

## Perspective

## Advancing cardiovascular medicine with machine learning: Progress, potential, and perspective

Joshua P. Barrios<sup>1,4</sup> and Geoffrey H. Tison<sup>1,2,3,\*</sup><sup>1</sup>Department of Medicine, Division of Cardiology, University of California, San Francisco, 555 Mission Bay Blvd South Box 3120, San Francisco, CA 94158, USA<sup>2</sup>Bakar Computational Health Sciences Institute, University of California, San Francisco, 555 Mission Bay Blvd South Box 3120, San Francisco, CA 94158, USA<sup>3</sup>Twitter: @GeoffTison<sup>4</sup>Twitter: @BarriosEtAl\*Correspondence: [geoff.tison@ucsf.edu](mailto:geoff.tison@ucsf.edu)<https://doi.org/10.1016/j.xcrm.2022.100869>

## SUMMARY

Recent advances in machine learning (ML) have made it possible to analyze high-dimensional and complex data—such as free text, images, waveforms, videos, and sound—in an automated manner by successfully learning complex associations within these data. Cardiovascular medicine is particularly well poised to take advantage of these ML advances, due to the widespread digitization of medical data and the large number of diagnostic tests used to evaluate cardiovascular disease. Various ML approaches have successfully been applied to cardiovascular tests and diseases to automate interpretation, accurately perform measurements, and, in some cases, predict novel diagnoses from less invasive tests, effectively expanding the utility of more widely accessible diagnostic tests. Here, we present examples of some impactful advances in cardiovascular medicine using ML across a variety of modalities, with a focus on deep learning applications.

## CARDIOVASCULAR MEDICINE IS RIPE FOR MACHINE LEARNING

Across many industries, there has been significant interest in machine learning (ML) in recent years, and biomedical science and medicine are no exception.<sup>1,2</sup> Broadly speaking, ML describes a variety of computational techniques that have expanded the traditional analytical toolkit beyond linear statistical models and generally comprise what is often referred to as “artificial intelligence” (AI). ML algorithms have made it possible to analyze higher dimensional data, and they have shown an ability to learn complex associations within these data without human-provided instructions. To do this, ML algorithms learn from many examples of data that have been labeled for a specific task, a process referred to as “training” or “learning.” This typically requires not only a large amount and variety of data but also diagnostic tests that capture information that are incrementally closer to the diagnostic ground-truth, to provide labels for the data. For several important reasons, cardiovascular medicine is ripe for advancement by ML.

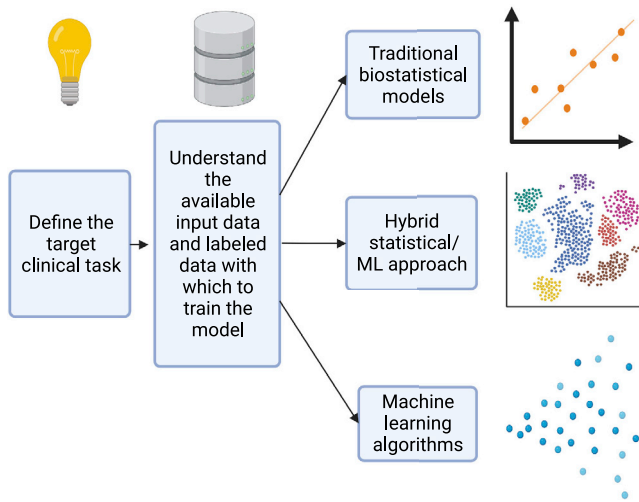
Modern medical practice is awash in data of numerous types. In cardiovascular medicine, recent decades have seen an expansion in the variety and quality of diagnostic tests, such as noninvasive imaging like computed tomography (CT) angiography, physiologic testing such a fractional flow reserve or serum biomarkers. These tests provide physicians with greater amounts of complementary information upon which to make diagnostic and therapeutic decisions but present a broad range

of accessibility, cost, and risk. This hierarchy of diagnostic tests is larger in cardiovascular medicine than other areas, with some tests providing rich labels upon which ML algorithms can be trained to use less costly/invasive tests to predict the results of more costly or invasive tests.

