European Journal of Cardio-Thoracic Surgery 60 (2021) 213–221
doi:10.1093/ejcts/ezab095 Advance Access publication 22 March 2021

REVIEW

Cite this article as: Ostberg NP, Zafar MA, Elefteriades JA. Machine learning: principles and applications for thoracic surgery. Eur J Cardiothorac Surg 2021;60:213–21.

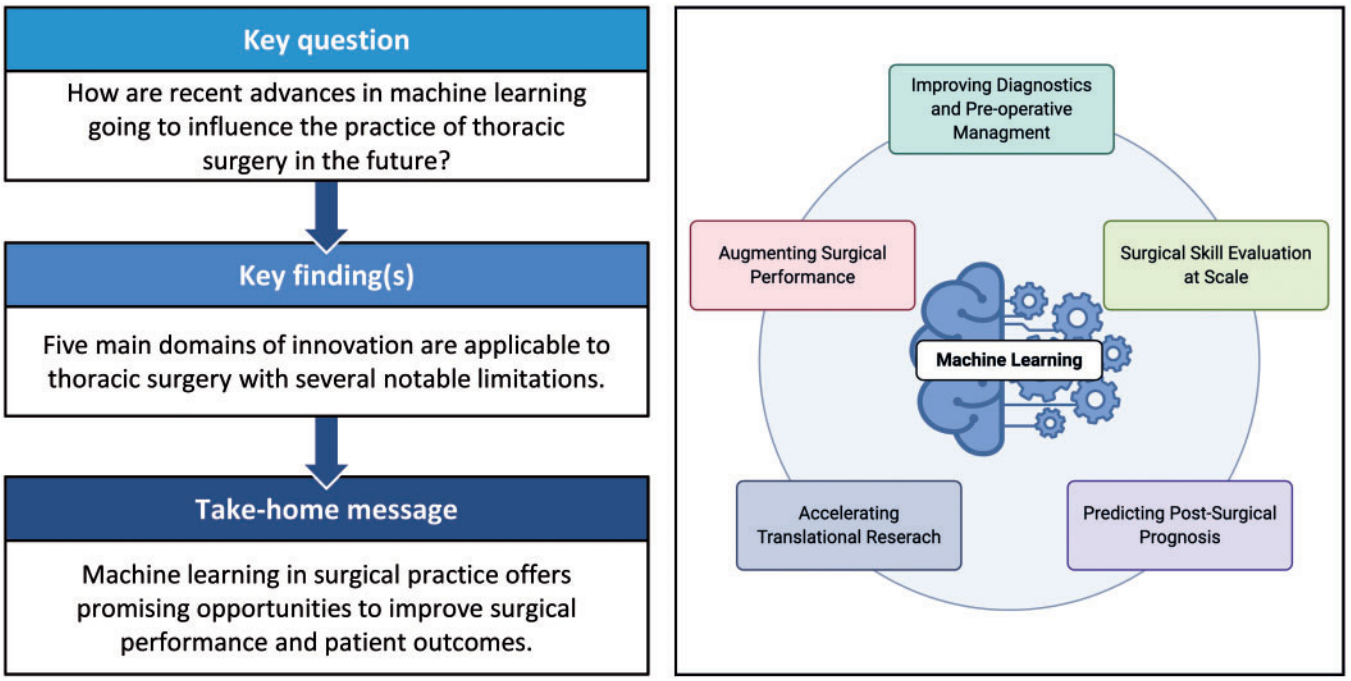
Machine learning: principles and applications for thoracic surgery

Nicolai P. Ostberg [a,b], Mohammad A. Zafar [a] and John A. Elefteriades [a,*]

[a] Aortic Institute at Yale-New Haven Hospital, Yale University School of Medicine, New Haven, CT, USA
[b] New York University Grossman School of Medicine, New York, NY, USA

* Corresponding author. Aortic Institute at Yale-New Haven, Yale University School of Medicine, Clinic Building CB 317, 789 Howard Avenue, New Haven, CT 06519, USA. Tel: 203-785-2551; fax: 203-785-3552; e-mail: john.elefteriades@yale.edu (J.A. Elefteriades).

Received 5 August 2020; received in revised form 25 January 2021; accepted 27 January 2021



Summary

OBJECTIVES: Machine learning (ML) has experienced a revolutionary decade with advances across many disciplines. We seek to understand how recent advances in ML are going to specifically influence the practice of surgery in the future with a particular focus on thoracic surgery.

METHODS: Review of relevant literature in both technical and clinical domains.

RESULTS: ML is a revolutionary technology that promises to change the way that surgery is practiced in the near future. Spurred by an advance in computing power and the volume of data produced in healthcare, ML has shown remarkable ability to master tasks that had once been reserved for physicians. Supervised learning, unsupervised learning and reinforcement learning are all important techniques that can be leveraged to improve care. Five key applications of ML to cardiac surgery include diagnostics, surgical skill assessment, postoperative prognostication, augmenting intraoperative performance and accelerating translational research. Some key limitations of ML include lack of interpretability, low quality and volumes of relevant clinical data, ethical limitations and difficulties with clinical implementation.

CONCLUSIONS: In the future, the practice of cardiac surgery will be greatly augmented by ML technologies, ultimately leading to improved surgical performance and better patient outcomes.

Keywords: Machine learning • Supervised learning • Deep learning • Predictive models • Prognostication

ABBREVIATIONS

ANN	Artificial neural network
AUROC	Area under the receiver operator curve
CNN	Convolutional neural network
ML	Machine learning
RL	Reinforcement learning

INTRODUCTION

There is seemingly endless interest in applying machine learning (ML) to medicine today. Almost daily, new research is published heralding the use of ML to improve clinical care, from improving disease prediction and screening to automated diagnosis across a variety of different specialties, particularly in the fields of cancer, neurology and cardiovascular medicine [1]. Some even believe that ML algorithms will one day replace the diagnostic thinking that physicians perform on a daily basis, including replacing specialists such as radiologists, although many experts believe that this fear is unfounded [2]. Regardless, almost all agree that ML stands poised to revolutionize the way that medicine is practiced across specialties in the coming decades.

