of longitudinal targeted maximum likelihood estimation. *Am J Epidemiol*. 2017;186(12):1370–1379.
25. Padula WV, Gibbons RD, Pronovost PJ, et al. Using clinical data to predict high-cost performance coding issues associated with pressure ulcers: a multilevel cohort model. *J Am Med Inform Assoc*. 2017;24(e1):e95–e102.
26. Arora P, Boyne D, Slater JJ, Gupta A, Brenner DR, Druzdzel MJ. Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Value Health*. 2019;22(4):439–445.
27. Liu M, Wu Y, Chen Y, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc*. 2012;19(e1):e28–e35.
28. Xu H, Fu Z, Shah A, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu Symp Proc*. 2011;2011:1564–1572.
29. Sohn S, Kocher JP, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc*. 2011;18(suppl 1):i144–i149.
30. Birnbaum B, Nussbaum N, Seidl-Rathkopf K, et al. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. Preprint. Posted online January 13, 2020. arXiv 2001.09765v1. https://doi.org/10.48550/arXiv.2001.09765
31. Hansen PW, Clemmensen L, Sehested TS, et al. Identifying drug–drug interactions by data mining: a pilot study of warfarin-associated drug interactions. *Circ Cardiovasc Qual Outcomes*. 2016;9(6):621–628.
32. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of Machine Learning Methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med*. 2016;44(2):368–374.
33. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015;7(299):299ra122.
34. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev*. 2019;11(1):31–39.
35. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6:26094.
36. Neugebauer R, Schmittdiel JA, van der Laan MJ. Targeted learning in real-world comparative effectiveness research with time-varying interventions. *Stat Med*. 2014;33(14):2480–2520.
37. Kempa-Liehr AW, Lin CY, Britten R, et al. Healthcare pathway discovery and probabilistic machine learning. *Int J Med Inform*. 2020;137:104087.
38. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*. 2021;592(7855):629–633.
39. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221–230.
40. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453.
41. Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A. A Machine Learning Approach for identifying gene biomarkers guiding the treatment of breast cancer. *Front Genet*. 2019;10:256.
42. Zucker DR, Ruthazer R, Schmid CH. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *J Clin Epidemiol*. 2010;63(12):1312–1323.
43. Bellman R. Dynamic programming. *Science*. 1966;153(3731):34–37.
44. Nilsson R, Peña JM, Björkegren J, Tegner J. Consistent feature selection for pattern recognition in polynomial time. *J ML Research*. 2007;8:589–612.
45. Liu H, Motoda H. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Berlin, Germany: Springer Science and Business Media; 1998.
46. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformatics*. 2015;2015:198363.
47. Kohavi R, Johnb GH. Wrappers for feature subset selection. *Artif Intell*. 1997;97(1-2):273–324.
48. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Med Res*. 2003;3:1157–1182.
49. Suto J, Oniga S, Sitar PP. Comparison of wrapper and filter feature selection algorithms on human activity recognition. IEEE Xplore. https://ieeexplore.ieee.org/document/7496749. Accessed May 1, 2022.
50. Talavera L. An evaluation of filter and wrapper methods for feature selection in categorical clustering. In: Famili AF, Kok JN, Peña JM, Siebes A, Feelders A, eds. *Advances in Intelligent Data Analysis VI. IDA 2005. Lecture Notes in Computer Science*. Vol 3646. Berlin, Germany: Springer; 2005.
51. Zhao Y, Long Q. Variable selection in the presence of missing data: imputation-based methods. Wiley Interdiscip Rev Comp Stat;9(5):e1402.
52. Seijo-Pardo B, Alonso-Betanzos A, Bennett K, et al. Analysis of imputation bias for feature selection with missing data. Paper presented at: Eur Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning; 25-27 April, 2018; Bruges, Belgium.
53. Van Calster B, Verbakel JY, Christodoulou E, Steyerberg EW, Collins GS. Statistics versus machine learning: definitions are interesting (but understanding, methodology, and reporting are more important). *J Clin Epidemiol*. 2019;116:137–138.