The widespread digitization of medical data over the past two decades makes medicine highly ripe for ML. With the move to electronic health record systems,<sup>3</sup> nearly all presently recorded medical data from imaging to labs are stored and accessed in digital formats. The ability to access most modern medical data electronically unlocks the potential of data-intensive algorithms for application in biology and medicine. Even so, a substantial amount of data processing and harmonization may still be required before these data can be used to train ML algorithms. Much of the data collected in medical practice may now be digitized but remains unstructured. For data formats that are less easily computable, such as free-text notes, recent ML-driven advances in fields such as natural language processing<sup>4–6</sup> can make it possible to either obtain structure from such unstructured data or to work with unstructured data formats directly.

The high prevalence of cardiovascular disease overall also translates to large amounts of cardiovascular-relevant patient data. Electrocardiograms (ECGs), for example, are frequently obtained many times over the course of a patient’s care, often by non-cardiovascular specialists or clinics, which reflects the relevance of cardiovascular concerns to medicine more broadly. This provides greater amounts of data to train ML models and offers opportunities for ML to assist in a greater number of clinical workflows.





**Figure 1. Understand the problem and the data to select the most appropriate analytic technique**

Clearly defining the clinical task of interest and understanding the nature of the available data and training labels is essential to selecting the most appropriate analytic technique. Created with [BioRender.com](https://app.biorender.com/illustrations/62e46a207b029dd42f0dcfb5). Link to BioRender illustration: <https://app.biorender.com/illustrations/62e46a207b029dd42f0dcfb5>.

Like many chronic diseases, cardiovascular disease diagnosis and management depend heavily on trends across data, providing another opportunity for assistance from ML algorithms. For diagnostic tests such as ECGs or echocardiograms (echo), understanding a given patient's *prior* study is critical to appropriately interpreting the current study. While physicians commonly compare a current study against the most recent prior study or several prior studies to identify changes, humans are generally less adept at evaluating subtle changes that arise over longer periods of time. Often it is simply not practical for a cardiologist to review more than a handful of prior studies prior to interpreting every current study. ML algorithms are positioned to assist with this in various ways, such as by providing quantification of previously subjectively determined binary diagnoses or by learning patterns of change in raw data that are associated with disease outcomes. ML can help to process large amounts of longitudinal data and discover patterns that may otherwise go unnoticed. Importantly, however, ML algorithms will not be the most appropriate analytical option for every clinical task. There are many tasks for which traditional statistical models or rule-based algorithms provide the best solution. Depending on the task of interest and the nature and quantity of the available input data and ground-truth labels, researchers should select the most appropriate analytic technique: traditional biostatistical models, ML algorithms, or a combination (Figure 1).

We will review some fundamentals of ML algorithms, then discuss some examples of specific applications to cardiovascular medicine.

## FUNDAMENTALS OF ML AND AI

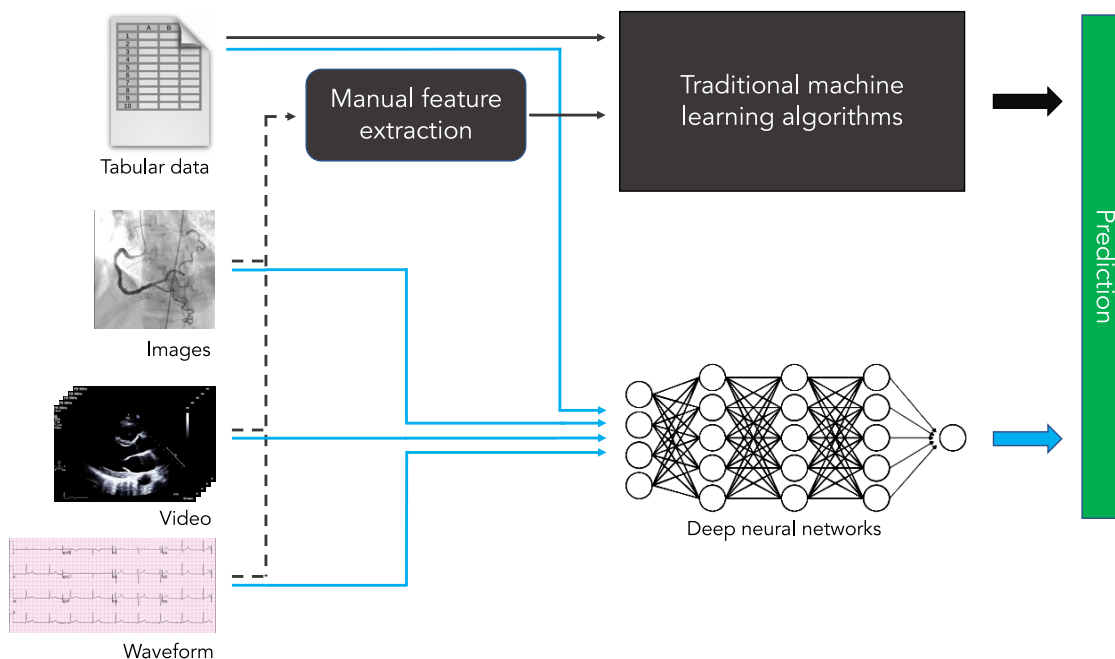
ML refers to a collection of computational techniques that learn patterns within data to accomplish certain tasks for which they