However, few physicians understand the implications that ML will have when applied to clinical practice. Particularly in the field of thoracic surgery, there have been relatively few direct applications of ML and several untapped areas for innovation. In this review, we hope to demystify the use of ML in medicine and inspire thoracic surgeons to embrace this powerful tool to improve clinical care and surgical outcomes.

We begin by reviewing 4 basic subtypes of ML: supervised learning, unsupervised learning, reinforcement learning (RL) and deep learning. We then lay out 5 domains in thoracic surgery that are ripe for innovation via ML. We finish by discussing some of the limitations of ML.

MACHINE LEARNING METHODS

There are 4 main ML domains that have been applied to medicine: supervised learning, unsupervised learning, RL and deep learning.

Supervised learning

A *supervised ML algorithm* maps a set of input variables ('features') to outcomes, where the outcomes are known. The form that these features and outcomes can take on varies across different applications. For example, features and outcomes can be binary, continuous or even vary temporally. If the outcome variable takes on a continuous range of values, it is known as a *regression* task and if the outcome variable takes on only 1 of 2 binary values, then the task is known as *classification*. Supervised learning algorithms produce different kinds of decision functions that will assigned predicted outputs based on input features (Table 1). The performance of these decision functions can then be compared to select the best performing algorithm.

Most physicians are familiar with *regression models*. While these models are simple, they can often provide great insight into linear relationships. Advanced forms of regression—known as LASSO and Ridge regression—will penalize large regression coefficients in a manner that limits the number of variables used to predict an outcome, which is known as 'regularization' or 'shrinkage'. Often times these linear models can perform nearly as well more advanced models and the impact of each input feature on the outcome is well defined, which can be particularly important in a medical contexts. If results from a linear model already offer sufficient improvement over the standard of care, further optimization might not be worthwhile.

A *support vector machine* is another common technique. Support vector machines are fast, relatively flexible and have been used in medicine for many years. The goal of a support vector machine is to find an optimal decision boundary between 2 or more classes that puts the most space (otherwise known as maximum margin) between the 2 groups. A useful analogy to remember is that a support vector machine finds the 'widest highway' between 2 groups of points.

Another set of models are known as *ensemble methods*, exemplified by the *random forest* method. A random forest will construct a series decisions trees using different combinations of explanatory variables to predict the outcome of interest. Each tree will predict an outcome and the mode or mean outcome of all of the decision trees will then be used as the final prediction for classification or regression. This method has been shown to perform well on numerous ML problems as the resulting classifier tends to generalize well to new data.

Several other kinds of supervised learning models exist such as a *Naïve Bayes models* or *linear discriminant analysis*. While these models have been utilized historically to solve some simple classification problems in medicine and have ample theoretical justification, their use has primarily been supplanted by some of the more advanced models mentioned above [3].

Deep learning

Although generally considered a subclass of supervised learning, *deep learning* deserves special mention due many transformative deep learning applications in medicine today. The core innovation of deep learning is the *artificial neural network* (ANN). ANNs are formed by a series of interconnected layers of neurons (known as hidden layers) which transform input data into scalar values based on a set of weights. The output value of each neuron is then passed through a non-linear transformation imitating the on–off nature of biological neurons into the next layer of neurons. This process continues until the final layer of the network where the output of the model is compared to the true value (Fig. 1A). The weights of each neuron are then optimized based off of the error of each prediction in a process known as backpropagation. Due to their flexible structure, ANNs are able to fit decision function to complex data patterns efficiently and their performance continually improves with larger amounts of data.

Table 1: Description of supervised learning algorithms used in medicine today

Method	Description	Use cases
Linear/logistic regression	Models a linear relationship between input features and output variables optimized using a least-squares approach.	Useful when interpretation is valued
Support vector machine	Find optimally separating hyperplane between data points to make a classification.	Useful with large numbers of input features
Random forest	Creates a series of decision tree classifiers on a subset of the data and features that are then ensembled together to create a prediction.	Robust to many kinds of data does not make any assumptions about the underlying data distribution, less influenced by outliers

