JAMA | Users' Guides to the Medical Literature

How to Read Articles That Use Machine Learning
Users' Guides to the Medical Literature

Yun Liu, PhD; Po-Hsuan Cameron Chen, PhD; Jonathan Krause, PhD; Lily Peng, MD, PhD

In recent years, many new clinical diagnostic tools have been developed using complicated machine learning methods. Irrespective of how a diagnostic tool is derived, it must be evaluated using a 3-step process of deriving, validating, and establishing the clinical effectiveness of the tool. Machine learning–based tools should also be assessed for the type of machine learning model used and its appropriateness for the input data type and data set size. Machine learning models also generally have additional prespecified settings called hyperparameters, which must be tuned on a data set independent of the validation set. On the validation set, the outcome against which the model is evaluated is termed the reference standard. The rigor of the reference standard must be assessed, such as against a universally accepted gold standard or expert grading.

JAMA. 2019;322(18):1806-1816. doi:10.1001/jama.2019.16489

Viewpoint page 1765 and Editorial page 1777

Supplemental content

CME Quiz at jamanetwork.com/learning

Author Affiliations: Google Health, Palo Alto, California.

Corresponding Author: Yun Liu, PhD, Google Health, 3400 Hillview Ave, Palo Alto, CA 94304 (liuyun@google.com).

Clinical Scenario

You are the chief medical officer of a large multifacility health care system. One of the medical staff committees of the organization reviewed guidelines from the American Academy of Ophthalmology recommending annual diabetic retinopathy screening for all adult patients with diabetes.[1] You determine that there is reasonably good evidence supporting this recommendation. Patients with diabetes are prone to developing retinopathy or macular edema and these diseases may progress to advanced stages before any symptoms occur. Screening allows for treatment of these diseases with anti–vascular endothelial growth factor (anti-VEGF) agents or laser photocoagulation in an early disease stage—before vision is compromised.

Despite the benefit of screening, your organization has very limited access to eye care. You have also found an article suggesting that such screening, using an automated system based in primary care clinics in a health system similar to yours, was effective for diabetic retinopathy screening.[2] In that study, nondilated digital retinal images were obtained in primary care clinics and automatically analyzed by artificial intelligence software. The system is proprietary, and you do not know how valid, reliable, and effective it might be. You perform a web search and find that there are several automated systems available that screen for diabetic retinopathy. You also find that it is currently believed that systems based on a machine learning method called convolutional neural networks (CNNs) seem to have the most promise for detecting diabetic retinopathy in clinical practice because these systems have the ability to manage very large amounts of information, high sensitivity, and high specificity.

A search of PubMed finds some articles that demonstrate the performance characteristics for automated systems for detecting eye disease. In one *JAMA* article, the ability of machine learning using modern CNNs to detect diabetic retinopathy was shown,[3] and in another, a CNN-based system was developed and validated using independent samples.[4] A third article described using a CNN-based system in a clinical setting.[5]

To assess this literature, you use the framework for assessing articles reporting the results of diagnostic tests (Users' Guide to the Medical Literature) (**Box 1**),[6,7] but you are unsure if the development of a diagnostic tool using machine learning differs from any other type of diagnostic test.

This article provides an overview of machine learning and how to assess the published literature describing the use of machine learning-based tools to establish medical diagnoses.

The literature regarding artificial intelligence, machine learning, or deep learning that supposedly reproduces human-level performance in clinical tasks is rapidly expanding (**Box 2**). Although the machinery used to implement these techniques is complex, once a machine learning system is developed, the system should be validated using similar rules for any system designed to aid clinician decision-making. Once derived, a model should be validated and its clinical effectiveness in real-world settings assessed.[8]

How machine learning methods work and how they are derived and validated should not remain a mystery to clinicians who rely on them to improve patient care. Just as radiologists understand the fundamental concepts of image acquisition as they review magnetic resonance images, clinicians who rely on machine learning models should likewise understand the major principles. This Users' Guide has 3 goals to facilitate clinicians' understanding of machine learning models: (1) to emphasize the importance of proper machine learning model validation and highlight any differences in this process relative to the validation of more traditional methods of statistical model development; (2) to review the basics of machine learning; and (3) to review how machine learning models should be implemented in clinical medicine.

Machine learning methods are not new in medicine. An example of a simple machine learning model is a rules-based system,[9] such as the Ottawa ankle rules, which determine the need for radiographs in the evaluation of ankle trauma.[10] The Ottawa ankle rules use a decision tree (eTable 2 in the Supplement). Complex machine learning methods provide a new way to derive models and are now possible because of the amount of data available and advanced computation resources. Regardless of how a model is built, it must be validated and its clinical effectiveness verified. Similar to the introduction and use of the Ottawa ankle rule,[8] machine learning studies need to have accurate predictions,[10,11] be validated in large and heterogeneous populations,[10-15] and demonstrate that their use improves clinical outcomes—ideally tested in randomized clinical trials in actual clinical practice.[16] To show that a model accurately differentiates one outcome from another, its discrimination and calibration must be assessed.[17,18] Discrimination metrics measure the model's ability to correctly distinguish different conditions from one another, such as determining from a retinal image if diabetic retinopathy is present or not. Some commonly used descriptive metrics are sensitivity, specificity, positive predictive values, and negative predictive values. The full range of possible results using different cut points for sensitivity and specificity of a model can be visualized by plotting the receiver operating characteristic curve. This curve can be summarized by calculating the area under the curve (AUC; also called the C statistic).[19] Calibration determines how well the model's predicted probability approximates the actual event probability. Calibration is best evaluated by plotting the actual observed event frequency against the average predicted probability for each decile of a population, and quantitatively and qualitatively assessing the deviation from a diagonal line having an intercept of 0 and slope of 1.[20] These and other validation considerations are presented as a checklist in the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines.[20]

The remainder of this Users' Guide covers additional considerations specific to machine learning studies, and these considerations are presented as a checklist in eTable 1 (Supplement), using as examples 2 machine learning studies[3,21] and a decision rule study.[22]

Machine Learning–The Basics

Machine learning methods use mathematical operations to process input data, resulting in a prediction. One commonly used approach for developing a diagnostic tool is logistic regression (**Figure 1**A). For each risk factor, logistic regression determines the relationship between parameters, which are numerical values (Box 2) and binary clinical outcomes such as the presence or absence of a disease entity (eg, retinopathy). When the parameters are greater than 0, the parameter is associated with an increase in risk of the outcome, and if the parameters are less than 0, the parameter is associated with reduced risk. Mathematically, the calculation of a diagnostic score involves multiplying each risk factor (eg, 1 or 0 for presence or absence of hypertension) with the corresponding numeric parameter and summing the results, yielding a probability that the outcome of interest is present.

