Advances in Laboratory Medicine through Artificial Intelligence

Emily R. Thompson and Liam J. Reynolds

Emily R. Thompson, Department of Pathology and Laboratory Medicine, University of California, San Diego, USA; Liam J. Reynolds, Centre for Bioengineering and Analytics Research, National University of Ireland, Galway, Ireland

Abstract: This review focuses on machine learning and on how methods and models combining data analytics and artificial intelligence have been applied to laboratory medicine so far. Although still in its infancy, the potential for applying machine learning to laboratory data for both diagnostic and prognostic purposes deserves more attention by the readership of this journal, as well as by physician-scientists who will want to take advantage of this new computerbased support in pathology and laboratory medicine.

Keywords: artificial intelligence; diagnostic aids; literature review; machine learning.

Introduction

In light of recent high-profile articles and editorials in high-impact journals (e.g. [1–3]), it appears that, with the decline of so-called "expert systems", machine learning (ML) has gained a place in medicine and captured the interest of medical researchers and practitioners in predictive methods within this subfield of computer science. The ever wider use of ML in clinical and basic medical research is reflected in the number of titles and abstracts of papers indexed on PubMed and published until 10 years ago (2006) as compared to the last 10 years (2007–2017), with a nearly 10-fold increase from 1000 to slightly more than 9000 articles (see Appendix for the detailed queries) in the past decade. In this short review, we will introduce what ML is in terms that we believe physicians can easily grasp, and we will then survey the most recent applications of this computational approach to laboratory medicine.

A gentle introduction to machine learning

Put in very general terms, ML is about *learning by machines*. More specifically, ML is an umbrella term for diverse computational methods by which machines can incrementally build an accurate data model according to a measure of how well the model supports a given task, which in medicine is usually of discriminative nature, i.e. classification or clustering. Here, two technical terms need to be distinguished: model and task. In an ML context, by model, we refer to the *functional representation* of a data set, i.e. the representation of any mapping that can be drawn to bind portions of the data set to a particular value on a specific measurement scale. The scale is either nominal or ordinal in a discriminative classification task, and the value is usually a label. In a discriminative regression task, on the other hand, the scale is either an interval or a ratio, and the predicted value is a number indicating some quantity, e.g. creatinine levels.

Most of the ML models described in the medical literature so far regard functional mapping between a set of values, likely associated with a single clinical case, and a single category (e.g. yes/no or one class out of a taxonomy) in order to support either a diagnosis or a prognosis. To illustrate, let's call this set of values x: in a prognostic decision task, the model is applied to answer questions like "does x represent (or are values pertaining to) a patient who is affected by a certain disease or not?" In a supervised ML context, the data are usually data sets that describe different cases along various dimensions or attributes, called features, and that human experts have already associated with "correct" values, which we shall call "y". Therefore, in the very concise terms that data scientists love, ML models work with functions like $y = f(x)$: the value of this function lies in its capability to yield the correct "y" also for some "x" that has not previously been classified by a human expert, thus providing an aid in the classification task.

Data scientists program machines, called "learners", to optimize a given model (i.e. make it more accurate in predicting y on the basis of a given unknown x taken from the target population) by autonomously and iteratively tweaking its parameters until no further improvement can (or should) be achieved: the process that the learners apply to a part of the available data, the "training data", without the direct intervention of human programmers, is called *learning*, in this case, machine learning. Mitchell [4] once put it in the following general terms: any machine (and, in fact, any system, be it software or human) is said *to learn in regard to a task T from experience E if, given a performance metrics P* (also called objective function, scoring function or loss function), *the latter's scores "improve" over time in carrying out T*. A common example of P is the number of errors made over the

number of attempts, whereby the system is said to learn from experience if P decreases over time. A sound ML approach needs the proper expression (i.e. not ambiguous and possibly formal) of three things: (1) the *task* (T) to be learned; (2) the *experience* (E) to be learned from the training data, and validate the model on using the testing data); and (3) the performance metrics (P) to evaluate the learning process so far: a PET. In similar concise terms, Domingos [5] proposed considering all ML problems as the "combination of just three components": (1) *representation*, basically of T and E, i.e. the model; (2) *optimization* of the model's parameters according to a portion of the Experience data called "training data" so as to achieve a better-scoring P; and (3) *evaluation*, i.e. computing P over a representation of T that is applied to the testing data other than the training data to see if the model generalizes well enough for some practical aim (Figure 1).

