Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Artificial Intelligence, Machine Learning, and Surgical Science: Reality *Versus* Hype

Majed El Hechi, MD,[a] *Thomas M. Ward, MD,*[b,c] *Gary C. An, MD, FACS,*[d] *Lydia R. Maurer, MD,*[b] *Mohamad El Moheb, MD,*[a] *Georgios Tsoulfas, MD, PhD, FACS,*[e] *and Haytham M. Kaafarani, MD, MPH, FACS*[a,*]

[a] *Division of Trauma, Emergency Surgery, and Surgical Critical Care, Massachusetts General Hospital, Boston, Massachusetts*

[b] *Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts*

[c] *Surgical Artificial Intelligence and Innovation Laboratory, Massachusetts General Hospital, Boston, Massachusetts*

[d] *Division of Acute Care Surgery, Department of Surgery, Robert Larner, MD College of Medicine, University of Vermont, Burlington, Vermont*

[e] *Department of Surgery, Aristotle University of Thessaloniki, Thessaloniki, Greece*

ARTICLE INFO

Article history:
Received 31 August 2020
Received in revised form
21 January 2021
Accepted 26 January 2021
Available online 18 March 2021

Keywords:
Artificial intelligence
Risk prediction
Machine-learning
Emergency surgery
General surgery

ABSTRACT

Artificial intelligence (AI) has made increasing inroads in clinical medicine. In surgery, machine learning–based algorithms are being studied for use as decision aids in risk prediction and even for intraoperative applications, including image recognition and video analysis. While AI has great promise in surgery, these algorithms come with a series of potential pitfalls that cannot be ignored as hospital systems and surgeons consider implementing these technologies. The aim of this review is to discuss the progress, promise, and pitfalls of AI in surgery.

Introduction

The era of artificial intelligence (AI) is here. From our smartphone applications guessing our home and work addresses based on the routes we take everyday, to the voice recognition devices playing our favorite music, to self-driving cars on streets, AI has drastically changed our lives. In surgery, AI offers potential new innovations that could shift practice. From the ability of machine learning (ML) algorithms to better predict the risk of surgery a priori and thus altering decision-making, to image recognition in the operating room, surgical research in AI is increasing rapidly. In the following manuscript, we describe the potential role of AI in surgical science, summarizing recent advances in risk prediction and computer

* *Corresponding author*. Division of Trauma, Emergency Surgery and Surgical Critical Care Massachusetts General Hospital, 165 Cambridge Street, Suite 810, Boston, MA, 02114, Tel.: +1617 726.2760; fax: +1617-726-9121.
E-mail address: HKAAFARANI@mgh.harvard.edu (H.M. Kaafarani).

0022-4804/$ – see front matter © 2021 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.jss.2021.01.046

vision, and also present criteria for critically appraising AI research. This work builds on a prior surgical review of AI, which nicely summarized the subfields of AI and provided a call to surgeons with clear examples of how to actively participate in this exciting and rapidly evolving field.[1]

AI and risk prediction in surgery

Postoperative complications occur in 15% of surgeries performed yearly in the United States,[2] incurring a total cost of 31 billion dollars.[3] This is to suggest that a potential for cost savings could come with the identification of modifiable risk factors, adjustment of surgical technique, and management based on identified proclivity to develop certain perioperative complications. For instance, work has been performed in this area looking at preoperative risk stratification for ventral hernia repairs with mesh. This has allowed for identifying and diminishing modifiable risk factors for recurrence.[4,5] Not all risk calculators include risk factors that are modifiable, but advances in the field hold promise to reduce cost in a similar way. Some of the most widely used preoperative risk stratification systems created over the last 2 decades to support surgeon experience and gestalt include the American Society of Anesthesiologists (ASA) classification system[6] and the APACHE score for intensive care unit mortality.[7] Additionally, the American College of Surgeons National Surgery Quality Improvement Program surgical risk calculator has evolved substantially over the last decade to include a hierarchical model to account for between-hospital variability, case mix, and risk adjustment.[8,9] While this does add complexity to the model's statistical underpinnings, the core ACS NSQIP calculator is similar to ASA and APACHE in that it includes the same variables for each patient. Specific risk calculators such as the Pancreas Club Fistula Risk Score Calculator[10] and the Mayo Clinic Postoperative Mortality Risk in Patients with Cirrhosis Calculator[11] offer context-specific risk prediction guidance that is useful to clinicians despite their linear and additive structures. Linearity (referring to the inclusion of the same variables, in the same ratio, every time) does have the benefit of making models more interpretable, and can confer highly performing models in targeted settings such as that of the Pancreas Club Fistula Risk Score Calculator[10] when relevant clinical factors are more limited; however, with more heterogenous data sets, these approaches do sacrifice important nonlinear patterns in the data.[12] In reality, the interaction between existing comorbidities and illness severity is not linear.[13] Some variables will weigh more (or less) toward an outcome based on the presence (or absence) of another variable.

To demonstrate the concept of nonlinearity, consider a patient undergoing emergency surgery for a cecal volvulus. Three independent risk factors that predict poor postoperative outcomes might include increased age, heart failure (HF), and chronic obstructive pulmonary disease (COPD). In a linear predictive model, each of these variables is marked as "present" or "absent" and given fixed weights, regardless of the presence of the other variables. Hypothetically, however, in a patient over 65 y old, it might be possible for COPD to play a role while HF does not, whereas in patients under 65 y, HF impacts outcome but COPD does not. Nonlinear ML models, including those discussed in this article, allow variables such as HF and COPD to impact the outcome to a variable degree depending on the value of another variable (i.e., age in this case).