Modern machine learning methods use greater numbers of mathematical operations than traditional regression techniques to better define complex relationships between risk factors and outcomes. In deep learning, for example, these operations are often performed in layers. Each layer resembles logistic regression because layer multiplies information from the previous layer by a set of parameters. The first layer often directly processes input from the data set (Figure 1B). Early layers perform mathematical operations to extract simple features, later (subsequent) layers build on the simple features to generate more complex ones, and the final layer uses these features to make predictions. For example, to distinguish between categories, including man-made objects and animals, the first few layers include simple patterns; the subsequent layer combines these patterns into more complex shapes and textures; and the final layers learn to recognize parts of buildings and animals, such as birds and dogs (Figure 1C).

Box 1. Evaluating and Applying the Results of Studies of Diagnostic Tests[a]

Are the results of the study valid?

- Primary guides
  - Was there an independent, blind comparison with a reference standard?
  - Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?
  - Was there a completely independent validation set?
- Secondary guides
  - Did the results of the test being evaluated influence the decision to perform the reference standard?
  - Were the methods for performing the test described in sufficient detail to permit replication?

What were the results?

- Are likelihood ratios, sensitivity, and specificity for the test results presented or data necessary for their calculation provided?

Will the results help me in caring for my patients?

- Will the reproducibility of the test result and its interpretation be satisfactory in my setting?
- Are the results applicable to my patient?
- Will the results change my management?
- Will patients be better off as a result of the test?

[a] Information in this box is based on Jaeschke et al.[6,7]

Development of Machine Learning Models

How Is the Specific Machine Learning Method Chosen?

The name machine learning is used because these methods learn from examples during a process called training. There are 2 commonly used machine learning schemes: supervised learning, and unsupervised learning (**Box 3**). In supervised learning, labeled data (eg, retinal fundus photographs read by expert graders for the presence or absence of diabetic retinopathy) are used for machine learning model development. In unsupervised learning, data are not explicitly labeled and are classified as to what the data might represent by some mathematical process. An example of this would be to identify features by clustering data into buckets that are similar to one another. By using labels, supervised learning generally requires less

Box 2. Glossary of General Terminology Associated With Machine Learning Methods

Feature: Features are the input variables to a machine learning model. For example, when developing a model predicting stroke risk, a feature would be a patient's height or weight. Features can be processed before they are entered into a model, such as combining height and weight into a body mass index. For an image, a feature may be some component of the image, such as an eye or a nose, when developing a facial recognition machine learning system.

Hyperparameter: Hyperparameters are parameters that are established before a model is trained and remain fixed through the training process. The hyperparameters generally affect the parameters that are learned during training and can have a large influence on the final accuracy. One of the difficulties in machine learning is in determining sets of hyperparameters that optimize the model fit.

Label: The label identifies what a collection of data (the model input) represents. For a stroke model, it would be stroke present or absent. When developing a machine learning system to identify diabetic retinopathy, the label for each fundus image would be present or absent, as determined by experts in interpreting such images.

Machine Learning, Artificial Intelligence, Deep Learning: Artificial intelligence is a loosely defined concept describing automated systems that can perform tasks considered to require "intelligence." Machine learning refers to the process of developing systems with the ability to learn from and make predictions using data. For example, a machine learning model can process an input (such as a retinal fundus photograph) and produce an output (such as the classification of the image showing that proliferative diabetic retinopathy is present). Deep learning is a more specific group of machine learning methods that uses many layers of arithmetic operations.[25,40]

Model, Algorithm: In the machine learning setting, model and algorithm are frequently used interchangeably to refer to the final ready-to-use machine learning method. These terms refer to the steps taken by the machine to assess input data and make a determination about what is shown in the data.

Overfitting: Overfitting is a scenario in which a machine learning model is trained to predict the training data too well, such that it does not generalize to new data sets. In theory, any set of data can be fit with a mathematical model if large numbers of parameters are entered into a mathematical model. This overfitting can occur even if there is no logical relationship between the data and the outcome. For example, a reasonably good fit can be obtained using regression to determine the relationship between age, cholesterol, and sex, to stroke because each of these variables has a physiological relationship with the development of atherosclerosis and subsequent stoke. The mathematical model relating these risk factors and stroke can have a better fit if more parameters than these are entered into the model, even if those parameters have nothing to do with stroke. The resulting model may not perform well clinically if its fit relies on these extra variables. When the model is applied to a different data set than the one on which it was developed, its predictive ability may fail.

Parameter: Parameters are the internal values of a machine learning model that are derived based on the training data. For example, the parameters in logistic regression include the weights that are multiplied with each input variable as part of the regression equation. If a logistic-regression model were developed to assess the need for a radiograph to evaluate an ankle trauma case, input features may include the presence of bone tenderness at anatomic sites A and B.

(continued)

Box 2. (continued)

The parameter associated with each site would be greater than 0, indicating a higher likelihood that radiographs are needed to rule out fracture. The overall score would be related to multiplying the presence or absence (1 or 0) of tenderness at A and B with their respective parameters. The values of the parameters are learned during a training process to optimize the fit between the available data and the machine learning model outputs.

Reference Standard: For a diagnostic test, a reference standard is the reference against which the proposed method is compared. The reference standard is often a widely accepted test or gold standard for the diagnosis, but it can also be based on diagnoses provided by expert clinicians.

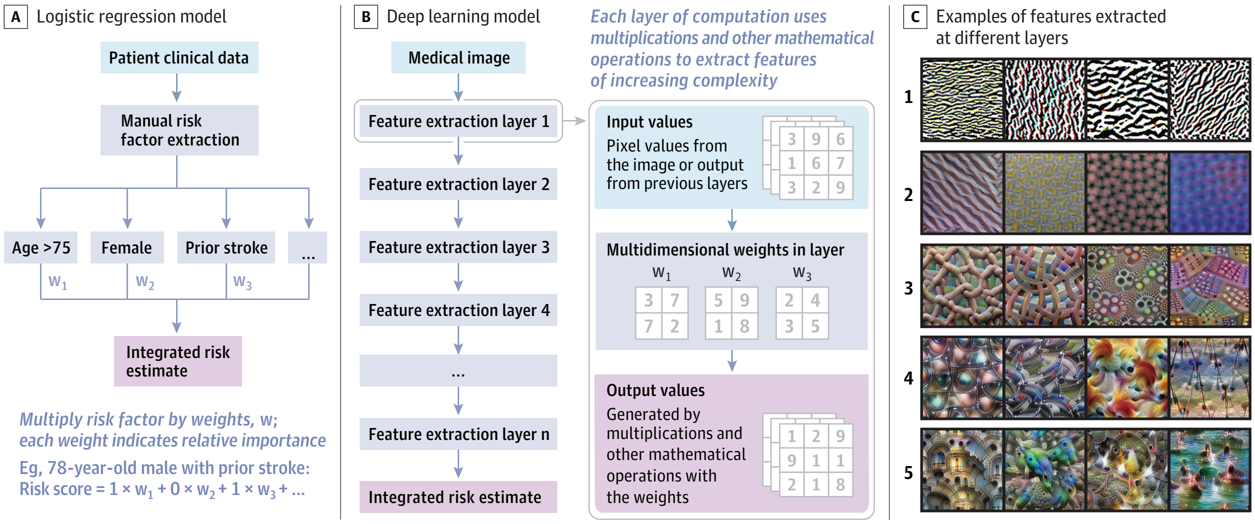
Training: The process of adjusting model parameters in a machine learning model to best match the model output with the reference standard label in the training set.