A commonly used technique to make full use of available data in both the optimization and the evaluation phase is k-fold cross-validation, which requires randomly partitioning the available data into a number of smaller and equally sized subsets. Each subset is held out, in turn, as a test set, whereas the others are used for training. In so doing, the average model performance on unseen data can be estimated more precisely.

This process is often depicted as being more computational than it actually is. In fact, training "PETS" in ML often requires substantial work by domain experts in activities to make ML effective. For instance, collect accurate and representative data to train the model with the old "garbage in, garbage out" principle that has full and, indeed, reinforced application in ML [6]; select the features for optimizing the model and transform them to facilitate the process (activities usually entailing feature selection and feature engineering with close collaboration between domain experts and data scientists); and choose the most suitable model family for the specific task (e.g. support vector machines, random forest, Bayesian classifiers, artificial neural networks [ANNs] to mention only the most common ones – a detailed description of these methods is beyond the scope and aims of this paper).

A recent survey has tested an impressive number of different classifiers (n = 179) on a likewise impressive number of different data sets (n = 121) and concluded that random forest and support vector machines (with Gaussian kernel) are the best performing models [7]. That said, empirical studies (e.g. [8]) have shown, and we concur, that, because the nature of the data is more important than the learning technique in many applications, the importance of the human-in-the-loop factor in ML cannot be undervalued [9]. Moreover, the choice of which ML model to apply is not only a matter of accuracy but also impacts on performance trade-offs. This is where domain experts are involved in deciding when the model is ready for application to data other than training data, i.e. to make predictions for medical decision making. Of note is that the learning process is not an asymptotical process that can continue until accuracy reaches 100% (hopefully). For such extreme accuracy would not be desirable owing to the high risk that it only reflects the phenomenon of over-fitting. Overfitting occurs when the model is very good in being "fit" to the "experience" data used in the training phase (or in the test phase when the training data and the test data share similar characteristics and both fail to fully represent the source population), but the model generalizes poorly on any new case that is presented to it.

Two compromises are key to the success and usefulness

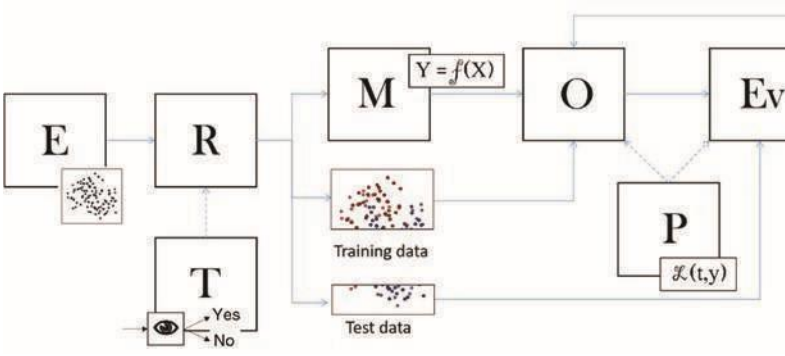

Figure 1: Diagram representing the main concepts of a typical supervised ML process. E denotes experience (data); R, their representation according to the task; T, the classifying classifier; O, the optimization step; P, the performance metrics; Ev, the evaluation, here re predicting y when the true value is t. The feedback loop iterating through the rightmost se by the machine.

of any ML model: the first regards the trade-off between bias and variance, which are the two main components of the expected (average) prediction error of an ML model (besides a portion of irreducible error). Bias is the error due to the difference between the average prediction of the model and the correct value that it is trying to predict, whereas variance is the amount that the prediction of the model will change if different training data are used, thus reducing the "generalization performance" of the model. Figure 2 depicts these dimensions by representing bias and variance as

functions of model complexity. Model complexity can be considered as being intuitively related to the number of features, the model type and the number of its parameters.

