

## MIT Open Access Articles

*Deep Learning for Cardiovascular Risk Stratification*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Schlesinger, Daphne E. and C.M. Stultz. "Deep Learning for Cardiovascular Risk Stratification." *Current Treatment Options in Cardiovascular Medicine* 22, 8 (June 2020): 15  
© 2020 Springer Nature

**Published Version:** <http://dx.doi.org/10.1007/s11936-020-00814-0>

**Publisher:** Springer Science and Business Media LLC

**Permanent Link:** <https://hdl.handle.net/1721.1/131714>

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of use:** Creative Commons Attribution





# Deep Learning for Cardiovascular Risk Stratification

Daphne E. Schlesinger<sup>1,2,3,4</sup>  
Collin M. Stultz, MD, PhD<sup>1,2,3,4,5,6,\*</sup>

## Address

<sup>1</sup>Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, 02139, USA

<sup>2</sup>Institute for Medical Engineering and Science, MIT, Cambridge, MA, 02139, USA

<sup>3</sup>Research Laboratory of Electronics, MIT, Cambridge, MA, 02139, USA

<sup>4</sup>Computer Science & Artificial Intelligence Laboratory, MIT, Cambridge, MA, 02139, USA

<sup>5</sup>Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139, USA

<sup>\*,6</sup>Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA  
Email: cmstultz@mit.edu

© The Author(s) 2020

This article is part of the Topical Collection on *State-of-the-Arts Informatics*

**Keywords** Risk stratification · Deep learning · Risk models

## Abstract

*Purpose of review* Although deep learning represents an exciting platform for the development of risk stratification models, it is challenging to evaluate these models beyond simple statistical measures of success, which do not always provide insight into a model's clinical utility. Here we propose a framework for evaluating deep learning models and discuss a number of interesting applications in light of these rubrics.

*Recent findings* Data scientists and clinicians alike have applied a variety of deep learning techniques to both medical images and structured electronic medical record data. In many cases, these methods have resulted in risk stratification models that have improved discriminatory ability relative to more straightforward methods. Nevertheless, in many instances, it remains unclear how useful the resulting models are to practicing clinicians.

*Summary* To be useful, deep learning models for cardiovascular risk stratification must not only be accurate but they must also provide insight into when they are likely to yield inaccurate results and be explainable in the sense that health care providers can understand why the model arrives at a particular result. These additional criteria help to ensure that the model can be faithfully applied to the demographic for which it is most accurate.

## Introduction

Accurate risk stratification remains a central theme in all stages of the care of patients with cardiovascular disease. Indeed, the likelihood that any patient will benefit from a given therapeutic intervention is a function, in part, of the risk associated with the intervention itself versus the risk that the patient will have an adverse event if no intervention is performed. Informed clinical decision making necessitates gauging patient risk using available clinical information.

A number of societal guidelines recommend the use of validated risk scores in the initial evaluation of patients with suspected coronary disease [1–3]. The use of accurate risk scores helps to ensure that patients who are at high risk of adverse outcomes are quickly identified and assigned a therapy that is appropriate for their level of risk. Nevertheless, risk stratification is far from a perfect science, and risk scores often fail to identify patients at high risk of inimical outcomes. This problem is made more apparent in light of the fact that a relative minority of patients with cardiovascular disease experience the gravest adverse outcomes. Moreover, while the prevalence of adverse events in high-risk populations is, by definition, large, the absolute number of events is also large in patients who are predicted to be low risk using traditional risk prediction metrics. This *low risk-*

*high number dilemma* is frequently encountered in many areas of cardiovascular clinical research [4]. As such, adequately identifying patient subgroups who are truly at high risk of adverse events remains a clear unmet clinical need. Novel methods are therefore needed to realize the full potential of clinical risk stratification from existing clinical observations. Machine learning and deep learning, in particular, holds the potential to robustly identify high-risk patient subgroups, suggest personalized interventions that can reduce a given patient's risk, and help ensure that appropriate resources are allocated to those patients who are in the most need.

In this review, we do not strive to review all of the relevant literature in the area of deep learning in cardiovascular medicine. Indeed, this review is written for the practicing clinician and strives to provide intuitive explanations for how deep learning models actually work and where they are most applicable. As the use of these models becomes ubiquitous in the clinical arena, it will be important for health care providers to critically evaluate them in order to determine the clinical usefulness of any given machine learning approach. Our goal is to provide a general framework for understanding what advantages these models hold and what considerations limit their broad applicability.

## Conventional approaches to risk stratification

The term machine learning is believed to have been originated by Arthur Samuel, an engineer and scientist who pioneered artificial intelligence in 1959 [5]. He described it as “programming computers to learn from experience.” There are diverse examples of machine learning in the clinical literature, including straightforward approaches like logistic regression and Cox proportional hazards modeling and more esoteric techniques like deep learning, which is described in the next section. Indeed, the former methods have actually been a part of the clinical literature for some time [6–8]. Therefore, while the term machine learning has only recently entered the medical lexicon, a number of existing clinical risk scores were developed and refined using approaches that fall under this umbrella term. The exorbitant list of such models is too lengthy to exhaustively review here. Instead, we focus on some approaches that are commonly used to assess patient risk.