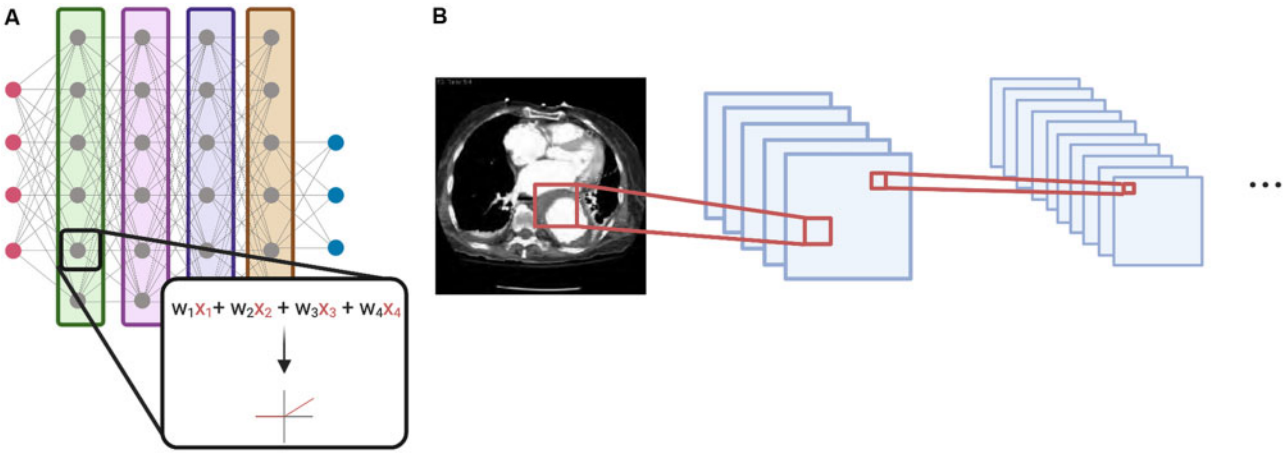


Figure 1: Deep learning methods. (**A**) An artificial neural network with 4 input features (red points), 4 hidden layers and 3 outputs (blue points). Each neuron in a neural network will take in input values (x_1–x_4) multiplied by a series of weights (w_1–w_4) which will then be fed into a non-linear transformation to be passed on to the next layer. The weights of the model are optimized via a process known as backpropagation. (**B**) Overview of convolutional neural network architecture. Features of an image are extracted by a series of convolutional filters that learn the underlying features of the image with no human intervention. The output of this model could be a binary classification or segmentation of a region of interest (i.e. lumen of the aorta).

Deep learning methods have been particularly successful when applied to images. Because spatial patterns and the sheer number of data points in an image often makes optimizing an ANN prohibitively slow, a variation of an ANN is generally applied, known as a *convolutional neural network* (CNN). A CNN uses convolution filters to extract features from images; as convolutional layers are stacked in a model, higher level features are extracted, such as shapes, with no *a priori* human intervention (Fig. 1B). This kind of feature extraction architecture can be applied to image classification, object localization within an image and segmentation.

Evaluating and optimizing supervised learning models

Once a supervised learning model is fitted, it must be evaluated in order to measure its effectiveness. This allows for comparison to other state-of-the-art models and gives physicians a sense of how the model will perform when implemented in a clinical setting. There are several metrics available to evaluate the performance of an ML model. One of the most common measures used in a classification task is the *area under the receiver operator curve* (AUROC). AUROC curves are constructed by plotting the false positive rate (1−specificity) against true positive rate (sensitivity) at given threshold probabilities produced by the model (Fig. 2A). An AUROC of 0.5 indicates a random classifier while an AUROC of 1.0 indicates a perfect classifier. The model with an AUROC closest to 1 is selected as the best performing model. Another useful metric is a precision recall curve, which plots the precision (positive predictive value) of a classifier against the recall (sensitivity) at different thresholds. Similar to AUROC curves, an *area under the precision recall curve* can be calculated as well and values closer to 1.0 indicate a superior classifier (Fig. 2A). Calculating area under the precision recall curve is particularly useful for classifying rare events, as AUROC curves can inflate performance. Regression tasks try to minimize performance metrics such as root mean squared error or maximize measures of fit such as R^2.

The goal of any ML model is to produce accurate predictions in the future. In order to estimate how well the model will perform on unseen data, available data are randomly split. Predictive models are fit on some fraction of the data available, usually 70–80% of the data, and then model performance will be evaluated on the remaining 20–30% of the data (Fig. 2B). In domains where a lot of data are available (e.g. electronic health record research or imaging), an additional split will be performed to produce a training, development and test set of data; models are evaluated on the development set of the data and once the best model is selected, it is evaluated once on the test set. If limited amounts of data are available, k-fold cross-validation can be employed where the data are split into k equally sized folds, models are trained on k-1 folds of the data and then evaluated

