Building the Model

Challenges and Considerations of Developing and Implementing Machine Learning Tools for Clinical Laboratory Medicine Practice

He S. Yang, PhD; Daniel D. Rhoads, MD; Jorge Sepulveda, MD, PhD; Chengxi Zang, PhD; Amy Chadburn, MD; Fei Wang, PhD

• ***Context.***—**Machine learning (ML) allows for the analysis of massive quantities of high-dimensional clinical laboratory data, thereby revealing complex patterns and trends. Thus, ML can potentially improve the efficiency of clinical data interpretation and the practice of laboratory medicine. However, the risks of generating biased or unrepresentative models, which can lead to misleading clinical conclusions or overestimation of the model performance, should be recognized.**

Objectives.—**To discuss the major components for creating ML models, including data collection, data preprocessing, model development, and model evaluation. We also highlight many of the challenges and pitfalls in developing ML models, which could result in misleading clinical impressions or inaccurate model performance, and provide suggestions and guidance on how to circumvent these challenges.**

Data Sources.—**The references for this review were identified through searches of the PubMed database, US Food and Drug Administration white papers and guidelines, conference abstracts, and online preprints.**

Conclusions.—**With the growing interest in developing and implementing ML models in clinical practice, laboratorians and clinicians need to be educated in order to collect sufficiently large and high-quality data, properly report the data set characteristics, and combine data from multiple institutions with proper normalization. They will also need to assess the reasons for missing values, determine the inclusion or exclusion of outliers, and evaluate the completeness of a data set. In addition, they require the necessary knowledge to select a suitable ML model for a specific clinical question and accurately evaluate the performance of the ML model, based on objective criteria. Domain-specific knowledge is critical in the entire workflow of developing ML models.**

(*Arch Pathol Lab Med.* 2023;147:826–836; doi: 10.5858/arpa.2021-0635-RA)

Machine learning (ML)[1] has emerged as a powerful tool for analyzing and interpreting laboratory test results as well as integrating clinical findings with laboratory data.[2] In recent years, there has been a surge of interest in using ML in the clinical laboratory to predict the accuracy of measured laboratory results,[2] identify preanalytical errors,[3–5] interpret complicated biochemical laboratory panels[6–8] and molecular results such the polymerase chain reaction (PCR),[9] establish population-specific reference intervals,[10,11] automate analysis of digital images,[12–14] improve test utilization,[15,16] enhance quality review,[17] and predict the onset and behavior of human diseases.[18–20] Selected examples are listed in the Table. The objective interpretation provided by ML can also be used as a tool to support decisions in laboratory medicine practice.[21–23] The clinical laboratory, at the interface of massive amounts of patient data and objective and subjective clinical practice, is the optimal area of medicine to lead the development, implementation, and integration of ML models for patient care.

Laboratory medicine is data rich owing to the enormous volume of laboratory test results produced by different sections of the clinical laboratory.[24] It is estimated that up to 70% of the data in the electronic health record (EHR) are derived from the clinical laboratory.[25] Most of these data are test results reported as individual numerical or categorical values in a structured format. Patient laboratory test profiles are high-dimensional data sets, as each patient usually has multiple individual laboratory test results generated from a single physician visit as well as longitudinal test results to monitor "wellness" status or to follow one or more disease processes.[2] The enormity of the data, including the number of tests and interdependent multidimensional relationships of different test results, is difficult for us, as humans, to

Accepted for publication May 24, 2022.
Published online October 10, 2022.

From the Departments of Pathology and Laboratory Medicine (Yang, Chadburn) and Population Health Sciences (Zang, Wang), Weill Cornell Medicine, New York, New York; the Department of Laboratory Medicine, Cleveland Clinic, Cleveland, Ohio (Rhoads); the Department of Pathology, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, Ohio (Rhoads); and the Department of Pathology, School of Medicine and Health Sciences, George Washington University, Washington, District of Columbia (Sepulveda).

The work of Wang and Zang is supported by NSF 1750326, NIH R01MH124740, and RF1AG072449.

The authors have no relevant financial interest in the products or companies described in this article.

Corresponding author: Fei Wang, PhD, Division of Health Informatics, Department of Population Health Sciences, Weill Cornell Medicine, 425 East 61st Street, New York City, NY 10065 (email: few2001@med.cornell.edu).

Representative Examples of Machine Learning Studies in Clinical Laboratory Medicine

Source, y	Task	Data Source and Size (Training and Test Sets)	ML Model	Evaluation Metric of Model Performance	Independent External Validation
Chabrun et al,[71] 2021	Interpretation of SPE	Retrospective SPE data set of 159 969 entries	Deep learning models	AUROC, accuracy, sensitivity, specificity, repeatability compared to human expert annotation	External test set from a different laboratory
Eisenhofer et al,[66] 2020	Identification and subtype classification of PA	Retrospective analysis of 462 patients tested for PA and 201 patients with hypertension	RF model for identification of PA SVM model for subtype classification	AUROC, F1 score, sensitivity, specificity	No additional external set
Luo et al,[2] 2016	Prediction of ferritin results from other laboratory test results	Retrospective analysis of laboratory test results in 5128 outpatient cases	MissForest imputation and Lasso regression for prediction of ferritin values MissForest imputation and LR model for prediction of ferritin classification	AUROC, regression, bias plot of measured ferritin versus predicted ferritin	No external validation set
Mathison et al,[13] 2020	Detection of intestinal protozoa in trichrome-stained stool specimens	Retrospective analysis of digital slide scanning of 11 classes of protozoa	DCNN	Precision-recall, confidence class chart, accuracy compared to manual microscopy examination	No external validation set
Rosenbaum and Baron,[3] 2018	Identification of "wrong blood in tube" errors	Retrospective analysis of 20 638 patient collections from 4837 patients	SVM model	AUROC, specificity, and positive predictive value	No external validation set
Than et al,[18] 2019	Prediction of the likelihood of acute myocardial infarction in ED patients	Prospective data collection from 11 001 patients in multiple centers	Gradient boosting model	AUROC, sensitivity, NPV, specificity, PPV compared to clinically adjudicated diagnosis	No additional external validation set
Wang et al,[12] 2021	Identification and classification of bacterial vaginosis	Retrospective analysis of 29 095 microscopic images and associated medical records	Convolutional neural network	AUROC, confusion matrix, accuracy	Independent test set of 1082 images
Wilkes et al,[7] 2018	Automated interpretation of urine steroid profiles	Retrospective data collection of 4619 urine steroid profile	RF model, WSRF model, and XGBT model	Nested CV, AUROC, confusion matrix; Boruta algorithm to assess feature importance	No external validation set
Wilkes et al,[6] 2020	Automated interpretation of PAA profile	Retrospective data collection of 2084 anonymized PAA profile	The ensemble of RF model, WSRF model, and XGBT model	Area under the precision-recall curve, confusion matrix, AUROC	No external validation set
Yang et al,[61] 2020	Prediction of SARS-CoV-2 infection	Retrospective data collection from 3356 SARS-CoV-2 RT-PCR–tested patients	GBDT model	AUROC, sensitivity, specificity, accuracy SHAP force plot to assess feature importance	External validation data set from a different hospital
Yang et al,[60] 2021	Longitudinal analysis of laboratory test result profiles in SARS-CoV-2–positive and negative patients	Retrospective data collection from 5785 SARS-CoV-2 RT-PCR–tested patients	UMAP analysis	NA	No external validation set
Yu et al,[17] 2019	Quality review of mass spectrometry data of THC-COOH results	Retrospective data obtained from gas chromatography in 1267 urine samples	SVM model	Recall and precision, confusion matrix, compared to manual result verification	No external validation set
Zhang et al,[15] 2020	Improve test utilization of PBFC	Retrospective analysis of 784 PBFC cases with concurrent or recent CBC/differential order	DT model LR model	AUROC, sensitivity, specificity, false-positive and false-negative rate	No external validation set

