Machine learning in cardiovascular medicine: are we there yet?

Khader Shameer,[1,2,3,4,5,6] Kipp W Johnson,[2,3,4,5] Benjamin S Glicksberg,[2,3,4,5,7] Joel T Dudley,[2,3,4,5] Partho P Sengupta[8]

[1]Departments of Medical Informatics and Research Informatics, Northwell Health, Great Neck, New York, USA
[2]Institute for Next Generation Healthcare, Mount Sinai Health System, New York City, New York, USA
[3]Icahn Institute for Genomics and Multiscale Biology, Mount Sinai Health System, New York City, New York, USA
[4]Department of Genetics and Genomic Sciences, Mount Sinai Health System, New York City, New York, USA
[5]Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, New York City, New York, USA
[6]Center for Research Informatics and Innovation, Northwell Health, New Hyde Park, NY, USA
[7]Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA
[8]Division of Cardiology, West Virginia Heart and Vascular Institute, Morgantown, West Virginia, USA

Correspondence to
Dr Partho P Sengupta, Heart & Vascular Institute, West Virginia University, Morgantown, WV 26506-8059, USA; Partho.sengupta@wvumedicine.org

Received 10 July 2017
Revised 19 December 2017
Accepted 21 December 2017
Published Online First 19 January 2018

Linked

► http://dx.doi.org/10.1136/heartjnl-2018-313377
► http://dx.doi.org/10.1136/heartjnl-2018-313303

To cite: Shameer K, Johnson KW, Glicksberg BS, *et al. Heart* 2018;**104**:1156–1164.

ABSTRACT

Artificial intelligence (AI) broadly refers to analytical algorithms that iteratively learn from data, allowing computers to find hidden insights without being explicitly programmed where to look. These include a family of operations encompassing several terms like machine learning, cognitive learning, deep learning and reinforcement learning-based methods that can be used to integrate and interpret complex biomedical and healthcare data in scenarios where traditional statistical methods may not be able to perform. In this review article, we discuss the basics of machine learning algorithms and what potential data sources exist; evaluate the need for machine learning; and examine the potential limitations and challenges of implementing machine in the context of cardiovascular medicine. The most promising avenues for AI in medicine are the development of automated risk prediction algorithms which can be used to guide clinical care; use of unsupervised learning techniques to more precisely phenotype complex disease; and the implementation of reinforcement learning algorithms to intelligently augment healthcare providers. The utility of a machine learning-based predictive model will depend on factors including data heterogeneity, data depth, data breadth, nature of modelling task, choice of machine learning and feature selection algorithms, and orthogonal evidence. A critical understanding of the strength and limitations of various methods and tasks amenable to machine learning is vital. By leveraging the growing corpus of big data in medicine, we detail pathways by which machine learning may facilitate optimal development of patient-specific models for improving diagnoses, intervention and outcome in cardiovascular medicine.

INTRODUCTION

Artificial intelligence (AI) and machine learning are umbrella terms for a set of algorithms, which allow computers to uncover patterns and make decisions from data. After emerging from a quiescent period, the capabilities and potential of AI in a variety of tasks from automated 'digital assistants' to self-driving cars are a ubiquitous component of popular culture. However, despite several promising developments and some progress in oncology, cardiovascular medicine has not yet experienced a similar AI revolution. Routine care of cardiovascular patients accumulates large amounts of data in electronic health records (EHR). The integration of significant amount of diverse data is challenging in a busy clinical setting, resulting in marked underutilisation of information that could influence clinical decisions. Furthermore, much of the research that drives biomedical care comes from conventional hypothesis-driven research studies that often explore a handful of preselected variables and their impact on cardiovascular phenotypes. In contrast, AI-based methods can use a multitude of variables in a hypothesis-free approach to enable data-driven discovery which can identify similarities and differences in patient phenotypes, standardise clinical diagnosis, improve existing therapies, find new drug targets, optimise care pathway modelling and help us deliver data-driven, high-quality precision care at an increased rate.[1 2]

AI-based methods are increasingly being applied to cardiology to interpret complex data ranging from advanced imaging technologies, EHR, biobanks, clinical trials, wearables, clinical sensors, genomics and other molecular profiling techniques. Advances in high-performance computing and the increasing accessibility of machine learning algorithms capable of performing complex tasks (eg, deep learning and reinforcement learning) have heightened clinical interest in applying these techniques in research and clinical care. In this article, we summarise the current opportunities and pitfalls of the application of machine learning in cardiovascular medicine. There is a growing need for robust, scalable workflows and connecting patient communities, providers, payers and healthcare technology industry stakeholders for the seamless adoption of machine learning approaches. There is also an imminent need to develop guidelines for data acquisition, data sharing, security and privacy protection, and standardisation of machine learning towards implementation of precision medicine. There remain significant challenges in the application of machine learning to cardiovascular medicine, and improved care will only result from enhanced dialogue and teamwork between clinicians, biomedical informatics scientists and machine learning experts.

Big data in cardiology

Cardiovascular medicine generates a plethora of biomedical, clinical and operational data as part of patient care delivery. Often these data are stored in diverse data repositories which are not readily utilisable for cardiovascular research due to challenges in automated abstraction and manual curation technical competency. Despite these challenges, the application of machine learning technologies in cardiovascular medicine is not new (figure 1). Scientists have long used computers and early techniques drawn from AI to analyse and interpret cardiovascular phenotyping data, for example, automated analysis of ECGs and imaging systems.[3] However, renewed interest in AI is emerging due to the

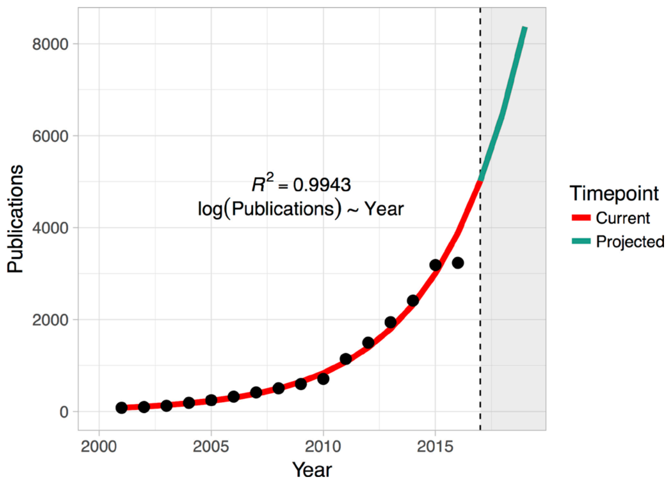

Figure 1 Projecting the growth of publications in PubMed with 'cardiology' and 'machine learning'. Data compiled using Medline (PubMed) trend (http://dan.corlan.net/medline-trend.html). Exponentiated regression of log number of publications on year is used to predict the future trend.

availability of a new generation of modern, scalable computing systems and algorithms capable of processing petabytes of data in real time. Furthermore, the high dimensionality of data increases analytical challenges and also simultaneously offers rich opportunities for improved discovery.[4] A few specific types of cardiovascular data sets with immediate needs for exploration with machine learning approaches are discussed in detail below.