ML models can be applied to essentially any domain where pattern recognition is useful, and are particularly effective when applied to very large amounts of source data or "big data." There are numerous ML models that have been applied to clinical medicine. This article will emphasize two broad types of models, black box models and interpretable models, and will focus on one example of each to illustrate their potential clinical applications to surgery.

A commonly used black box ML method is the artificial neural network (ANN). ANNs in medicine have found fruitful applications in diagnosis, prognosis, and risk stratification. ANNs are composed of many computational units which encode data, perform calculations in hidden steps, and transmit output (Fig. 1). If we were to enter the clinical characteristics of the patient undergoing surgery for cecal volvulus into an ANN to estimate mortality risk, the analysis would be conducted in its many hidden layers, and a mortality risk percentage estimate would be the output. The output produced is based on a large equation derived from training on thousands of patients, but because this equation often involves numerous variables interacting in nonlinear ways, the end user cannot view the internal workings of the model, therefore making it a black box.[14] The black box dilemma has been cited as one of the major criticisms of many methods used in ML for clinical applications.[15,16] Critics of black box models have raised concerns about "trusting" the model's output when they have no ability to check what variables are being included and how they are being used to determine the risk. The algorithm cannot be held accountable for the risk estimate it provides. In other words, if a surgeon wishes to understand why an ANN estimated a particular mortality risk to guide decision-making or counsel a patient, the model cannot be probed for an answer. While this can be an appropriate model type for image interpretation for instance, clinicians appreciate a more interpretable model when using it for clinical decision-making.

However, another category of ML exists, composed of methodology that does not compromise interpretability. Optimal Classification Trees are an example of interpretable methodology and are faithful to the nonlinear structure in the data without losing interpretability, making them ideal for preoperative risk assessment.[17] This novel Optimal Classification Tree technology was recently leveraged to create a nonlinear risk calculator of 30-day outcomes for patients undergoing emergency surgery, which out-performed the American College of Surgeons National Surgery Quality Improvement Program risk calculator and the ASA classification system (Fig. 2).[18] Furthermore, to make this technology easy to use, the algorithms were implemented in a smartphone application, called Predictive OpTimal Trees in Emergency Surgery Risk (Fig. 3).[19]

In this era of big data, standardized information is being collected for millions of patients. ML models have the potential to exploit the constant stream of data to make more accurate predictions. Furthermore, these models have the

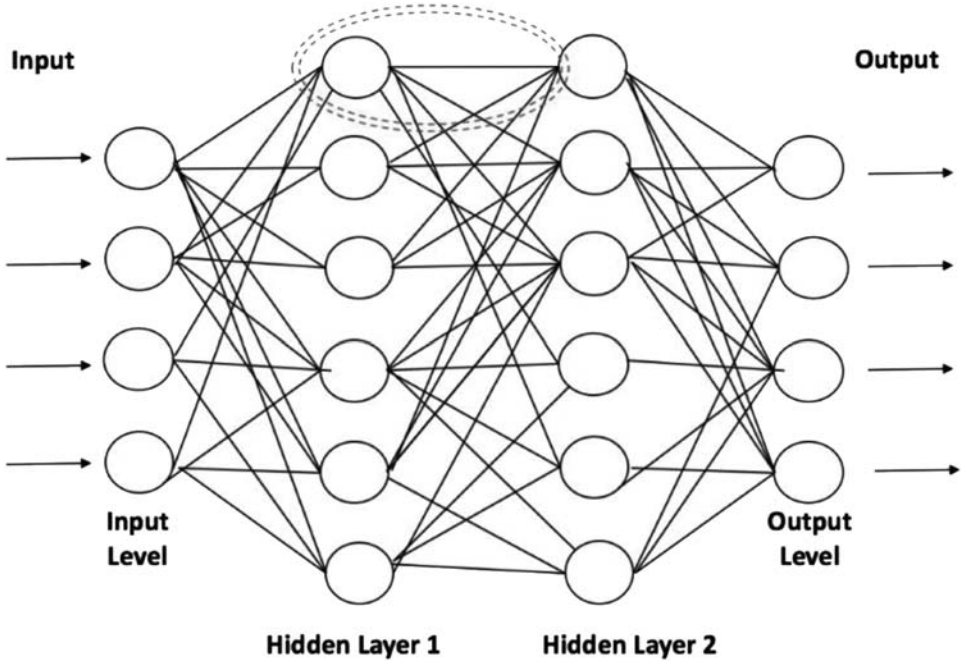

Fig. 1 – A visual representation of neural networks. Computational units at the input level receive data inputs. Input-level neurons receive data, whereas hidden layer neurons conduct the calculations necessary to analyze the complex relationships in the data. The hidden layer then sends the data to an output layer which shows the final analysis. *Original Figure Source (Modified Here):* Hashimoto DA, Rosman G, Rus D, and Meireles OR. Artificial intelligence in surgery: promises and perils. Annals of surgery. 2018 Jul; 268(1):70.

potential to integrate into a hospital's electronic medical record,[20,21] where the impact is twofold. First, preoperative risk will be automatically generated using the patient variables in the electronic medical record, allowing prediction to merge seamlessly into physician workflow. Second, integration will ensure closed-loop learning, whereby models will have access to a patient's postoperative outcome and can continuously learn to improve prediction accuracy. In fact, this integration has already started nationwide, two examples of which include MySurgeryRisk[22] and Pythia.[23] Another exciting future prospect for AI is the potential ability to identify actionable breakpoints, where a medical intervention can alter a patient's clinical course toward a more favorable outcome. Optimal Prescriptive Trees[24] are a promising tool that will learn from existing data to recommend or prescribe the best personalized care intervention for each patient that can effectively reduce the risk of postoperative complications or mortality. Nevertheless, the surgeon will always remain the ultimate decision maker and communicator with the patient, and thus carries the responsibility of understanding the capabilities and limitations of AI-generated risk predictions.