Abbreviations: AUROC, area under the receiver operating characteristic curve; CBC, complete blood count; DCNN, deep convolutional neural network; DT, decision tree; ED, emergency department; GBDT, gradient-boosted decision tree; LR, logistic regression; ML, machine learning; NA, not applicable; nested CV, nested, stratified, k-fold repeated cross-validation; NPV, negative predictive value; PA, primary aldosteronism; PAA, plasma amino acid; PBFC, peripheral blood flow cytometry; PPV, positive predictive value; RF, random forest; RT-PCR, reverse transcription–polymerase chain reaction; SHAP, Shapley additive explanations; SPE, serum protein electrophoresis; SVM, support vector machine; THC-COOH: 11-Nor-9-carboxy-Δ^9-tetrahydrocannabinol; UMAP, unified manifold approximation and projection; WSRF, weighted-subspace random forest; XGBT, extreme gradient boosted tree.

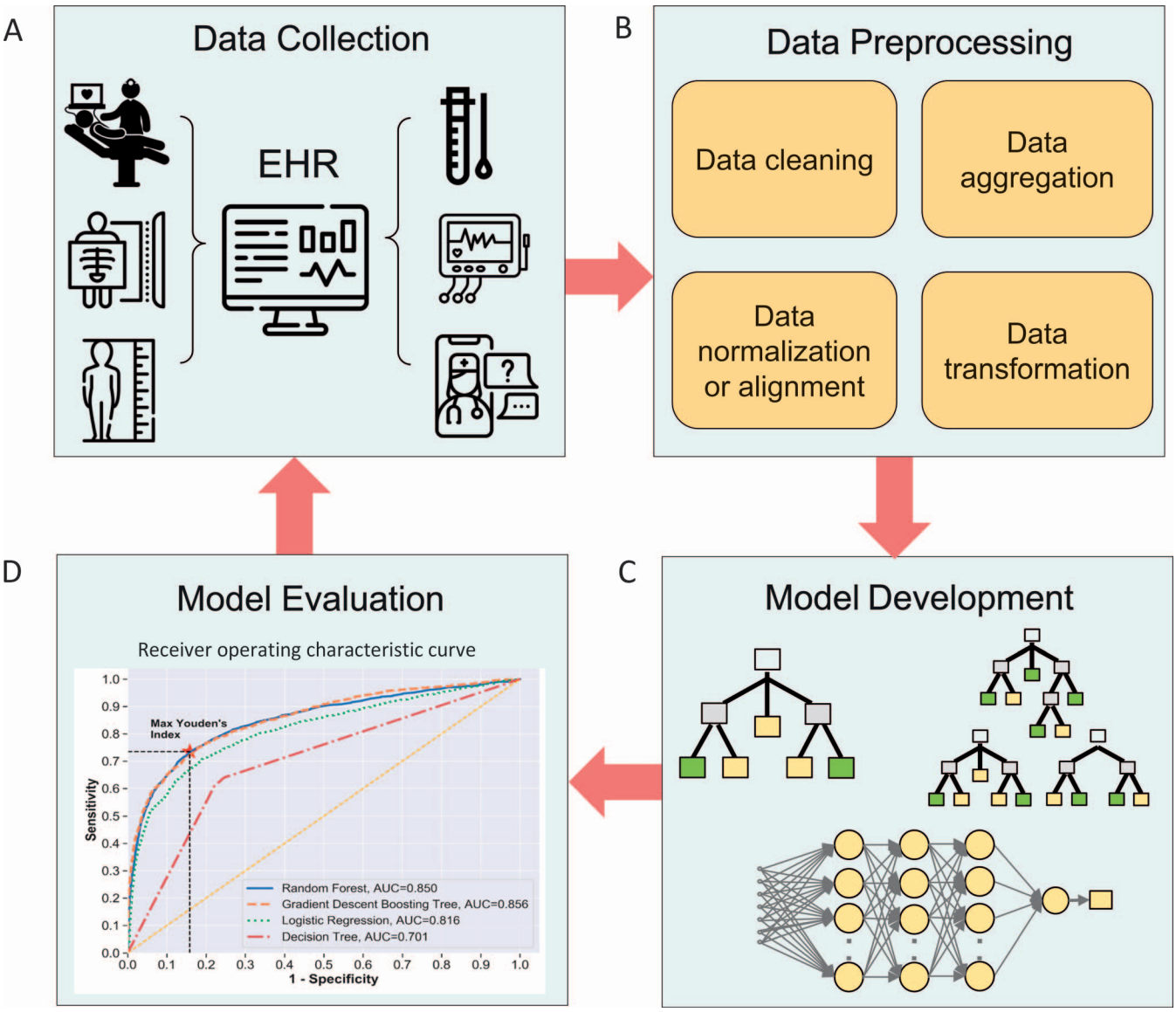

Figure 1. *Four major components in the workflow of developing machine learning (ML) models: data collection (A), data preprocessing (B), ML model development (C), and ML model evaluation (D). The analysis of laboratory data may require several iterations of the above steps. Abbreviations: AUC, area under the receiving operating characteristic curve; EHR, electronic health record.*

interpret without computational assistance. To compensate for this complexity, medicine often uses data simplification methods, including scoring tools, such as APACHE[26] and MELD.[27] However, optimal ML algorithms, which can evaluate larger data sets, have the potential to more accurately and automatically predict disease acuity, diagnosis, and prognosis. An ML approach could also identify vitally important minor variables that are not incorporated in current manual scoring models. Thus, ML technology not only offers tools to analyze massive quantities of clinical data but also can discover "hidden" patterns behind the data.

The development of ML models for laboratory data analysis requires 4 major components: data collection, data preprocessing, model development, and model evaluation (Figure 1, A through D). Although there is an increasing number of reports on a variety of applications for ML in the clinical laboratories,[23,28] only a few publications have focused on the challenges and pitfalls of each of these components. While ML holds tremendous potential to improve laboratory medicine practice efficacy, there is also a risk of generating biased or unrepresentative models, leading to misguided conclusions or overoptimistic estimations of model performance. Furthermore, despite familiarity with traditional data approaches, many laboratory professionals are not familiar with the workflow of ML analysis, resulting in a knowledge gap with respect to the development, understanding, and use of ML models. To address these issues, this review will discuss key considerations for each of the 4 major components of ML model development, highlight the challenges of developing and evaluating an ML model, and discuss the steps for creating a new ML algorithm for use in the clinical laboratory setting.

