Curr Treat Options Cardio Med (2020) 22:15
DOI 10.1007/s11936-020-00814-0

State-of-the-Arts Informatics (C Stultz, Section Editor)

Deep Learning for Cardiovascular Risk Stratification

Daphne E. Schlesinger[1,2,3,4]
Collin M. Stultz, MD, PhD[1,2,3,4,5,6,*]

Address
[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, 02139, USA
[2]Institute for Medical Engineering and Science, MIT, Cambridge, MA, 02139, USA
[3]Research Laboratory of Electronics, MIT, Cambridge, MA, 02139, USA
[4]Computer Science & Artificial Intelligence Laboratory, MIT, Cambridge, MA, 02139, USA
[5]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139, USA
[*,6]Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA
Email: cmstultz@mit.edu

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 "programing 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, stoke, 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

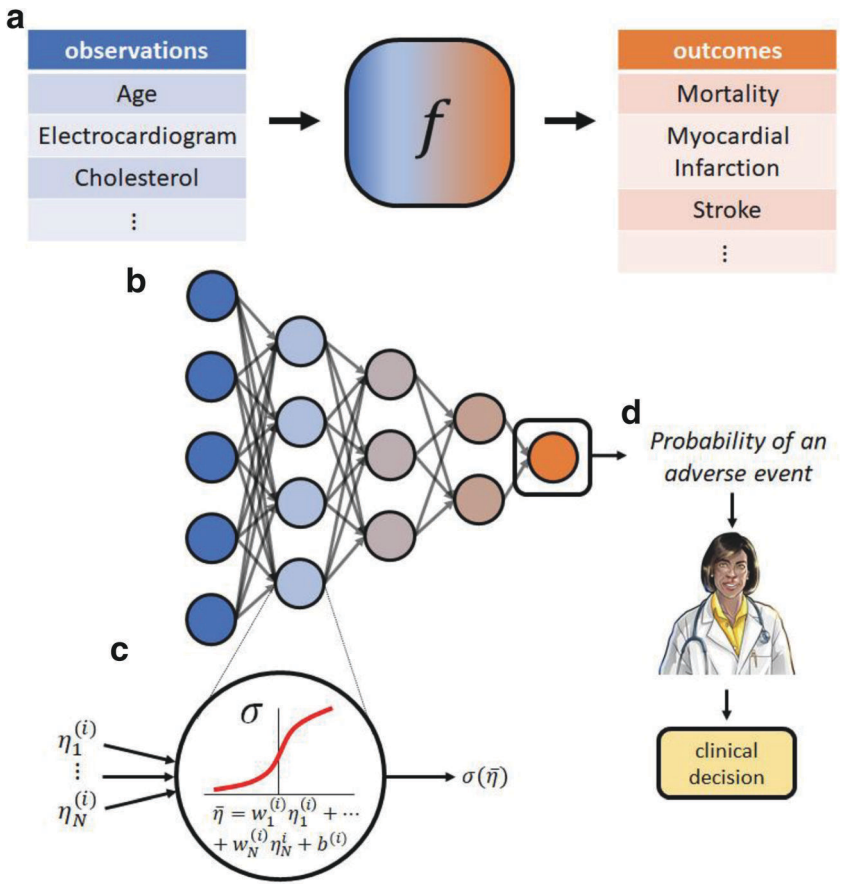

Fig. 1. In our applications, a neural network acts as a function that takes some observations as input and produces some prediction of outcomes as the output (**a**). This function is generated by adding many simple functions (represented by circular nodes that process information), each of which takes 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 ($\vec{w}^{(i)}, b^{(i)}$ for each node), which are chosen by training the network (**c**). Each layer can be though of 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;

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.