Figure 2 also suggests that there is an optimal complexity in which over- or underfitting can be avoided or minimized. Overfitting occurs when the model represents the training data so well that it probably also represents the intrinsic noise in it, which means that it will likely exhibit a relatively low accuracy when presented with new data other than the training data. Underfitting, in contrast, is when the model is too simplistic and hence of little predictive value. A common way to reduce variance, prevent overfitting, and find the optimal "middle" complexity is to evaluate the performance of the ML model with a loss function that is corrected with a so-called regularization term, usually a sort of penalty for complexity. The loss function can be seen as a way to quantify the difference between good and bad models, where the idea is to represent with such a function both a cost term and a regularization term and to optimize the sum of both. Each model family has its own way to be evaluated (where simpler solutions are usually preferable); however, this is beyond the scope of the present review. Suffice it to add that, with reference to Figure 3, regularization techniques applied to ML models make decision boundaries smoother and more regular.

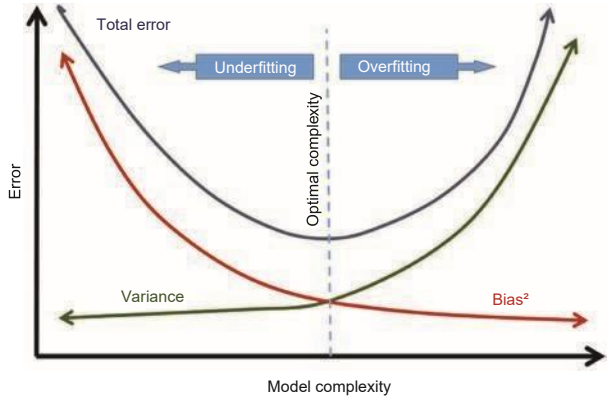

Figure 2: Bias and variance of a model contributing to the total prediction error.
Adapted from a diagram by Fortmann-Roe in [10].

the receiver operating characteristic (AUROC) curve, which describes the discriminative power of a classifier model for multiple cut-off points independent of both class distribution and unequality of prediction error costs.

The second subtle trade-off occurs between accuracy and interpretability. Accuracy regards how good the model is at predicting y, given x. Measuring this kind of performance is not trivial, and simple methods that count the number of misses with respect to the total number of tries should be avoided (although they are still common in the specialist literature). This is due to the simplistic assumption that all errors are equal and the vulnerability of ML to the "accuracy paradox". The accuracy paradox occurs when one predictive class is much more frequent in the training data set than the others, so that the model "cheats" by predicting that class for any given x. Rather than plain (overall) accuracy (which is based on a single cut-off point or pattern separating positive from negative cases), other measures are recommended to evaluate model performance [11], including nosologic indexes of specificity, sensitivity (or recall), precision and the harmonic mean of the latter two (also called the F_1 score), mainly for their quality of being independent of prevalence or class unbalance or, even better, the area under

Irrespective of the evaluation metrics used, rule-based models (e.g. decision trees, naïve-Bayes) usually perform worse than black-box models (e.g. random forests, support vector machines, deep convolutional neural networks), with some striking exceptions, as reported by Kumar and Sahoo [12] where laboratory data were used to predict different types of liver diseases. However, the former models can be inspected for higher interpretability of the predictions generated, whereas for the latter models, as their name suggests, it is much more difficult to understand why they make a particular prediction (unless some linear local-

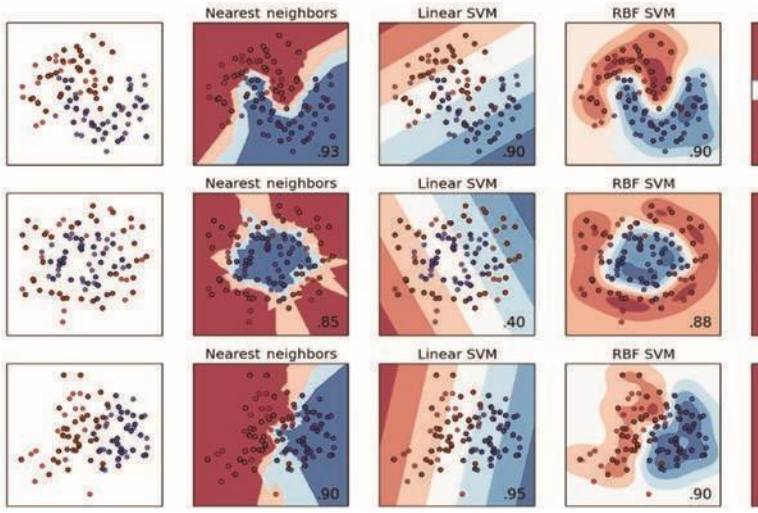

Figure 3: Visual comparison of the decision boundaries of several common ML classifie
The plots show the training data as points in darker shaded colors and the testing data a
is given in the bottom right corner. Adapted from an image by scikit-learn.org © 2010–