54. Boulesteix AL, Schmid M. Machine learning versus statistical modeling. *Biom J*. 2014;56(4):588–593.
55. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist Sci*. 2001;16(3):199–231.
56. Dietterich TG. *Ensemble methods in machine learning*. In: *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*. Vol 1857. Berlin, Germany: Springer; 2000.
57. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247.
58. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216–1219.
59. DREAM7. DREAM7 challenge. NCI. http://dreamchallenges.org/. Accessed May 1, 2022.
60. COVID-19 DREAM challenge — syn21849255. Synapse. https://www.synapse.org/#!Synapse:syn21849255/wiki/601865. Accessed May 13, 2020.
61. Shmueli G. To explain or to predict? *Statist Sci*. 2010;25(3):289–310.
62. Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC Med Inform Decis Mak*. 2021;21(1):54.
63. Van der Laan MJ, Rose S. *Targeted Learning in Data Science*. Cham, Switzerland: Springer International Publishing; 2018.
64. Pearl J. *Causality*. Cambridge, United Kingdom: Cambridge University Press; 2009.
65. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. *Int J Epidemiol*. 2021;49(6):2058–2064.
66. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*. 2014;25(3):418–426.
67. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res*. 2012;21(1):7–30.

68. Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. *Rev Econ Stud*. 2014;81(2):608–650.
69. Díaz I. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*. 2020;21(2):353–358.
70. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*. 2011;22(1):42–52.
71. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
72. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9(4):403–425.
73. Ramsahai R, Grieve R, Sekhon J. Extending iterative matching methods: an approach to improving covariate balance that allows prioritisation. *Health Serv Outcomes Res Methodol*. 2011;11(3):95–114.
74. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185(1):65–73.
75. Van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer-Verlag; 2011.
76. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol*. 2011;173(7):761–767.
77. Van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):11.
78. Zheng W, Van der Laan MJ. Cross-validated targeted minimum-loss-based estimation. In: *Targeted Learning. Springer Series in Statistics*. New York, NY: Springer; 2011:459–474.
79. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econ J*. 2018;21(1):C1–C68.
80. Luedtke AR, van der Laan MJ. Evaluating the impact of treating the optimal subgroup. *Stat Methods Med Res*. 2017;26(4):1630–1640.
81. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Statist*. 2019;47(2):1148–1178.
82. Scarpa J, Bruzelius E, Doupe P, Le M, Faghmous J, Baum A. Assessment of risk of harm associated with intensive blood pressure management among patients with hypertension who smoke: a secondary analysis of the systolic blood pressure intervention trial [published correction appears in *JAMA Netw Open*. 2019;2(4):e193146]. *JAMA Netw Open*. 2019;2(3):e190005.
83. Bress AP, Greene T, Derington CG, et al. Patient selection for intensive blood pressure management based on benefit and adverse events. *J Am Coll Cardiol*. 2021;77(16):1977–1990.
84. Murphy SA. Optimal dynamic treatment regimes. *J R Stat Soc B*. 2003;65(2):331–366.
85. 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.
86. Padula WV, Sculpher MJ. Ideas about resourcing health care in the United States: can economic evaluation achieve meaningful use? *Ann Intern Med*. 2021;174(1):80–85.
87. Davis C, Naci H, Gurpinar E, Poplavska E, Pinto A, Aggarwal A. Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009-13. *BMJ*. 2017;359:j4530.
88. Marshall DA, Burgos-Liz L, IJzerman MJ, et al. Applying dynamic simulation modeling methods in health care delivery research-the SIMULATE checklist: report of the ISPOR simulation modeling emerging good practices task force. *Value Health*. 2015;18(1):5–16.
89. Briggs AH. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics*. 2000;17(5):479–500.
90. Wennberg J, Gittelsohn. Small area variations in health care delivery. *Science*. 1973;182(4117):1102–1108.
91. Hougham GW, Ham SA, Ruhnke GW, et al. Sequence patterns in the resolution of clinical instabilities in community-acquired pneumonia and association with outcomes. *J Gen Intern Med*. 2014;29(4):563–571.
92. Paschke A, Dimancesco D, Vian T, Kohler JC, Forte G. Increasing transparency and accountability in national pharmaceutical systems. *Bull World Health Organ*. 2018;96(11):782–791.