are trained. The most common approach used to train ML algorithms in most settings, including in medicine, is called supervised learning. Training an algorithm with supervised learning involves presenting to the algorithm many examples of input data that have been labeled (often by a human expert) for the task of interest. The algorithm makes predictions each time, and learns patterns within the data by minimizing error when comparing its predictions against the labeled data. This error is calculated mathematically using a customizable equation called a loss function. The parameters of the algorithm are adjusted on each round to minimize this loss, causing the algorithm to learn patterns in the data. For example, in order to distinguish a normal sinus rhythm ECG from an atrial fibrillation ECG, an ML algorithm would learn from previously categorized raw 12-lead ECG waveforms what ECG features differentiate atrial fibrillation from normal sinus rhythm. These features might be irregular R-R intervals or absence of a P-wave or other similar ECG features, but the features identified and the relative importance of each feature (and the interactions between features) are learned by the algorithm during the training process.

To examine the ability of the algorithm to generalize to unseen data, the algorithm is continually tested during training on a separate subset of the data called the development, or validation, dataset. The algorithm's parameters are typically adjusted by investigators based on the development dataset results to maximize generalization performance and prevent overfitting. Overfitting describes the situation in which the model learns the training dataset perfectly, memorizing training dataset-specific features at the expense of features that can allow the model to perform well on unseen data. Final algorithm performance is then reported on a third held-out dataset, the test dataset, that had no role in the training of the algorithm. In medical applications, it is also increasingly important to report performance in one or more additional test datasets from external institutions. This provides additional evaluation of the robustness of the algorithm, helping to ensure that the algorithm performance is not dependent on specific characteristics of the training population or dataset.

Typical tasks of medical ML models include classification (e.g., to predict "disease" versus "no disease") or regression (e.g., to predict the continuous value of a diagnostic test) although many other relevant tasks can be performed such as drawing a bounding box around an item of interest in an image (e.g., localizing a pulmonary nodule on an X-ray). Based on the task, the relevant metrics of performance may vary. Relevant metrics can include accuracy, specificity, and sensitivity, much like any other diagnostic test, or the F1 score, which is the harmonic mean of precision (positive predictive value) and recall (sensitivity). ML reports will often also include the area under the receiver operating characteristic curve (AUROC), which is a metric of discrimination performance across all possible thresholds of its output score.

Traditional ML algorithms, such as support vector machines or tree-based models, are best suited for structured data resembling numbers on a spreadsheet, also called "tabular data." For example, these data might contain a series of lab values, vital signs, or measurements derived from more complex data modalities, such as cardiac chamber dimensions from a cardiac



**Figure 2. Example data types for traditional ML versus DNN algorithms**

Traditional ML algorithms such as regression models, tree-based methods (e.g., random forest or gradient boosted models), or support vector machines typically require tabular input data formats. These tabular data can be extracted from structured data such as electronic health records, or they can be derived from more complex data types through manual extraction of human-defined features such as manual measurement of chamber volumes from an echo. DNNs allow for more high-dimensional and complex input data types.

MRI. In the case of derived measurements, humans are usually required to interpret the raw data before they can be input into the ML model in a process called feature engineering (Figure 2). In general, these traditional ML algorithms can capture complex relationships and have shown promise in medical applications. However, although these algorithms can process higher dimensional input data compared with traditional statistical models, they are still limited in the types of data that they can reasonably accept. For high-dimensional data modalities such as images or videos, these traditional ML algorithms are not able to handle raw data formats and require humans to first derive summary measurements, which are subject to bias in the selection of which measurements to derive and human error during measurement.