AI and computer vision (CV): A video is worth a thousand words

The field of CV teaches computers to understand the contents of images and videos. Initially viewed as a simple problem solvable in a summer's work,[25] optimism rapidly declined: CV could only reliably understand written text.[26] In 2012, Krizhevsky revolutionized the field with a novel ANN architecture that finally attained human-like image comprehension.[27] Since then, we have seen an explosion of the application of CV in daily life, including its use in autonomous-driving vehicles.[28] In addition, in the medical literature, there is promising work on automated chest x-ray and dermatologic lesion interpretation.[29,30] CV applications in surgery are more limited, with current efforts tackling intraoperative step identification in cholecystectomy,[31] sleeve gastrectomy,[1] sigmoidectomy,[32] peroral endoscopic myotomy,[33] and cataract surgery.[34]

Why are CV applications in medicine so limited compared with other fields? The difficulties arise from how machines learn and the inputs they require. In CV, ANNs typically take labeled input data (e.g., annotated images) and self-tune millions of parameters until they can accurately identify the labels for each image given the image alone. For example, if given a picture of a cat, they will output "cat." This type of learning is called supervised learning. CV in medicine faces similar challenges to other fields; however, the problem of image annotation is a special consideration for CV in medicine. Minor variations in an image that would not change a human observer's understanding can fool CV models to interpret an "STOP" sign as a "Speed Limit 45" sign (Fig. 4).[35] Similarly, hand-crafted images that look like pure noise to a human are interpreted wrongly, yet with high confidence, as animals (Fig. 5).[36] Thus, data must be meticulously curated for accurate results. If images contain confounding, visually distinct cues, the network will learn these cues rather than the more subtle diagnostic cues. This pitfall likely occurred in early models trained on radiology and dermatologic pathology data sets, where the ANNs likely identified pneumothoraces

based on the thoracostomy tubes present on chest radiographs[37] and malignant dermatologic lesions by the rules placed next to the resected specimen, which are more commonly used in malignant lesions for accurate determination of size.

Aside from careful consideration of curating input images, supervised learning methods for CV also require annotation of the images, which poses a workforce challenge. One common solution is outsourcing. Google's reCAPTCHA is an image-verification system that prompts users to identify images meeting certain criteria; for example, click all images that contain traffic lights.[38] Unbeknownst to users, some of the verification images are in fact unlabeled data from CV data sets, and in completing reCAPTCHA tests, they are simultaneously assisting in the annotation of large data sets. While harnessing the power of online users is convenient for this purpose, the annotation of surgical data sets requires annotators familiar with the field, and such domain-specific knowledge is rare in the populace. Shortcuts, such as using natural language processing to generate automated labels from radiology reports, have led to large discrepancies between these automated labels and human labeling, leading to incorrect training resulting in encoding errors.[37] Nothing has yet surpassed a human teacher in training computers to encode images such as humans. Despite these difficulties, CV and surgery remains promising. Data sets must be representative and meticulously curated to have no confounding factors. Surgical annotators must annotate to well-defined standards, and more importantly, annotate to be explainable. These explainable annotations will begin to expose some inner workings of the black box aspects of CV. Instead of annotating an overall label (e.g., gallbladder dissection), we must annotate multiple phenomena, such as "gallbladder," "dissection tool," and "blunt dissection," from which the AI can say "I believe this step is gallbladder dissection because I recognized the appropriate tools and anatomy typically used in this step." Annotations will also become more explainable once we incorporate anatomical information into the training of CV models. Automated recognition of anatomy is a hard problem: soft tissue, due to its deformability, is inherently more difficult to reliably recognize compared with rigid defined structures (such as laparoscopic images).[39] Fortunately, with modern ANN designs, models are rapidly improving their ability to recognize soft tissue structures,[40,41] which was demonstrated in a recent article that showed reliable identification of anatomical landmarks during laparoscopic cholecystectomy.[42]

The application of CV in surgery will likely not be a revolutionary process but an evolutionary one. Gradual deployment is essential. With our current ability to identify operative steps, we could automate operative note dictation (sorely needed, with 47.5% of notes inadequately describing cases)[43] and optimize operating room workflow. Soon, real-time video analysis could provide intraoperative decision support (e.g., verification that it is safe to clip), and in the more distant