Tuning: The process of adjusting the hyperparameters of a trained model to increase the model's fit to the tuning set. When tuning a machine learning model, hyperparameters are repeatedly adjusted, each time training a new machine learning model on the training set and evaluating that machine learning model on the tuning set. The optimal hyperparameter configuration is typically the configuration that leads to the best tuning set accuracy.

data than unsupervised learning. Thus, most recent machine learning methods achieving human-level performance that correctly classify clinical information, such as identifying when retinopathy is present in a fundus photograph, use supervised learning.[3,4,21] Unsupervised learning is still an active area of research and claims of human-level performance without the use of any labels in data sets should be carefully validated.

Even with supervised learning, the type of method used should be appropriate for the type and amount of input data. A list of methods that are appropriate for use with various data types and data set sizes is shown in eTable 2 (Supplement). Generally, more recent machine learning methods (eg, convolutional neural networks and recurrent neural networks) work better than traditional methods (eg, logistic regression or support vector machines) in assessing complex data like medical images or text and large data sets. For example, image classification problems, which determine what visual findings are present in an image, such as detecting diabetic retinopathy in retinal fundus images, generally require the use of artificial neural networks (eg, deep learning). Older machine learning methods require experts to predefine known discriminative features and actively help the algorithm identify them. However, images are represented in a computer as 2-dimensional grids of numerical pixel values, and it is very difficult to describe relevant features as patterns of numbers. For example, how does one describe in grids of numbers the characteristic features of proliferative diabetic retinopathy? By being shown a large number of examples, artificial neural networks automatically learn from executing complex mathematical functions that describe discriminative visual features and use the presence and extent of these features to interpret the image.[24-26] Conversely, simpler machine learning systems (eg, logistic regression) that use limited input variables such as age and hypertension status have less information with which to predict an outcome.

Figure 1. Comparison of How Traditional Decision Rules Work vs Recent Machine Learning Methods


A Logistic regression model
Patient clinical data
Manual risk factor extraction
Age >75
Female
Prior stroke
...
w_1
w_2
w_3
Integrated risk estimate
Multiply risk factor by weights, w; each weight indicates relative importance
Eg, 78-year-old male with prior stroke: Risk score = 1 × w_1 + 0 × w_2 + 1 × w_3 + ...
B Deep learning model
Medical image
Feature extraction layer 1
Feature extraction layer 2
Feature extraction layer 3
Feature extraction layer 4
...
Feature extraction layer n
Integrated risk estimate
Each layer of computation uses multiplications and other mathematical operations to extract features of increasing complexity
Input values
Pixel values from the image or output from previous layers
3 9 6
1 6 7
3 2 9
Multidimensional weights in layer
w_1
w_2
w_3
3 7
7 2
5 9
1 8
2 4
3 5
Output values
Generated by multiplications and other mathematical operations with the weights
1 2 9
9 1 1
2 1 8
C Examples of features extracted at different layers
1
2
3
4
5

A, Decision rules and machine learning methods both use mathematical functions to process input data and make a prediction. Many decision rules are based on multiplying risk factors with weights that represent the relative importance of each risk factor. The weights are frequently determined by training a logistic-regression model on data from a patient cohort. For ease of use without a calculator, these weights can be converted to point scores, and the sum of the point scores can be looked up in a risk table. B, Instead of a single set of multiplications, more sophisticated machine learning methods can leverage millions or billions of multiplications and other mathematical operations to extract descriptive features of complex input data, such as images. The weights or parameters of these operations are also derived using the data. C, Each layer can be inspected and visualized for having an intuitive understanding of the patterns being identified.[23] Although this example focuses on the specific problem of image interpretation, the general concept of learning complex features through many layers of mathematical operations is applicable to many recent methods.

How Much Data Are Required for Recent Machine Learning Methods?

For simple machine learning methods such as regression techniques, at least 5 to 10 outcome events per input variable have been recommended.[27] Developing accurate deep-learning models that have millions of parameters frequently require many fewer events per parameter because of various regularization techniques that are commonly used simultaneously (Box 3). Regularization is a technique similar to curve smoothing. One regularization method is called parameter regularization, which essentially smooths the fit of the model to avoid overfitting to any given data set. Overfitting occurs when the model perfectly fits data to the available parameters but not in a way that has any relationship to the clinical outcome that is being modeled. One characteristic of mathematical modeling is that a perfect fit relating data to an outcome can be achieved if a sufficiently large number of parameters are available to fit the data. A method called early stopping achieves similar results to regularizing the parameters by stopping the training process before the model overfits to the data set. Another method of regularization is to initialize the parameters from a deep neural network that has already been trained on another task. By avoiding the need to start anew, the network can learn more quickly and use fewer examples. Yet another technique increases the effective data set size by augmenting the model with data that have been artificially modified, such as by rotating an image slightly or changing the overall brightness of an image. Artificial modification of data is a way of teaching the model that the exact orientation or brightness does not matter in determining what visual finding the model is attempting to identify. Ensembling combines learning from multiple models by averaging their predictions, which improves the accuracy of the final model (Box 3). Though these regularization techniques are helpful and indeed essential for modern machine learning methods, tens of thousands of example data such as retinal images may still be required in the training set to attain high accuracy.[3]

How Does Regularization Influence Machine Learning Model Development?

Most regularization techniques, such as those described in the previous section, influence the learned parameters of the machine learning model. However, the use of these techniques involves setting additional hyperparameters. Hyperparameters are analogous to adjusting knobs of an amplifier to fine-tune the bass and treble of an audio production—tuning the knobs affects the final result. In machine learning, if randomness is controlled for in the training process, fixing the hyperparameter settings results in complete determination of the final numerical values of the learned parameters. However, changing the hyperparameters and training a new machine learning model results in different values of the learned parameters (Box 3). Because hyperparameters have a large effect on the model performance, manually tuning these hyperparameters is an important part of machine learning studies. This tuning process generally requires the use of a tuning set (often a subset of the development data set) that is independent of the final validation set. This is done by repeatedly trying different hyperparameter configurations, training the machine learning model on the training set, and evaluating the machine learning model on the tuning set.