Perhaps the most medically relevant ML advances in recent years have been driven by algorithm architectures that can accept raw data modalities such as free text, images, and videos directly as inputs, led primarily by a category of algorithms called deep neural networks (DNNs).<sup>1</sup> DNNs are highly flexible and customizable algorithms composed of many layers of model parameters or weights that can be added or removed depending on the complexity of the input data and the nature of the task of interest. Augmented by additional data-transforming filters (such as convolutional filters), which help further distill the data, the multi-layered architecture of a DNN is thus able to learn very complex relationships within the input data to accomplish the target task, without requiring (or being limited by) pre-specified human-defined features (Figure 2). This combination of a DNN's ability to learn complex features and the ability to accept high-dimensional, even multi-modal, raw input data makes it possible to

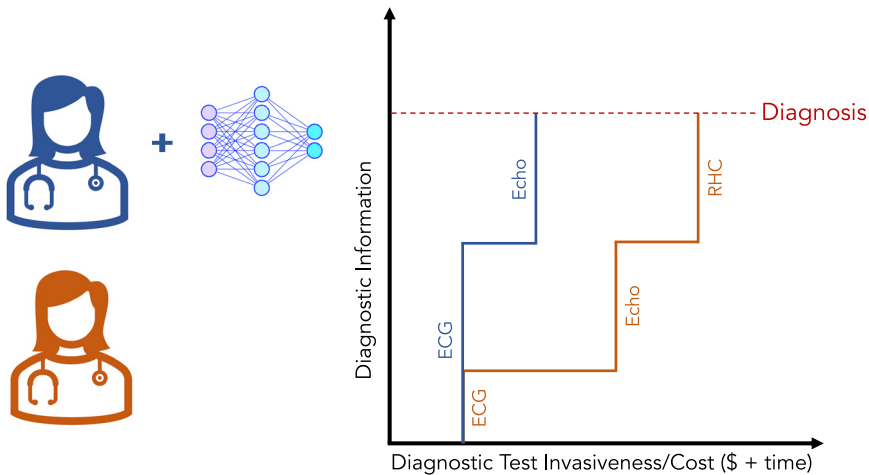
accomplish novel tasks and potentially enable data-driven discovery of new physiologic associations. These advances fundamentally changed the types of applications in which ML could provide value to medicine, making it possible to algorithmically analyze complex, raw medical data without requiring physician review.

### RECENT ADVANCES AND APPLICATIONS OF ML IN CARDIOVASCULAR MEDICINE

ML algorithms have been applied to data modalities across the spectrum of the cardiovascular workflow. Various groups have trained DNNs to support fully automated analysis of ECGs, echos, radiologic studies, angiograms, and others. In addition to performing standard medical tasks with a specific type of data, ML algorithms have demonstrated the ability to perform novel tasks not previously associated with specific diagnostic tests, effectively expanding the utility of existing tools (Figure 3). Here, we will review some recent advances in ML applications in cardiovascular medicine.

#### Electrocardiography

The ECG is the most common diagnostic test in cardiology, providing widely accessible information about cardiac electrical function and structure. While traditional ML algorithms have long been applied to derived measurements from the ECG, more recently DNNs have demonstrated the capability to analyze raw ECG waveforms to perform a variety of tasks. Automated ML algorithms have replicated standard ECG diagnoses,



**Figure 3. Using ML to extract more diagnostic information at lower cost**

Ideally, a physician aided by ML algorithms can obtain more information per test, at a lower cost. This would increase the sensitivity of less invasive and more accessible diagnostic tests, possibly decreasing the need for more specific testing or time to diagnosis. The diagnostic tests shown are representative examples of cardiovascular tests that can provide incrementally more information for certain diagnoses. The threshold for diagnosis (red line) will vary based upon clinical disease or application. ECG, electrocardiogram; RHC, right heart catheterization.

performed novel tasks not typically performed by cardiologists using ECG alone, and assisted in analyzing longitudinal trends in disease status and response to drug therapy.

Most commercial ECG algorithms apply what are called rule-based algorithms to analyze an ECG; for example, to detect atrial fibrillation or to measure the QT interval. These rule-based models rely on previously defined diagnostic criteria or measurements typically used by human readers and have enabled the automated ECG diagnosis algorithms that have assisted clinicians for decades. ML algorithms provide complementary strengths to rule-based algorithms since they do not depend on previously defined criteria, tend to improve performance as available training dataset sizes increase, and are capable of learning patterns for themselves from the raw data, some of which may have been previously unknown to humans.