Imaging and high-density phenotyping data

Cardiac imaging modalities like echocardiography, CT, MRI, single-photon emission CT, near-infrared spectroscopy, intravascular ultrasound and optical coherence tomography allow for visual assessment of structural changes. Since these structural changes are related to underlying disease aetiology and pathophysiological mechanisms, imaging is obviously widely used in cardiovascular medicine.[5] Imaging data sets are complex and stored in a variety of formats (JPEG, MPEG, DICOM, and so on) and are of varying dimension and scale (two dimensional, three dimensional, four dimensional, and so on), thereby representing the leading edge of cardiovascular big data. For example, in a recent study around 20% of Medicare patients underwent echocardiography, accounting for approximately 7.07 million echocardiography examinations. Given that a single echocardiography examination generates 2 GB data, this means about 14 petabytes of echocardiography results are collected annually. Traditional statistical methods cannot efficiently handle and learn from such elaborate data sets to develop diagnostic and predictive models for assisting clinical decision-making. Machine learning, on the other hand, could play an important role in automating cardiovascular imaging workflows through faster reading, interpretation and diagnosis.[6–9]

High-throughput molecular profiling data

Different omics technologies may be applied to perform deep profiling of particular molecular entities (eg, genes -> genomics, RNAs -> transcriptomics, proteins -> proteomics). The genetic basis of cardiovascular disease (CVD) is a long-standing and fruitful theme in research and discovery. Due to the influx of low-cost, affordable genomic testing technologies, genomics in particular is emerging in cardiology. For example, polygenic risk scores are currently available that can predict and provide 2, 5 or 10-year forecasts of cardiovascular events like myocardial infarction, heart failure and stroke. In addition to the discovery of at-risk patients, genomic results can also help customise therapy selection (eg, warfarin dosing). Since 2007 and 2010, several large genome-wide association studies and phenome-wide association studies, respectively, have demonstrated that CVD pathophysiology results from the complex interaction of many genes, non-coding regions and regulatory proteins.[10–12] Typically, a patient at risk is assessed for a potential cardiovascular event incidence (eg, heart attack, stroke, heart failure, and so on) using risk scores. Contrast these scores to more conventional scoring methods such as the Framingham Risk Score (FRS), a widely used risk score for predicting 10-year coronary heart disease risk based on age, gender and smoking status: notably, the FRS lacks any genomic component. However growing evidence from multiple studies suggests that incorporating genetic components and computing personalised genomic risk scores could play an important role in CVD risk estimation.[13] In addition to genomics, emerging molecular profiling technology including transcriptomics, proteomics, epigenomics, metabolomics and compiling longitudinal microbiome data could provide new insights into the pathophysiology of CVDs (see table 1 and figure 2). In the future, such pan-omics data sets could be generated from patient samples and become a routine part of cardiovascular care. While high-throughput technologies provide insight into a large number of genes, proteins and metabolites pertinent to CVD, the integration and interpretation of these data often requires a combination of statistical and machine learning approaches. For example, using high-throughput differential gene expression analyses from microarrays or RNA sequencing using statistical methods provides a list of genes that are altered in the patients but not controls. This may provide insight into disease processes, but it is unlikely to be directly clinically applicable. Using a machine learning framework, a signature-based predictive model could be built and used to develop a new biomarker-based diagnostic test. Molecular profiling and downstream analyses using statistical and machine learning approaches can also aid in identifying new target molecules for drug discovery and drug repositioning.[5 14] Emerging studies suggest that combining pan-omics technologies with imaging will be routine in the future to compile a stratified profile of patient population.[5] For example, recently, Li *et al* demonstrated that layering genomic data on EHR-based phenomics can be used to find and characterise patient subtypes (in this case, for type 2 diabetes mellitus).[15]

Data from clinical trials, population studies and disease registries

Current guidelines developed by various cardiovascular societies are implicitly based on an 'average patient'.[16] However, the complexity of human pathophysiology often cannot be adequately modelled by assuming patients conform to the same general rubric, and consequently, the guidelines may not be applicable for some patients. Machine learning tools can help develop standardised predictive models that could aid cardiologist with patient-specific guidelines. Such approaches can be considered a cognitive extension that augments rather than replaces physicians with AI for making patient-specific decisions. Importantly, accelerated and automated clinical decision-making would help providers to reclaim time and improve patient–provider interactions.

Table 1 High-dimensional and high-throughput molecular profiling experiments that require statistical and machine learning approaches for biological and clinical analyses and inferences

Biomolecular profiling technologies	Definition	Example in cardiovascular medicine	PMID
Epigenomics	Cataloguing the epigenetic modifications of genome	Emerging role of DNA methylation, histone density and post-translational modifications in the setting of cardiovascular diseases	25408699
Genomics	Assessing all genomic variations in an individual in the setting of a disease, response to therapy, characterising rare diseases or for molecular subtyping of common diseases	Genome-wide association studies of peripheral artery disease reveal ATXN2-SH2B3 locus.	25009551
Glycomics	Subtheme of metabolomics that aim to characterise the glycan molecules	Regulation of cardiac mitochondrial function using O-linked beta-N-acetylglucosamine modification (O-GlcNAcylation)	26446791
Immunomics	Flow cytometry-based phenotyping of immune and inflammation markers	Role of immune cell types including natural killer T (NKT) cells in promoting atherosclerosis	28127028
Lipidomics	Subtheme of metabolomics that aim to characterise the lipid molecules	Molecular lipid species (triacylglycerols, cholesterol esters and phosphatidylethanolamines) associated with cardiovascular diseases	24622385
Metabolomics	Study of the metabolites (targeted and untargeted) and their regulation due to diseases and other perturbations	Metabolite biomarkers improve the predictive value of Framingham Risk Score.	28255100
Metagenomics	Genomics of the environment	Metagenomic analyses using next-generation sequencing revealed microbial species in the setting of endocarditis.	24485222
Microbiome	Assess the microbial dysbiosis in the gut microbiome in the setting of a disease and treatment	Intestinal microbiota and metabolism of L-carnitine accelerates atherosclerosis in mouse.	23563705
Phenomics	Assessing phenotypes and impact of phenotypes due to perturbations at scale; where a phenotyping could be contextual (cellular, tissue, whole body or electronically derived)	A phenome-wide association study using variants influencing platelet traits is associated with myocardial infarction.	24026423
Proteomics	Estimating the list of proteins (targeted or untargeted) and their differential expression in the setting of a disease, therapy or other experimental conditions	Urinary proteome as a biomarker in coronary artery disease	20811296
Transcriptomics	Assessing the global gene expression signature in the setting of a particular disease or therapeutic intervention	Differential expression analyses of genes in the setting of high-dose statin therapy	27989886

PMID refers to the unique identifier number used in PubMed.