DATA COLLECTION

The collection of sufficiently large high-quality data sets is of paramount importance for building an ML model. An

ideal training data set should cover the variability of samples across the demographic and geographic spectrum of the patients served by the laboratory, as well as the different aspects of their diseases, including associated comorbidities and temporal variations. An example of such a comprehensive training data set, collected by Cohen et al,[10] was to build personalized models consisting of 2.1 billion laboratory results obtained from 92 different laboratory tests performed for 2.8 million adults during a span of 18 years. The multivariate longitudinal analysis, based on trajectories of the patients' within-normal laboratory test results, predicted individual patient risks of future laboratory abnormalities and subsequent related diseases. The model had the added value of establishing age-adjusted reference ranges for a variety of laboratory tests. In contrast, insufficient or nonrepresentative training data may lead to discrepant model performance between the training set and an independent test data set. The size of the data set required by a specific ML model depends on its complexity, that is, the number of parameters to be included in the model. Complex ML models, such as deep neural networks,[29] built upon insufficient training samples, tend to be sensitive to changes in data distribution. As such, the model cannot be generalized to the real-world setting, as it would lead to unintended bias in decision-making. For example, a model predicting acute kidney injury would be biased if a disproportionate amount of data was collected from White patients without sufficient data from other racial groups. Similarly, a prediction model of COVID-19 disease progression would not perform well in the intensive care unit setting if the training set was collected from outpatients. Therefore, the completeness, quality, and appropriateness of the data set should be carefully evaluated before training an ML model.

Equally important is the transparent reporting of data set characteristics from which an ML model is trained, as these data directly affect the reproducibility, generalizability, and interpretation of an ML model.[30,31] However, to the best of our knowledge, there are no specific guidelines for the development and application of ML in laboratory medicine. Reporting guidelines, such as the 25-item Consolidated Standards of Reporting Trials (CONSORT)[32] and the 33-item checklist on the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT),[33] which mainly apply to clinical trials or health care studies in general, were recently extended to include clinical trials evaluating interventions with an artificial intelligence (AI) component.[34] In addition, the MINimum Information for Medical AI Reporting (MINIMAR) guideline has been proposed for general AI models developed for enhancing "clinical decision-making for diagnosis, treatment and prognosis."[30] The MINIMAR guideline outlines the essential components that should be reported, including the study population and setting, data source and cohort selection criteria, and patient demographic characteristics such as age, sex, race, ethnicity, and socioeconomic status. These factors are also crucial in reporting the ML models developed for clinical laboratory medicine. Moreover, the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement also provides guidelines,[35] in the form of a 22-item checklist, for the standardization of ML model reporting. The TRIPOD guidelines were used by Than et al[18] in their study using ML to predict myocardial infarction.

Collection of a sufficiently large and comprehensive data set from a single institution is sometimes challenging. Thus, integration of clinical laboratory data from multiple sites may be necessary. The establishment of clinical collaborative consortia involving multiple institutions, such as the Observational Health Data Science and Informatics (OHDSI)[36] and the National Patient-Centered Clinical Research Network (PCORNET),[37] is vitally important. While data from multiple institutions could improve the performance of ML models, the variations among the different laboratories, including instrument platforms, test methodologies, test reagents, and sample handling, pose technical difficulties for data aggregation. For example, the mean test result of neonatal total bilirubin from the same proficiency testing sample performed on several different chemistry analyzers using diverse methodologies ranged from 14.96 to 19.77 mg/dL.[38] Therefore, before combining interinstitutional data, it is important to understand the distribution, that is, mean and variation, of each laboratory test as well as the number of samples from each platform to align the results across multiple platforms.

The issue of laboratory data normalization is complex, and there is no universally agreed-upon standardized approach. If there is no proportional bias between assays, this can be accomplished by using data normalization (eg, with z-score, standard deviation units, multiple of median, reference change values) with respect to each assay before aggregation. If there are parallel comparison studies and the assays are linear, linear regression can be used, with the slope of the regression line and the intercept adjusting for proportional bias and systematic bias, respectively.[39] If there are no method comparison studies, normalization can be achieved by harmonizing distributions and reference intervals,[40] although these normalization methods may not preserve all the information content of the original results.[41]

Vital sign names, laboratory test names, and measurement units should also be normalized between all facilities from which the cohorts are derived, using appropriate ontologies, such as the Logical Observation Identifier Names and Codes (LOINC; https://loinc.org, accessed June 15, 2022),[42,43] Ontology of Units of Measure,[44] or Disease Ontology.[45] In the data collection step, domain expertise is necessary to determine if the results of an analyte determined by different methodologies can be combined. For example, tacrolimus levels in transplant patients, measured by immunoassay, may require complex normalization to be combined with results obtained by mass spectrometry, as the latter assay is specific to the tacrolimus molecule, whereas the former assay cross-reacts with tacrolimus metabolites.[46,47]

Differences in ordering practices can introduce bias when aggregating data from multiple institutions. One approach is to analyze ordering practices separately to ascertain bias by using mean abnormal result rates[48] or Bayesian test utilization statistics.[49] Another potential source of bias is relying on labeling designed for billing, such as the International Classification of Diseases, which may not be accurate or contain enough details for the purposes of the ML project. Manual review of health records and automated natural language processing tools to extract the relevant information from the EHR[50] may result in better data for ML development.

An algorithm for the interpretation of primary data, such as digital images, can be developed by using data from the laboratory in which the algorithm will be deployed.[9,13]

Interpretations may be different in different laboratory practices, such as the microbiology laboratory.[9] If the ML tool is intended to be used as an aid in the interpretation in multiple laboratories, then additional preprocessing steps are usually needed in the workflow to ensure data inputs are interpreted similarly and accurately.[12] In addition, uncommon findings can confound model interpretation if the model has not been trained how to interpret them.[51] Uncommon inputs can be amplified in the training set by augmenting the limited training inputs that are available.[9,13]

DATA PREPROCESSING

The quality of data is also critical for the development of an optimal ML model.[52] Sensible model performance stems primarily from high-quality training data with a minimum number of missing values and outliers. Missing values can result from systematic missingness and/or random missingness. For instance, laboratory tests are not always ordered on a regular basis for all patients in clinical practice, especially with different institutional practices, leading to systematic missing values. This "missingness" can be mitigated by ML approaches. Systematic missing values, however, can be informative in identifying temporal patterns of laboratory ordering missingness.[53] In contrast, random missing values occur when a laboratory test cannot be performed or reported owing to preanalytical and analytical issues, such as the wrong tube type, insufficient sample volume, substance interference, or sample out-of-stability range (time or temperature). Thus, the missing value rate of each laboratory test should be assessed before training an ML model. For laboratory tests with a missing rate exceeding a certain threshold (eg, 30%), an investigation into potential reasons for the high missing rate and an evaluation of the reliability of the remaining available data are necessary to determine if the data are suitable for model training. Our investigation of laboratory test result profiles in patients with SARS-CoV-2 infection showed that the missing data rate of specific laboratory tests changed over time from the initial COVID-19 outbreak in March and April 2020 to a period of declining infections in May and June 2020. Particularly in our hospital, inflammatory markers, such as ferritin, C-reactive protein, and procalcitonin, were ordered frequently for patients with COVID-19 in March and April but were ordered less frequently in May and June. Thus, the missing rate of these tests was higher in the post-apex phase of COVID-19. One should also determine whether it is reliable to impute the missing values from the remaining results by using techniques such as multiple imputation.[54]

In addition, outlier laboratory results, specifically those exceeding 3 times the standard deviation from the mean of that specific test if the distribution is normal, can occur owing to (1) pathophysiologic reasons, for example, exceptionally elevated creatinine kinase level in rhabdomyolysis cases[55] or unusually high troponin level in the setting of septic shock[56]; (2) preanalytical errors, for example, wrong tube type or sample out of stability[57]; or (3) analytical errors such as those caused by instrument or reagent issues. These outliers can skew the ML process. It is therefore recommended to check the distributions of laboratory results, through statistical means such as a box plot or a violin plot, to identify the outliers before initiating the process of developing an ML model. The causes of the outliers should be investigated to determine whether to include or exclude specific results. Thus, while some outliers may be informative, requiring separate treatment using, for example, the robust loss function,[58] others are erroneous and should be removed.