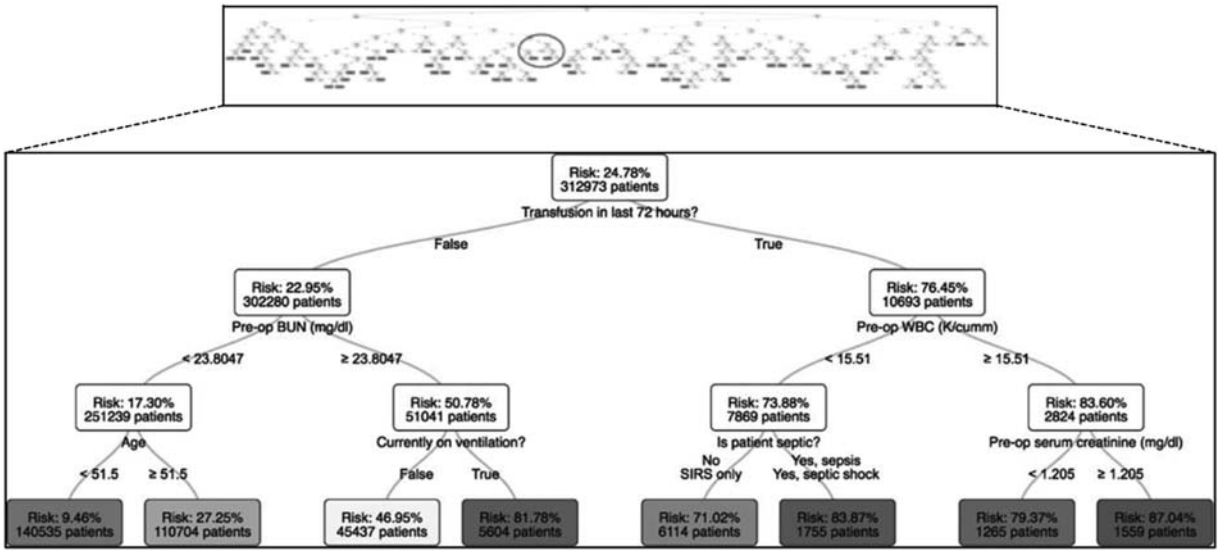

Fig. 2 – A comprehensive Optimal Classification Tree to predict postoperative 30-day mortality, with a zoom-in on one tree branch. The first node shows that there are ~313,000 patients in the data set, and the overall risk of mortality is around 25%. The next decision-tree split refers to transfusion in the 72 h before surgery. If none occurred, the algorithm leads to the left branch of the tree. If a transfusion occurred, the algorithm leads to the right branch of the tree with a different risk estimate, and the tree proceeds to split further. The variables used by the tree are not the same at each level but change based on the responses at the prior node. This is a demonstration of how a model can capture nonlinear interactions between variables. Moreover, the variables at every node and their cutoff values were determined by the algorithm, as the algorithm deemed these variables and cutoff points to the most important in estimating patient's risk at that every level. *Original Figure Source:* Bertsimas D, Dunn J, Velmahos GC, and Kaafarani HMA. Surgical risk is not linear: Derivation and validation of a novel, user-friendly, and machine learning–based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator. Annals of Surgery. 2018; 268(4):574-583.