Caution should be exercised during the training and tuning of recent machine learning models. For example, a system can be tuned to 100% accuracy on the training set but may only have random accuracy on the validation set, indicating the machine learning model's

Box 3. Glossary of Terms Associated With Machine Learning Methods: Types of Machine Learning Schemes, Data Set Names, and Regularization

Types of Machine Learning Schemes

Supervised Learning: Training a model with input data and its corresponding labels. The machine learning model attempts to determine a relationship between the input data and the label associated with the data. Examples include developing a machine learning system that can take a retinal image (input) and identify whether it contains retinopathy (the label).

Unsupervised Learning: Training a model to identify patterns within the input data without the use of labels. The most common unsupervised learning method is clustering, which groups data into similar subgroups.

Data Set Names

Development Set: A data set used for developing the machine learning model, frequently further split into the training and tuning sets.

K-fold Cross-Validation: This is a technique that uses multiple splits within the development set to reduce the effects of randomness of the split. For example, if k = 2, the development set is split evenly into A and B. Two models are developed: one trained using A and tuned on B, and one trained using B and tuned on A. The cross-validated evaluation is usually the average of the 2 performance estimates using A and B. An independent validation set should be used to evaluate the performance of the final model trained on the entire development set. A leave-one-out cross validation is when k is the total number of data points in the data set.[a]

Training Set: A subset of the development set that is used to develop the machine learning model where training is performed by updating the model parameters iteratively until the model optimally fits the data.

Tuning Set: A subset of the development set that is used to tune the hyperparameters of a model. In the machine learning community, this may be referred to as the *validation set*. In this guide, we will use *tuning* for consistency, and in medical research, a model must be validated using a data set that is completely independent of the training or tuning set.

Validation Set: A data set that is independent from the training or tuning set. Validation sets are used to evaluate the model performance before a machine learning model can be applied clinically. The validation set should not be used to train or tune the machine learning model, including hyperparameters or choice of machine learning method. In the machine learning community, the validation set may be referred to as the *test*, *holdout*, or *evaluation* set.

Regularization[b]

Data Augmentation: Computationally modifies the input data during the training process to increase the effective data set size and improve both overfitting and final accuracy. This is particularly helpful for neural networks applied to images where the image orientation, scale/magnification, color, brightness, saturation, contrast, and other aspects can be extensively modified. For example, when a machine learning system is trying to identify a nose in a facial recognition system, it does not matter where the nose is in the image or in what direction the nose is facing. To help the system learn what a nose looks like, the same image may be used several times, rotated at various angles or otherwise altered to facilitate recognition of discriminative visual features that are independent of these modifications.

(continued)

Box 3. (continued)

Early Stopping: This technique is most relevant for neural networks, where training is generally done by gradual adjustment of the parameters. To help avoid overfitting, the training process is terminated before the model fits too well to the training set. Typically, performance on the tuning set is monitored throughout the training process, and early stopping is done at the point that maximizes tuning set performance.

Ensemble: The technique of combining multiple outputs of machine learning models to improve stability of the final prediction and hence, overall performance by a few percents. This can be done by developing multiple machine learning models and averaging their outputs given the same input data. Another method is to run the same machine learning model on multiple input images, which can be multiple images from the same patient (such as fundus images from both eyes) or the same image after artificial perturbations (such as those used for data augmentation).

Fine-Tuning, Preinitialization, Warm Start: This technique uses a machine learning model that was previously trained on another data set to initialize the parameters of the desired machine learning model. This can help develop accurate neural networks with smaller data sets. Although more helpful when the other data set is similar in terms of data type or prediction task, the use of unrelated data sets can still be helpful.

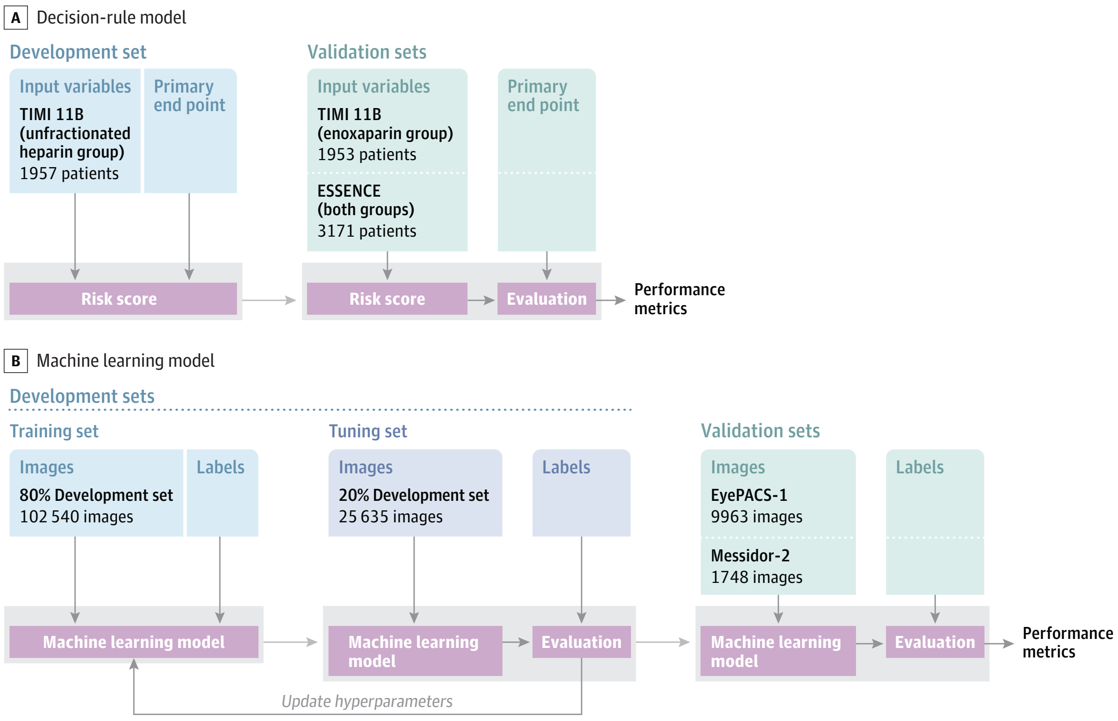
Parameter Regularization: Helps to prevent parameters from becoming too large (shrinkage) and thus overfitting. Ways of doing this include L1 (also called *lasso*) and L2 (also called *ridge*), and the combination (*elastic net*). L1 has the advantage of incorporating feature selection, which is helpful in determining the most important input features. For neural networks, another parameter regularization technique is termed *weight decay*, which prevents the parameters from becoming too large by subtracting the weights by a predetermined factor.