Hannun et al.<sup>7</sup> was among the first to train an end-to-end DNN using raw ECG waveforms to detect 12 arrhythmia diagnoses. They demonstrated performance similar to cardiologists for set of 12 arrhythmia diagnoses, providing proof of concept that DNNs can be trained to analyze raw ECG data. The average F1 score was similar to or higher than the average F1 score of a committee of cardiologists. When matching specificity to the specificity of the cardiologists, the model showed similar or higher sensitivity for all 12 arrhythmia diagnoses.

Attia et al.<sup>8</sup> demonstrated that a clinically relevant novel task could be performed through ML analysis of an ECG: prediction of asymptomatic left ventricular (LV) dysfunction. Using paired ECGs and echoes, they trained a DNN to identify LV ejection fraction  $\leq 35\%$  from standard 12-lead ECG data. Others have similarly used DNN-based approaches to accomplish a variety of novel tasks using ECGs. Examples include detection of hypertrophic cardiomyopathy (HCM),<sup>9,10</sup> pulmonary hypertension,<sup>9,11</sup> amyloid,<sup>9,12</sup> mitral valve prolapse,<sup>9</sup> mitral and aortic regurgitation,<sup>13</sup> aortic valve stenosis,<sup>13,14</sup> hyperkalemia,<sup>15</sup> and mortality risk estimation.<sup>16,17</sup>

Recent work by Hughes et al.<sup>18</sup> trained a DNN to detect 38 different ECG diagnoses across a range of diagnostic classes, including arrhythmia, conduction, infarct, and chamber enlargement. In addition to comparing the DNN against a

committee of electrophysiologists, they also compared against automated diagnoses from the commercially available GE MUSE system. The DNN outperformed the rule-based MUSE diagnoses for all classes except supraventricular tachycardia. Compared with clinical cardiologist reads, the DNN demonstrated comparable or better performance on 34 out of 38 classes. Human readers outperformed the MUSE system for 30 out of 38 classes. Hughes et al. also applied a technique, broadly called AI explainability, that helps illuminate what portions of the ECG waveform the DNN learned as being the most important for each diagnosis it makes. In addition to illuminating well-understood ECG regions associated with diagnoses, such as the delta-wave for Wolff-Parkinson-White syndrome, such techniques provide the possibility to discover new physiologic associations in a data-driven manner using ML. Such techniques may be even more important when ML algorithms successfully make novel diagnoses based on a given data modality, providing an opportunity for us to learn from the algorithms what patterns in the data—and their physiologic underpinnings—are consistently associated with the novel diagnosis.

More recently, DNN analysis of ECGs has been used to track longitudinal disease status and response to drug therapy, potentially providing a new paradigm for longitudinal monitoring. In a cross-institutional/industry collaboration, Tison et al.<sup>19</sup> deployed two DNNs separately trained to detect HCM to clinical trial patients receiving a new drug for HCM, who also underwent serial ECG, echo, and lab measurements. Both DNNs' predictions of HCM risk not only decreased over time in HCM patients receiving drug therapy but also tracked decreases in cardiac hemodynamics (measured by serial echos) and serum lab values, providing the first demonstration of ECGs not only capturing this type of information but also tracking changes in these measurements over time.

Other demonstrations further illustrate the use of ML to make predictions beyond what can typically be done by human experts using standard ECG data. Tison et al.<sup>9</sup> demonstrated that ECG data alone can predict continuous cardiac structural and functional metrics as quantified by echoes such as LV mass (in  $\text{g}/\text{m}^2$ ), LA volume (in  $\text{mL}/\text{m}^2$ ), and mitral annulus ( $e'$ ). While ECG-based rules exist for clinicians to predict some of these metrics in a binary yes/no manner from ECGs, the possibility of