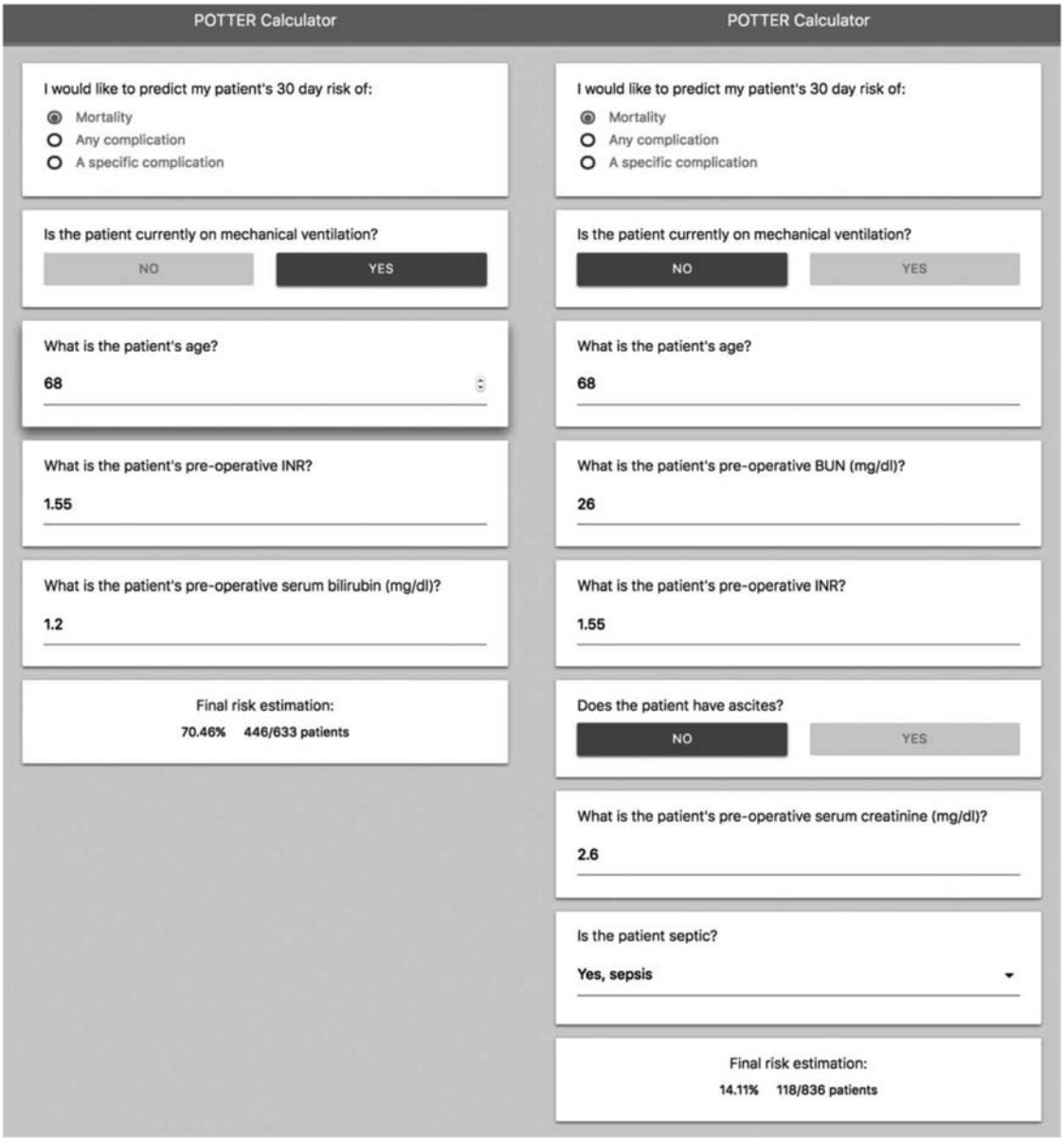

Fig. 3 – An illustration of the interactive POTTER application, where the answer to a question dictates the next question. In this particular example, whether the provider answers yes or no to the question regarding mechanical ventilation takes the algorithm and questions in a different direction. *Original Figure Source:* Bertsimas D, Dunn J, Velmahos GC, and Kaafarani HMA. Surgical risk is not linear: derivation and validation of a Novel, user-friendly, and machine learning–based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator. Annals of Surgery. 2018; 268(4):574-583.

future, we may one day even obtain a global positioning system-like guidance system for operating. Some CV systems can even "see" more than the human eye (e.g., identifying sex from retinal images alone),[44] so CV systems may even be able to provide suprahuman visual support to the surgeon. Intraoperative kinematics already can predict some postoperative variables,[45] so video analysis promises even more assistance in predicting complications. These systems, although, must serve to guide and help the surgeon, rather than to replace the surgeon, such as the lane departure warning feature in many cars today. Ultimately, the surgeon will still have the final say when it comes to operative decisions.

With the amalgamation of numerous surgical videos and the annotation knowledge of surgical experts across the world, CV can offer a source of collective expertise. This collective experience can be thought of as a "Collective Surgical Consciousness" that will help and guide surgeons everywhere to perform safer surgery.

Caveats for assessing ML and AI in biomedicine

"Red teaming" is the practice of rigorously challenging plans, policies, systems, and assumptions by adopting an adversarial

Fig. 4 – A stop sign with slight modifications that some computer vision models will mistakenly interpret as a speed limit sign. *Original Figure Source:* Eykholt K, Evtimov I, and Fernandes E *et al.* Robust Physical-World Attacks on Deep Learning Models. arXiv:170708945 [cs]. April 2018. http://arxiv.org/abs/1707.08945. Accessed January 28, 2020. . (Color version of figure is available online.)

approach. The purpose of "red teaming" is to overcome cognitive errors such as group think and confirmation bias that impair the ability for critical thinking.[46] The biomedical research community is currently inundated with applications of ML and AI. It is critical for practitioners to be able to look through the hype and critically assess these projects and their claims as they would any other scientific claim in their field. However, the technical nature of much of ML/AI can overwhelm nonspecialists with jargon and technical detail, limiting one's ability to apply their usual means of evaluating evidence. This section is intended to be a brief primer on how to cut through the technical jargon and help researchers establish a minimal baseline literacy in ML/AI, so that they can have their own internal "red team" when assessing these projects' claims. In fact, we advise that peer review journals have every submitted biomedical article on ML/AI address and respond to the content of this section.

As relayed earlier in this article, ML encompasses a wide range of analytical methods that seek to provide insight into a large data set. However, this section will focus on the assessment of ANNs trained using ML: the reason for this is that, to a great degree, the current explosion of interest in ML and AI involves systems that train ANNs. ANNs, defined earlier in this editorial, are specifically addressed in this section in the way that they are governed by what is called the Universal Approximation Theorem. The simplest interpretation of this daunting term is that an ANN can approximate a mathematical function to generate any set of data; in other words, given enough data, an ANN can be fit to any data set. This immediately leads to the following insights:

1. ML and AI are data-centric and thus are subject to all limitations of correlation (which is not causation). It almost seems trivial to note (but unfortunately is not) that this means that nothing can be identified via ML/AI outside of what is already in the data. This is because data are the output of the behavior of a particular system. These data are generated by the mechanisms that govern that system, and ML attempts to create an ANN that approximates the functions of that system. If, however, a change to those mechanisms is imposed (for instance, by the application of a new drug), and then the data outputted from that system would be different than the data used to train the ANN, and therefore the ANN would not be able to prospectively evaluate the effect of that novel drug. Thus, the function that is represented by the trained ANN (as per the Universal Approximation Theorem) would not reflect the impact of the new intervention on the system. Therefore, ML/AI cannot predict the effect of a novel drug; it can hypothesize a novel mechanism, but a potential drug that utilizes that mechanism would still need to go through all standard