[a] Note that cross-validation can also be used by splitting the entire data set into multiple development sets and validation sets. The reader should be aware that this procedure evaluates the average performance of many machine learning models. Afterwards, the final machine learning model that is trained using the entirety of the data set will require further validation using an independent data set.

[b] Techniques to reduce overfitting, such as by reducing the number of parameters in a model or avoiding an overly precise fit of the model to the data set. Smoothing a noisy curve is an example of regularization and can be achieved in a regression analysis by reducing extremely high parameter values or setting the parameters of unimportant features to 0 in the regression equation.

complete memorization of the training data when a tuning set is not used.[28] Though clinical decision rules are developed using one data set for derivation and one or more unique data sets for validation (**Figure 2**A; see Box 3 for an explanation of the frequently inconsistent data set terminology),[18] machine learning, by contrast, typically requires 2 data sets for the development stage alone: a training set, from which to learn parameters and a tuning set, to adjust hyperparameters (Figure 2B). When the training and tuning sets are small, randomness in the partitioning can reduce the reproducibility of the tuning process. To improve the reproducibility, tuning may be repeated using multiple random partitions of training and tuning sets within the development set (cross-validation; Box 3). If neither a tuning set nor cross-validation is described in a publication describing the results of an machine learning process, the reader should

Figure 2. Comparison of the Development and Validation of a Decision Rule vs a Machine Learning Model



A, A decision rule typically has a small number of parameters (eg, 5-10), such as the weights or points for each risk factor. These parameters are generally derived using a single development set and evaluated on 1 or more validation sets. B, Although the parameters of a machine learning model are similarly derived from the data, there are typically additional hyperparameters, such as learning rate, that affect the final derived parameters. These hyperparameters need to be tuned using a tuning set that is independent of the validation set to avoid overfitting.

assess whether the validation set was inadvertently used for tuning hyperparameters. A seemingly benign choice is that of selecting an operating point, also called a threshold, cut point, or cutoff. For example, if the output of machine learning is above some cutoff value, the feature the machine learning system is trying to identify is considered to be present. Cutoff selection using the validation set may hide calibration issues, such as when a machine learning model is trained using an enriched data set that has half of all its data containing the finding of interest. If used without further modification, the model may result in false-positive output when validated on a general patient population with only a small number of patients having the clinical entity the machine learning system is trying to identify.

Validation of Machine Learning Models

Is the Reference Standard High Quality?

Because many machine learning studies intend to demonstrate comparable performance of a clinically relevant task to clinicians, such as reading a radiograph or pathology slide, a key consideration is the quality of the reference standard. However, determination of the reference standard often requires subjective clinical judgement, which results in intrarater and interrater variability. This variability can be reduced by adjudication by a panel of experienced experts, for example, ensuring that retinal fundus photographs are graded and adjudicated by a panel of experienced retina specialists. Krause et al[29] showed the effect the quality of reference standard had on the reliability of the evaluation metrics. Using the majority vote of 3 retinal specialists as the reference standard, their machine learning model yielded an error (measured by 1−AUC) of 6.6%. However, when evaluated against a reference standard defined by the adjudicated grades from 3 specialists, error by the same machine learning model decreased to 4.6%,[29] a 30% relative reduction in errors (6.6%-4.6%). This difference in measured error rate was due solely to validating the machine learning model against a more rigorous reference standard (adjudicated vs majority vote). Thus, a high-quality reference standard is especially important for precise estimation of model and human performance to support model performance claims. To avoid bias, the reference standard must be determined independently—the clinicians grading images should be blinded to the machine learning predictions. These considerations are especially critical in studies that propose to use machine learning models to expand access to health care services; even a small difference in model performance can potentially affect a large number of patients.

Are the Results Unexpected?

If the study design is high quality, with respect to neither training nor tuning being performed on the validation set, the final consideration is a qualitative assessment of whether the reported performance is too good to be true on an absolute scale. Given sufficient high-quality training data and appropriate tuning, recent machine learning models can generally classify images with performance

Box 4. Using Convolutional Neural Networks to Detect Diabetic Retinopathy—Evaluating the Results[a]

How serious is the risk of bias?

Primary guides

Was there an independent, blind comparison with a reference standard?

To develop and test a machine learning algorithm, Gulshan et al,[3] worked with 54 ophthalmologists who were licensed in the United States or were final year (postgraduate year 4) ophthalmology residents to grade all the images used in their study. The graders were given a 19-image test to ensure they were proficient at reading retinal images, and as the study progressed, the graders' intragrader and intergrader consistency was determined. The graders then used a software system that presented the images to be graded along with a scale regarding the image quality and, if the image was of sufficiently high quality, the grade for diabetic retinopathy or diabetic macular edema. Each image in the development set was graded 3 to 7 times.

Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?

The data sets used in the study by Gulshan et al[3] were derived from hospitals and from clinics using the EyePACS system. Three eye hospitals in India contributed images (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya). In the United States, EyePACS was used. EyePACS clinics serve higher percentages of the Latino populations such that the EyePACS data set was enriched for Hispanic patients (≈55%), with white, black, and Asian patients each comprising approximately 5% to 10% of the population. A variety of different camera systems were used to obtain the images. The development set consisted of 128 175 macula-centered images of which 33 894 were from India and the rest from EyePACS sites. Thus, the retinopathy system developed by Gulshan et al[3]may not generalize to all populations and would need to be tested against non-Indian, non-Hispanic patients.

Was there a completely independent validation data set (and for machine learning prediction models was tuning reported)?

In the study by Gulshan et al,[3] 80% of the data were used to optimize the machine learning algorithm parameters, and 20% were used for tuning. The tuning set was used to determine when to stop the training process by terminating it when the area under the curve for the algorithm's performance reached a peak in the separate tuning data set.

Initially, Gulshan et al[3] used a subset of the development set of images from EyePACS for a validation study. This was not optimal since it is almost ensured in any modeling study that a statistical model or, in this case, a model derived by machine learning, will yield a very good fit from data derived from the same source from which the derivation data came. From a statistical and mathematical perspective, fitting data derived from the same population, irrespective of how it is sampled, is exactly the same process. For these reasons, split samples or cross-validation methods do not actually reflect true validation of any model. Thus, Gulshan et al,[3] also tested their machine learning system against a completely independent data set, the Messidor-2 publicly available fundus image database.