interpretable model is used in a *post hoc* manner, as described by Ribeiro et al. [13]). Although many physicians would concur with Kuhn and Johnson [14] that "as long as the model can be appropriately validated, it should not matter whether it is a black box or a simple, interpretable model", others could make a strong argument that this would be unethical in medicine. In light of these contentions, we believe that more questions could be asked after the deployment of an ML-driven diagnostic aid in actual practice: has this aid brought value to this practice? Has it had a positive impact on clinical outcome? Also, further research and a stronger evidence-based approach should be applied to ML assessment in medicine, and to laboratory medicine in particular, as various authors have strongly advocated [15].

A short review of ML applied to laboratory medicine

Machine learning (ML) is routinely used in biochemical development and for evaluating and interpreting data in genomics, transcriptomics and proteomics pathways (e.g. [16, 17]), whereas in clinical laboratory medicine, it has been applied to classical biomarker testing of biological materials. Several so-called "expert systems" have been recently described, patented and commercialized for clinical laboratory purposes. Designed to evaluate specific data in hematology, urinalysis or clinical chemistry, they are traditionally based on a predefined decision tree encompassing logic rules and checks to exclude diagnostic hypotheses or define them or suggest further analysis to complete the diagnosis and support decision making. By contrast, ML is a completely different approach, where "rules" are learned by the machine. More often than not, speaking of explicit rules is inappropriate, as the prediction is somehow hidden in the model's non-linear parameters that bend the decision boundaries around the data, literally (Figure 3).

In order to explore how widespread the current use of ML is in laboratory medicine, we performed a database search using the search terms *laboratory medicine* and *machine learning* in the whole article content (see Appendix), which retrieved 781 papers in Scopus and only 57 papers in PubMed. More specific queries using the search terms *laboratory medicine*, *laboratory tests* and *machine learning* in either the title or abstract identified 34 papers in Scopus and only three in PubMed, one of which was a research journal article [18]. The difference between the Scopus and PubMed databases can be easily explained by the inclusion of keywords in the search query in the Scopus database. Further analysis of the papers retrieved from Scopus showed that few were pertinent to the aim of this literature review. Finding so few articles on the application of ML to traditional laboratory parameters was somewhat unexpected, as the laboratory is regarded as the main supplier of quantitative, structured and codified data in clinical medicine. Nonetheless, we expect that ML methods will become more extensively used in the analysis of laboratory parameters, and especially for data that can be easily grouped and compared across different groups. The application of ML in laboratory medicine should be supported as a means to enhance laboratory organization and expand the core skills set of laboratory experts, within a broader process of change and innovation (e.g. [3, 19]).

We claim this for a number of reasons. First, laboratories are a major part of today's healthcare systems. However, despite high throughput with low turnaround times, the capacity to screen data for results of special interest has decreased and few tests are directly diagnostic [20]. Second, technological advances have enabled the integration of expert system capabilities and software applications, including autoanalyzers and modules of laboratory information systems (LIS) [21]. Since this kind of support is usually based on dichotomous thresholds or rigid mutual exclusion of data, it can be difficult if not impossible to obtain precise or personalized results [22], suggesting an obvious margin for improvement.

Third, because patients can now directly access their laboratory test results by downloading them from the Web portal (and increasingly through smartphone apps) of their diagnostic provider, there is an increasing demand for meaningful, possibly personalized reference limits and the need to interpret precision asterisks [20], the conventional signs indicating abnormal or borderline values. Finally, with the convergence of smartphones and innovative biosensors based on microfluidics and microelectronics, the vision of the lab-on-a-chip (LOC) and related models for laboratory medicine has opened opportunities, "in which a smartphone-enabled portable laboratory is brought to the patient instead of the patient being brought to the laboratory" [23].