Furthermore, when interpreting images in the laboratory using ML tools, it may be appropriate to preprocess the data to ensure all data inputs are normalized in size, color shift, and magnification.[12] It may also be appropriate to downsize, flatten, or remove the complexity of image data in order to decrease the computational burden needed to develop the model.

Missing values and outliers are often not easily recognized by manual inspection of the data. Visualization of the data structure before training and tuning ML models allows for a quick evaluation for possible problems in data preprocessing. Visualizing the structure or distribution of the data also facilitates better understanding of the data and an opportunity for inspiring insights into the selection of ML models. The high-dimensional data obtained from clinical tests are challenging to visualize directly. As such, dimension reduction and embedding techniques can be used to map high-dimensional data to 2D or 3D low-dimensional spaces while exhibiting their local and global structures. Commonly used visualization methods include principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and unified manifold approximation and projection (UMAP) analysis. Specifically, PCA is a linear transformation that projects original higher-dimensional data into a lower orthonormal space in which each dimension tries to preserve as much of the data's variation as possible. t-SNE directly solves for low-dimensional coordinates of high-dimensional data samples, which preserves the local neighborhood structures of the data samples in their original high-dimensional space. UMAP learns an explicit mapping function to achieve the same goal as t-SNE such that the embeddings can be conveniently extended to unseen testing data. A simplified example is shown in Figure 2, A through D, where a mixture of four 3-dimensional Gaussian distributed data is projected onto a 2-dimensional space by using PCA, t-SNE, and UMAP analysis. Overall, these visualization methods show to some extent the underlying structure of data distribution and allow insights into the complexities of the clinical questions that the ML model will be designed to address. For instance, t-SNE analysis revealed intrinsic distinctions between the clotted and no-clot-detected samples, based on the results of a panel of coagulation tests, which supported the development of backpropagation neural networks to automatically identify clotted specimens.[59] In addition, UMAP analysis visualized the distinct differences in laboratory test result profiles between SARS-CoV-2 reverse transcription-polymerase chain reaction (RT-PCR)–positive and RT-PCR–negative patients, which was used to improve the understanding of an ML model performance in predicting SARS-CoV-2 infection in emergency department patients.[60]

After missing values and outliers in a data set are properly accounted for, an appropriate mathematical representation form should be constructed that collects all laboratory test results for the downstream ML task. The "appropriateness" of such representation is dependent on the clinical problem and the specific ML model. For example, the results of a collection of laboratory tests for a particular patient are represented as a vector, with each dimension of the vector corresponding to the value of a specific laboratory test. This vector can be used to build an

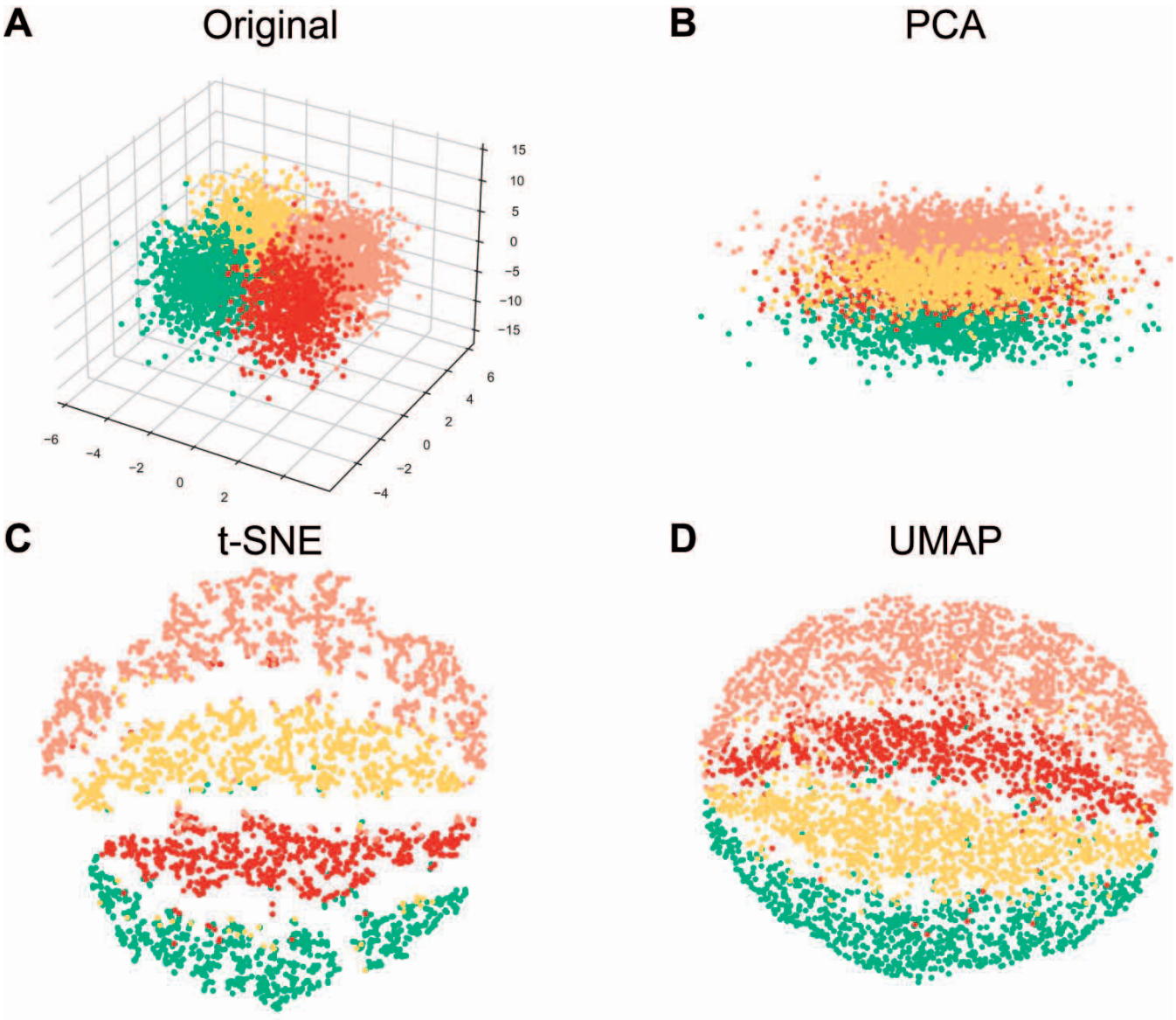

Figure 2. *Illustration of different dimension reduction algorithms on the same synthetic data set. A, The synthetic data points were sampled from a mixture of four 3-dimensional Gaussian distributions. High-dimensional data in the real world may not be Gaussian or with clear clusters. Here, different colors are used to represent different Gaussian distributions as a simplified example. B, The first 2 principal components of the principal component analysis (PCA) on the original data. In this example, PCA is not able to distinguish original data in the linearly transformed 2-dimensional space. C, The first 2 dimensions of the original data transformed by t-distributed stochastic neighbor embedding (t-SNE) analysis, which is a nonlinear dimension reduction technique. PCA analysis is used for initialization before t-SNE to keep the global structure of the original data. Here, t-SNE analysis can distinguish different data clusters in a 2-dimensional space. D, The first 2 dimensions of the original data transformed by unified manifold approximation and projection (UMAP) analysis, another nonlinear dimensionality reduction technique. UMAP analysis successfully distinguishes 4 data clusters in a lower space after PCA initialization.*

ML model to predict clinical outcomes for patients in particular clinical settings, for example, SARS-CoV-2 infection.[61] There are other clinical scenarios where the temporal trends of specific laboratory tests are investigated. In these situations, the values of laboratory tests over time are concatenated in longitudinal sequences.[62] Time series analysis and ML have also been used to calculate the "shelf-life" of a laboratory result[63] and to optimize repeated ordering practices.[64,65] Some commercial autoML software or cloud service automates the data preprocessing steps, including data cleaning or representation construction, in order to simplify the workflow and save users' effort. However, it is recommended to inspect the inputs and outputs of the autoML software for optimizing ML model performance as well as debugging errors.