(continued)

Box 4. (continued)

Gulshan et al[3] used the most reliable graders from the derivation process to grade the validation images. Like the derivation process, each image in the validation set was graded multiple times (7 times on average).

Secondary guides

Did the results of the test being evaluated influence the decision to perform the reference standard?

Grading of the data sets used in the study by Gulshan et al[3] for developing the reference and validation studies was performed independently and for the purposes of the study and was not in any way influenced by the patient's clinical care.

Were the methods for performing the test described in sufficient detail to permit replication?

The methods described for obtaining the data sets and how they were analyzed by Gulshan et al[3] were extensively described and could be repeated by other investigators.

[a] Using convolutional neural networks to detect diabetic retinopathy is based on assessment of Gulshan et al.[3] Information in this box is based on Jaeschke et al.[6,7]

comparable with humans[26] (eg, accurate diagnosis of diabetic retinopathy on par with retinal specialists[29] and highly sensitive detection of individual tumors in large pathology images).[21] Notably, in these situations, the diagnostic performance of clinicians is limited by interrater and intrarater variability that results from factors such as subjectivity of image interpretation (eg, assessment of lesion size or severity), fatigue from grading many images or large images, and particularly in real clinical scenarios, limited time to assess the image. By contrast, machine learning methods have 2 major advantages: absolute consistency in performance without variability due to fatigue or external factors and by extension, the ability to exhaustively review every part of large images. More generally, the numerical precision of computer approaches may be advantageous for tracking subtle changes such as lesion size over time.

When machine learning results seem too good to be true, recall that machine learning methods can only be as good as the information in the training set. Therefore, machine learning methods should not be able to exceed the performance of extremely careful and experienced clinicians who have been given sufficient time to make a decision. However, because machines do not fatigue like people do, machines can outperform clinicians because they can rigorously examine large amounts of data and consistently arrive at the same result, whereas a clinician might overlook something.

There have been unexpected claims such as detecting previously unknown correlations, for example, the association between the cardiovascular risk factors of age and sex and retinal image findings.[30,31] In this instance, the machine learning tool correctly identified the self-reported sex of patients with a near-perfect AUC of 0.97. Because clear sex-specific anatomic differences had not been previously reported in retinal fundus photographs, this finding was particularly surprising. Independent researchers validated these results in another population,[31] increasing the confidence in the robustness of the machine learning model to find nuances in images that were previously not recognized by humans. When new unexpected associations are found by machine learning systems, the new

Box 5. Using Convolutional Neural Networks to Detect Diabetic Retinopathy—Applying the Results[a]

What were the results?

Are likelihood ratios, sensitivity, and specificity for the test results presented or data necessary for their calculation provided?

When the machine learning algorithm was optimized for high specificity, the specificity, when tested against the validation EyePACS data set, was 98%, and the sensitivity was 90%. The sensitivity for the Messidor-2 validation set was 87% and the specificity was 99%, showing the importance of validating any model against a completely independent data set. The independent validation set will have patients with different characteristics from those included in the derivation data set, resulting in a more realistic assessment of how the machine learning model will perform in actual clinical practice.

When the machine learning retinopathy algorithm was optimized for high sensitivity and tested against the EyePACS validation set, the sensitivity was 98% and the specificity was 93%. For the Messidor-2 validation set, sensitivity was 96% and specificity was 94%.

Will the results help me in caring for patients?

Will the reproducibility of the test result and its interpretation be satisfactory in my setting?

Gulshan et al,[3] developed a machine learning algorithm that when optimized for high specificity, proved to have high sensitivity and specificity when tested against a validation set developed from France. These results were promising given that the machine learning algorithm was derived from very large numbers of images in the derivation and validation sets. Because of the large numbers, the algorithm should consistently identify retinopathy findings on fundus photographs. However, the algorithm may not perform as well when images are derived from different photographic systems and from different patient populations than those used in derivation and validation sets.

In another study about independently developing a machine learning algorithm to detect diabetic retinopathy, the sensitivity was 91% and the specificity was 92% for detecting retinopathy in a multiethnic cohort of patients, suggesting that machine learning systems for detecting diabetic retinopathy are probably applicable to patient cohorts from a diverse range of racial and ethnic backgrounds.[4]

In another study Gulshan et al[3] used for developing the machine learning retinopathy screening algorithm (setting was 2 hospitals in India: Aravind Eye Hospital and Sankara Nethralaya), the automated retinopathy system performed reasonably well. Trained graders had a sensitivity of 73% to 90% and specificity ranged from 84% to 99%, as compared with grading by an expert retinal specialist who evaluated the same images. The automated diabetic retinopathy system's sensitivity was 89% at Aravind Eye Hospital and 92% at Sankara Nethralaya; the system's specificity was 92% at Aravind Eye Hospital and 95% at Sankara Nethralaya.[3]

Collectively, these studies show that an automated diabetic retinopathy screening system could work as effectively as having trained graders read fundus photography images when the algorithm was developed with the same patient population. It remains to be seen if this same algorithm will be as effective in patient populations other than those for which it was developed.

(continued)

Box 5. (continued)

Are the results applicable to my patient?

Unless patients have the same characteristics as those in the populations for which this particular algorithm was developed, it may not perform as well as was reported in the literature reviewed.

Will the results change my management?

Because the automated diabetic retinopathy screening system has not been validated in a population like the one you are managing, you cannot conclude it will change management.

Will patients be better off as a result of the test?

In theory, because many more patients require retinopathy screening than are resources available to achieve this screening, implementation of an automated system could benefit your patients. Whereas, the currently available systems may not work for your patient population, you conclude that if one could be validated against your patients, implementation of such a system might benefit them.

[a] Using convolutional neural networks to detect diabetic retinopathy is based on assessment of Gulshan et al.[3] Information in this box is based on Jaeschke et al.[6,7]

observations should be validated in additional patient cohorts to ensure that the results are not due to artifacts in the machine learning system, confounding factors, or flaws in the study design.

How to Detect Overly Optimistic Estimation of Model Performance

Readers should be aware of the potential of machine learning to overfit to the development set by learning patterns that appear only in that data set or by learning parameters that are too specific to the development set. This overfitting will manifest as low accuracy on new data sets, suggesting a lack of generalizability to other data sets. One way to detect overfitting is to compare the performance of the machine learning model in the tuning and validation sets, if both are reported. A large gap in performance between the tuning and validation sets may be indicative of overfitting to the tuning set. However, a variety of other factors, such as differences in patient populations (eg, age or disease subtype) or data source (eg, different imaging instruments or configurations) may also be responsible. As such, assessment of overfitting is an evaluation involving both technical machine learning expertise (eg, qualitative assessment of tuning-validation performance gap) and clinical intuition (eg, qualitative assessment of patient population differences between development and validation sets). Thus, a discussion with an experienced machine learning scientist about any flaws in the machine learning development may be helpful, in addition to a clinical assessment of the validation procedure.