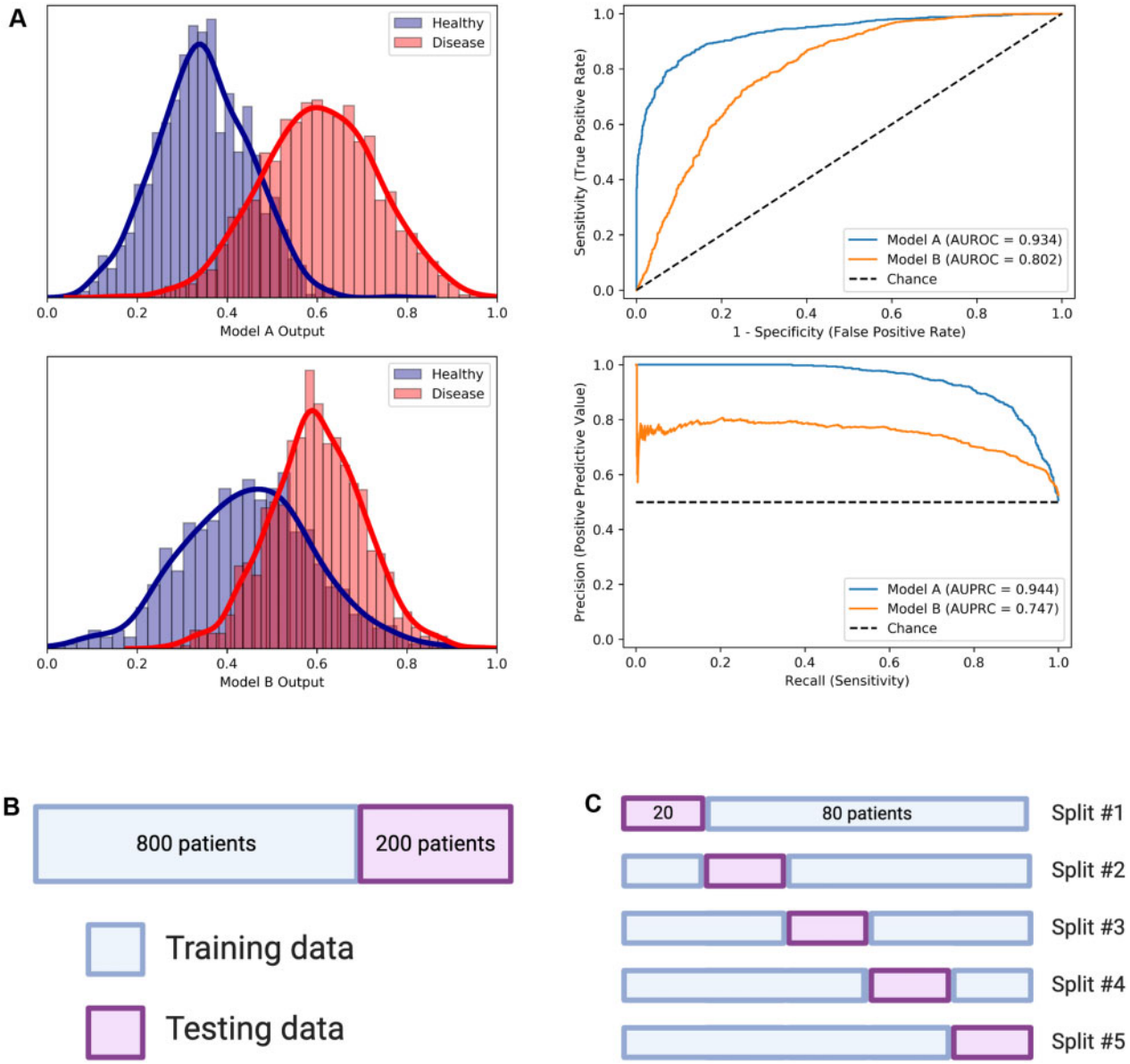


Figure 2: Machine learning evaluation. (**A**) Machine learning model A produces output probabilities that more clearly segregate healthy and disease patients compared to model B. This is reflected by higher area under the receiver operator curve and area under the precision recall curve measures. (**B**) For a relatively large dataset of 1000 patients, an 80–20 train test split leaves 800 patients to fit model parameters and 200 patients to test the model. (**C**) For smaller datasets of 100 patients, cross-validation is a useful technique. Here, five-fold cross-validation results in 5 performance metrics which can then be averaged to determine model performance.

on the remaining held-out fold of the data (Fig. 2C). Typically, 5- or 10-fold cross-validation is used in practice.

Unsupervised learning

Unsupervised learning is a method that can be used to uncover hidden patterns in data with no human intervention. The difference from supervised learning is that the only inputs to the algorithm are the raw features and the outcomes are unknown. Importantly, this means that evaluation metrics such as AUROC cannot be calculated for unsupervised models; hence, it is difficult to objectively evaluate the outputs of these models and expert input is often needed to determine clinical efficacy.

There are 2 main methods of unsupervised learning: *principle component analysis* and *cluster analysis* (Table 2). In high-dimensional data, where there are many variables captured for each example, it can be difficult to find clusters. Principle component analysis seeks to summarize data in lower dimensions, projecting data from k dimensions down to two or three dimensions that humans can interpret while best representing the underlying data distribution These low-dimensional projections will often reveal clusters of data that would not be visible or interpretable in higher dimensions.

Cluster analysis, as the name suggests, is focused on finding groups of similar examples within data based on a similarity metric. Some examples include k-means clustering, which can be used to find k clusters of similar data points within a distribution. The k-means algorithm will first randomly define k centroids within the data distribution and then iteratively improve upon these centroid definitions, which are then used to assign cluster identities to data points within the distribution.

Table 2: Description of unsupervised learning algorithms used in medicine today

Method	Description	Use cases
K means clustering	Defines k centroids and closest examples are assigned to the centroid. Centroid locations are iteratively improved until convergence.	Finding new clinical phenotypes
Principle component analysis	Projects high-dimensional data into lower dimensions that explain most of the variation in the data.	Visualizing high-dimensional data, feature engineering

Table 3: Summary of machine learning studies applicable to thoracic surgery

Reference	Objective	Sample size and modality	Methods	Application to thoracic surgery
Wang *et al.*, [5]	Detect 8 common chest X-ray pathologies	108 948 frontal chest X-rays	CNN	Rapid detection and screening of post-surgical complications
Kusunose *et al.* [6]	Detection of wall motion abnormalities on echocardiograms	300 echocardiograms	CNN	Rapidly diagnose or screen cardiac abnormalities
Bai *et al.* [7]	Segmentation of ascending and descending aorta	500 aortic magnetic resonance images	ANN + RNN	Objective and consistent aortic diameter measurement
Ouyang *et al.* [8]	Segment the left ventricle to continuously measure ejection fraction	10 030 echocardiograms	CNN	Objectively and rapidly measure cardiac function in real time
Li *et al.* [9]	Diagnose abdominal aortic aneurysm using both electronic health record and genomic data	401 patients with computed tomography scans and whole genome sequencing	Logistic regression	Screening high-risk populations and basic research
Wu *et al.* [10]	Predicting in-hospital rupture of type A aortic dissection	1133 patients	Random forest	Stratifying patient with type A aortic dissection for closer monitoring and follow-up
Chang Junior *et al.* [11]	Predicting 30 day mortality for patients undergoing surgery for congenital heart defects	2240 patients	Random forest	Stratifying patients for closer monitoring and follow-up
Wang and Majewicz Fey [12]	Objectively evaluate 3 surgical skills—suturing, needle passing and knot tying—using ML	40 trials (8 participants)	CNN	Rapid and continuous surgical skill evaluation without need for manual annotation
Jin *et al.* [13]	Detect surgical instruments in laparoscopic videos	15 videos	CNN	Real-time technique feedback in the operating room
Kilic *et al.*, [14]	Estimate operative mortality risk of cardiac surgery	11 190 patients	XGBoost	Inform surgical risks of high-risk patients
Czerny *et al.* [15]	Predict 30-day mortality rate for patients undergoing surgery for type A aortic dissection	2537 patients	Logistic regression	Risk stratify post-operative patients for closer follow-up
Pirruccello *et al.* [16]	Identify patients with an aortic aneurysm via ML in order to identify causative genes	33 420 patients	CNN	Increase the scale and speed of basic research projects by eliminating manual diameter measurement

ANN: artificial neural network; CNN: convolutional neural network; ML: machine learning.