Need for machine learning in cardiovascular medicine

A primary difference between statistical methods and machine learning methods is that the former primarily help to understand relationships between a limited and a number of variables, but the latter contribute to identify and engineer features from the data and perform prediction. Machine learning methods thus complement and extend existing statistical methods—providing tools and algorithms to understand patterns from large, complex and heterogeneous data. Although classical statistical methods are capable of both discovery and prediction, machine learning methods are suitable and generalisable across a variety of data types and offer analyses and interpretation across complex variables.[17] Additionally, machine learning techniques typically rely on fewer assumptions and provide superior and more robust predictions. Several major types of machine learning algorithms are described below:

Supervised learning

Supervised learning is a machine learning approach where the investigator uses a database of observations with labelled outcomes or classes. These data are generally used to develop a model to predict or classify future events, or to find which variables are most relevant to the outcome. Examples of supervised learning algorithms include ordinary least squares regression,[18] logistic regression,[19] least absolute shrinkage and selection operator (LASSO) regression,[20] ridge regression,[21] elastic net regression,[21] linear discriminant analysis,[22] Naïve Bayes classifiers,[9] support vector machines,[23] Bayesian networks,[24] a variety of decision trees[25] especially Random Forests[26] and AdaBoost or gradient boosting classifiers,[27] artificial neural networks and ensemble methods.[7] Some of the examples of supervised machine learning tasks include regression, classification, predictive modelling and survival analysis. Supervised machine learning can help circumvent classification problems in phenotypically difficult patients (figure 3). For example, we recently described the use of supervised learning for differentiating athlete's heart and hypertrophic cardiomyopathy.[7] In another similar example, we developed a cognitive machine learning-based classifier to distinguish between constrictive pericarditis and restrictive cardiomyopathy.[8]

Unsupervised machine learning

Unsupervised learning is a machine learning approach suitable for a data set with no prior label or annotation available. Instead, the goal is to learn the relationship between variables and uncover hidden structure in a data set. Examples of unsupervised learning include clustering methods (hierarchical or K means),[28] principal component analysis, information maximising component analysis,[29] self-organising maps,[30] topological data analysis and deep learning. Specifically, deep learning is an emerging subdiscipline of machine learning that leverages an artificial neural network with many hidden layers of neurons. Deep learning algorithms can take a large number of features and derive neural network-based 'representations' that are capable of fast learning across a large number of samples.[31] Deep learning algorithms are particularly well suited for computer vision. This enables computers to perform reasoning and interpretation of imaging, often using convolutional neural network-based representation.[32] Currently, deep learning architectures are being applied to a variety of imaging-based calssifications.[33–35] Developing deep learning models is computationally expensive, and this field is rapidly developing due to the increasing availability of high-performance computing infrastructure via cloud computing and graphical processing unit-based model building.[31]

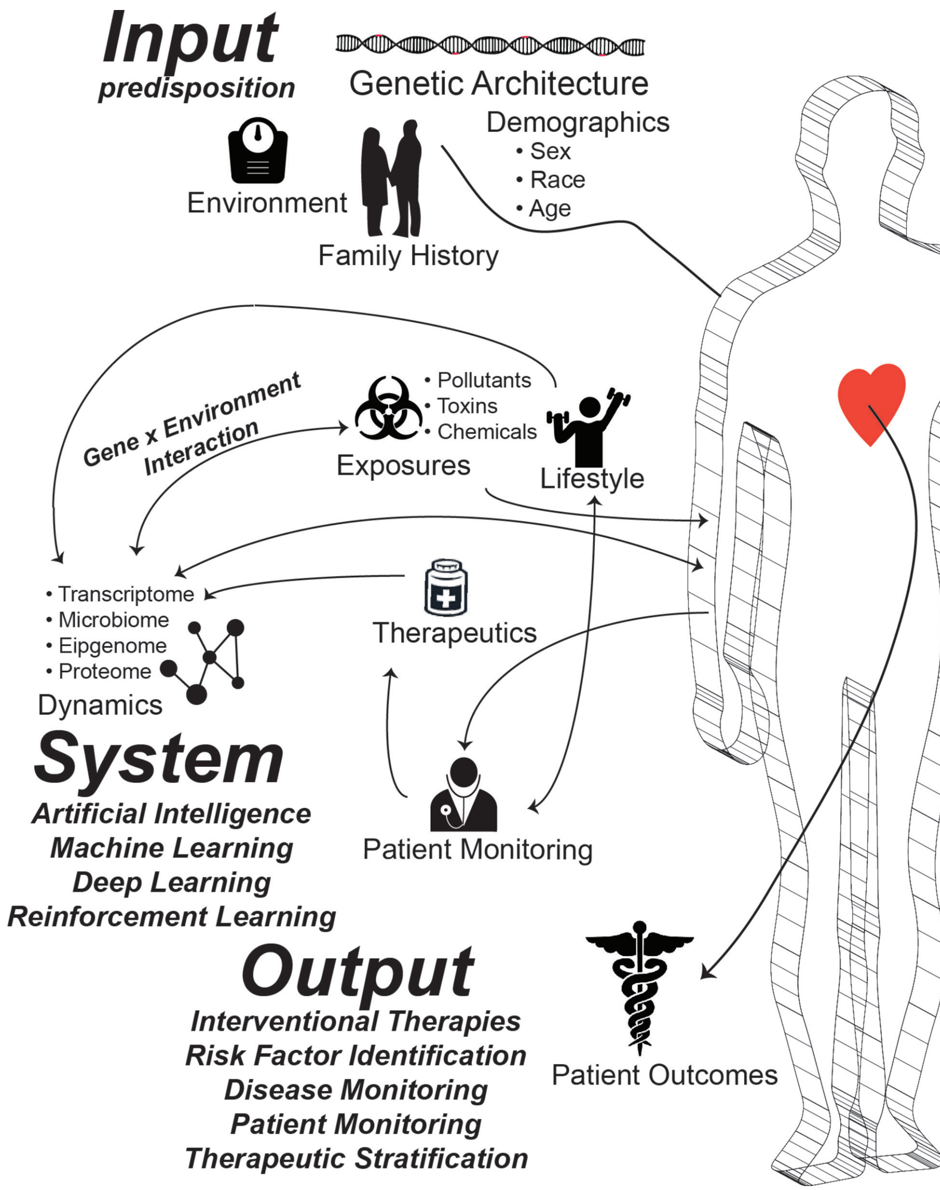

Figure 2 An infographic depicting various elements amenable to machine learning in cardiovascular medicine.