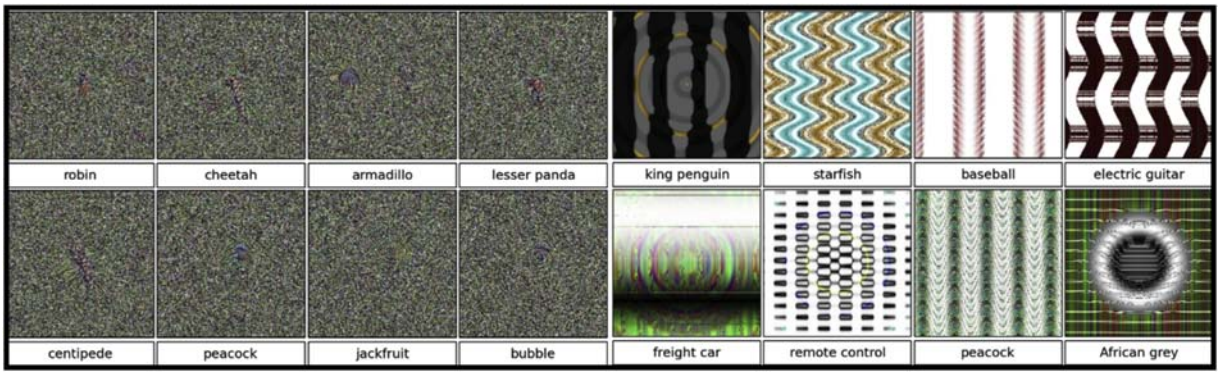

Fig. 5 – Examples of images that, while completely unrecognizable to human observers, are confidently interpreted by some CV models as everyday items and animals. *Original Figure Source:* Nguyen A, Yosinski J, and Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).; 2015:427-436. https://doi.org/10.1109/CVPR.2015.7298640. (Color version of figure is available online.)

steps involved in developing and testing a new drug. ML/AI cannot predict "what if" scenarios, unless someone has already tried those scenarios out to generate the data that would reflect the dynamics that would be captured by the ANN. More importantly, even if someone has tried out a particular intervention, if that sample is not large (i.e., less than 1000s of instances), it can be very misleading, as the sparsity of data would not reflect the total space of possibilities produced by the actual system, which leads to the second point in the following.

2. ANNs can be fit to anything, almost too well. The greater the number of hidden layers in an ANN (termed a deep network), the more readily it will fit any training data set. The standard method of evaluating a predictive tool is to divide the available data into a training set and test set; this is usually a 70%-80% training *versus* 30%-20% testing proportion, which is termed internal validation. Assuming an appropriately randomized division, a sufficiently deep ANN will almost always pass this test; therefore, internal validation performed in the classical fashion could be meaningless. The intent behind any AI system is that it provides a benefit in the real world, with much more noise, heterogeneity, and confounding factors than an intrinsically limited training set: this is the concept of *external validation*. Because ANNs are so good at fitting their training set, they are intrinsically brittle, and often rapidly fail external validation. The tech community realizes this, which is why they have adopted the principle that their predictive algorithms are constantly being improved through routine use.[47] While this paradigm manifests to some degree in traditional biomedicine, where "approved" therapies and devices do go through assessment and some potential refinement when placed in clinical practice, the tolerance for the tech mantra of "Fail Early, Fail Often" is not acceptable. This paradigm, which is predicated on a rapid cycle of iterative improvement, is crucial in modern ML/AI given the brittle nature of ANNs and the recognition that their performance will immediately degrade once they are released into the wild (because they are seeing data they were not trained on). Tech companies can tolerate rapid iteration because the results of their predictive systems are not mission critical per use instance; i.e., Amazon does not suffer unduly each time their book-suggestion algorithm does not result in a purchase. The failure of that predictive algorithm in a particular instance is of no real consequence and is just used to refine their algorithm. However, this lack of mission criticality cannot be translated, in many cases, to biomedicine where there is little to no tolerance for ongoing error that affects an individual patient. For instance, there would be little tolerance for missed cancer diagnoses on AI-interpreted imaging studies if those missed diagnoses were justified as "Fail Early, Fail Often."
3. There is no insight into the function the ANN is approximating, and this makes them subject to unintended and unanticipated behavior. This manifests in two critical areas: 1) the data the researcher thinks an ANN is using is not what it is actually using, for example, identifying cancer on histology slides based on batching by index of suspicion and 2) a flawed objective function for optimization. A popular illustrative example of point 2) is the creation of a paper clip-making AI overtaking the universe.[48,49] This fanciful example posits the creation of an AI that is tasked with optimizing the creation of paper clips; however, without constraints on how resources can be accessed and reallocated toward this goal, the AI ends up taking over the universe by repurposing all machinery to the creation of paper clips. These issues of unintended consequences are particularly accentuated when there is no ground truth by which to judge annotation. Consider the difference between an ML/AI classifier for cancer identification on CT scan, where there is ground truth in knowing which of the nodules did actually have cancer (histology), *versus* an ML/AI classifier that relies on "expert" opinion intrinsic to diagnostic criteria (i.e., pneumonia or the View of Safety). In the former, there is ground truth biological criteria (was there cancer or not?) as opposed to a more nebulous means of confirmation in clinical settings (e.g., are all radiographic pneumonias confirmed with bronchoalveolar lavage cultures? How precisely is the lower 1/3 of the gallbladder determined free from the cystic plate?). Accepting that there are always epistemic limitations on "absolute truth," certain circumstances allow for greater degrees of trustworthiness in terms of determining ground truth. In particular, in cases that heavily rely on annotation as a means of judging "truth," it must be recognized that there will always be variation among annotators, consensus does not necessarily mean truth, and increasing the number of annotators only increases noise in the system.