In this context, apomediation refers to progressive disintermediation whereby traditional intermediaries, such as healthcare professionals who give "relevant" information to their patients, are functionally replaced by apomediaries, i.e. network/group/collaborative filtering processes [24]. ML systems can be seen as new and "smarter" apomediaries that act as gap fillers that analyze the increasing amount of diagnostic data a patient can access without mediation by a general practitioner or laboratory specialist, and then refer the patient to a

specialist only in case of likely positive or anomalous results. This can be done by factoring together the diverse phenotypic attributes of a patient (i.e. in addition to body mass index [BMI], age, gender and ethnicity) or, better yet, of the patient's history of past basal values associated with a healthy condition. In this case, the very notion of reference limits would change, and ML, by leveraging and improving other statistical approaches, could help limit the misinterpretation of values outside of reference limits or of apparently normal data but also diagnostic for some conditions (e.g. [25]).

Furthermore, some envision an ML-based clinical decision support that, by predicting correlated test results and enhancing the diagnostic value of multianalyte sets of test results, could help to reduce redundant laboratory testing [26] and, hence, lower healthcare costs, which are estimated to total $5 billion yearly in the United States alone [27]. Finally, the growing number of available and affordable types of diagnostic tests, different measurement methods and patient phenotypes (e.g. ethnic subtypes) has produced an unprecedented complexity of data interpretation and integration that calls for novel management technologies. In the following section, we report on research into the potential of ML models to address these challenges in laboratory medicine.

ML models in laboratory medicine

Diri and Albayrak [28] evaluated the performance of four classifiers applied to the data from five lab tests for thyroid dysfunction (distinguishing between a diagnosis of euthyroidisim, hypothyroidism and hyperthyroidism). The tests were as follows: T3-resin uptake test, total serum thyroxin, total serum triiodothyronine, basal thyroid-stimulating hormone (TSH) and maximal absolute difference of TSH value after injection of 200 μg of thyrotropin-releasing hormone as compared to basal value. The authors also used cobwebs to visualize classifier performance when the data had more than two classes. A Bayesian classifier showed the best overall performance, with an average accuracy of 96%.

In a study involving 757 patients, Nelson et al. [29] applied logistic regression and an ML model they called "a relevance vector machine" and found that creatinine level was a clear predictor of outcome in traumatic brain injury, whereas glucose, albumin and osmolality levels were predictors depending on the model used.

Lin et al. [30] mined concepts from clinical narratives and lab values gleaned from electronic medical records to automatically discover rheumatoid arthritis. After experimenting with a range of ML algorithms, they found that the linear kernel support vector machines performed best, with an AUROC curve of 0.83, after which also inflammatory markers were considered (with a 6% increase as compared to no laboratory test).

Given the complicated characteristics of warfarin, Liu et al. [31] used two well-known lab tests, alanine aminotransferase (ALT) and serum creatinine (SCr), in combination with data about warfarin dose, gender, age and weight, to build a classification model that could predict adequate or inadequate warfarin therapy and minimize the odds of drug-to-drug interactions. In an analysis of 377 inpatients, they compared the performance of seven classification techniques and found that C4.5 decision tree and random forest scored best and predicted the adequacy of warfarin "more accurately than does the clinical physicians' subjective decision". This result, the authors claimed, showed the importance of making the best use of lab test results in clinical practice, especially in virtue of the relative simplicity and low cost of collecting accurate lab data.

Razavian et al. [32] collected administrative claims, pharmacy records, healthcare utilization and lab test results of 4.1 million individuals between 2005 and 2009 to evaluate a prediction ML model for type 2 diabetes. Among the different variables associated with the development of diabetes, high ALT concentrations were associated with the highest odds ratio. Among the lab tests, the best prediction variables were glycated hemoglobin (HbA_{1c}), glucose, high-density lipoprotein cholesterol, carbon dioxide and glomerular filtration rate (GFR). The authors remarked that the study also showed how administrative data can be a powerful tool for population health management and clinical hypothesis generation for risk factor discovery, and that these data can help guide interventions in at-risk populations.

Putin et al. [33] applied ML to laboratory parameters to predict chronological age via an ensemble of 21 deep neural networks developed and applied to more than 50,000 samples. They found that albumin concentration, followed by glucose, best identified chronological age. The ensemble identified five markers (albumin, glucose, alkaline phosphatase [ALP], urea and erythrocytes) as the most valuable for predicting subject age.

Yuan et al. [34] built and evaluated three classifiers based on supervised ML methods to discriminate between positive and negative urine samples. Based on a classification and regression tree (CART), the model showed the best results on the test set, with a sensitivity of 86.0%, a specificity of 98.0%, an AUC of 94.3% and an overall accuracy of 95.6%. The results suggested that ML is a valuable method to

construct classifiers for urine microscopic review rules that can supplement other reported microscopic review rules.