Reinforcement learning

Reinforcement learning is another emerging subdiscipline of machine learning based on behavioural psychology. In classical machine learning, a model is trained and tested. Future predictions are based on this static model. Variants of this approach exist such as real-time machine learning. However, reinforcement learning uses an alternate approach where a software agent acts in a prespecified environment to maximise a reward. The agent thus discovers the appropriate behaviour using some 'reward' criteria to handle the decision-making function (policy). Currently, reinforcement learning is being used for medical image analytics,[36] disease screening[37] and personalised prescription selection.[38] One particularly exciting example of reinforcement learning was then this method was used to select the optimal sequence of medications in non-small cell lung cancer.[39] Because many problems in clinical medicine can be formatted to fit the format of a reinforcement learning problem, we envision that this method will be used, for example, to select personalised blood pressure drugs and dosages, intelligently segment medical imaging data, or to interview patients and extract relevant diagnostic information.

Design and evaluation of a machine learning-based classifier

Figure 3 provides a generic outline of the various steps involved in the process of building predictive models. Supervised machine learning projects begin with formulating a predictive task to solve. Data can be aggregated prospectively with clinical trials or retrospectively from EHRs, clinical registries and other databases. After finalising the source databases, the data have to be checked, filtered and transformed depending on the data types and nature of the prediction task. Similar to a clinical trial, inclusion and exclusion criteria can be recommended and samples can be filtered out to ensure appropriateness, reproducibility and optimal performance. Typically, data sets are split into separate training and testing data sets. A predictive model is constructed using training data sets, and its performance is then evaluated on the held-out test data. Related but more sophisticated techniques for evaluating test/training splits such as K-fold or leave-one-out

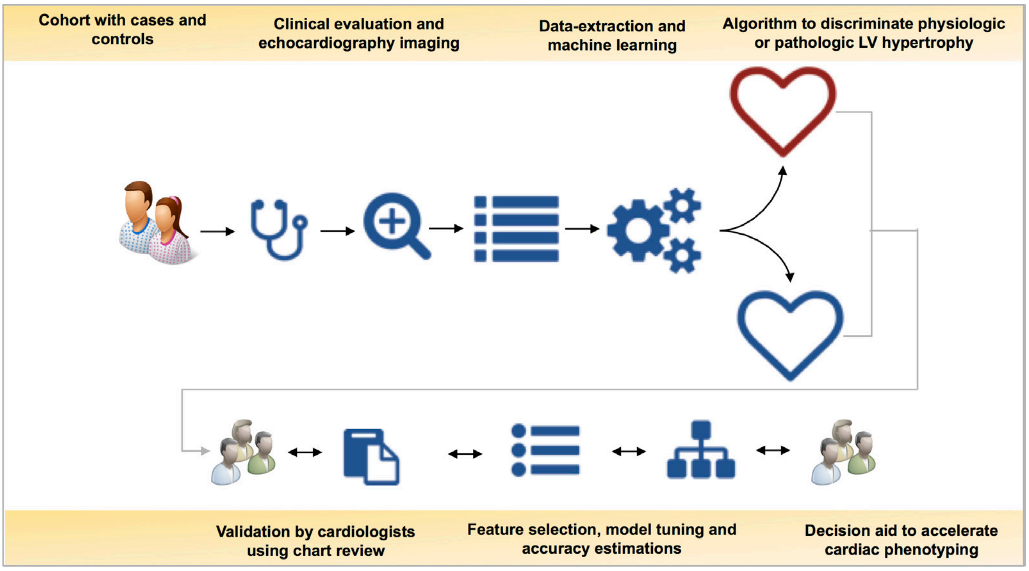

Figure 3 Cardiology informatics workflow used to develop a machine learning-based predictive model discriminates physiological function (athlete's heart) from pathological (hypertrophic cardiomyopathy) function. LV, left ventricular.

cross-validation can be used. Specific supervised learning performance statistics like Cohen's kappa statistics, Matthew's correlation coefficient, or F1 score augment traditional statistics like true positive rate, false positive rate, positive prediction value, negative prediction value, and so on (figure 3). In binary classification problems, overall model classification performance is often evaluated by computing the area under the receiver operating characteristic curve. Further refinement of a model can be achieved using the technique of feature selection. Feature selection consists of identifying the subset of predictor variables most relevant to a prediction task and removing the remaining features from the model. This helps to avoid overfitting a particular model to the training data set.

Delivering machine learning algorithms at the point of care

Predeveloped machine learning models can be stored in a computer to automatically perform precision phenotyping or patient acuity monitoring. The machine learning module may exist as standalone software; incorporated into a software suite for imaging, echo or other cardiovascular imaging modalities; as a cloud computing service provided by service providers such as Amazon Web Services, Microsoft Azure, Google Cloud Platform, or in a virtual private server-based system. Machine learning-based predictive models may also be embedded in EHRs or mobile device applications.

Challenges and opportunities

Need for orthogonal evidence from AI in cardiovascular medicine

Designing predictive and prescriptive models in cardiovascular medicine could help enable risk stratification and thus have significant implications in quality of healthcare delivery and impact on patient outcomes.[40] However, machine learning algorithms are not a panacea and also have limitations. One especially important challenge is the interplay between overfitting and underfitting the data sets. Hence, assessing bias-variance trade-off is a key step for predictive model validation. In short, the success of a machine learning project depends on the number of observations, number of features, selection and parameterisation of features and algorithm chosen for the model. Most importantly, an optimal predictive model can only be as good as the signal within the input data sets. This limitation can be addressed by performing orthogonal validation with independent data or different data modality to further validate the machine learning-based feature selection and predictive analytics.