One useful application of unsupervised learning in medicine is discovery of new subtypes of a complex disease. Once this new subgroup is identified, different treatment regimens can be explored that may provide better outcomes to patients. This was demonstrated in a novel study that utilized hierarchical clustering to discover 3 distinct phenotypes of heart failure with preserved ejection fraction [4]. Still, supervised learning remains the dominant form of ML utilized today.

Reinforcement learning

RL is a fundamentally different framework compared to supervised learning. RL algorithms do not need extensive training data of input–output pairs as is needed in supervised learning; instead, RL algorithms will train an agent (e.g. surgical robot) to perform a series of actions (e.g. suture a wound) that incentivizes positive behaviours (e.g. closing the wound) and disincentives negative behaviours (e.g. bleeding) as formalized in a reward function. This is the kind of core technology at the heart of surgical robotics. Algorithms will iteratively try different series of actions until the reward function has been sufficiently optimized and the system is able to achieve appropriate performance.

RL has seen a number of remarkable achievements in the past decade. However, applications to surgery are relatively infrequent and clinical implementation is difficult, given clinical and anatomic variations. Additionally, imaging, monitoring data and 'tactile' inputs need to be integrated. Practical surgical applications remain crude and are currently far from implementation in the OR.

APPLICATIONS TO THORACIC SURGERY

Equipped with an understanding of the principles of ML, we can begin to explore how this technology can be used by thoracic surgeons in a variety of settings. Broadly, there are 5 fundamental areas in which ML can aid thoracic surgeons: improving *diagnosis* and preoperative management, augmenting *surgical performance* in the OR, *skill assessment, post-procedure prognostication* and *translational research*. We will discuss each setting separately, show specific examples of what has already been accomplished, and examine work that needs to be done in the future (Table 3).

Aiding diagnosis and preoperative management

Due to its ability to draw inferences from the complex, high-dimensional and often multimodal data needed to make a diagnosis, ML has been extensively investigated as a tool to improve diagnosis. There are 2 imaging-based applications of ML applicable to thoracic surgery: automated diagnosis of cardiac pathology and segmentation of the aforementioned cardiac pathology. On the diagnostic front, CNNs are able to detect subtle patterns in biomedical images in order to quickly and accurately detect pathology. One classic example is an ML algorithm that was able to detect 8 distinct pathologies from chest X-rays [5]. Another study used CNNs to better detect wall motion abnormalities in echocardiographic images, achieving an AUROC of 0.99, outperforming physicians at the same task [6].

Segmentation is another fruitful application of CNNs, particularly for tasks reliant upon measuring the dimensions of organs and vessels [7]. There are numerous examples applicable to thoracic surgery, including aortic diameter [17] and volumetric segmentation of the left ventricle to measure cardiac function [8]; in all cases, CNNs are able to match or outperform human-level performance. These studies show how ML can augment a thoracic surgeon's practice and quickly calculate clinically relevant cardiac specific parameters, allowing for more time for direct patient care.

Additionally, early diagnosis of thoracic and abdominal aneurysms would be particularly beneficial due to the fact that as many as 95% of patients have no symptoms prior to life threatening complications [18]. Recently, an ML approach that integrated genomic and electronic health record data also demonstrated remarkable ability to diagnose abdominal aortic aneurysms while also elucidating some of the underlying genetic mechanisms [9]. Additionally, a random forest classifier trained on 1133 patients was able to predict in-hospital rupture of the ascending aorta for patients with thoracic ascending aortic aneurysms with an AUROC of 0.752 and sensitivity of 0.99 [10].

Augmenting intraoperative surgical performance

Surgical robotics has advanced greatly in the past few decades, with examples including the da Vinci Surgical System (Intuitive Surgical, Sunnyvale, CA, USA) for a variety of minimally invasive surgical procedures and the Sensei X robotic catheter system (Hansen Medical Inc., Mountain View, CA, USA) for cardiac catheter insertion [19]. However, these systems currently require continuous or nearly continuous human intervention. Because thoracic surgeons are capable of performing a wide array of procedures with high complexity and dynamic patients, surgical robots will likely never be fully independent of human control. However, RL powered robotics are already capable of performing simple surgical subtasks, such as simple suturing and precise surgical cutting [20, 21].

Creating a general purpose surgical robot for thoracic surgery that is able to perform all parts of even a single procedure is very challenging; creating a robot that is able to perform all of the tasks that a single human thoracic surgeon can do is likely impossible. However, a long-term goal might be to get ML robotics could advance to the point where simple surgical tasks are automated and surgeons primarily play the role of decision-maker, much like a pilot using auto-pilot controls in an airliner. A surgeon can set an objective for a robot to perform and then observe the procedure, intervening as necessary, while not worrying about the psychomotor nuances of the procedure.

Outside of the realm of robotics, ML can also help integrate signals derived from patient monitoring equipment during a surgery in order to give early warnings to surgeons during a procedure. A recent randomized trial investigated the implementation of an ML system to detect intraoperative hypotension during elective non-cardiac surgeries with remarkable success, decreasing the median time of hypotension from 32.7 min to just 8 min [22]. Particularly in high-risk operations, similar early warning systems can be implemented specifically for thoracic surgery procedures, decreasing rates of adverse events.