Dermici et al. [18] used a commercial software program to train an ANN for application in the central laboratory of a large university hospital for the efficient, rapid and reliable evaluation of biochemical test results. The ANN was applied to more than 250,000 samples to evaluate a set of routine parameters (sodium, potassium, calcium, magnesium, glucose, uric acid, chloride, urea, creatinine, aspartate aminotransferase, ALT, gamma-glutamyl transferase [GGT], ALP). Evaluation by ANN was compared with evaluation by seven pathologists of different expertise. The sensitivity of the model was 91% and specificity was 100%, with a K score of 0.95. The K score analysis revealed that five out of seven pathologists gave very high agreement scores in the evaluation of model judgment (0.81–1.00). When a reassessment of the specialists' decision was requested, after comparison with the ANN evaluation, the pathologists changed their reports significantly in many cases, so as to increase agreement between the human and the automatically generated report. The time between receipt of the data and release of the reports was clearly lower in the case of ANN. The authors concluded that a decrease in time and related costs, at similar quality and appropriateness levels, can be expected from the introduction of similarly accurate automatic supports.

Proponents of ML methods assert that ML may be useful for prognosis, i.e. predicting disease evolution and progression, early detection, when a disease is still in its early, asymptomatic stage and primary prevention to reduce the risk of development of disease. The classic predictive approach has been based on regression models, e.g. the logistic model to predict 30-day mortality risk for patients with ST-segment elevation myocardial infarction (STEMI), the Weibule model for the SCORE (systematic coronary risk evaluation) model and a Cox model applied to the Framingham Risk Score for cardiovascular diseases. Goldstein et al. [35] described an ML method for cardiovascular risk prediction that was trained with the data of 1944 patients with a primary diagnosis of acute myocardial infarction. The authors used 13 lab parameters measured in at least 80% of the patients (calcium, carbon dioxide, creatinine, creatine kinase-MB, hemoglobin, glucose, mean corpuscular volume, mean corpuscular hemoglobin concentration, platelets, potassium, red cell distribution width [RDW], sodium, leukocytes) and calculated the median and the minimal and maximal values of these parameters to obtain 43 predictor variables of hospital mortality. The ML model trained on this data set showed that there is a non-linear relationship between calcium and hemoglobin and postinfarction mortality. The authors applied five ML approaches to build models with different characteristics and performance: the variables were similarly relevant and the models detected the high impact of carbon dioxide (minimum value), calcium (all measures), hemoglobin (median value), potassium (all measures) and leukocytes (maximum value). Chen et al. [36] developed an ML model for predicting changes in GFR in Chinese patients with type 2 diabetes. Because current GFR equations (Cockcroft and Gault, Modification of Diet in Renal Disease formula, Chronic Kidney Disease Epidemiology Collaboration) are known to be inaccurate in persons with diabetes, a model including sex, age, serum creatinine and BMI provides optimal modification of these equations in such patients.

Another example of ML models that evaluate the effectiveness and efficacy of well-established lab tests is the evaluation of the role of tumor markers in cancer diagnosis in asymptomatic subjects. Tumor marker testing is currently recommended for diagnostic assessment and especially during follow-up after chemotherapy but not during screening. Multiple tumor marker applications could be used even in the screening phase, with the probabilistic power of a group of molecules reaching the proper edge, sufficient to identify asymptomatic disease. This approach is now supported by new metabolomics and proteomic approaches. Surinova et al. [37] found that five proteins not routinely measured in laboratories were selected by an ANN from among approximately 300 secreted and cell surface candidate glycoproteins, which could represent a panel for the early diagnosis of colorectal cancer before clinical symptoms appear. Classical tumor markers, however, even when grouped by increasing sensitivity and specificity, are not useful for cancer screening in apparently healthy subjects, as reported by Wang et al. [38] who used an ML method to study its diagnostic power in screening with the tumor markers AFP, CEA, CA 19.9, CYFRA 21.1 and SCC, in addition to PSA for men and CA 15.3, CA 125 for women. Evaluation of tumor marker screening in approximately 21,000 individuals showed an inadequate positive prediction value, a reduction in absolute risk and an increase in absolute risk. The authors concluded that the combined tests should not be proposed for cancer screening.