Are Machine Learning Model Predictions Repeatable and Reproducible?

Repeatability and reproducibility are 2 critical aspects of measuring the consistency of machine learning model performance. When given the same image twice, the outputs of a given machine learning model should be identical. In the case of repeat imaging however, despite visual similarity, subtle changes in the numerical pixel values will alter the machine learning predictions. Before they are used clinically, the machine learning predictions to slight changes in pixel values between images taken via the same imaging hardware and protocol should be measured (ie, repeatability). More crucially, the machine learning predictions to differences in imaging hardware, operators, and protocol between institutions should be quantified (ie, reproducibility). In other words, outside of controlled laboratory conditions, there needs to be understanding about real-world conditions that affect the performance of machine learning models.

Considerations for Clinical Implementation

For What Purpose Can the Machine Learning Model Be Used?

Similar to how a diagnostic test can be used (in principle) for triaging, screening, or diagnostic purposes, a machine learning model, developed to perform a specific task, can be used for several purposes. For example, in a diagnostic application, machine learning may be helpful in 3 distinct phases: prediagnosis, peridiagnosis, and postdiagnosis. Before a diagnosis is made, the machine learning model may help prescreen patients to select only the highest-risk patients for further evaluation, reducing clinical workload.[32,33] In this manner, machine learning may expand the access of health care to underserved patient populations, such as by increasing availability of diabetic retinopathy screening to patients with diabetes in rural areas. During diagnosis, a machine learning model might improve the accuracy or efficiency of diagnosis by assisting clinicians with image review in real time for faster or more consistent detection of abnormalities in radiology, ophthalmology, or pathology images.[34-36] After diagnosis, machine learning models can be used for quality improvement by overreading images to detect diagnostic errors before patient care is affected.[37,38] Notably, regardless of exact purpose, the combination of evaluation by clinicians and machine learning can be more efficient and accurate than either alone.[34] Therefore a further worthwhile consideration is how to best leverage the complementary strengths of machine learning methods and clinician gestalt and experience.

In particular, the different uses influence how machine learning predictions should be presented to clinicians, also termed *user interface design*. For example, for detection of diabetic retinopathy, showing additional information about the part of the image that the machine learning model used to make predictions can be especially helpful for retina specialists.[35] In another example, for detection of metastatic breast tumors in sentinel lymph node biopsy, showing the raw predictions of each region of the image slowed pathologists down due to too much information. Conversely, highlighting only the most suspicious regions substantially expedited image review.[34] More generally, even simpler aspects, such as whether to consider the machine learning–predicted probabilities as opposed to a final classification like referable diabetic retinopathy, will require careful thought and clinical studies to measure the effect on diagnostic variability and patient care.

How Can the Machine Learning Model Be Implemented in Clinical Practice?

Unlike decision rules, the implementation of machine learning models into routine clinical workflows may be more complicated. Whereas decision rules can be applied by consulting a risk table, calculator, or even mental counting of risk factors, machine learning methods require computer programs. Because computers and electronic health records are now commonplace in routine clinical settings, the need for computers is not a barrier. However, whether the machine learning computation is performed on a local computer or remotely in the "cloud" has implications for patient privacy, workflow integration, and maintenance of these programs, and requires careful thought.

Measuring and Monitoring Clinical Effect

Even if a machine learning model has been thoroughly validated in different studies and the logistical, technical, and regulatory hurdles have been overcome for integration into the clinical workflow, the system still requires further research to measure the system's clinical effectiveness. Several aspects of clinical effectiveness can be measured and tracked, including patient outcomes and costs. More subtly, adverse effects on clinician workloads and behaviors must be assessed to avoid increasing fatigue that might result from clinicians needing to respond to false-positive machine learning reports that could blunt human responsiveness to real problems identified by machine learning systems. In addition, machine learning models may result in clinician overreliance on automated systems, resulting in errors caused by faulty machine learning processes. The overall usefulness and safety of machine learning systems is ideally assessed through large randomized controlled trials, such as those that were performed to evaluate the Ottawa ankle rules.[16] However, like after-market surveillance for drugs, continued monitoring of machine learning systems is essential to help detect unexpected problems that may arise from changes in practice or patient populations.

Updating the Machine Learning Model Over Time

Machine learning models differ from decision rules because the accuracy of machine learning models can be improved over time, as exemplified by an improvement in diabetic retinopathy grading from being comparable with ophthalmologists[3] to being on par with retina specialists.[29] These improvements were due to better machine learning methods and data such as adjudicated labels from retina specialists. In addition, an increase in data set size also substantially improves machine learning model performance.[3] Thus, in addition to updating the machine learning model over time, as a response to changes in practice or patient population, ongoing data collection will lead to improved machine learning models, though with gradually diminishing returns. Partially due to recognition that updates to improve accuracy of machine learning models can increase the quality of care, the US Food and Drug Administration is testing the Digital Health Software Precertification (Pre-Cert) Program to facilitate faster approvals where appropriate.[39]

Resolving the Clinical Scenario—Using the Guide

Machine learning is a powerful new tool that greatly expands the ability to understand the relationship between data and some clinical features such as retinopathy lesions on a fundus photograph. Even though machine learning greatly expands the ability to analyze data, its implementation in clinical practice should follow the same rules that previously existed for assessing diagnostic tests (Box 1).[6,7] The chief medical officer seeking to assess the literature on using machine learning to diagnose retinopathy finds 2 articles on the topic. The first article developed and validated a machine learning system to automatically read retinal photographs to determine if retinopathy was present.[3] In **Box 4** and in **Box 5**, we evaluate the article by Gulshan et al[3] using the Users' Guide to Assessment of Diagnostic Studies.[6,7]

Conclusions

Machine learning is not new in medicine and has been used productively in simpler incarnations as clinical decision rules. Clinicians should verify the validity and impact of machine learning methods just like any other diagnostic or prognostic tool. Readers of studies reporting the results of machine learning systems should assess the most crucial elements of machine learning model validation, such as whether the study design over-represents model performance through inappropriate hyperparameter tuning or a poor-quality reference standard. Crucially, the machine learning model has to be validated on an independent data set not used for training or tuning the model. Finally, clinical gestalt plays a crucial role in evaluating whether the results are believable: because one of the biggest strengths of machine learning models is consistency and the lack of fatigue, a useful check for believable machine learning results is whether an experienced expert could reproduce the claimed accuracy given an abundance of time. Results that substantially exceed what even such a hypothetical expert is capable of should be scrutinized and validated carefully.