One of the earliest models for quantifying the risk of adverse cardiovascular outcomes was developed by Killip et al. in 1967, where 250 patients were divided into four simple classes of increasing severity of illness, ranging from no clinical signs of heart failure to cardiogenic shock [9]. The primary goal of this study was to trial an improved workflow for cardiac intensive care, but the

data collected over the course of study revealed patterns in patient survival based on their class (now called the Killip class). The utility of these classes for identifying high risk patients has been born out in a number of studies, and these classes remain a part of the clinical assessment of patients who present with an acute myocardial infarction.

Over time, more sophisticated statistical techniques have been used to develop more sophisticated risk stratification models. Both the Framingham risk score—which quantifies the risk of adverse events (death from coronary heart disease, nonfatal MI, angina, stroke, transient ischemic attack, intermittent claudication, and heart failure) in patients who had no prior history of cardiac disease—and the Global Registry of Acute Coronary Events (GRACE) score—which quantifies all-cause mortality in patients who present with an ACS—were developed using Cox proportional hazards regression [10, 11]. Another class of risk scores, developed from and named for the Thrombolysis in Myocardial Infarction (TIMI) study groups, was developed specifically for patients who present with symptoms consistent with an acute coronary syndrome. Here, features that were discriminatory with respect to the combined outcome of all-cause mortality, new or recurrent MI, or severe recurrent ischemia in their cohort were selected using logistic regression. Seven features were selected in the final model. To use the risk score itself, the physician simply counts the number of features that are present to estimate the short-term risk of either mortality after a myocardial infarction post ST segment elevation MI or a combined outcome of all-cause mortality, new or recurrent MI, or severe recurrent ischemia requiring revascularization post non-ST segment elevation ACS [12, 13].

Regression modeling has found a role for quantifying patient risk in other disorders apart from ischemic heart disease. Pocock et al., for example, performed a meta-analysis of heart failure patients from 30 different studies, amounting to 39,372 patients. They used multivariable piecewise Poisson regression methods to identify features that are predictive of mortality at 3 years. These features were then converted into an integer risk calculator, called the Meta-analysis Global Group in Chronic Heart Failure (MAGGIC) score, with higher values corresponding to greater risk [14]. Similarly, the Seattle Heart Failure Model was developed on a cohort of 1125 patients, using a multivariate Cox proportional hazards model. This model provides estimates for 1-, 2-, and 3-year mortalities [15, 16].

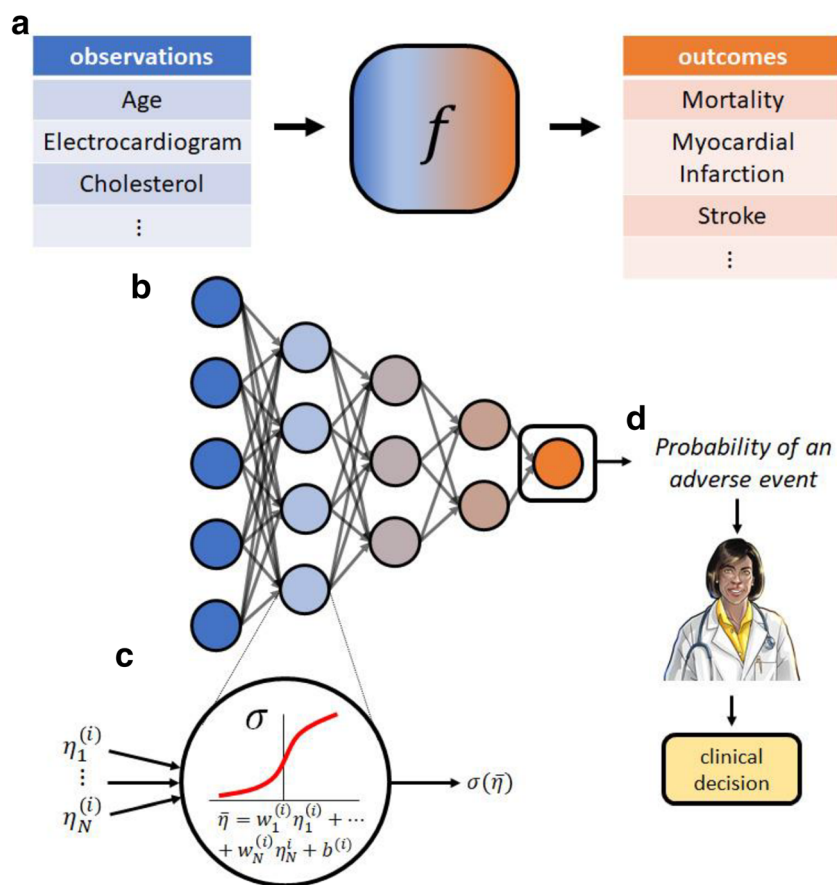
Logistic regression and proportional hazard models are advantageous because they are easy to interpret: each clinical feature in the model has an associated weight that corresponds to how important that feature is for the model arriving at a particular result. However, such models are relatively simple and cannot necessarily capture complex mechanisms relating observations and outcomes of interest.