An Australian epidemiological expert group reported on examples of ML applied to biochemical and hematological tests, including the demonstration of an interrelationship between GGT and liver function tests (ALP, albumin, lactadehydrogenase and aminotransferases), enhanced prediction of hepatitis B and

C through the use of hepatitis C virus and the correlation between RDW and hemoglobin in anemia diagnosis [39].

Luo et al. [40] investigated the feasibility of an automated clinical decision support to predict test results using the results from other tests. As a proof of concept, they showed that ML models based on patient demographics (age and gender) and results of other lab tests (each collection had a median of 23 of the 40 tests) can discriminate normal from abnormal ferritin results with a high degree of accuracy (AUC 0.97, held-out test data) and even predict numerical results for ferritin (by regression) with moderate accuracy. They also claimed that predicted ferritin results could better reflect underlying iron status than measured ferritin in some cases. Their results were shared by other studies, like that of Waljee et al. [41], who found that the missForest model outperforms other methods for imputing missing laboratory results.

Somnay et al. [42] used serum levels of preoperative calcium, phosphate, parathyroid hormone, vitamin D and creatinine as potential predictors of primary hyperparathyroidism in a sample of 11,830 patients. Among the ML algorithms tested, the Bayesian network models proved most accurate, correctly classifying 95% of all primary hyperparathyroidism patients (AUROC 0.99). Interestingly, omitting parathyroid hormone from the model did not substantially decrease its accuracy. The study concluded that ML can accurately diagnose primary hyperparathyroidism without human input even in cases of mild disease.

Lastly, as mentioned above, ML may be applied to identify reference limits for lab parameters. Reference intervals should be defined for each specific test by the laboratory, considering the method used, the preanalytical phase and the type of population accessing the diagnostic service. However, because setting reference ranges is difficult, expensive and time consuming [43], reference ranges are generally collected from the literature or adopted from those suggested by the laboratory test manufacturers. The methods described in the specialist literature are usually based on traditional, descriptive statistical approaches and used to obtain reference limits directly from laboratory data. This is feasible, especially when a large amount of outpatient data are available or when the population is currently known as healthy or has a low prevalence of disease.

Conclusions

The potential application of ML models to laboratory data is relevant but not yet fully realized. Although it is reasonable to expect that as ML methods become better known they will be applied to reduce costs, support clinical decision making and improve outcomes, we also advocate for further research to address the main obstacles to a wider adoption and exploitation of these methods in laboratory medicine. Further study is also needed to understand whether and how the best ML practices can be advantageously transferred to laboratory medicine from other areas that pioneered this computational approach, like cardiology, oncology and radiology, to tap into the related opportunities and strengths and avoid threats and weaknesses [44].

References

1. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. J Am Med Assoc 2016;315:551–2.
2. Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30.
3. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216.
4. Mitchell TM. Machine learning. McGraw Hill series in computer science, 1997.
5. Domingos P. A few useful things to know about machine learning. Commun ACM 2012;55:78–87.
6. Cuda J, Seigh L, Clark K, Monaco S, Pantanowitz L. Utilizing computerized provider order entry (CPOE) to reduce the garbage in garbage out effect in the cytology laboratory. J Am Soc Cytopathol 2016;5:S85.
7. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems. J Mach Learn Res 2014;15:3133–81.
8. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Medinfo 2004;11(Pt 1):736–40.
9. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 2016;3:119–31.
10. Fortmann-Roe S. Understanding the Bias-Variance Tradeoff. http://scott.fortmann-roe.com/docs/BiasVariance.html. Date last accessed 31 Mar 2017. Archived at: http://archive.is/z4cf.
11. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Tech 2011;2:37–63.
12. Kumar Y, Sahoo G. Prediction of different types of liver diseases using rule based classification model. Technol Health Care 2013;21:417–32.
13. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: explaining the predictions of any classifier. In KDD 2016, the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM, 2016:1135–44.

14. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer, 2013.
15. Ammenwerth E, Rigby M, editors. Evidence-based health informatics: promoting safety and efficiency through scientific methods and ethical policy. Amsterdam: IOS Press, 2016.
16. Camaggi CM, Zavatto E, Gramantieri L, Camaggi V, Strocchi E, Righini R, et al. Serum albumin-bound proteomic signature for early detection and staging of hepatocarcinoma: sample variability and data classification. Clin Chem Lab Med 2010;48:1319–26.
17. Madabhushi A, Doyle S, Lee G, Basavanhally A, Monaco J, Masters S, et al. Integrated diagnostics: a conceptual framework with examples. Clin Chem Lab Med 2010;48:989–98.
18. Demirci F, Akan P, Kume T, Sisman AR, Erbayraktar Z, Sevinc S. Artificial neural network approach in laboratory test reporting. Am J Clin Pathol 2016;146:227–37.
19. Forsting M. Machine learning will change medicine. J Nucl Med 2017;58:357–8.
20. Horowitz GL. The power of asterisks. Clin Chem 2015;61: 1009–11.
21. Connelly DP. Embedding expert systems in laboratory information systems. Am J Clin Pathol 1990;94(4 Suppl 1):S7–14.
22. Lippi G, Bassi A, Bovo C. The future of laboratory medicine in the era of precision medicine. J Lab Precis Med 2016;1:7.
23. Komatireddy R, Topol EJ. Medicine unplugged: the future of laboratory medicine. Clin Chem 2012;58:1644–7.
24. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. J Med Internet Res 2008;10:e22.
25. Poole S, Schroeder LF, Shah N. An unsupervised learning method to identify reference intervals from a clinical database. J Biomed Inform 2016;59:276–84.
26. Lindbury BA, Richardson AM, Badrick T. Assessment of machinelearning techniques on large pathology sets to address assay redundancy in routine liver function test profiles. Diagnosis 2015;2:41–51.
27. Jha AK, Chan DC, Ridgway AB, Franz C, Bates DW. Improving safety and eliminating redundant tests: cutting costs in U.S. hospitals. Health Aff 2009;28:1475–84.
28. Diri B, Albayrak S. Visualization and analysis of classifiers performance in multi-class medical data. Expert Syst Appl 2008;34:628–34.
29. Nelson DW, Rudehill A, MacCallum RM, Holst A, Wanecek M, Weitzberg E, et al. Multivariate outcome prediction in traumatic brain injury with focus on laboratory values. J Neurotrauma 2012;29:2613–24.
30. Lin C, Karlson EW, Canhao H, Miller TA, Dligach D, Chen PJ, et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PLoS One 2013;8:e69932.
31. Liu KE, Lo CL, Hu YH. Improvement of adequate use of warfarin for the elderly using decision tree-based approaches. Methods Inf Med 2014;53:47–53.
32. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prection of type 2 diabetes from claims data and analysis of risk factors. Big Data 2015;3:277–87
33. Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, et al. Deep biomarkers of human aging: application of deep neural networks to biomarker development. Aging 2016;8:1021.
34. Yuan C, Ming C, Chengjin H. UrineCART, a machine learning method for establishment of review rules based on UF-1000i flow cytometry and dipstick or reflectance photometer. Clin Chem Lab Med 2012;50:2155–61.
35. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017;38:1805–14.
36. Chen J, Tang H, Lv L, Wang Y, Liu X, Lou T. Development and validation of new glomerular filtration rate predicting models for Chinese patients with type 2 diabetes. J Transl Med 2015;13:300–17
37. Surinova S, Choi M, Tao S, Schuffler PJ, Chang CY, Clough T, et al. Prediction of colorectal cancer diagnosis based on circulating plasma proteins. EMBO Mol Med 2015;7:1166–78.
38. Wang HY, Hsieh CH, Wen CN, Wen YH, Chen CH, Lu JJ. Cancers screening in an asymptomatic population by using multiple tumour markers. PLoS One 2016;11:e0158285.
39. Richardson A, Signor BM, Lidbury BA, Badrick T. Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data. Clin Biochem 2016;49:1213–20.
40. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. Am J Clin Pathol 2016;145:778–88.
41. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 2013;3:e002847.
42. Somnay YR, Craven M, McCoy KL, Carty SE, Wang TS, Greenberg CC, et al. Improving diagnostic recognition of primary hyperparathyroidism with machine learning. Surgery 2017;161:1113–21.
43. Henny J, Vassault A, Boursier G, Vukasovic I, Mesko Brguljan P, Lohmander M, et al. Recommendation for the review of biological reference intervals in medical laboratories. Clin Chem Lab Med 2016;54:1893–900.
44. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. J Am Med Assoc 2017;318:517–8.