People working in AI are well versed in the limitations of their tools.[50-52] However, there is a tendency today for some to become advocates for their particular approach, and this can lead to a culture of advocacy that is in direct opposition to the skeptical foundations of the scientific method. Therefore, having a literate internal "red team" is of critical importance for any responsible practitioner. The following are suggested questions to use for assessing an ML/AI project (adapted from ref 50):

1. Is there external validation, in which the developed ANN is tested against a data set not only distinct from its training set, but from a collection context more reflective of a clinical population?
2. If the task is a classifier (i.e., disease identification, prognosis prediction, etc), how well does that ANN classifier perform against standard regression (ideally performed with the same training and test set)? This is an important assessment because for many cases classical logistic regression performs as well as ML.[53]
3. If the task is a classifier, what is the basis of "ground truth" in the training criteria? Is there objective criteria (i.e., mortality and tissue-proven cancer) for that ground truth?
4. Could there be hidden confounders in the training data? Always be wary of hidden confounders in the training data, particularly selection bias, when trying to interpret the result of an ML/AI project. A classic example of this is the inherent bias in sentencing guidelines.[54] Manifestations in biomedicine could occur from labeling and batching practices and preselection of collateral patient features. It is important to remember that any bias in the original data set will be reflected by the AI.

The intent of listing these limitations and caveats regarding ML and AI is not to prompt the rejection of these technologies. Rather, the goal is to foster reasonable expectations from methods currently subject to a whirl in the hype machine. "Irrational exuberance" (to borrow a phrase from Alan Greenspan) leads to inevitable crashes when unrealistic expectations are not met, which can set back real development by years and even decades by providing an easy target for nay-sayers. There is no question that ML and AI represent potential game-changers in how biomedical science is performed and what can be achieved in the future, but it is imperative that, as responsible scientists, we continue to evaluate these newer technologies with the same evaluative rigor that is applied to all aspects of surgical science.

Conclusion

AI is a critical part of the future of medicine and surgery in a variety of areas, such as screening, diagnosis, treatment, multidisciplinary patient-targeted care, and health care management. Special emphasis was placed in this review on the areas of preoperative risk prediction, which is a paramount aspect of patient safety, as well as the role of AI and CV in shaping the evolution of daily surgical practice. With all promises of AI and ML comes a feeling of unease given how little we truly know about the nature, capabilities, and pitfalls of AI. We have also seen in this review that there are several potential limitations that should be carefully taken into consideration, when evaluating the role of AI. The key here is to welcome this technology, but at the same time seek to learn from and understand it, especially the challenges and obstacles involved in its application in medicine and surgery. Once we do that successfully, then we will be able to better integrate this promising technology into the future of medicine.

Acknowledgment

Author's contributions: Drs. El Hechi, Ward, An, and Tsoulfas contributed to the conception and design, writing the article, and approval of the final version of the manuscript. Drs. Maurer and El Moheb contributed to the conception and design, critical revision of the article, and approval of the final version of the manuscript. Dr. Kaafarani contributed to the conception and design, writing of the article, critical revision of the article, and approval of the final version of the manuscript.

Disclosure

Dr. Ward reports research support from Olympus Corporation. Dr. Kaafarani reports research support from the CRICO/Risk Management Foundation. The other authors report no proprietary or commercial interest in any product mentioned or concept discussed in this article.

REFERENCES

1. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268:70–76.
2. Wanderer JP, Gratch DM, Jacques PS, Rodriquez LI, Epstein RH. Trends in the prevalence of intraoperative adverse events at two academic hospitals after implementation of a mandatory reporting system. *Anesth Analg*. 2018;126:134–140.
3. Desebbe O, Lanz T, Kain Z, Cannesson M. The perioperative surgical home: an innovative, patient-centred and cost-effective perioperative care model. *Anaesth Crit Care Pain Med*. 2016;35:59–66.
4. Howard R, Thompson M, Fan Z, et al. Costs associated with modifiable risk factors in ventral and incisional hernia repair. *JAMA Netw Open*. 2019;2:e1916330.
5. Fischer JP, Basta MN, Mirzabeigi MN, et al. A risk model and cost analysis of incisional hernia after elective, abdominal surgery based upon 12,373 cases: the case for targeted prophylactic intervention. *Ann Surg*. 2016;263:1010–1017.
6. Wolters U, Wolf T, Stutzer H, Schroder T. ASA classification and perioperative variables as predictors of postoperative outcome. *Br J Anaesth*. 1996;77:217–222.
7. Knaus WA, Wagner DP, Draper EA, et al. The Apache III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults. *Chest*. 1991;100:1619–1636.
8. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg*. 2013;217, 833-842 e831-833.
9. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *J Am Coll Surg*. 2013;217:336–346.
10. Callery MP, Pratt WB, Kent TS, Chaikof EL, Vollmer Jr CM. A prospectively validated clinical risk score accurately predicts pancreatic fistula after pancreatoduodenectomy. *J Am Coll Surg*. 2013;216:1–14.
11. Kamath PS, Wiesner RH, Malinchoc M, et al. A model to predict survival in patients with end-stage liver disease. *Hepatology*. 2001;33:464–470.
12. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based predictive OpTimal trees in emergency surgery risk (POTTER) calculator. *Ann Surg*. 2018;268:574–583.
13. Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N Engl J Med*. 2017;376:2507–2509.
14. Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ*. 2019;364:l886.
15. Yu K-H, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf*. 2019;28:238–241.
16. Buranyi S. Rise of the racist robots—how AI is learning all our worst impulses. *Guardian*. 2017;8.
17. Bertsimas DDJ. Optimal classification trees. *Mach Learn*. 2017;106:1039–1082.
18. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMJ. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and Machine-learning-based predictive optimal trees in emergency surgery risk (Potter) calculator. *Ann Surg*. 2018;268:574–583.
19. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear: derivation and validation of a novel, user-