## What is deep learning?

The diverse, nonuniform terminology in the medical literature unfortunately tends to obfuscate the meaning of the term “deep learning.” Deep learning is a subfield of machine learning that strives to find powerful abstract representations of data using complex artificial neural networks (ANNs) that are then used to accomplish some prespecified task. While these abstract data representations are powerful ways to describe clinical data, they are difficult to comprehend and explain; that is why they are, indeed, “abstract.”

ANNs correspond to a class of machine learning algorithms whose algorithmic structure is inspired by structure of the human brain and how it is believed that humans compute [17, 18]. A neural network consists of interconnected artificial neurons that pass information between one another. A typical ANN contains an input layer, which contains several artificial neurons that take clinically meaningful data as input. The input layer then passes the clinical data to other inner, or “hidden,” layers, each of which performs a series of relatively simple computations. At each layer, more abstract representations of the input data are obtained. Eventually, the information is passed to an output layer that yields a clinically meaningful quantity (Fig. 1).

Deep learning models, in practice, correspond to neural networks that contain several hidden layers. These models, originally referred to as multilayer perceptrons, were popularized in the early 1980s for applications such as image and speech recognition, then receded in popularity in favor of simpler, easier to



**Fig. 1.** In our applications, a neural network acts as a function that takes some observations as input and produces some prediction of outcomes as the output (a). This function is generated by adding many simple functions (represented by circular nodes that process information), each of which takes all the outputs of the previous layer as its input, which renders a network “fully connected” (b). These simple functions are strictly increasing and include parameters ( $w^{(i)}, b^{(i)}$  for each node), which are chosen by training the network (c). Each layer can be thought of as an abstraction of the data, which is eventually separable in the last layer if the model works well. The output of the last layer is the probability of an adverse event, which a clinician may use to inform her clinical decisions (d).

train, and perhaps more explainable models [19, 20]. In recent years, however, deep neural network (DNN) learning has resurged dramatically both because of the availability of so-called “big data” and the development of computational methods that facilitate the training of large neural networks. In many of today’s applications, these networks can be quite large, having on the order of  $10^5$ – $10^6$  artificial neurons and millions of modifiable parameters. Parenthetically, as the size of clinical datasets is typically much smaller, care must be taken when implementing these models to ensure that they are not overtrained.

While the structure of ANNs, and DNNs in particular, are inspired by the structure of neurons in the human brain, these models are best thought of as universal function approximators. Indeed, it has been mathematically proven that any continuous function on compact spaces can be represented by a neural network, under certain constraints [21, 22]. These models therefore form an efficient platform for generating functions that model complex relationships between patient characteristics/features and outcomes. This highlights an important difference between DNNs and simpler methods like logistic regression, which models the relationship between outcomes (i.e., the logarithm of the odds ratio) and patient features as a linear function. By contrast, a DNN corresponds to a complex, highly nonlinear function that takes patient information as input (including medical images) and outputs the corresponding outcome. An additional advantage of DNNs is that they can use input data in “raw” form, with little preprocessing.

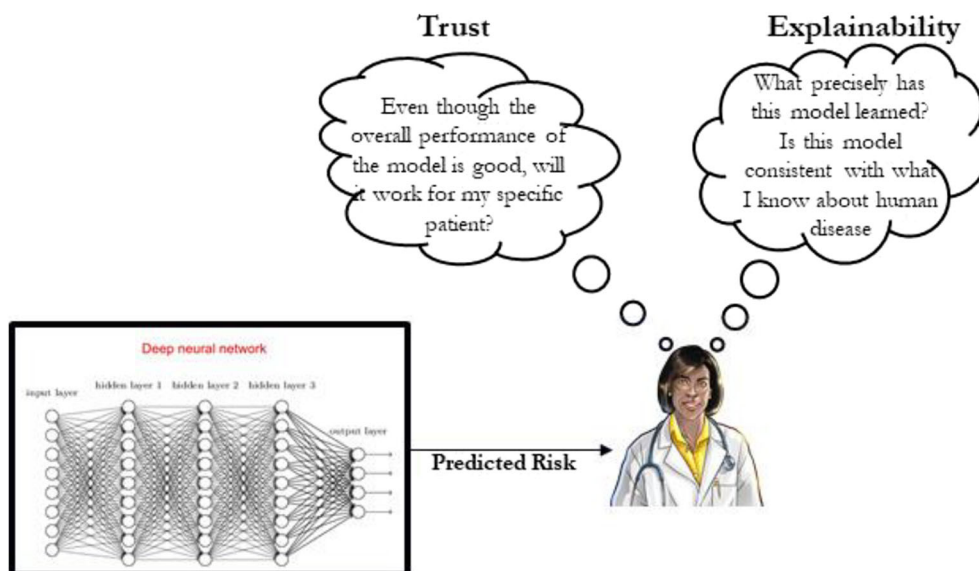
Deep learning models can, in principle, capture complex, nonlinear, relationships between patient features and outcomes and therefore necessarily meet the first criteria. However, because these models generate abstract representations of the input data, it can be very difficult to understand what the model has learned and consequently why the model arrives at a particular result. Moreover, understanding when the model will fail—i.e., which patients are most likely to be associated with an incorrect prediction—can be just as challenging.