ARTICLE INFORMATION

Author Contributions: Dr Liu had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs Liu and Chen contributed equally to this article.
Concept and design: All authors.
Acquisition, analysis, or interpretation of data: Liu, Chen, Peng.
Drafting of the manuscript: Liu, Chen.
Critical revision of the manuscript for important intellectual content: All authors.
Administrative, technical, or material support: Liu, Chen, Peng.
Supervision: Liu, Chen, Peng.
Other - machine learning expertise: Liu, Chen, Krause.

Conflict of Interest Disclosures: Dr Liu reported holding a patent in machine learning to each of the following: analyzing retinal fundus photographs and another for analyzing histopathology slides (status pending or granted), analyzing skin conditions (status pending), and analyzing physiological signals (status granted); and being an employee of Google and holding Alphabet stock (part of the compensation package). Dr Chen reported being an employee of Google and holding Alphabet stock (part of the compensation package) . Dr Krause reported receipt of personal fees from Stanford University outside the submitted work and being an employee of Google and holding Alphabet stock (part of the compensation package) . Dr Peng reported holding a patent to the each of the following: predicting cardiovascular risk factors in retinal fundus photographs using deep learning, fundus imagery machine learning systems, health predictions from histopathology slides, and pathology heatmap predictions (status pending for all); and being an employee of Google and holding Alphabet stock (part of the compensation package).

REFERENCES

1. American Academy of Ophthalmology. Diabetic Retinopathy PPP—Updated 2017. https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017. Published December 18, 2017. Accessed September 16, 2019.

2. Walton OB IV, Garoon RB, Weng CY, et al. Evaluation of automated teleretinal screening program for diabetic retinopathy. *JAMA Ophthalmol*. 2016;134(2):204-209. doi:10.1001/jamaophthalmol.2015.5083

3. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

4. Ting DSW, Cheung CY-L, 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. doi:10.1001/jama.2017.18152

5. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M-L, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. *JAMA Netw Open*. 2018;1(5):e182665. doi:10.1001/jamanetworkopen.2018.2665

6. Jaeschke R, Guyatt G, Sackett DL; Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA*. 1994;271(5):389-391. doi:10.1001/jama.1994.03510290071040

7. Jaeschke R, Guyatt GH, Sackett DL; Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA*. 1994;271(9):703-707. doi:10.1001/jama.1994.03510330081039

8. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS; Evidence-Based Medicine Working Group. Users' guides to the medical literature, XXII: how to use articles about clinical decision rules. *JAMA*. 2000;284(1):79-84. doi:10.1001/jama.284.1.79

9. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391

10. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Ann Emerg Med*. 1992;21(4):384-390. doi:10.1016/S0196-0644(05)82656-3

11. Lucchesi GM, Jackson RE, Peacock WF, Cerasani C, Swor RA. Sensitivity of the Ottawa rules. *Ann Emerg Med*. 1995;26(1):1-5. doi:10.1016/S0196-0644(95)70229-6

12. Kelly AM, Richards D, Kerr L, et al. Failed validation of a clinical decision rule for the use of radiography in acute ankle injury. *N Z Med J*. 1994;107(982):294-295.

13. Stiell I, Wells G, Laupacia A, et al. Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries: Multicentre Ankle Rule Study Group. *BMJ*. 1995;311(7005):594-597. doi:10.1136/bmj.311.7005.594

14. Auleley G-R, Kerboull L, Durieux P, Cosquer M, Courpied J-P, Ravaud P. Validation of the Ottawa ankle rules in France: a study in the surgical emergency department of a teaching hospital. *Ann Emerg Med*. 1998;32(1):14-18. doi:10.1016/S0196-0644(98)70093-9

15. Papacostas E, Malliaropoulos N, Papadopoulos A, Liouliakis C. Validation of Ottawa ankle rules protocol in Greek athletes: study in the emergency departments of a district general hospital and a sports injuries clinic. *Br J Sports Med*. 2001;35(6):445-447. doi:10.1136/bjsm.35.6.445

16. Auleley GR, Ravaud P, Giraudeau B, et al. Implementation of the Ottawa ankle rules in France: a multicenter randomized controlled trial. *JAMA*. 1997;277(24):1935-1939. doi:10.1001/jama.1997.03540480035035

17. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: Users' Guides to the Medical Literature. *JAMA*. 2017;318(14):1377-1384. doi:10.1001/jama.2017.12126

18. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63. doi:10.7326/M14-0697

19. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36. doi:10.1148/radiology.143.1.7063747

20. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

(TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-73. doi:10.7326/M14-0698

21. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al; the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

22. Antman EM, Cohen M, Bernink PJ, et al. The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. *JAMA*. 2000;284(7):835-842. doi:10.1001/jama.284.7.835

23. Olah C, Mordvintsev A, Schubert L. Distill website.Feature visualization: how neural networks build up their understanding of images. https://distill.pub/2017/feature-visualization/. Accessed October 25, 2019.

24. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. Red Hook, NY: Curran Associates; 2012:1097-1105.

25. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

26. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211-252. doi:10.1007/s11263-015-0816-y

27. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. 2007;165(6):710-718. doi:10.1093/aje/kwk052

28. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. Paper presented at: 5th International Conference on Learning Representations, ICLR 2017; April 24-26, 2017; Toulon, France. https://dblp.org/db/conf/iclr/iclr2017. Accessed October 11, 2019.

29. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125(8):1264-1272. doi:10.1016/j.ophtha.2018.01.034

30. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0

31. Ting DSW, Wong TY. Eyeing cardiovascular risk factors. *Nat Biomed Eng*. 2018;2(3):140-141. doi:10.1038/s41551-018-0210-5

32. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep*. 2016;6:26286. doi:10.1038/srep26286

33. Raumviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;10(2):25. doi:10.1038/s41746-019-0099-8

34. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018;42(12):1636-1646. doi:10.1097/PAS.0000000000001151

35. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. (December 2018). doi:10.1016/j.ophtha.2018.11.016

36. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. *Invest Radiol*. 1990;25(10):1102-1110. doi:10.1097/00004424-199010000-00006

37. Halligan S, Mallett S, Altman DG, et al. Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: multiobserver study. *Radiology*. 2011;258(2):469-476. doi:10.1148/radiol.10100354

38. White CS, Pugatch R, Koonce T, Rust SW, Dharaiya E. Lung nodule CAD software as a second reader: a multicenter study. *Acad Radiol*. 2008;15(3):326-333. doi:10.1016/j.acra.2007.09.027

39. US Food and Drug Aministration. Digital health software precertification (Pre-Cert) program. https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default.htm. Accessed January 27, 2019.

40. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA*. 2018;320(11):1101-1102. doi:10.1001/jama.2018.11100