Surgical observation and evaluation

Objective, real-time evaluation of surgical skills in trainees is a particularly difficult task considering the dynamic nature of the operating room and the variety of surgical procedures and environments in which surgeons operate. However, ML algorithms have been shown to accurately assess surgical performance and provide quantitative and actionable feedback to surgeons on 3 simple performance tasks—suturing, needle passing and knot tying—based on short video clips [12]. Another area of active research is surgical phase recognition, which consists of automatically detecting the temporal phase of a surgery to improve scheduling and throughput. Limited work has been done specifically for thoracic surgery and instead has been primarily focused on common ophthalmology or laparoscopic procedures [23,24].

One final application used deep learning to recognize surgical tools that are being used during laparoscopic surgeries for the purpose of tracking tools during a surgery and using this as a proxy for surgical skill and quality [13]. Aspirational goals of these kinds of technologies include alerting surgeons if they are deviating from the performance of other surgeons in the database or providing real-time instructive feedback on technique while in the operating room. Objective evaluation of surgical skill, both real time and retrospective, has long been a goal of the surgical community; leveraging ML can help achieve this goal [25].

Post-surgical prognosis

Much of clinical medicine involves using patient data to make predictions about future outcomes and then managing patients based on these predictions. While historically these decisions were made based on clinical experience and medical literature, ML has opened up the possibility of making highly accurate predictions of patient outcomes that allows for highly individualized

Table 4: Limitations of machine learning and potential solutions

Limitation	Potential solutions
Lack of model interpretability	• Rely on simpler models • Perform careful model auditing to discover blind spots
Barriers to clinical implementation	• Study clinician workflow • Partnerships with ML practitioners • Improve user interfaces
Low data quality and quantity	• Establish data sharing agreements between institutions • Audit databases consistently for accuracy
Limited model evaluation post-deployment	• Implement mechanisms to monitor model performance in clinical settings • Evaluate magnitude of practice changes post-deployment
Ethical considerations	• Increase patient access to ML by expanding implementation to community and smaller hospitals • Ensure that input data are representative of the treated patient population • Appropriately inform patients about the use of ML in clinical care

ML: machine learning.

patient management. This is reflected by the exponential rise of published clinical scoring systems, with approximately 250 000 such publications since 1965 [26].

Predictive models are particularly important in thoracic surgery, where surgical complication rates are higher than other surgical specialties [27]. A recent publication used XGBoost, a type of ensemble model, to predict the operative mortality rate of 11 190 patients from a single institution undergoing cardiac surgery demonstrated improved AUROC, calibration, accuracy and F1 score over the state of the art Society of Thoracic Surgeons Predicted Risk of Mortality (STS PROM) score [14]. Another study applied logistic regression on 1000 patients, improving prediction of in-hospital mortality following cardiac surgery compared to other measures such as Acute Physiology and Chronic Health Evaluation II (APACHE II) and Parsonnet score [28]. A third study was able to predict 30-day mortality for patients with acute type A aortic dissection using logistic regression with an AUROC of 0.728 [15]. Still another study predicted 30-day mortality for patients undergoing surgery for congenital heart disease with an AUROC of 0.902 [11].

However, it is important to note that advanced ML techniques are not a panacea for predicting outcomes; one report noted that ML did not outperform simple logistic regression when predicting in hospital mortality after cardiac surgery [29]. Continued expansion of large postoperative databases may improve the predictive power of ML techniques [30].

Accelerating translational research

Translational research innovations have always shaped the quality of thoracic surgery in a number of ways, from the use of deep hypothermia for patients under circulatory arrest to a detailed understanding of the impact of genetics on aortic aneurysms and dissections [31–33]. Surgical innovation in the future will continue to rely upon similar kinds of translational research. ML techniques are poised to accelerate this process in diverse.

Genetic studies in particular can be revolutionized by ML. ANNs have also been utilized in genomics to predict pathogenicity of mutations or genetic regulatory mechanisms. One study used a neural network to identify SNPs associated with inheritable cardiac disease [34]. Another recent study used deep learning techniques on magnetic resonance images of the aorta from the UK Biobank to measure the aortic diameter of over 30 000 individuals. Then, GWAS study ultimately identified that the gene *SVIL*, a gene highly expressed in vascular smooth muscle, as significantly associated with both ascending and descending dilation [16]. Only 116 images needed manual assessment; deep learning was able to segment the remaining images with high accuracy. Deep learning on biomedical imaging and electronic health record data coupled with genetic sequences can revolutionize the study of aortic disease—and other cardiac domains—by increasing the sample size of studies and the speed at which discoveries are made.

LIMITATIONS OF MACHINE LEARNING IN MEDICINE

Although much hype has surrounded the field of ML, it is not a panacea for all of the diagnostic and management challenges facing surgeons today. Few of the promising applications discussed above are commonly used by physicians during everyday practice. There are several limitations of ML that deserve mention, which are summarized in Table 4.

Perhaps the biggest shortcoming of ML in medicine revolves around a lack of interpretability regarding the outcomes produced by the ML model at hand. The major benefit of ML—uncovering highly complex and non-linear associations between features—also means that humans are unable to understand what is going on behind the 'black box' used to make these associations. This differs from other modelling methods in medicine. For example, in linear regression, examining the weights (i.e. β) of the regression model gives the user a very straightforward interpretation: increasing input feature x_k by 1 increases or decreases the output y by β_k. Several attempts have been made to allow for the same kind of convenient interpretation in ML, yet al.l make assumptions that are frequently violated, are computationally expensive, or are very sensitive to perturbations in the model [35]. More effort should be invested in giving clinicians a look behind the 'black box' to improve confidence in the outputted results.

Clinical implementation is a challenge faced by both clinicians and ML practitioners alike. While a well implemented ML model fits seamlessly into pre-existing digital infrastructure, a poorly implemented model may significantly hinder the workflow of physicians by requiring manual input of from the patient's digital record. A poorly structured user interface may draw physicians away from time spent with patients [36]. Because much of the innovation in ML is centred on algorithm and dataset development, less focus has been placed on ensuring the ML models are easy to use by physicians. Shifting this perspective could greatly accelerate ML adoption in clinical practice. Ideally, ML models should work in the background, automating tasks that physicians find mundane or frustrating (e.g. data entry) and augmenting clinical care whenever possible. Yet few studies have examined the nuances of implementation [37]. In addition, overreliance

upon ML and automation bias may pose real future risks if ML becomes widely adopted in medicine; even the most reliable ML model can never replace clinical experience [38].