estimating the continuous value of a measure such as LV mass in  $\text{gm/m}^2$  from an ECG alone demonstrates the broader range of possibilities ML can offer. Ulloa Cerna et al. developed an ML model incorporating an ECG DNN score to predict valvular disease, reduced EF, and increased interventricular septal thickness.<sup>20</sup> Attia et al.<sup>21</sup> trained a DNN to identify patients with atrial fibrillation (Afib) using normal sinus rhythm ECG, and Raghunath et al. trained a DNN to predict new-onset Afib within 1 year after the ECG in patients with no prior history of Afib.<sup>22</sup> Together, these results suggest that sinus ECGs contain signals to predict Afib that are not currently appreciated by human experts. Kwon et al. developed a DNN to detect anemia from the ECG,<sup>23</sup> whereas Bos et al.<sup>24</sup> trained a DNN to diagnose concealed long QT syndrome (LQTS), a genetically defined disorder in which 40% of patients present with a normal QT at rest. The DNN was not only able to identify patients with this concealed LQTS but it was also able to distinguish between the three main genotypic subgroups using ECG alone.

In all cases, the AI/ML algorithms exhibit some degree of error. Therefore, prior to clinical deployment of ML algorithms, the next critical step in evaluation must be a real-world prospective assessment of the clinical benefit of deploying the algorithm and what degree of error is acceptable, which is specific to each disease and clinical context. Properly designed randomized controlled clinical trials (RCTs) provide the highest-level of evidence in this regard, and RCTs examining ML algorithm deployment will certainly be forthcoming in the near term. One important example is the EAGLE trial, one of the first RCTs to evaluate DNN analysis of 12-lead ECGs to predict low ejection fraction.<sup>25</sup> In the EAGLE trial, 120 primary care teams were cluster randomized to access to the DNN's analysis, and the primary outcome was new diagnosis of low ejection fraction ( $\leq 50\%$ ) within 90 days. The intervention arm exhibited increased diagnosis of low ejection fraction, suggesting that application of the DNN algorithm for ECG analysis can increase the earlier diagnosis of low EF in a primary care setting.

### Echo

Echo is a central diagnostic test in cardiology, providing invaluable information about cardiac anatomy and real-time hemodynamics that are essential for the diagnosis and management of most cardiac diseases. Although it is noninvasive, does not use ionizing radiation, and has lower cost, echo does require greater specialized training to both obtain and interpret compared with other tests such as the ECG. ML algorithms can provide value across the echo workflow, from acquisition to pre-processing, anatomic measurement, and interpretation. As reviewed below, the bulk of AI applications in echos have focused on one or more of three main goals: image pre-processing, automating echo measurements, and automating disease diagnosis.

Echo studies consist of multiple two-dimensional videos of the heart from various standard views. The first task a cardiologist in training must learn is to recognize each of the standard echo views, before progressing to learn how to interpret each view. In a similar way, since echo views are typically not currently recorded in echo video metadata, the first step in automated AI echo analysis is view classification. Early work showed some

success using traditional ML approaches relying on features hand-derived from echo images<sup>26–29</sup>; however, in the past 5 years, DNNs have achieved reliable echo view classification directly using the raw data. Gao et al.<sup>30</sup> was among the first to apply DNNs in this manner, showing that DNNs could be trained to analyze raw image frames directly and outperforming hand-engineered ML approaches. Subsequent work expanded the number of views that could be identified,<sup>31</sup> while Zhang et al.<sup>32</sup> additionally averaged a DNN's predictions across multiple video frames to achieve high accuracy in classifying 23 distinct views.

Once echo views have been identified, the next step in the clinical echo work flow is performing standard measurements to quantify size and volume of cardiac structures. This is usually performed manually by either the sonographer or the interpreting cardiologist. AI-based automation may increase accuracy, reproducibility, and speed of performing these measurements. Various efforts have developed ML-based approaches to replicate specific echo measurements, such as LV size and volume,<sup>33–35</sup> or leaflet dimensions.<sup>36</sup> Zhang et al.<sup>32</sup> was the first to employ a specific DNN architecture called a U-Net to segment multiple cardiac chambers simultaneously, similar to an echocardiographer tracing cardiac chambers, and then use these to calculate standard measurements of various chambers such as volume, mass, ejection fraction and longitudinal strain. In contrast, Ghorbani et al.<sup>37</sup> showed that a DNN could be trained to directly predict chamber volumes, such as end-systolic LV volume or end-diastolic LV volume, and even ejection fraction from raw images without first segmenting the chamber. In subsequent work, Ouyang et al.<sup>38</sup> trained a video-based DNN to perform segmentation of the LV and estimation of ejection fraction, and showed that this could be done across multiple cardiac cycles captured during the echo video. The more cardiac cycles examined, the lower the error in the DNN-estimated ejection fraction. When measurements are performed manually, only several frames (often from a single cardiac cycle) are chosen and measured by the sonographer on account of the time required to perform the manual measurements. While this is the standard of care, it does raise the possibility that the frames selected may be not representative or that measurement error may more greatly influence the final assessment. The work by Ouyang et al. thus demonstrates an important manner in which ML can be used to increase reproducibility and decrease measurement error within the clinical workflow, since there is minimal to no incremental cost in deploying the AI algorithms across a large number of cardiac cycles or frames. In a similar manner, strategically trained AI algorithms could decrease variability and increase reproducibility at various other points in the echo work flow from echo acquisition, measurement of other echo parameters, and disease interpretation.