## Evaluating deep learning risk models

Standard performance metrics, such as the area under the receiver operating characteristic curve (AUC), accuracy, and the sensitivity/specificity, provide useful information for gauging how a risk model will perform, on average. Nevertheless, these metrics do not by themselves offer any interpretative insights, nor do they help the user understand how the model will perform on any individual patient. The upshot being that conventional statistical metrics of success are not always sufficient to determine the clinical utility of a deep learning model.

When evaluating applications of machine learning to medical problems, there are particular criteria that must be considered given our current understanding of human physiology and the reality of medical practice (Fig. 2). In addition to having a level of performance that ensures that it will perform well, on average, on the population of interest, ideally a good algorithmic solution should also:

1. Provide information about potential failure modes; i.e., indicate when it is likely to yield a false result;



**Fig. 2.** Issues that hinder the clinical acceptance of deep learning models.

2. Be explainable in the sense that clinicians can understand why the model arrives at a particular result.

Although determining when a model will fail is challenging, it is an essential task. Formally, this can be understood as finding, a priori, patient characteristics or subgroups that are associated with incorrect predictions. The development of methods that identify such “failure modes” are also a nascent area of research within the machine learning community, with most of the published research appearing in specialized machine learning conferences or non-peer-reviewed online printed archives, with little associated work appearing in the clinical literature. Nevertheless, insights into when a model will fail can often be garnered if the model itself is explainable; i.e., understanding how/why a model arrives at a particular result often provides clues as to how the model can yield an incorrect result.

Recently, a new method was described for identifying when a given clinical risk score will yield unreliable results [23••]. The approach identifies, a priori, patient cohorts associated with reduced model accuracy, discriminatory ability, and poor calibration. Application to the GRACE risk model correctly identifies patient cohorts where the GRACE score has reduced performance. Advantages of the method are that it is straightforward to implement and that it can be applied to any risk model, regardless of how the risk model was developed—thereby making the approach appropriate for deep learning models. General methods along these lines will likely play an increasingly important role in determining when complex risk models are expected to yield useful predictions.

In addition to deciphering when a given model is likely to fail, developing methods that “explain” what a model has learned is an important part of any comprehensive strategy that strives to maximize clinical acceptance. Nevertheless, conceptions of explainability or interpretability of machine learning

models are diverse, and it is difficult to determine exactly what this term means in the context of machine learning models. In his article, "The Mythos of Model Interpretability," Zachary Lipton identifies five types of interpretability for machine learning models: trust, causality, transferability, informativeness, and fairness [24]. Of particular interest for medical algorithms are causality and informativeness. Causality describes if the relationships discovered by the model are truly causal or merely correlative. While causality in machine learning is an active area of research, it is always very difficult to tease out causal relationships from a retrospective analysis of any dataset [25]. An informative deep learning model provides some intuition to support how it arrives at a given result. In order to impart useful intuitions, however, one needs to translate the abstract representations learned by a deep learning model into language that is easily understood by the health care practitioner. In short, in the medical context, we ideally need models that yield insights that are translatable into the language of physiology (Fig. 2).

There are a limited number of tools that have been used to provide interpretations/explanations of what a deep neural network has learned. Shapley values, Gradient-weighted Class Activation Mapping (Grad-CAM) methods, and saliency maps represent a class of methods that can provide insight into what input features are most responsible for the risk model making a prediction [26–28]. Grad-CAM and saliency maps, in particular, are typically used with convolutional neural networks (described below) and provide insight into the relative importance of different parts of an image for a specific prediction [29]. For example, consider a model trained to distinguish between different objects, such as dogs and humans. A saliency map may reveal that pixels corresponding to the legs (four for a dog and two for a human) are most dispositive. Hence, for such a simple task as differentiating humans from dogs, saliency maps provide easily understood "explanations." However, for more complex classification tasks, saliency maps may not yield such readily interpretable insights. Indeed, these methods generally do not provide information about how the data in these regions were used to arrive at a particular decision, nor do they necessarily provide any causal insights. More generally, it has been argued that the attempts to explain deep models are inherently flawed because such post hoc explanations can never have true fidelity with respect to the original complex model [30]. In this vein, the use of interpretable models have an advantage in that they are designed to yield explanations that can be understood by domain experts. Nevertheless, it is not clear that commonly used interpretable models can capture the complex nonlinear relationships described above in manner that yields clear explanations. A compromise may be to build models that combine both mechanistic/physiologic models and deep learning models to enhance both model explainability and predictive performance. This is an active area of research.