93. NHS constitution for England. Department of Health & Social Care. https://www.gov.uk/government/publications/the-nhs-constitution-for-england. Accessed May 3, 2020.
94. FDA transparency initiative overview. U.S. Food and Drug Administration. https://www.fda.gov/about-fda/transparency. Accessed May 4, 2020.
95. How we work. European Medicines Agency. https://www.ema.europa.eu/en/about-us/how-we-work/transparency. Accessed May 4, 2020.
96. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of Machine Learning in medicine. *JAMA*. 2017;318(6):517–518.
97. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in Machine Learning Algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544–1547.
98. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med*. 2018;378(11):981–983.
99. Orsini LS, Berger M, Crown W, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing-why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Value Health*. 2020;23(9):1128–1136.
100. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320–1324.
101. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–1363.
102. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–1374.
103. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. 2021;27(2):186–187.
104. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
105. European Union General Data protection regulation. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1552577087456&uri=CELEX:32018R1725. Accessed May 4, 2020.
106. Turek M. Explainable artificial intelligence (XAI). https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed May 4, 2020.
107. AI for healthcare: creating an international approach together. Global Digital Health Partnership. https://gdhp.nhp.gov.in/assets/pdf/WhitePapers2020/GDHP_PolicyWorkStreams_AIWhitePaper_December_2020.pdf. Accessed January 4, 2021.
108. Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence. https://ec.europa.eu/futurium/en/ai-alliance-consultation. Accessed May 4, 2020.
109. Draft analysis grid for the evaluation of medical devices embedding artificial intelligence. Clôture de la consultation Le. https://www.has-sante.fr/jcms/p_3118247/en/draft-analysis-grid-for-the-evaluation-of-medical-devices-embedding-artificial-intelligence. Accessed January 4, 2021.
110. Buston O, Nika S, Fenech M. Guidance for princeple 7 of the NHS code of conduct for data-drive health and care technology: explainability in data-driven health and care technology. Future Advocacy. https://futureadvocacy.com/wp-content/uploads/2019/10/Explainability-in-data-driven-health-and-care-technology-Future-Advocacy-edited.pdf. Accessed May 4, 2020.
111. A guide to good practice for digital and data-driven health technologies. Department of Health & Social Care. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology. Accessed January 24, 2021.
112. Project ExplAIn interim report. Information Commissioner's Office. https://ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/. Accessed January 24, 2021.
113. Evidence standards framework for digital health technologies. National institute for Health and Care Excellence. https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies. Accessed May 4, 2020.
114. Artificial intelligence and MLin software as a medical device. US Food & Drug Administration. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. Accessed May 4, 2020.
115. An overview of clinical applications of artificial intelligence. Canadian Agency for Drugs and Technologies in Health. https://cadth.ca/dv/ieht/overview-clinical-applications-artificial-intelligence. Accessed January 24, 2021.
116. Artificial intelligence for health in New Zealand. AI Forum of New Zealand. https://aiforum.org.nz/wp-content/uploads/2019/10/AI-For-Health-in-New-Zealand.pdf. Accessed January 24, 2021.
117. Chakraborty B, Murphy SA. Dynamic treatment regimes. *Annu Rev Stat Appl*. 2014;1:447–464.
118. Degeling K, Koffijberg H, IJzerman MJ. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. *Expert Rev Pharmacoecon Outcomes Res*. 2017;17(1):17–25.
119. Marshall DA, Grazziotin LR, Regier DA, et al. Addressing challenges of economic evaluation in precision medicine using dynamic simulation modeling. *Value Health*. 2020;23(5):566–573.
120. Padula WV, McQueen RB, Pronovost PJ. Can economic model transparency improve provider interpretation of cost-effectiveness analysis? Evaluating tradeoffs presented by the second panel on cost-effectiveness in health and medicine. *Med Care*. 2017;55(11):909–911.
121. Malone DC, Ramsey SD, Patrick DL, et al. Criteria and process for initiating and developing an ISPOR good practices task force report. *Value Health*. 2020;23(4):409–415.