Generalisability of machine learning models

In a recent, methodologically sound study, Frizzell *et al* showed that machine learning approaches performed poorly compared with traditional statistical methods in the context of congestive heart failure readmission prediction.[41] As the authors themselves conclude, results of their large-scale machine learning attempt to predict readmissions are largely not concordant with both earlier and emerging findings.[9 42] Here, we would like to highlight the conceptual limitations of readmission prediction and challenges with modelling approaches presented. First, there is a philosophical difference between predictive logistic regression models and machine learning-based modelling strategies. All modelling strategies begin with assumptions, and structural limitations and successful modellers must remain cognisant of these. Clinical context, local factors (including physician preferences, federal, regional and local care standards), selection of medications and other clinical decisions influence the final design of the model. Heart failure is a complex, heterogeneous disease. Readmission, layered on top of this, adds an enormous amount of complexity due to biological, clinical, socioeconomic and psychological interactions. One solution to avoid the impact of such bias is the generalisation of prediction task and building of different predictive models for various regions, cities or even individual hospitals. We took such an approach to model readmission at our hospital, obtaining a significantly higher C-statistic (0.78) than those shown by Frizzell *et al* (AUCs in the range of 0.6–0.7). There is no *a priori* reason to expect that a model built with data from, for example, a hospital in New York City should be able to predict readmission in Mobile, Alabama. Bias due to variation in hospital-to-hospital care pathways is not accounted in the model by Frizzell *et al*; we assume these rates would be different across various hospitals in the cohort as the authors also did not perform a feature selection or accuracy assessment of a reduced representation model.[43] It is reasonable to assume that the factors driving readmission in the different regions are different.[43] In general, missing information will be a challenge when trying to draw clinical inferences from a data set. However, for several reasons, machine learning is often superior to traditional statistical methods in the context of missing data. First, machine learning methods tend to make fewer assumptions about the underlying data, which means that missing data that might challenge the assumptions of statistical models are less of an issue. Such approaches are being particularly helpful in 'messy' problems such as EHR-based inference. Second, modelling missing data using imputation technique is the most common method to deal with missing data. Empirical evidence shows that machine learning methods can be used to define, interpret and handle missing data as part of the predictive modelling in biomedicine than traditional statistical methods.

Causality versus association in machine learning

One notable feature of machine learning algorithms is that highly predictive variables are not necessarily causative for the outcome of interest. Hence, mere automated inference using machine learning would not be an ideal application in the setting of cardiovascular medicine; instead, leveraging interpretable models and features may help in democratising machine learning in cardiovascular medicine. Consider the example of a drug which correlates with increased mortality—it may instead be the case that the drug is merely a marker for an underlying disease, which causes mortality, and not the causative agent itself. A classic example that highlights the causality versus association in the setting of cardiology is the association of high-density lipoprotein (HDL) cholesterol—higher levels—with low CVD risk, but drugs that increase HDL (such as the trapibs) do not appear to markedly reduce cardiovascular risk. In this context, HDL could be an indirect or surrogate molecular marker at least in some of the disease subtypes, which is associated with coronary artery disease, but it does not participate in causing or alleviating it. To model the causality versus association problem, some solutions for causal inference have been developed such as Mendelian randomisation (MR) or statistical analysis based on directed acyclic graphs. MR is an emerging technique, which relies on the fact that genetic material is randomly inherited. A particular risk allele (eg, a single-nucleotide polymorphism (SNP) that changes low-density lipoprotein (LDL) cholesterol levels) can act as an instrumental variable for the exposure studies. As the prior knowledge suggests the SNP is causally affecting LDL levels, and the SNP is randomly inherited, if LDL levels are associated with the outcome we can assume a causal relationship. MR has been used in a number of cardiovascular studies to causally implicate factors like alcohol intake[44] and LDL cholesterol.[45] MR has recently been used to demonstrate that a biomarker associated with CVD (cystatin C) was not causally related.[46] Teasing such complex causal and association effects is crucial to developing personalised therapies in cardiology. New methods are now being designed to integrate MR with machine learning for systematic and automated causal inference using large-scale population and phenomic data sets.[47]

Standardisation of machine learning using a minimum information model in the setting of biomedicine

Summarising evidence from multiple studies and meta-analyses are vital for developing guidelines and policy implementations in public health and medicine. However, to achieve such concordance using machine learning, a new policy on the dissemination and reporting of machine learning models in medicine is required. Publications should report minimum sets of data elements: for example, all model parameters should be provided, and the manuscript should also communicate variables using a standard ontology (see table 2). Any data transformations applied to the data, sampling methods and random number generator seeds should be disclosed. Further, all data, version numbers of software and associated code should be released in a public domain software archive system to enable replication. Journals and researchers could adopt such a minimum information model for

Table 2 Open access biomedical and healthcare ontologies and big data resources in cardiovascular medicine for developing machine learning resources

Resource	Description	URL
Cardiac Electrophysiology Ontology (EP)	Ontology describing electrophysiological experiments in cardiology (232 terms)	https://bioportal.bioontology.org/ontologies/EP
Cardiovascular Disease Ontology (CVDO)	An ontology to describe entities related to cardiovascular diseases	https://github.com/OpenLHS/CVDO
Cardiovascular Gene Ontology (CVGO)	Cardiovascular disease genes and their annotations (4000 human proteins)	http://www.ebi.ac.uk/GOA/CVI
ClinicalTrials.gov	Database of public and private clinical trials with 3165 cardiology studies	https://clinicaltrials.gov/ct2/results?term=cardiology
Congenital Heart Defects Ontology	Ontology describing congenital heart defects data	https://bioportal.bioontology.org/ontologies/CHD
database of Genotypes and Phenotypes (dbGaP)	Compendium of genome-phenome data sets (155 540 variables, 6935 analyses, 2892 documents and 1878 data sets in 167 studies)	https://www.ncbi.nlm.nih.gov/gap/?term=cardiovascular
Electrocardiography Ontology (ECG)	Ontology describing ECGs, their capture method(s) and their waveforms	http://www.cvrgrid.org/
Genome-Wide Association Studies (GWAS) Catalog	430 studies, 3214 associations and 252 catalogue traits	https://www.ebi.ac.uk/gwas/search?query=cardiovascular
Heart Failure Ontology (HFO)	Ontology describing heart failure-relevant information including the causes and risk factors, signs and symptoms, diagnostic tests and results, and treatment	https://bioportal.bioontology.org/ontologies/HFO
NHLBI GO Exome Sequencing Project (ESP)	Genes and variants contributing to heart, lung and blood disorders (6503 samples)	http://evs.gs.washington.edu/EVS/
Ontology of Cardiovascular Drug Adverse Events (OCVDAE)	Ontology of cardiovascular disease drug-associated adverse events (2948 terms)	https://bioportal.bioontology.org/ontologies/OCVDAE
Phenome-Wide Association Studies (PheWAS) Catalog	Catalogue of phenome-wide associations from multiple studies	https://phewascatalog.org/

machine learning and mandate use in the cardiovascular medicine community. Clinician investigators and data scientists need to come together for defining data sharing and usage policy. Compiling data for various disease phenotypes and therapeutic stratifications amenable to machine learning from healthcare providers and health systems is crucial to improve the generalisability and predictive provenance of clinical algorithms for learning and prediction. Such efforts would help reduce ambiguity and promote the adoption of machine learning as a standard mode of data interpretation in the cardiology community.