MODEL DEVELOPMENT

The suitability of an ML approach for a particular application depends on how the data and labels are collected as well as the availability of the data during the ML model training process. There are 3 basic approaches in ML: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning. (1) Supervised learning builds a model by mapping the input feature variables to a target variable, which could be numerical (a.k.a. regression) or

categorical (a.k.a. classification). Supervised learning algorithms are suitable for clinical situations where the training data and their corresponding labels can be collected offline with relationships that are assumed to be consistent with the unknown test data. Thus, the ML model trained by training data can be used to predict the value or classification of unknown observations. For example, a random forest model built on plasma concentrations of a panel of steroids resulted in an accurate classification of patients with or without primary aldosteronism.[66] Additionally, an ensemble of 3 supervised ML models predicted the interpretation of a plasma amino acid profile, thereby supporting the diagnosis of inherited metabolic disorders.[6] (2) Unsupervised learning, such as clustering or unsupervised dimensionality reduction, aims to build models to better characterize data when specific outcomes are not available or are unknown beforehand. Unsupervised learning is a discovery type of analysis to better understand the subgroups in a data set.[67] For example, Su et al[62] identified subphenotypes of patients with COVID-19 by performing hierarchical agglomerative clustering on laboratory test profiles at their onsets.[62] (3) Reinforcement learning aims to learn a sequence of actions toward a specific goal by maximizing certain cumulative rewards. Reinforcement learning is suitable for dynamic online learning scenarios where the data and labels come in gradually and their distributions may change over time.[68] Komorowski et al[20] applied reinforcement learning to adjust the dosage of vasopressors in intravenous fluids for treating septic patients in the intensive care unit, with the goal of maximizing patient survival. Therefore, it is necessary to first specify the problem formulation and then select the appropriate learning model candidates.

The intended audience and intended use should also be considered when generating an ML system. For example, an ML model could be designed to alert a nurse on the floor that the patient might be at the beginning stages of sepsis,[69] or to interpret whether a laboratory testing is positive or not.[61] Depending on the intended audience and use, the performance target of the algorithm may differ. For example, if alerting the nurse of possible sepsis, the acceptable limit of falsely positive results may be higher than for interpreting a PCR fluorescent curve in a laboratory setting.

The data input also needs to be considered, including sample size and number of independent variables, that is, feature dimensions, in the sample representations. A general rule is that the sample size needs to be at least on the same level as feature dimensions for the ML model learning in order to be properly fit. However, some deep learning models[70] are particularly "data hungry" and may require orders of magnitude more samples than feature dimensions. For example, a deep learning model built on a data set of 159 969 expert-annotated serum protein electrophoresis entries resulted in highly accurate identification and quantification of monoclonal gammopathies.[71] However, if the amount of data is not sufficiently large, strategies such as dimension reduction[72] and model pretraining[73] could be considered. Dimension reduction methods reduce the number of features so the model can be more reliably trained.[7] For example, PCA[74] maps the original high-dimensional feature space to a low-dimensional space with maximal information preservation through linear transformation, and deep autoencoders[75] achieve the same goal with nonlinear transformation. Eventually, an appropriate ML model should be chosen to match the clinical task and the size of available data sets. Choosing such a model may well result in a trade-off among sensitivity, specificity, and accuracy, as well as computational cost. When developing an algorithm attempting to identify rare errors or uncommon results, it may be necessary to augment the data set with available real error data or uncommon results.[76]

Most ML models involve hyperparameters, which are parameters defining the model architecture, controlling the learning process, and determining the values of model parameters, such as the number of trees in a random forest model or the number of clusters in a K-means model. The values of model hyperparameters cannot be estimated from training data; instead they are usually specified by the users. The process of searching for model hyperparameters is often referred to as hyperparameter tuning. For predictive modeling, cross validation can be performed to determine the best model hyperparameter settings. For clustering, quantitative model selection criteria, such as the Bayesian information criterion[77] or Akaike information criterion,[78] can be used to select the optimal number of clusters.

MODEL EVALUATION

Last but not least, model performance is evaluated with a diverse set of criteria, such as accuracy, sensitivity or recall, precision or positive predictive value, specificity, and area under the receiver operating characteristic curve (AUROC) for predictive models, while normalized mutual information[79] or the Silhouette index[80] is used for clustering models. The receiver operating characteristic (ROC) curve is used to depict the performance of a binary classification model, based on its sensitivity and specificity as the threshold changes (Figure 3, A and B). The ROC curve helps to visualize the trade-off between the true-positive rate and the false-positive rate of an ML model using different thresholds: sensitivity increases with the compromise of specificity, and vice versa. Higher sensitivity may be needed for ML-based classification models that generate screening-type alerts (yellow star in Figure 3), whereas high specificity may be needed for models used for the purposes of disease confirmation (green star). The ROC curve is more appropriate when positive cases and negative cases are relatively balanced. AUROC represents the ability of a classifier to distinguish between 2 classes. The higher the AUROC, one can assume better performance of the classifier. However, no AUROC threshold can guarantee successful model application in clinical practice. For rare diseases, an AUROC value could be misleading because of the extreme imbalance between patients with and without the outcome event. In such a situation, the precision-recall plot, which is a plot of precision on the y-axis and the recall on the x-axis, is recommended to visualize model performance for an imbalanced data set (Figure 3, B). The precision-recall plot illustrates the trade-off between the true-positive rate and positive predictive value for a predictive model using different thresholds.