It has been argued that clinicians should embrace black box models rather than strive to develop explanations that provide insight into how the model arrives at a particular result [31]. Proponents of this thesis argue that clinical decision-making is frequently rooted in an incomplete understanding of the disease process in question and how the potential intervention actually works. Hence requiring deep learning models to be explainable holds them to a higher standard than other methods used to inform clinical decision making and further stymie innovation in this space.

While there is merit to this argument, there is little doubt that clinical decisions are grounded in some understanding of the disease process. Indeed, it is precisely this, albeit imperfect, understanding that guides our therapeutic choices. By contrast, deep learning models represent an unprecedented level of opaqueness with respect to clinical understanding. In the setting of black models and only statistical measures of the model's overall performance, additional information are needed to determine when a model prediction is appropriate for a specific patient. While the identification of model failure modes and explainability are distinct concepts, they are related. Failure mode analyses strive to identify patient subgroups where the model has reduced performance, and a comprehensive understanding of how a complex model arrives at a particular result provides further assurance that the model is appropriate for a given patient, who has a given set of clinical characteristics. Explanations that are inconsistent, for example, with our understanding of the underlying pathophysiology should not be trusted.

In sum, it is our view that deep learning models for any clinical application should be evaluated using these metrics, in addition to standard statistical measures of performance. In what follows we discuss several recent applications of deep learning methods for cardiovascular risk stratification and evaluate them relative to the metrics discussed above.

## Deep learning for risk prediction

Deep learning for image classification has a relatively extensive literature. Indeed, the Imagenet challenge—a worldwide competition for classifying millions of curated images—has led to the development of many sophisticated algorithms for image classification [32]. In a number of applications, these image classification algorithms have been modified and fruitfully applied to clinical images to quantify patient risk. However, these methods have mainly been used for automatic disease diagnosis from pathology slides and radiological scans [33–38]. These algorithms are usually implemented using a class of DNNs called convolutional neural networks (CNNs). CNNs are inspired by the structure of the mammalian visual cortex, where each neuron “sees” a small region of the visual field, called the receptive field of that neuron [39]. In a CNN, the information contained in adjacent groups of pixels of an image, analogous to the receptive field, is summarized, using a mathematical operation called a convolution to create an abstraction of the information in the image [40].

In cardiology, deep learning work has been focused on the automatic interpretation of cardiac images, with few applications to the development of models that directly quantify patient risk [41]. Recent studies have highlighted the ability of CNNs to identify echocardiographic windows using the images alone [42, 43], correctly segment the left ventricle in both cardiac CT images and cardiac MRIs [44, 45], and accurately detect cardiac MR motion artifacts [46]. The use of CNNs to garner insights into the risk of future adverse outcomes, however, is still a nascent area of investigation.

A recent study that purports to use medical image data for assessing cardiovascular risk was published by Poplin et al. [47••]. In that work, the authors used a CNN to predict age, gender, smoking status, systolic blood pressure,

diastolic blood pressure, and, most importantly, major adverse cardiovascular events (MACE) within 5 years from the time that retinal fundus images were acquired. The dataset used to develop and validate the model was obtained from the UK Biobank and EyePACS (a retinal image database consisting of images obtained during routine diabetic screening in clinics in the USA). They report an AUC of 0.70 for predicting MACE after 5 years using their deep algorithm. This performance exceeds that of predictions made based on single risk factors such as age and systolic blood pressure. However, they do not outperform an existing, simpler proportional hazards model, SCORE (Systematic COronary Risk Evaluation), proposed by Conroy et al. in 2003 [48]. In addition to predicting risk, they utilized saliency maps, described above, to attempt to explain their algorithm. Saliency maps highlight portions of the retinal images that contributed significantly to the predictions their models produced. However, the usefulness of these saliency maps is limited because they give us no information about the mechanism by which certain features of the retina relate to cardiovascular risk and if the deep learning model has recapitulated that mechanism.

Recently, there have been attempts to extend classification algorithms, which were originally designed to analyze medical images, to different types of data in the Electronic Medical Record (EMR). The EMR can be divided into two types of data: structured data and unstructured data. Structure medical data refers to what can be found in the pre-existing fields with the electronic medical record; e.g., lab results, vital signs, and demographic information. Unstructured data refers to what appears in medical notes written by health care practitioners. In a recent study, Mayampurath et al. assembled structured data from the electronic health record into a visual format that could then be used to train a CNN to predict in-hospital outcomes [49]. Essentially, the EMR is converted to a two-dimensional medical image, which enables the use of standard machine learning techniques appropriate for medical image processing. The image itself maps time on one axis and 156 clinical variables (including vital signs, laboratory results, medications, diagnostic tests, and nurse examinations), recorded over the first 48 h of admission, on the other axis. Overall, the discriminatory ability of the best performing CNN (the authors considered more than one) was 0.91, suggesting that the method holds considerable promise.