Accessing high-quality data remains a central challenge faced by ML in clinical practice. A common aphorism among ML practitioners is 'garbage in, garbage out'; this is to say, the quality of the data dictates the quality of the model. If the data that are used to train ML models are not representative of the patient data in clinical practice, the model is useless and will generalize poorly to new data. This can occur when training data systematically overrepresents or underrepresents a particular patient population (i.e. cancer patients) or the database draws from only a specific segment of the population. As a general rule of thumb, the larger and more diverse a database is, the greater its utility for ML. Yet producing large, heterogenous datasets in medicine is necessarily difficult due to privacy and regulatory requirements (e.g. HIPAA, IRB). Physicians and researchers should strive to ensure that clinical data sets are heterogenous and capture the full range of presentations that they might encounter by establishing data sharing agreements between institutions, capturing as much relevant clinical data as possible, remaining vigilant against inclusion and exclusion bias, and frequently auditing databases to ensure high-quality data input.

Finally, ML models often do not receive the same level of scrutiny after deployment in a clinical setting. Much like postmarketing surveillance of pharmaceuticals, ML models should be continually audited post-deployment for both accuracy and efficacy. For example, although computer-aided mammography has increased in prevalence in the past 2 decades, there has been no appreciable increase in diagnostic accuracy [39]. In addition, models should be subjected to evaluation post-deployment in order to assess the quality of the input data and ensure that the model is not underperforming on novel data.

Ethical considerations

There are several ethical issues that need to be considered when implementing ML algorithms in clinical practice. One key tenets is that all patients should receive equitable access to the benefits provided by ML models. There is an apparent geographic bias in current implementations of ML applications in the USA; a recent study showed that a disproportionate number of ML studies used patient cohorts from just 3 states—California, New York and Massachusetts—with no representation from 34 of 50 states [40]. A more conscious effort needs to be made to diversify both the populations that have access to ML and training physicians to advocate for ML integration at a variety of clinical settings.

It is also critical that the data that are inputted into ML models are broadly representative of kinds of patients being treated. Models will perform poorly if patient populations are not represented in the training data. In addition, if racially biased care is reflected in the training data, ML may recapitulate these biases, as recently discovered in an algorithm used nationwide to stratify patients into high-risk care management programmes [41]. To combat this bias, data should be continually examined to make sure that the diversity of patients seen in clinical practice is represented in the training data and that a model is not consistently underperforming on subgroups of patients.

As with other medical interventions, patients should be informed when ML is being utilized in their care, as well as any being made aware of potential shortcomings. This will become more difficult as artificial intelligence becomes integrated into everyday medical practice via ambient intelligence and integrated clinical decision support tools [42]. Explicit informed consent discussions should take place about the role of human oversight in any ML process, how patient data are being captured and used, and privacy measures in place to protect health information.

CONCLUSIONS

When the words 'machine learning' are uttered to a thoracic surgeon, reactions can range from unrestrained optimism about the future of ML in medicine to confusion and fear about complicated mathematics and even to resentment about automation. The proper response likely falls somewhere between all of these sentiments. ML certainly has a remarkable future in surgery if implemented properly. However, in order to fully realize the potential of ML in thoracic surgery, surgeons need to be able to interface with ML practitioners in a way that drives meaningful collaboration. In order for these collaborations to be successful, surgeons need a high-level understanding of different ML techniques, awareness of current ML efforts in thoracic surgery and recognition of some potential shortcomings of ML. We hope to have covered all 3 of these domains in this review.

Conflict of interest: The disclosures for Dr John A. Elefteriades are as follows: Coolspine—Principal; Terumo—Data and Safety Monitoring Board; Cryolife—Consultant; DuraBiotech—Consultant. The other authors have nothing to disclose.

Author contributions

Nicolai P. Ostberg: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Validation; Visualization; Writing—original draft. **Mohammad A. Zafar:** Conceptualization; Data curation; Formal analysis; Investigation; Project administration; Visualization; Writing—original draft. **John A. Elefteriades:** Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Visualization; Writing—review & editing.

Reviewer information

European Journal of Cardio-Thoracic Surgery thanks Mitsuru Asano, Tomislav Kopjar and Nikolay O. Travin for their contribution to the peer review process of this article.

REFERENCES

[1] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S *et al.* Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230–43.
[2] Chan S, Siegel EL. Will machine learning end the viability of radiology as a thriving medical specialty? Br J Radiol 2019;92:20180416.
[3] Langarizadeh M, Moghbeli F. Applying naive Bayesian networks to disease prediction: a systematic review. Acta Inform Med 2016;24:364–9.
[4] Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M *et al.* Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation 2015;131:269–79.
[5] Wang X, Peng, Y Lu, L Lu, Z Bagheri, M and Summers, RM *ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Honolulu, HI, USA, 2017, pp. 3462–71, doi:10.1109/CVPR.2017.369.

[6] Kusunose K, Abe T, Haga A, Fukuda D, Yamada H, Harada M *et al.* A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. JACC Cardiovasc Imaging 2020; 13:374–81.

[7] Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W *et al.* Deep learning for cardiac image segmentation: a review. Front Cardiovasc Med 2020;7:25.

[8] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP *et al.* Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580:252–6.

[9] Li J, Pan C, Zhang S, Spin JM, Deng A, Leung LLK *et al.* Decoding the genomics of abdominal aortic aneurysm. Cell 2018;174:1361–72.e10.