To measure the sensitivity and specificity of a model, an operating point on the ROC curve corresponding to a particular threshold should be determined. The operating point is the optimized threshold chosen to maximize model performance in the training set. The selection of an operating point could be based on clinical needs, or compared with a gold standard, such as human manual

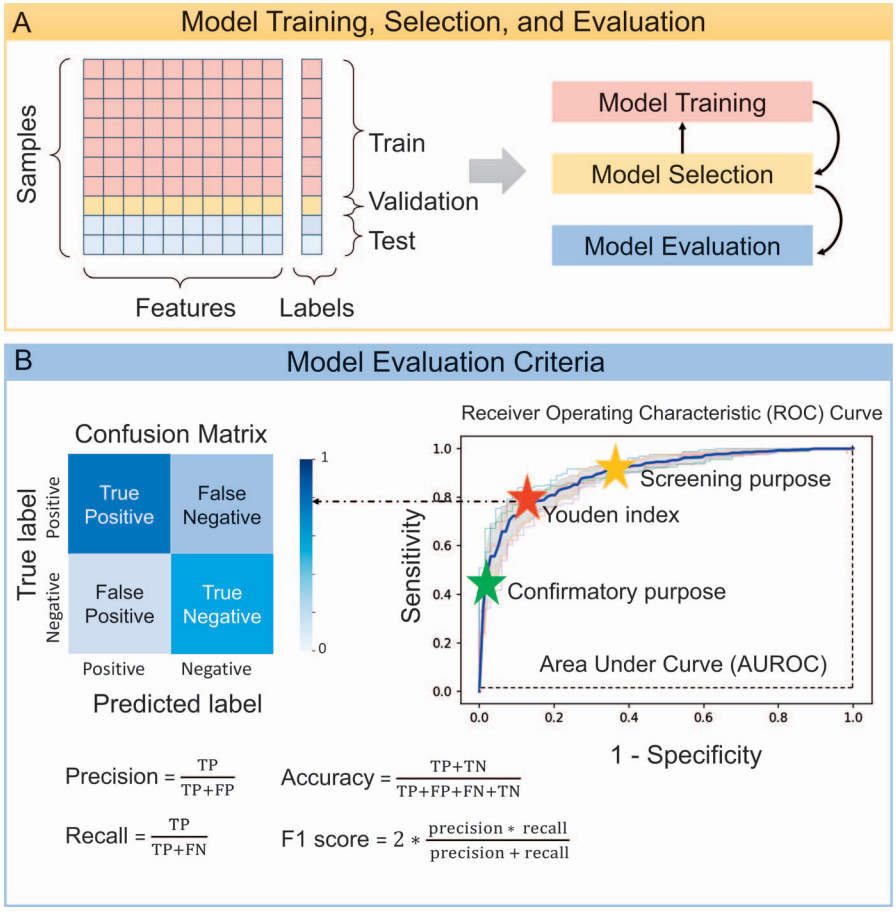

Figure 3. *Evaluation criteria of ML models. A, The workflow of training, selection, and evaluation of ML models. When selecting ML models, it is common to split data into training, validation, and test sets. The model is trained on the training set, and specific hyperparameters are selected based on the model performance on the validation set. Final model evaluation is performed on the unseen test set. B, Evaluation criteria for the binary ML classification model. The ROC curve and AUROC are the most commonly used metrics for model evaluation. The x-axis of the ROC curve is the false-positive rate (1 − specificity), and the y-axis is the true-positive rate (sensitivity). The red star represents the operating point determined by the maximum Youden index where the sum of sensitivity and specificity is the highest on the curve. The yellow and green stars represent the operating points used for screening (high sensitivity) and confirmatory (high specificity) purposes, respectively. Once an operating point is selected, a confusion matrix as well as precision, recall, accuracy, and F1 score can be calculated to summarize the performance of a model. Abbreviations: AUROC, area under the receiver operating characteristic curve; FN, false negative; FP, false positive; ML, machine learning; ROC, receiver operating characteristic; TN, true negative; TP, true positive.*

adjudication. Alternatively, the maximum Youden index[81] can be chosen, where the sum of sensitivity and specificity is the highest on the ROC curve (red star in Figure 3, B), as shown in Yang et al.[61] Once the operating point of an ML model is chosen, a confusion matrix, as well as precision, recall, and accuracy, can be calculated to summarize the model performance. Precision in this context is the same as positive predictive value, which indicates the proportion of positive cases that were correctly identified. Sensitivity or recall indicates the proportion of actual positive cases that are identified correctly (Figure 3, B). It is noteworthy that some of these measures, when used alone, may be misleading in certain clinical scenarios. For example, prediction accuracy can lead to over-optimistic results when predicting rare clinical outcomes.[82] Therefore, it is strongly recommended to use multiple criteria to comprehensively evaluate model performance rather than a single criterion.

When conducting a quantitative evaluation, the entire data set should be partitioned into nonoverlapping training, validation, and testing sets (Figure 3, A). This is to avoid the issue of model overfitting, that is, the learned model fits the training data perfectly, but its performance does not generalize well on unseen testing data. Another potential issue is model underspecification,[83] which can result in the model trained in hospital A not "working" for hospital B because of the "dataset shift" between the hospitals.[84] Therefore, at least one independent external testing data set is needed for understanding model generalizability. Moreover, as the ML model is typically developed from retrospective data, it is important to evaluate the model in the prospective setting, since laboratory test results evolve over time.

Besides quantitative performance metrics, model interpretability is also essential as clinicians prefer to use models that they can understand, including how conclusions are reached, and models that align well with their experience and knowledge.[82] In this context, most of the conventional ML models, such as logistic regression and decision tree, are self-interpretable owing to their linear or rule-based nature. In contrast, more recent ML models, such as deep neural networks, are largely "black boxes." Although there has been research suggesting accuracy could be more important than interpretability,[85] the assumption is that such accuracy should be widely tested and proven to be generalizable, which is challenging in many clinical scenarios. Model interpretation can transform the quantitative model into explainable and understandable data, increasing the adoption and generalization of these models. For "black-box" models, post-hoc interpretation approaches could be applied for distilling their dark knowledge. Specifically, after learning, a black-box model can generate an output given any specific input. Then all model input-output pairs can be collected and used to learn another more explainable model, such as the decision tree model or linear model to map these inputs to their corresponding outputs. The Shapley additive explanations (SHAP) technique is an approach of this kind,[86] decomposing the model prediction for each sample as an additive integration of the contributions from each individual feature.

It is important to realize the workflow of developing an ML model is not necessarily a one-way, one-time process. It is usually an iterative process, re-refining every step to improve the accuracy and generalizability of the model. When an ML model is deployed in real world of clinical practice, its performance should be closely monitored, as there may be changes due to disease evolution, patient population drift, or testing platform changes.[87] Thus, data need to be continuously collected with constant model adaptation, that is, fine tuning and retraining of the ML model over time. For this to be achieved, flexible IT infrastructure is also needed to facilitate continuous model improvement.

IMPLEMENTATION OF MACHINE LEARNING MODELS IN CLINICAL PRACTICE

After optimization and evaluation using criteria as described above, the ML model is ready for implementation. Typically, implementation of the model requires less computational resources than those required for derivation of the model. Even with large deep neural networks, which have thousands of feature inputs, the final model may require only a handful of extracted features for use. Therefore, relatively simple models can be implemented by using calculations and rule-based approaches in the laboratory information systems and/or EHR. Alternatively, external ML software can be incorporated into the EHR by using the "app store" concept popularized with cell phone applications, which is available for major commercial EHR systems. In addition to allowing third-party ML software to interact with their system, major EHR developers have begun to incorporate ML modules, most of which are concerned with predicting future outcomes and events. Although these modules may incorporate laboratory data, specific applications in laboratory medicine are not yet routinely available. Less desirable is to use external software that is not connected to the EHR data. However, decision support systems in the EHR may be able to identify use cases and provide links and custom HL7/Fast Healthcare Interoperability Resources (FHIR) interfaces to the external ML software for further processing.

Despite these advances, many challenges remain to be overcome before there will be widespread implementation of ML in the health care system.[88,89] Some of these challenges include high cost, privacy and security concerns, lack of explainability of the model outputs (particularly the deep learning "black box" systems), unintended bias, brittleness (susceptibility to irrelevant inputs), and lack of reproducibility. In addition, many laboratorians and clinicians are uncomfortable with the paucity of external validations as well as the lack of clinical accuracy expressed as sensitivity and specificity at various cutoff points for practical clinical application. It is appropriate for laboratories to consider ML models as test systems and validate these in-silico tools in a manner similar to how in vitro assays are validated. Reviewing this validation process before implementation can identify unexpected failures and potentially prevent implementing a solution that does not perform as anticipated.