Engaging cardiovascular medicine stakeholders in machine learning

Patients regularly pose questions that may be difficult to answer without data-driven analytics. For example, 'Are there many patients similar to me treated here?', 'What was the direct outcome of my treatment?', 'Are there any alternative treatment options?', or 'Are there other medications to reduce my symptom burden?'. Providers are keen to improve their overall quality of patient care. Payers are interested in providing optimal care to a patient with minimal investment, thus ensuring maximum return on investment. The pharmaceutical industry is engaged in development and delivery of therapies that suit a patient. As healthcare is driven by these four related, yet diverse stakeholders, engaging them to need an all-encompassing strategy satisfies the multitude of requirements. Thus, tools to identify population-level risk and infer the individual risk of a patient are key. We envisage that machine learning-based, hyperlocal analytics would allow a primary care physician or cardiologist to answer these questions with clarity. Collectively, engaging patients, providers, payers and pharma to leverage modern analytical approaches as part of routine clinical care may improve outcomes and overall survival. To encourage the use of machine learning-based methods, medical students should be trained in the application of both statistical and machine learning in medicine and healthcare (table 3). The physician of the future must be both a clinician and a data scientist. Indeed, the ability to apply advanced data science concepts to clinical care may ultimately prove to be a decisive differentiator between physicians and mid-level providers. Initiating machine learning contests and challenging students, residents and trainees, as part of major cardiovascular conferences, would further improve the adoption of machine learning in cardiovascular medicine.[48]

Table 3 Differentiating data-driven, machine learning approaches compared with statistical approaches

Modalities	Medicine	Data-driven medicine
Analytical methods	Statistical methods	Machine learning or deep learning
Analytics strategy	Hypothesis driven	Data driven
Data complexity	Low dimensional	High dimensional
Data model	Stable data model	Evolving, flat data model
Data size	Megabytes or gigabytes	Terabytes, petabytes, exabytes
Data source	Centralised	Decentralised
Data source	Structured	Structured, semistructured or unstructured
Data storage	Excel or MySQL	MySQL, NoSQL or Graph databases
Data types	Traditional data types	Evolving data types
Example method	Student's t-test	Random Forests
Example software	Excel or SAS	R, Python
Sample size	Cohort size (~10 K)	Large cohort size (>100 K)

Limitations of machine learning in cardiovascular medicine

Statistical methods have been applied to biomedical problems for decades, from old and ubiquitous inventions such as survival analysis with Cox's proportional hazard regression, to the design of LASSO regression for the analysis of genotyping data. In general, methods like these aim to identify, quantify and interpret the relationship between variables. In abstract, this is necessarily the same goal as most machine learning and AI methods when applied to cardiology. In our view, there is not a sharp distinction between statistics and machine learning; but instead, the respective sets of methods exist on a continuum defined primarily by motivations for their creation and application. LASSO, in particular, is often interpreted as a machine learning method, although it is an extension of regression techniques. For example, parametric statistical methods assume distributions for each variable and then attempt to model the probability of a given set of observations. In both supervised and unsupervised machine learning, we are often interested in solving the same class of problem. Thus, classical statistics methods like logistic regression are essentially forms of 'machine learning'. However, machine learning has emerged because in many cases more classical methods are either inefficient or less accurate in real-world situations. As we discussed in the previous example, logistic regression relies on assumptions that are often unrealistic, requiring variables to have log odds linearly related to an outcome, continuous decision boundaries in classification problems, independence of observations and errors, and so on. In contrast, many machine learning methods require far fewer assumptions of the given data: methods like Random Forests have virtually no statistical assumptions of the data; we only aim to apply an algorithm and 'see how it works.' This approach is particularly useful because, in reality, no complex problem is linear—Random Forests and more sophisticated AI methods like deep learning can learn highly complex relationships from the data, which often results in far better performance in prediction, clustering and other tasks.

Similar to statistical and mathematical methods, careful design and evaluation of machine learning models is an important aspect. Various factors including data quality, data integration strategy, choice of machine learning algorithm, validation methods, orthogonal evidence, the clinical and biological relevance of the machine learning model and comparative effectiveness to the standard of care affect a machine learning model in the setting of cardiovascular medicine.[2] Choice of methods used for data and integrating data from various database systems and normalising to a single data model is critical. Assigning appropriate case–control labels and validating manual or automated abstraction algorithms are also crucial. Without extensive evaluation, models are often overfit to training data, and hence following a structured machine learning workflow (see figure 4) defined and proposed by the cardiovascular medicine community is a need of the hour. In this context, we want to emphasise the relevance of Dr Box's quote on statistical models, 'All models are wrong, but some are useful.'[49]

CONCLUSIONS

In this review article, we reviewed the basics of machine learning algorithms and what potential data sources exist in medicine; examined the need for AI in medicine; and evaluated the potential limitations and challenges of implementing AI in medicine. With the increasing availability of big data sourced from different avenues of biomedicine and healthcare delivery and the progression towards precision medicine approaches, the application of