A significant advantage here is that they can leverage methods used to “interpret” what CNNs have learned about images to help explain why their deep learning model arrives at a particular result. In their work, the authors used a standard method—Gradient-weighted Class Activation Mapping or Grad-CAM—to understand what clinical features are most important for discriminating between patients who die in-hospital and those who do not [28]. Not surprisingly, the method finds that vital signs, interventions (e.g., mechanical ventilation), and administered medications were important for distinguishing between those who would have an in-hospital event and those who would not. Of interest, the model does suggest that simple nursing examinations, represented by Morse and Braden scores, may be important for predicting in-hospital mortality. Moreover, it is noteworthy that there are many different ways to organize data arising from the EMR into two-

dimensional representations and not all visual representations will have the same prognostic information. The authors of this study only experiment with three different ways to organize the data.

While these results are encouraging, the problem of predicting in-hospital mortality using 48 h of admission data may be, relatively speaking, not that difficult. For example, one would likely do fairly well predicting in-hospital mortality using a simplified set of input features that includes where the patient is admitted (ICU vs. hospital floor), vital signs trajectories during the first 48 h (higher death rates are expected in patients who become hypotensive soon after admission), and whether the patient requires mechanical ventilation or inotropic support soon after admission. As the authors do not compare their method to what would be obtained using a simple method such as logistic regression model using a rich set of clinical features, it is not clear whether a CNN is truly necessary for this task.

One very popular data source for machine learning is the electrocardiogram because it is routinely measured, cheap to administer, and apparently rich in information, some of which may not be easily discernable by humans. In addition, a variety of deep learning methods exist that can effectively deal with time series data, much like that arising from a single lead and multiple lead ECGs. Many of these approaches have already been applied to the interpretation and classification of electroencephalographic signals [50].

Attia et al. also mined the ECG for new information by attempting to predict left ventricular systolic dysfunction from the 12-lead ECG and transthoracic echocardiogram (TTE) using a convolutional neural network [51]. As LV dysfunction itself is a powerful predictor of subsequent heart failure, the resulting network indirectly identifies patients at elevated risk of adverse events [52]. By traditional statistical metrics (e.g. AUC) their classifier performed extremely well, with some exceptions (positive predictive value). The low positive predictive value (PPV) tells us that the model has many false positives, but, crucially, this does not help us predict *when* the model will fail; i.e., for which type of patients. The work also does not provide insights on the details of the relationship between the ECG and ALVD. For example, some determination about what segments of the ECG contribute to the prediction would be highly informative and of scientific interest.

Myers et al. applied a recurrent neural network (RNN)—a structure used to analyze time-series data—to continuous ECG data, along with a set of patient features, to predict the risk of death 1 year after non-ST segment elevation myocardial infarction (NSTEMI) [53]. For these studies, samples from the ST segments of each beat were identified and extracted in an automated fashion and then used as input to the RNN. The resulting neural network, which incorporates information from approximately 1 min of continuous ECG data, had improved predictive and discriminatory ability relative to a logistic regression model that used the same patient features and summary information from the admission 12-lead ECG. Nevertheless, the complexity of the model makes it difficult to understand precisely how and why the model arrives at a particular result. Consequently, while the model itself

has improved performance relative to existing methods, the ultimate clinical utility of the method remains to be determined.

## “All models are wrong, but some are useful”

The recent, notable successes of deep learning approaches argue that they will have place in the pantheon of methods used to build risk stratification models. However, it is not always clear when these approaches should be chosen over more standard methods such as logistic regression modeling or Cox proportional hazards regression.

In general, given a powerful set of independent patient characteristics, good results will be obtained regardless of the model used. In this instance, simple regression models may indeed be preferred over deep learning models as they can offer additional insights into the relative importance of the chosen features. The main advantage of deep learning, in terms of model performance, is that it can be used when one does not have good insight into what clinical features have the greatest predictive value. For example, the degree of ST segment elevation/depression is known to have prognostic significance in patients who present with symptoms consistent with an acute coronary syndrome. However, there are likely other ECG characteristics that are associated with adverse outcomes, so methods that use the entire ECG signal are likely to capture subtle ECG features that have prognostic significance. By design, deep learning models can utilize raw data, such as the raw ECG signal (i.e., samples) and therefore does not require insight into the type of features that may be important for a particular prediction task. Hence, there is a role for deep learning when one has incomplete knowledge of the important prognostic information that may be most relevant for quantifying patient risk. Nevertheless, it is important to keep in mind that the performance of deep learning models depends heavily on the size and quality of the data set which one uses.