friendly, and machine-learning-based predictive OpTimal trees in emergency surgery risk (POTTER) calculator. *Ann Surg*. 2018.
20. Amrock LG, Neuman MD, Lin HM, Deiner S. Can routine preoperative data predict adverse outcomes in the elderly? Development and validation of a simple risk model incorporating a chart-derived frailty score. *J Am Coll Surg*. 2014;219:684–694.
21. Anderson JECD. Using electronic health records for surgical quality improvement in the era of big data. *JAMA Surg*. 2015:24–29.
22. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann Surg*. 2019;269:652–662.
23. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med*. 2018;15:e1002701.
24. Bertsimas D, Dunn J, Mundru N. Optimal prescriptive trees. *INFORMS J Optimization*. 2019;1:164–183.
25. Papert S. *The summer vision project. Memo AIM-100*. MIT AI Lab; 1966. Available at: http://hdl.handle.net/1721.1/6125. Accessed May 20, 2020.
26. LeCun Y The MNIST database of handwritten digits.
27. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM*. 2017;60:84–90.
28. Autopilot. Available at: https://www.tesla.com/autopilot. Accessed March 18, 2020.
29. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP. *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning*. arXiv preprint arXiv:1711.05225; 2017. Available at: https://arxiv.org/abs/1711.05225. Accessed May 20, 2020.
30. Estava A, Kuprel B, Novoa R. Dermatologist level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115.
31. Yengera G, Mutter D, Marescaux J, Padoy N. *Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks*. arXiv preprint arXiv:1805.08569; 2018. Available at: https://arxiv.org/abs/1805.08569. Accessed May 20, 2020.
32. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg Endosc*. 2019:1–8.
33. Ward TM, Hashimoto DA, Ban Y, et al. Automated operative phase identification in peroral endoscopic myotomy. *Surg Endosc*. 2020:1–8.
34. Zisimopoulos O, Flouty E, Luengo I, et al. Deepphase: surgical phase recognition in cataracts videos. *Int Conf Med Image Comput*. 2018:265–272.
35. Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification. *Proc IEEE Conf Comput Vis Pattern Recogn*. 2018:1625–1634.
36. Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *Proc IEEE Conf Comput Vis Pattern Recogn*. 2015:427–436.
37. Oakden-Rayner L. CheXNet: an in-depth review. Luke Oakden-Rayner (PhD Candidate/Radiologist) Blog. 2018 Aug. Available at: https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnetan-in-depth-review. Accessed 20 May 2020.
38. Von Ahn L, Blum M, Langford J. Telling humans and computers apart automatically. *Comm ACM*. 2004;47:56–60.
39. Mountney P, Lo B, Thiemjarus S, Stoyanov D, Zhong-Yang G. A probabilistic framework for tracking deformable soft tissue in minimally invasive surgery. *Int Conf Med Image Comput*. 2007:34–41.
40. He Q, Bano S, Ahmad OF, et al. Deep learning-based anatomical site classification for upper gastrointestinal endoscopy. *Int J Comput Assist Radiol Surg*. 2020.
41. Pfeiffer M, Riediger C, Weitz J, Speidel S. Learning soft tissue behavior of organs for surgical navigation with convolutional neural networks. *Int J Comput Assist Radiol Surg*. 2019;14:1147–1155.
42. Tokuyasu T, Iwashita Y, Matsunobu Y, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surg Endosc*. 2020:1–8.
43. Oudard S, Latorzeff I, Caty A, et al. Effect of adding docetaxel to androgen-deprivation therapy in patients with high-risk prostate cancer with rising prostate-specific antigen levels after primary local therapy: a randomized clinical trial. *JAMA Oncol*. 2019;5:623–632.
44. Ting DSW, Wong TY. Eyeing cardiovascular risk factors. *Nat Biomed Eng*. 2018;2:140–141.
45. Hung AJ, Chen J, Gill IS. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA Surg*. 2018;153:770–771.
46. Rouse M, Wigmore I. Red teaming. Available at: https://whatis.techtarget.com/definition/red-teaming. Accessed March 7, 2020.
47. Lomonaco V. Why continual learning is the key towards machine learning. Available at: https://medium.com/continual-ai/why-continuous-learning-is-the-key-towards-machine-intelligence-1851cb57c308. Accessed March 10, 2020.
48. Bostrom N. Ethical issues in advanced artificial intelligence. *Sci Fiction Philos*. 2003:277–284.
49. Gans JS. *Self-regulating artificial general intelligence*. *National Bureau of Economic Research*; 2018. Available at: https://www.nber.org/papers/w24352. Accessed May 20, 2020.
50. Davis, Ernest., Marcus, Gary. *Rebooting AI: Building Artificial Intelligence We Can Trust*. United States: Knopf Doubleday Publishing Group; 2020. Available at: https://www.penguinrandomhouse.com/books/603982/rebooting-ai-by-gary-marcus-and-ernest-davis/. Accessed May 20, 2020.
51. L'heureux A, Grolinger K, Elyamany HF, Capretz MA. Machine learning with big data: challenges and approaches. *IEEE Access*. 2017;5:7776–7797.
52. Riley P. Three pitfalls to avoid in machine learning. *Nature*. 2019;572:27–29.
53. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
54. Angwin J, Larson J. Machine bias. ProPublica. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed March 7, 2020.