The regulatory environment is another obstacle impeding the implementation of ML algorithms in clinical practice. Just recently the US Food and Drug Administration has proposed a regulatory framework for addressing the use of ML algorithms in medical devices, which includes in vitro diagnostic testing.[90] This framework is particularly focused on adaptive ML technologies that may continually or iteratively update and improve underlying algorithms as further inputs and human feedback are obtained. The regulatory framework relies on premarket review and approval of prespecified requirements for "Quality Systems and Good Machine Learning Practices," including the prespecification of the types of changes allowed when the ML system is in use and algorithm change protocols in place to appropriately control the risks of the predicted modifications, as well as continuous postmarket review of the ML performance.[91]

SUMMARY

The application of ML tools in laboratory medicine is rapidly expanding, as demonstrated by the exponential increase in publications during the past decade.[28] Mining laboratory big data has limitless potential to improve the efficiency of laboratory workflows as well as in assisting in the interpretation of clinical and laboratory data, and as such is likely to expand significantly in the near future. With the growing interest in building ML models, laboratorians and clinicians will be able to properly collect data, to combine data from multiple institutions, and to correctly handle missing data and outliers. In addition, they will require the necessary knowledge to select a suitable ML model and properly evaluate model performance, based on objective criteria. The role of laboratorians is not just to provide data, but also to use their clinical knowledge with the data to guide model development, to correctly interpret the model, and to evaluate its performance in the patient care setting. The future of personalized and generalized medicine requires interdisciplinary collaboration between laboratory medicine and data science experts to create innovative, accurate ML learning models, which will advance the medical field, provide needed support in periods of health care crisis, and best treat individual patients.

The authors would like to thank Ming Yang, PhD, and Zehra Abedi, MS, for their efforts on proofreading and editing the language of this paper.

References

1. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–260.
2. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. *Am J Clin Pathol*. 2016;145(6):778–788.
3. Rosenbaum MW, Baron JM. Using machine learning-based multianalyte delta checks to detect wrong blood in tube errors. *Am J Clin Pathol*. 2018;150(6):555–566.
4. Mitani T, Doi S, Yokota S, Imai T, Ohe K. Highly accurate and explainable detection of specimen mix-up using a machine learning model. *Clin Chem Lab Med*. 2020;58(3):375–383.
5. Benirschke RC, Gniadek TJ. Detection of falsely elevated point-of-care potassium results due to hemolysis using predictive analytics. *Am J Clin Pathol*. 2020;154(2):242–247.
6. Wilkes EH, Emmett E, Beltran L, Woodward GM, Carling RS. A machine learning approach for the automated interpretation of plasma amino acid profiles. *Clin Chem*. 2020;66(9):1210–1218.
7. Wilkes EH, Rumsby G, Woodward GM. Using machine learning to aid the interpretation of urine steroid profiles. *Clin Chem*. 2018;64(11):1586–1595.
8. Ganetzky RD, Master SR. Machine learning for the biochemical genetics laboratory. *Clin Chem*. 2020;66(9):1134–1135.
9. Alouani DJ, Rajapaksha RRP, Jani M, Rhoads DD, Sadri N. Specificity of SARS-CoV-2 real-time PCR improved by deep learning analysis. *J Clin Microbiol*. 2021;59(6):e02959–20.
10. Cohen NM, Schwartzman O, Jaschek R, et al. Personalized lab test models to quantify disease potentials in healthy individuals. *Nat Med*. 2021;27(9):1582–1591.
11. Obstfeld AE, Patel K, Boyd JC, et al. Data mining approaches to reference interval studies. *Clin Chem*. 2021;67(9):1175–1181.