Data availability is a significant barrier to the deployment of the type of models described in this review. For medical problems, data must be fully de-identified to be shared widely, but there is no useful standard for de-identification—given that perfection is impossible on large scale data, what percentage of missed protected health information (PHI) is permitted? Large scale projects like the Medical Information Mart for Intensive Care (MIMIC) have had success with automating de-identification, but they too expect errors and have an explicit protocol for reporting PHI in the data [54]. In general, there are few large, general medical datasets, and a significant fraction of studies (including many of those discussed here) use MIMIC or the UK-Biobank dataset. Those who do develop their own unique datasets often clutch that data close to their chests for a variety of reasons, not the least of which is the immense amount of work required to reliably de-identify that data.

We believe, from our own and others' clinical experience, that deep models will not be adopted into medical practice if they are not powerful, explainable, and capable of providing detailed failure mode information. None of the work discussed here hits all three marks, and the

wider literature is largely barren in this regard. Future work must emphasize these factors should it wish to impact patient outcomes in the clinic. Through this effort, perhaps we can design, if not “true,” then truly useful models.

## Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References and Recommended Reading

Papers of particular interest, published recently, have been highlighted as:

- Of importance
- Of major importance

1. Amsterdam EA, Wenger NK, Brindis RG, Casey DE, Ganiats TG, Holmes DR, et al. 2014 AHA/ACC Guideline for the management of patients with non-ST-elevation acute coronary syndromes: Executive summary. *Circulation*. Lippincott Williams and Wilkins; 2014. p. 2354–94.
2. Roffi M, Patrono C, Collet JP, Mueller C, Valgimigli M, Andreotti F, et al. 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST segment elevation: Task force for the management of acute coronary syndromes in patients presenting without persistent ST segment elevation of. *Eur. Heart J*. Oxford University Press; 2016. p. 267–315.
3. Levine GN, Bates ER, Blankenship JC, Bailey SR, Bittl JA, Cercek B, et al. 2015 ACC/AHA/SCAI focused update on primary percutaneous coronary intervention for patients with ST-elevation myocardial infarction: An update of the 2011 ACCF/AHA/SCAI guideline for percutaneous coronary intervention and the 2013 ACCF/AHA guideline for the. *Circulation*. Lippincott Williams and Wilkins. 2016;133:1135–47.
4. Myerburg RJ, Reddy V, Castellanos A. Indications for Implantable Cardioverter-Defibrillators Based on Evidence and Judgment. *J Am Coll Cardiol*. 2009;747–63.
5. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. *IBM J Res Dev*. 1959;3:211–29.
6. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol*. 2001;54:979–85.
7. Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. *Perspect Clin Res*. Medknow. 2011;2:145.
8. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic Regression. *Crit Care*. 2005;9:112–8.
9. Killip T, Kimball JT. Treatment of myocardial infarction in a coronary care unit: A Two year experience with 250 patients. *Am J Cardiol*. Excerpta Medica. 1967;20:457–64.
10. D’Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*. 2008;117:743–53.
11. Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, Van De Werf F, et al. A validated prediction model for all forms of acute coronary syndrome estimating the risk of 6-month postdischarge death in an international registry. *J Am Med Assoc*. 2004;291:2727–33.
12. Antman EM, Cohen M, Bernink PJLM, McCabe CH, Horacek T, Papuchis G, et al. The TIMI Risk Score for Unstable Angina/Non–ST Elevation MI. *JAMA*. American Medical Association. 2000;284:835.