ML algorithms have been trained to detect various cardiovascular diseases from raw echo images and from echo-derived parameters. Examples include pulmonary arterial hypertension,<sup>32</sup> HCM,<sup>32,39</sup> cardiac amyloidosis,<sup>12,32</sup> coronary artery disease,<sup>40,41</sup> LV hypertrophy,<sup>42</sup> regional wall motion abnormalities,<sup>43</sup> and pediatric congenital heart disease.<sup>44</sup> Several studies have shown the ability to detect heart failure defined by low ejection fraction,<sup>32,37,38</sup> as well as heart failure with preserved ejection fraction,<sup>45–48</sup> and to differentiate stress cardiomyopathy from acute



myocardial infarction.<sup>49</sup> Ulloa Cerna et al. trained an algorithm that demonstrated the ability to predict all-cause mortality from echo videos, a task not typically performed by cardiologists interpreting echos.<sup>50</sup> Similarly, Hughes et al. developed a DNN capable of estimating lab values such as brain natriuretic peptide (BNP), troponin I, blood urea nitrogen (BUN), and others from the echo video alone, a task clearly outside the normal range of diagnostic ability of the echo.<sup>51</sup> Clearly, this is far from the comprehensive set of echo diagnoses that would be needed to support fully automated echo interpretation. These studies, however, provide a valuable proof of concept that ML algorithms can be trained across a range of clinically relevant diagnoses. Future work will be needed to continue to expand the list of diagnoses for which accurate reproducible algorithms exist.

ML algorithms have also been applied to assist with the acquisition of echo images that is typically performed by a specially trained echo sonographer. Becoming a proficient echo sonographer is a years-long process. ML algorithms could assist less experienced operators to obtain basic echo images. Narang et al.<sup>52</sup> conducted a prospective study with such a DNN-based system to test how well it could guide untrained operators to obtain echo images. In this study, the ML system provided guidance to nurses who had not previously performed echo regarding how to acquire and improve specific echo views in real time. They demonstrated that the system could guide the novice sonographers to obtain diagnostic-quality echo images. This shows how ML algorithms can be deployed to extend the ability to obtain echos to those without prior experience, potentially expanding the accessibility of this specialized test. Future integration of DNNs into the clinical work flow in similar ways may offer clinicians and health systems an opportunity to reconsider what can be achieved with standard medical tests and who can help accomplish it.

### Other data types in cardiovascular medicine

Other modalities in cardiovascular medicine also provide fertile ground for future ML efforts. ML may offer the greatest potential for data types that have traditionally been more difficult to analyze in an automated manner, such as diagnostic tests recorded in video formats such as coronary angiography, high-dimensional genetic data, and free-text notes from electronic health records.

Coronary angiography is the definitive diagnostic test for the evaluation of coronary artery disease, providing the basis for recommending both medical and procedural therapy (including percutaneous coronary intervention [also known as stents], and coronary bypass surgery). Coronary angiograms are acquired and stored as video data consisting of fluoroscopic (X-ray) clips obtained during the injection of radiocontrast dye into the various coronary arteries for visualization of the coronary artery lumen and potential stenoses. Prior research has suggested that there is substantial variability and possibility for operator bias in the process of manual interpretation of angiograms.<sup>53–55</sup> Until recently, the video-based nature of raw angiogram data had made it difficult to develop automated analysis approaches. Recent work from Avram et al.<sup>56</sup> and Zhou et al.<sup>57</sup> demonstrated that DNNs can be trained to accurately perform various steps required to analyze coronary angiograms, such

as viewpoint classification or coronary artery localization, providing the basis for automated angiogram interpretation. Avram et al.<sup>56</sup> deployed multiple DNN's together in a pipeline to achieve accurate automated estimation of coronary artery stenosis severity. Future AI-driven progress with coronary angiography may eventually lead to assisted or automated interpretation, which could help to reduce human error and increase the reproducibility of measurements, and possibly even expand the range of diagnoses that are possible to make through angiography, similar to what has been achieved with ML analysis of ECG and echo.