12. Wang Z, Zhang L, Zhao M, et al. Deep neural networks offer morphologic classification and diagnosis of bacterial vaginosis. *J Clin Microbiol*. 2021;59(2):e02236–20.
13. Mathison BA, Kohan JL, Walker JF, Smith RB, Ardon O, Couturier MR. Detection of intestinal protozoa in trichrome-stained stool specimens by use of a deep convolutional neural network. *J Clin Microbiol*. 2020;58(6):e02053–19.
14. Huang L, Wu T. Novel neural network application for bacterial colony classification. *Theor Biol Med Model*. 2018;15(1):22.
15. Zhang ML, Guo AX, Kadauke S, Dighe AS, Baron JM, Sohani AR. Machine learning models improve the diagnostic yield of peripheral blood flow cytometry. *Am J Clin Pathol*. 2020;153(2):235–242.
16. Lidbury BA, Richardson AM, Badrick T. Assessment of machine-learning techniques on large pathology data sets to address assay redundancy in routine liver function test profiles. *Diagnosis (Berl)*. 2015;2(1):41–51.
17. Yu M, Bazydlo LAL, Bruns DE, Harrison JH Jr. Streamlining quality review of mass spectrometry data in the clinical laboratory by use of machine learning. *Arch Pathol Lab Med*. 2019;143(8):990–998.
18. Than MP, Pickering JW, Sandoval Y, et al. Machine learning to predict the likelihood of acute myocardial infarction. *Circulation*. 2019;140(11):899–909.
19. Tomasev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116–119.
20. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24(11):1716–1720.
21. De Bruyne S, Speeckaert MM, Van Biesen W, Delanghe JR. Recent evolutions of machine learning applications in clinical laboratory medicine. *Crit Rev Clin Lab Sci*. 2021;58(2):131–152.
22. Baron JM, Kurant DE, Dighe AS. Machine learning and other emerging decision support tools. *Clin Lab Med*. 2019;39(2):319–331.
23. Harrison JH, Gilbertson JR, Hanna MG, et al. Introduction to artificial intelligence and machine learning for pathology. *Arch Pathol Lab Med*. 2021;145(10):1228–1254.
24. Badrick T, Banfi G, Bietenbeck A, Cervinski MA, Loh TP, Sikaris K. Machine learning for clinical chemists. *Clin Chem*. 2019;65(11):1350–1356.
25. Dark Daily. EHR Systems continue to cause burnout, physician dissatisfaction, and decreased face-to-face patient care. https://www.darkdaily.com/2017/12/22/ehr-systems-continue-to-cause-burnout-physician-dissatisfaction-and-decreased-face-to-face-patient-care-1222/. Accessed March 15, 2022.
26. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med*. 1985;13(10):818–829.
27. Wiesner R, Edwards E, Freeman R, et al. Model for end-stage liver disease (MELD) and allocation of donor livers. *Gastroenterology*. 2003;124(1):91–96.
28. Lippi G. Machine learning in laboratory diagnostics: valuable resources or a big hoax? *Diagnosis (Berl)*. 2019;8(2):133–135.
29. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
30. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27(12):2011–2015.
31. Campbell JP, Lee AY, Abramoff M, et al. Reporting guidelines for artificial intelligence in medical research. *Ophthalmology*. 2020;127(12):1596–1599.
32. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA*. 1996;276(8):637–639.
33. Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann Intern Med*. 2013;158(3):200–207.
34. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–1374.
35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
36. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574–578.
37. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578–582.
38. College of American Pathologists. Neonatal Bilirubin Proficiency Testing Survey 2020. https://www.cap.org/laboratory-improvement/proficiency-testing. Accessed May 23, 2022.
39. Martin RF. General deming regression for estimating systematic bias and its confidence interval in method-comparison studies. *Clin Chem*. 2000;46(1):100–104.
40. Karvanen J. The statistical basis of laboratory data normalization. *Drug Inf J*. 2003;37(1):101–107.
41. Chuang-Stein C. Some issues concerning the normalization of laboratory data based on reference ranges. *Drug Inf J*. 2001;35(1):153–156.
42. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem*. 2003;49(4):624–633.
43. Stram M, Gigliotti T, Hartman D, et al. Logical observation identifiers names and codes for laboratorians. *Arch Pathol Lab Med*. 2020;144(2):229–239.
44. Gkoutos GV, Schofield PN, Hoehndorf R. The Units Ontology: a tool for integrating units of measurement in science. *Database (Oxford)*. 2012;2012:bas033.
45. Schriml LM, Mitraka E, Munro J, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 2019;47(D1):D955–D962.
46. Annesley TM, McKeown DA, Holt DW, et al. Standardization of LC-MS for therapeutic drug monitoring of tacrolimus. *Clin Chem*. 2013;59(11):1630–1637.
47. Agrawal YP, Cid M, Westgard S, Parker TS, Jaikaran R, Levine DM. Transplant patient classification and tacrolimus assays: more evidence of the need for assay standardization. *Ther Drug Monit*. 2014;36(6):706–709.
48. Hauser RG, Shirts BH. Do we now know what inappropriate laboratory utilization is: an expanded systematic review of laboratory clinical audits. *Am J Clin Pathol*. 2014;141(6):774–783.
49. Hauser RG, Jackson BR, Shirts BH. A bayesian approach to laboratory utilization management. *J Pathol Inform*. 2015;6:10.
50. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform*. 2020;8(3):e17984.
51. Durant TJS, Dudgeon SN, McPadden J, et al. Applications of digital microscopy and densely connected convolutional neural networks for automated quantification of Babesia-infected erythrocytes. *Clin Chem*. 2021;68(1):218–229.
52. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. *JAMA Intern Med*. 2019;179(3):293–294.
53. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. *J Biomed Inform*. 2014;51:24–34.
54. Rubin D. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473–489.
55. Luckoor P, Salehi M, Kunadu A. Exceptionally high creatine kinase (CK) levels in multicausal and complicated rhabdomyolysis: a case report. *Am J Case Rep*. 2017;18:746–749.
56. Matsunaga N, Yoshioka Y, Fukuta Y. Extremely high troponin levels induced by septic shock: a case report. *J Med Case Rep*. 2021;15(1):466.
57. Stankovic AK, Smith S. Elevated serum potassium values: the role of preanalytic variables. *Am J Clin Pathol*. 2004;121(suppl):S105–S112.
58. Mukherjee D, Guha A, Solomon J, Sun Y, Yurochkin M. Outlier-robust optimal transport. *Proc Mach Learn Res*. 2021;139:7850–7860.
59. Fang K, Dong Z, Chen X, et al. Using machine learning to identify clotted specimens in coagulation testing. *Clin Chem Lab Med*. 2021;59(7):1289–1297.
60. Yang HS, Hou Y, Zhang H, et al. Machine learning highlights downtrending of COVID-19 patients with a distinct laboratory profile. *Health Data Sci*. 2021;2021:7574903.
61. Yang HS, Hou Y, Vasovic LV, et al. Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. *Clin Chem*. 2020;66(11):1396–1404.
62. Su C, Xu Z, Hoffman K, et al. Identifying organ dysfunction trajectory-based subphenotypes in critically ill patients with COVID-19. *Sci Rep*. 2021;11(1):15872.
63. Levy-Fix G, Gorman SL, Sepulveda JL, Elhadad N. When to re-order laboratory tests: learning laboratory test shelf-life. *J Biomed Inform*. 2018;85:21–29.
64. Yu L, Li L, Bernstam E, Jiang X. A deep learning solution to recommend laboratory reduction strategies in ICU. *Int J Med Inform*. 2020;144:104282.
65. Baron JM, Huang R, McEvoy D, Dighe AS. Use of machine learning to predict clinical decision support compliance, reduce alert burden, and evaluate duplicate laboratory test ordering alerts. *JAMIA Open*. 2021;4(1):ooab006.
66. Eisenhofer G, Duran C, Cannistraci CV, et al. Use of steroid profiling combined with machine learning for identification and subtype classification in primary aldosteronism. *JAMA Netw Open*. 2020;3(9):e2016209.
67. Haymond S, McCudden C. Rise of the machines: artificial intelligence and the clinical laboratory. *J Appl Lab Med*. 2021;6(6):1640–1654.
68. Coronato A, Naeem M, De Pietro G, Paragliola G. Reinforcement learning for intelligent healthcare applications: a survey. *Artif Intell Med*. 2020;109:101964.
69. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46(4):547–553.
70. Jackups R Jr. Deep learning makes its way to the clinical Laboratory. *Clin Chem*. 2017;63(12):1790–1791.
71. Chabrun F, Dieu X, Ferre M, et al. Achieving expert-level interpretation of serum protein electrophoresis through deep learning driven by human reasoning. *Clin Chem*. 2021;67(10):1406–1414.
72. Van Der Maaten L, Postma E, Van Den Herik J. Dimensionality reduction: a comparative review. *J Mach Learn Res*. 2009;10(1-41):66–71.
73. Hendrycks D, Lee K, Mazeika M. Using pre-training can improve model robustness and uncertainty. *Intern Conf Mach Learn*. 2019:2712-2721.
74. Herve A, William LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433–459.
75. Tschannen M, Bachem O, Lucic M. Recent advances in autoencoder-based representation learning. Posted online December 12, 2018. *arXiv preprint arXiv: 181205069*.
76. Nardelli P, Estepar RSJ. Targeting precision with data augmented samples in deep learning. *Med Image Comput Comput Assist Interv*. 2019;11769:284–292.

77. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461–464.
78. Aho K, Derryberry D, Peterson T. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*. 2014;95(3):631–636.
79. Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J Mach Learn Res*. 2002;3:583–617.
80. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53–65.
81. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.
82. Wang F. Machine learning for predicting rare clinical outcomes—finding needles in a haystack. *JAMA Netw Open*. 2021;4(5):e2110738.
83. AACC CLN Stat. How underspecification undermines artificial intelligence. 2020. https://www.aacc.org/cln/cln-stat/2020/december/17/how-underspecification-undermines-artificial-intelligence. Accessed March 15, 2022.
84. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283–286.
85. van der Veer SN, Riste L, Cheraghi-Sohi S, et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *J Am Med Inform Assoc*. 2021;28(10):2128–2138.
86. Lundberg S, Lee S. A unified approach to interpreting model prediction. In: *Proceedings of the 31th International Conference on Neural Information Processing System*. Long Beach, CA: Publisher; 2017:4768–4777.
87. Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res*. 2021;5(1):1.
88. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
89. Paranjape K, Schinkel M, Hammer RD, et al. The value of artificial intelligence in laboratory medicine. *Am J Clin Pathol*. 2021;155(6):823–831.
90. US Food and Drug Administration. Artificial intelligence and machine learning (AI/ML) software as a medical device action plan. 2021. https://www.fda.gov/media/145022/download. Accessed March 15, 2022.
91. US Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) - discussion paper and request for feedback. https://www.regulations.gov/document/FDA-2019-N-1185-0001. Accessed March 15, 2022.