13. Morrow DA, Antman EM, Parsons L, De Lemos JA, Cannon CP, Giugliano RP, et al. Application of the TIMI risk score for ST-elevation MI in the National Registry of Myocardial Infarction 3. *J Am Med Assoc. American Medical Association.* 2001;286:1356–9.
14. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB, et al. Predicting survival in heart failure: A risk score based on 39,372 patients from 30 studies. *Eur Heart J.* 2013;34:1404–13.
15. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, et al. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation.* 2006;113:1424–33.
16. Mozaffarian D, Anker SD, Anand I, Linker DT, Sullivan MD, Cleland JGF, et al. Prediction of mode of death in heart failure: the Seattle Heart Failure Model. *Circulation.* 2007;116:392–8.
17. Sze V, Chen Y-H, Yang T-J, Emer J. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc IEEE.* 2017;105:2295–329.
18. Wasserman PD, Schwartz T. Neural networks. II. What are they and why is everybody so interested in them now? *IEEE Expert.* 1988;3:10–5.
19. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern. Springer-Verlag.* 1980;36:193–202.
20. Dechter R. Learning While Searching in Constraint-Satisfaction-Problems. *Proc 5th Natl Conf Artif Intell.* 1986. p. 178–83.
21. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst.* 1989;2:303–14.
22. Hornik K. Approximation capabilities of multilayer feedforward networks. *Neural Networks.* 1991;4:251–7.
- 23.●● Myers P., Ng K, Severson K, Kartoun U, Dai W, Huang W, et al. Identifying Unreliable Predictions in Clinical Risk Models. *npj Digit Med.* 2020;3. This article presents a novel method for evaluating model reliability in its prediction on a specific patient, a key step towards bringing deep models into clinical practice.
24. Lipton ZC. The Mythos of Model Interpretability. *Commun ACM.* 2018;61:35–43.
25. Kalisch M, Bühlmann P. Causal structure learning and inference: A selective review. *Qual Technol Quant Manag. Chung Hua University.* 2014;11:3–21.
26. Shapely LS. A value for n-person games. *Contrib Theory Games.* 1953;2:307–17.
27. Lundberg SM, Allen PG, Lee S-I. A Unified Approach to Interpreting Model Predictions. *Adv Neural Inf Process Syst* 30. 2017.
28. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proc IEEE Int Conf Comput Vis. Institute of Electrical and Electronics Engineers Inc.;* 2017. p. 618–26.
29. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps. *Int Conf Learn Represent.* 2013.
- 30.● Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell. Springer Science and Business Media LLC;* 2019;1:206–15. In this perspective, Cynthia Rudin make a compelling argument regarding common approaches to making deep models, or “black boxes,” explainable. She suggests that those working in the field should focus on developing tools which are *inherently interpretable*, contrary to dominant trends in machine learning research.
31. Wang F, Kaushal R, Khullar D. Should Health Care Demand Interpretable Artificial Intelligence or Accept “Black Box” Medicine? *Ann Intern Med.* 2020;172:59.
32. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis. Springer New York LLC.* 2015;115:211–52.
33. Cruz-Roa AA, Arevalo Ovalle JE, Madabhushi A, González Osorio FA. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics).* 2013. p. 403–10.
34. Sirinukunwattana K, Raza SEA, Tsang YW, Snead DRJ, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Trans Med Imaging. Institute of Electrical and Electronics Engineers Inc.* 2016;35:1196–206.
35. Albarqouni S, Zurich E, Achilles F, Belagiannis V, Demirci S, Baur C, et al. AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Trans Med Imaging.* 2016;35:1321.
36. Sun W, Zheng B, Qian W. Computer aided lung cancer diagnosis with deep learning algorithms. In: Tourassi GD, Armato SG, editors. *Proc SPIE.* 2016. p. 97850Z.
37. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017;
38. Jo T, Nho K, Saykin AJ. Deep Learning in Alzheimer’s Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Front Aging Neurosci. Frontiers Media S.A.;* 2019;11.
39. Hubel DH, Wiesel T. Receptive fields of single neurones in the cat’s striate cortex. *J Physiol.* 1959;148:574–91.
40. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput. MIT Press - Journals.* 1989;1:541–51.
41. Tsay D, Patterson C. From Machine Learning to Artificial Intelligence Applications in Cardiac Care. *Circulation.* 2018;138:2569–75.
42. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med. Springer Nature.* 2018;1.

43. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*. Lippincott Williams and Wilkins. 2018;138:1623–35.
44. Dormer JD, Fei B, Halicek M, Ma L, Reilly CM, Schreibmann E. Heart chamber segmentation from CT using convolutional neural networks. *SPIE-Intl Soc Optical Eng*. 2018;100.
45. Avendi MR, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med Image Anal*. Elsevier B.V. 2016;30:108–19.
46. Oksuz I, Ruijsink B, Puyol-Antón E, Clough JR, Cruz G, Bustin A, et al. Automatic CNN-based detection of cardiac MR motion artifacts using k-space data augmentation and curriculum learning. *Med Image Anal*. Elsevier B.V. 2019;55:136–47.
- 47.●● Poplin R, Varadarajan A V., Blumer K, Liu Y, McConnell M V., Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. Nature Publishing Group; 2018;2:158–64. An interesting study that uses deep neural network models to predict cardiovascular outcomes from retinal images. Unlike many approaches, the authors also use methods to help understand why the model arrives at a given result.
48. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J*. 2003;24:987–1003.
49. Mayampurath A, Sanchez-Pinto LN, Carey KA, Venable L-R, Churpek M. Combining patient visual timelines with deep learning to predict mortality. Raza M, editor. *PLoS One*. 2019;14:e0220640.
50. Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: A review. *J Neural Eng*. Institute of Physics Publishing; 2019.
51. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat. Med*. Nature Publishing Group; 2019. p. 70–4.
52. Echouffo-Tcheugui JB, Erqou S, Butler J, Yancy CW, Fonarow GC. Assessing the Risk of Progression From Asymptomatic Left Ventricular Dysfunction to Overt Heart Failure: A Systematic Overview and Meta-Analysis. *JACC Hear Fail*. Elsevier Inc. 2016;4:237–48.
53. Myers PD, Scirica BM, Stultz CM. Machine Learning Improves Risk Stratification After Acute Coronary Syndrome. *Sci Rep*. Nature Publishing Group. 2017;7:12692.
54. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. Nature Publishing Groups; 2016. p. 3.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.