[10] Wu J, Qiu J, Xie E, Jiang W, Zhao R, Zafar MA *et al.* Predicting in-hospital rupture of type A aortic dissection using random forest. J Thorac Dis 2019;11:4634–46.

[11] Chang Junior J, Binuesa F, Caneo LF, Turquetto ALR, Arita E, Barbosa AC *et al.* Improving preoperative risk-of-death prediction in surgery congenital heart defects using artificial intelligence model: a pilot study. PLoS One 2020;15:e0238199.

[12] Wang Z, Majewicz Fey A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int J Comput Assist Radiol Surg 2018;13:1959–70.

[13] Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A *et al.* Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. arXiv e-Prints 2018; arXiv:1802.08774. pp. 691–9, doi:10.1109/WACV.2018.0081.

[14] Kilic A, Goyal A, Miller JK, Gjekmarkaj E, Tam WL, Gleason TG *et al.* Predictive utility of a machine learning algorithm in estimating mortality risk in cardiac surgery. Ann Thorac Surg 2020;109:1811–19.

[15] Czerny M, Siepe M, Beyersdorf F, Feisst M, Gabel M, Pilz M *et al.* Prediction of mortality rate in acute type A dissection: the German Registry for Acute Type A Aortic Dissection score. Eur J Cardiothorac Surg 2020;58:700–6.

[16] Pirruccello JP, Chaffin MD, Fleming SJ, Arduini A, Lin H, Khurshid S *et al.* Deep learning enables genetic analysis of the human thoracic aorta. bioRxiv 2020; 2020.05.12.091934.

[17] Bai W, Suzuki, H Qin, C Tarroni, G Oktay, O Matthews, PM, *et al.* Recurrent Neural Networks for Aortic Image Sequence Segmentation with Sparse Annotations. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Cham: Springer International Publishing, 2018.

[18] Elefteriades JA, Sang A, Kuzmik G, Hornick M. Guilt by association: paradigm for detecting a silent killer (thoracic aortic aneurysm). Open Heart 2015;2:e000169.

[19] Peters BS, Armijo PR, Krause C, Choudhury SA, Oleynikov D. Review of emerging surgical robotic technology. Surg Endosc 2018;32:1636–55.

[20] Schulman J, Gupta A, Venkatesan S, Tayson-Frederick M, Abbeel P. *A case study of trajectory transfer through non-rigid registration for a simplified suturing scenario. In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Tokyo, Japan, 2013, pp. 4111–17, doi: 10.1109/IROS.2013.6696945.

[21] Thananjeyan B, Garg A, Krishnan S, Chen C, Miller L, Goldberg K. *Multilateral surgical pattern cutting in 2D orthotropic gauze with deep reinforcement learning policies for tensioning. In 2017 IEEE International Conference on Robotics and Automation (ICRA)*. 2017.

[22] Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P *et al.* Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA 2020;323:1052.

[23] Kitaguchi D, Takeshita N, Matsuzaki H, Takano H, Owada Y, Enomoto T *et al.* Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. Surg Endosc 2020;34:4924–31.

[24] Zisimopoulos O, Flouty E, Luengo I, Giataganas P, Nehme J, Chow A *et al.* DeepPhase: surgical phase recognition in CATARACTS videos. arXiv e-Prints 2018; arXiv:1807.10565.

[25] Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. Surg Endosc 2011;25:356–66.

[26] Challener DW, Prokop LJ, Abu-Saleh O. The proliferation of reports on clinical scoring systems: issues about uptake and clinical utility. JAMA 2019;321:2405–6.

[27] Crawford TC, Magruder JT, Grimm JC, Suarez-Pierre A, Sciortino CM, Mandal K *et al.* Complications after cardiac operations: all are not created equal. Ann Thorac Surg 2017;103:32–40.

[28] Turner JS, Morgan CJ, Thakrar B, Pepper JR. Difficulties in predicting outcome in cardiac surgery patients. Crit Care Med 1995;23:1843–50.

[29] Benedetto U, Sinha S, Lyon M, Dimagli A, Gaunt TR, Angelini G *et al.* Can machine learning improve mortality prediction following cardiac surgery? Eur J Cardiothorac Surg 2020;58:1130–6.

[30] D'Agostino RS, Jacobs JP, Badhwar V, Fernandez FG, Paone G, Wormuth DW *et al.* The society of thoracic surgeons adult cardiac surgery database: 2018 update on outcomes and quality. Ann Thorac Surg 2018;105: 15–23.

[31] Ostberg NP, Zafar MA, Ziganshin BA, Elefteriades JA. The genetics of thoracic aortic aneurysms and dissection: a clinical perspective. Biomolecules 2020;10:182.

[32] Haverich A, Hagl C. Organ protection during hypothermic circulatory arrest. J Thorac Cardiovasc Surg 2003;125:460–2.

[33] Parolari A, Tremoli E, Songia P, Pilozzi A, Bartolomeo RD, Alamanni F *et al.* Biological features of thoracic aortic diseases. Where are we now, where are we heading to: established and emerging biomarkers and molecular pathways. Eur J Cardiothorac Surg 2013;44:9–23.

[34] Burghardt TP, Ajtai K. Neural/Bayes network predictor for inheritable cardiac disease pathogenicity and phenotype. J Mol Cell Cardiol 2018; 119:19–27.

[35] Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019;19:146.

[36] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347–58.

[37] Escobar GJ, Turk BJ, Ragins A, Ha J, Hoberman B, LeVine SM *et al.* Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. J Hosp Med 2016;11:S18–24.

[38] Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Aff (Millwood) 2014;33:1148–54.

[39] Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015; 175:1828–37.

[40] Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. JAMA 2020;324:1212–3.

[41] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366:447–53.

[42] Haque A, Milstein A, Fei-Fei L. Illuminating the dark spaces of healthcare with ambient intelligence. Nature 2020;585:193–202.