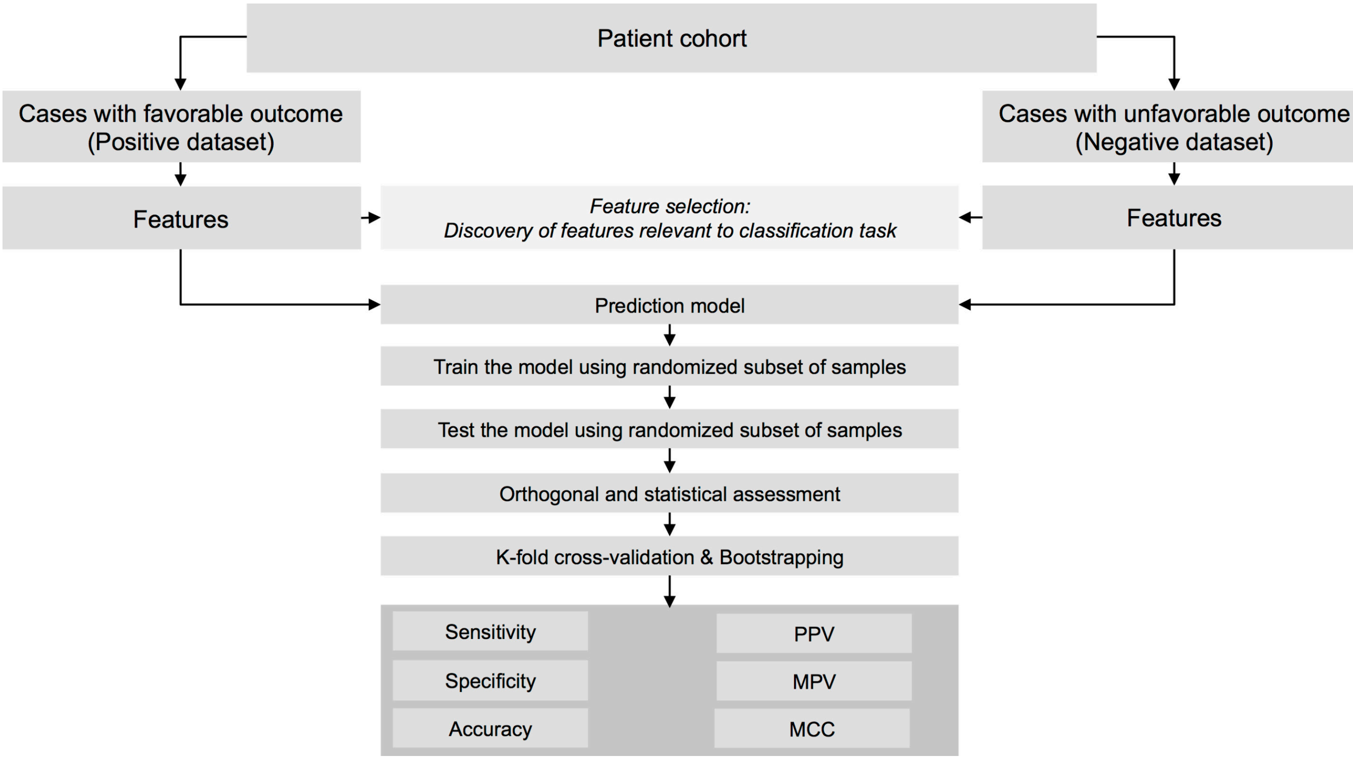

Figure 4 Architecture of a machine learning model to predict a binary outcome. MCC, Mathew's Correlation Coefficient; NPV, Negative Predictive Value; PPV, Positive Predictive Value.

AI for analyses and interpretation of medical data will continue to grow. CVDs represent groups of diseases that can benefit greatly from pre-emptive prediction, prevention and proactive management, and thus AI methodologies. Various types of AI algorithms will be essential for understanding the nuanced individual risk factors, behavioural drivers and therapeutic pathways predictive of disease outcomes in specific patient cohorts and also for instituting early therapeutic interventions. The application of machine learning algorithms in prospective clinical trials would allow comparison with current standard of care practices with a goal of implementing precision diagnostics, risk stratification and personalised therapeutics. Various machine learning methods like supervised learning, cognitive learning and unsupervised learning methods including deep learning could uncover hidden structure in big data in cardiology and can help subtype chronic complex diseases common in cardiovascular medicine; such insights may also lead to novel therapeutic discoveries and help improve delivery of personalised cardiovascular care. While there exist compelling debates in medicine about whether AI will replace doctors in the near future, we believe it will instead primarily augment and extend the current practice of healthcare delivery, thus improving the lives of patients, providers and society at large.

Contributors PPS and JTD planned the outline of the article. KS, KWJ and BSG conducted the empirical analyses and literature survey, and compiled the figures. All authors contributed to the manuscript writing and editing.

Funding KS, KWJ, BSG and JTD acknowledge support from the National Institutes of Health: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK, grant number R01-DK098242-03); National Cancer Institute (NCI, grant number U54-CA189201-02); Illuminating the Druggable Genome (IDG); Knowledge Management Center sponsored by NIH Common Fund; National Center for Advancing Translational Sciences (NCATS, grant number UL1TR000067); and Clinical and Translational Science Awards (CTSA) grant.

Competing interests PPS is a consultant for TeleHealthRobotics, HeartSciences and Hitachi-Aloka. JTD has received consulting fees or honoraria from Janssen Pharmaceuticals, GlaxoSmithKline, AstraZeneca and Hoffman-La Roche. JTD is a scientific advisor to LAM Therapeutics and holds equity in NuMedii, Ayasdi and Ontomics. KS has received consulting fees or honoraria from Philips Healthcare, Google, LEK Consulting and Parthenon-EY. All other authors declare no competing interests.

Provenance and peer review Commissioned; externally peer reviewed.

REFERENCES

1 Johnson KW, Shameer K, Glicksberg BS, *et al*. Enabling precision cardiology through multiscale biology and systems medicine. *JACC Basic Transl Sci* 2017;2:311–27.
2 Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N Engl J Med* 2017;376:2507–9.
3 Tajik AJ. Machine learning for echocardiographic imaging: embarking on another incredible journey. *J Am Coll Cardiol* 2016;68:2296–8.
4 Nagueh SF. Unleashing the potential of machine-based learning for the diagnosis of cardiac diseases. *Circ Cardiovasc Imaging* 2016;9:e005059.
5 Kini AS, Vengrenyuk Y, Shameer K, *et al*. Intracoronary imaging, cholesterol efflux, and transcriptomes after intensive statin treatment: the YELLOW II study. *J Am Coll Cardiol* 2017;69:628–40.
6 Slomka PJ, Dey D, Sitek A, *et al*. Cardiac imaging: working towards fully-automated machine analysis & interpretation. *Expert Rev Med Devices* 2017;14:197–212.
7 Narula S, Shameer K, Salem Omar AM, *et al*. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. *J Am Coll Cardiol* 2016;68:2287–95.
8 Sengupta PP, Huang YM, Bansal M, *et al*. Cognitive machine-learning algorithm for cardiac imaging: A pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circ Cardiovasc Imaging* 2016;9.
9 Shameer K, Johnson KW, Yahi A, *et al*. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: A case-study using mount sinai heart failure cohort. *Pac Symp Biocomput* 2016;22:276–87.
10 Shameer K, Tripathi LP, Kalari KR, *et al*. Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Brief Bioinform* 2016;17:841–62.

11 Deloukas P, Kanoni S, Willenborg C, *et al*. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013;45:25–33.
12 Shameer K, Denny JC, Ding K, *et al*. A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* 2014;133:95–109.
13 Kullo IJ, Jouni H, Austin EE, *et al*. Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels. *Circulation* 2016;133:1181–8.
14 Shameer K, Glicksberg BS, Hodos R, *et al*. Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. *Brief Bioinform* 2017.
15 Li L, Cheng WY, Glicksberg BS, *et al*. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* 2015;7:311ra174.
16 Douglas PS, Garcia MJ, Haines DE, *et al*. ACCF/ASE/AHA/ASNC/HFSA/HRS/SCAI/SCCM/SCCT/SCMR 2011 Appropriate Use Criteria for Echocardiography. A Report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, American Society of Echocardiography, American Heart Association, American Society of Nuclear Cardiology, Heart Failure Society of America, Heart Rhythm Society, Society for Cardiovascular Angiography and Interventions, Society of Critical Care Medicine, Society of Cardiovascular Computed Tomography, Society for Cardiovascular Magnetic Resonance American College of Chest Physicians. *J Am Soc Echocardiogr* 2011;24:229–67.
17 Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;349:255–60.
18 O'Connor AM, Wray J, Tomlinson RS, *et al*. Impact of surgical complexity on health-related quality of life in congenital heart disease surgical survivors. *J Am Heart Assoc* 2016;5:e001234.
19 Damen JA, Hooft L, Schuit E, *et al*. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
20 van Loo HM, van den Heuvel ER, Schoevers RA, *et al*. Sex dependent risk factors for mortality after myocardial infarction: individual patient data meta-analysis. *BMC Med* 2014;12:242.
21 Ojeda FM, Müller C, Börnigen D, *et al*. Comparison of cox model methods in a low-dimensional setting with few events. *Genomics Proteomics Bioinformatics* 2016;14:235–43.
22 Alonso-Betanzos A, Bolón-Canedo V, Heyndrickx GR, *et al*. Exploring guidelines for classification of major heart failure subtypes by using machine learning. *Clin Med Insights Cardiol* 2015;9:57–71.
23 Shameer K, Pugalenthi G, Kandaswamy KK, *et al*. Insights into protein sequence and structure-derived features mediating 3d domain swapping mechanism using support vector machine based approach. *Bioinform Biol Insights* 2010;4:BBI.S4464–42.
24 Tylman W, Waszyrowski T, Napieralski A, *et al*. Real-time prediction of acute cardiovascular events using hardware-implemented Bayesian networks. *Comput Biol Med* 2016;69:245–53.
25 Ruiz-Fernández D, Monsalve Torra A, Soriano-Payá A, *et al*. Aid decision algorithms to estimate the risk in congenital heart surgery. *Comput Methods Programs Biomed* 2016;126:118–27.
26 Shameer K, Pugalenthi G, Kandaswamy KK, *et al*. 3dswap-pred: prediction of 3D domain swapping from protein sequence using Random Forest approach. *Protein Pept Lett* 2011;18:1010–20.
27 Xiong G, Kola D, Heo R, *et al*. Myocardial perfusion analysis in cardiac computed tomography angiographic images at rest. *Med Image Anal* 2015;24:77–89.
28 Almeida VG, Borba J, Pereira HC, *et al*. Cardiovascular risk analysis by means of pulse morphology and clustering methodologies. *Comput Methods Programs Biomed* 2014;117:257–66.
29 Zhang X, Ambale-Venkatesh B, Bluemke DA, *et al*. Information maximizing component analysis of left ventricular remodeling due to myocardial infarction. *J Transl Med* 2015;13:343.
30 Lagerholm M, Peterson C, Braccini G, *et al*. Clustering ECG complexes using hermite functions and self-organizing maps. *IEEE Trans Biomed Eng* 2000;47:838–48.
31 Narula S, Shameer K, Salem Omar AM, *et al*. Reply: deep learning with unsupervised feature in echocardiographic imaging. *J Am Coll Cardiol* 2017;69:2101–2.
32 Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017;19:221–48.
33 Gao X, Li W, Loomes M, *et al*. A fused deep learning architecture for viewpoint classification of echocardiography. *Information Fusion* 2017;36:103–13.
34 *A combined multi-scale deep learning and random forests approach for direct left ventricular volumes estimation in 3D echocardiography*: 2016 Computing in Cardiology Conference (CinC); 2016 11-14 Sept, 2016.
35 *A left ventricular segmentation method on 3D echocardiography using deep learning and snake*: 2016 Computing in Cardiology Conference (CinC); 2016 11-14 Sept, 2016.
36 *A reinforcement learning framework for medical image segmentation*: The 2006 IEEE International Joint Conference on Neural Network Proceedings, 2006.
37 Lee EK, Kurokawa YK, Tu R, *et al*. Machine learning plus optical flow: a simple and sensitive method to detect cardioactive drugs. *Sci Rep* 2015;5:11817.
38 *Post cardiac surgery recovery process with reinforcement learning*: 2015 19th International Conference on System Theory, Control and Computing (ICSTCC); 2015 14-16 Oct, 2015.
39 Zhao Y, Zeng D, Socinski MA, *et al*. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* 2011;67:1422–33.
40 Goldstein BA, Navar AM, Pencina MJ, *et al*. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198–208.
41 Frizzell JD, Liang L, Schulte PJ, *et al*. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol* 2017;2:204–9.
42 Mortazavi BJ, Downing NS, Bucholz EM, *et al*. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes* 2016;9:629–40.
43 Janecek A, Gansterer W, Demel MA, *et al*. *On the relationship between feature selection and classification accuracy*. Workshop and Conference Proceedings: Journal of Machine Learning Research, 2008;4:90–105.
44 Holmes MV, Dale CE, Zuccolo L, *et al*. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ* 2014;349:g4164.
45 Holmes MV, Asselbergs FW, Palmer TM, *et al*. Mendelian randomization of blood lipids for coronary heart disease. *Eur Heart J* 2015;36:539–50.
46 van der Laan SW, Fall T, Soumaré A, *et al*. Cystatin C and cardiovascular disease: a mendelian randomization study. *J Am Coll Cardiol* 2016;68:934–45.
47 Hemani G, Bowden J, Haycock PC, *et al*. Automating mendelian randomization through machine learning to construct a putative causal map of the human phenome. *bioRxiv* 2017.
48 Shameer K, Badgeley MA, Miotto R, *et al*. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Brief Bioinform* 2017;18:105–24.
49 George EPB. *Science and statistics*. Journal of the American Statistical Association, 1976;71:791–9.