Several deep learning (DL) applications have been developed for cardiovascular CT applications. Recent ML work has expanded the information that can be derived from various types of CT scans; for example, enabling automatic estimation of coronary artery calcium score from CT angiography<sup>58</sup> or low-dose chest CT scans,<sup>59</sup> or quantifying epicardial and thoracic adipose tissue from non-contrast CT scans.<sup>60</sup> DL algorithms have accurately estimated coronary artery stenosis<sup>61</sup> and fractional flow reserve<sup>62</sup> using coronary CT angiography. These efforts demonstrate the potential for DL algorithms to broaden the scope of current clinical tools such as CT and potentially reduce the need for more invasive procedures such as coronary angiography.

Cardiac magnetic resonance imaging (CMR) provides high-resolution images of high diagnostic value, but is relatively expensive, slow, and labor intensive. Many DL applications in the CMR domain focus on alleviating these pain points by automating manual processes and improving efficiency. Schlemper et al.<sup>63</sup> developed a DL-based 3D reconstruction pipeline that can interpolate slices with long inter-slice distance, increasing the speed of acquisition and providing fast real-time reconstruction. Blanser et al.<sup>64</sup> present a DL system for automated prescription of imaging planes by using a U-Net to localize anatomic landmarks. A series of recent contributions have demonstrated automated segmentation of anatomical structures in CMR images,<sup>65–67</sup> with future similar disease-specific work sure to follow. The automated cardiac diagnosis challenge (ACDC) dataset is a publicly available dataset of 150 CMR studies with manual segmentation labels as well as five diagnostic labels. The availability of these dataset has enabled multiple efforts that have produced DL-based methods for automated diagnosis of these five diagnostic categories: normal, HF with infarction, dilated cardiomyopathy, HCM, and abnormal right ventricle.<sup>68–71</sup> This highlights the importance of labeled public datasets for the rapid development of algorithms.

ML algorithms have been developed for other modalities of large-scale data relevant for cardiovascular medicine, including genetic data and free-text data from the electronic health record. Free-text data present a substantial analytic challenge due to the complexity of natural language. The analytical field known as natural language processing (NLP) has developed over decades, which aims to accurately analyze and process free-text data. In the past 5–10 years, DNN-related architectures have revolutionized the NLP field by providing exponential progress compared with prior techniques, mirroring advances made in the image processing and computer vision fields. For example, the DNN architecture class known as the transformer is widely applied

in NLP due to its ability to learn very long-range associations across large bodies of text.<sup>4–6</sup> As AI-based text-analysis capabilities continue to advance outside of medicine, there will be significant opportunity to extrapolate these advances to better analyze free-text medical data, such as from the electronic health record, diagnostic test reports, and possibly patient-reported symptoms and outcomes. Moon et al.<sup>72</sup> recently demonstrated the extraction of sudden cardiac death risk factors from free-text notes, and Blecker et al. identified heart failure patients at large scale from the electronic health record.<sup>73</sup> As the field matures, medical implementations of transformers<sup>6</sup> and future such developments in free-text analysis will undoubtedly be applied to advance cardiovascular applications.

Analysis of genetic data has long benefitted from the robust bioinformatics analysis community and library of techniques that have driven the countless advances in modern genetics. Cardiovascular genetics has recently exhibited increasing examples of ML-driven analysis, including for the investigation of coronary artery calcium,<sup>74</sup> pulmonary hypertension,<sup>75</sup> and multiple clinically relevant variant detection<sup>76,77</sup> from next-generation sequencing or proteomic data. The greatest potential for AI in genetic analysis may lie in the analysis of polygenic disorders, since algorithms such as DNNs have enormous capacity to examine high-level and very complex interactions within the genetic data, if provided adequate quantities of training data.

### THE CRITICAL NEED FOR PROSPECTIVE CLINICAL EVALUATION OF AI IN CARDIOLOGY

As with any diagnostic tool, development and validation of AI algorithms are only the first step toward clinical application. The next critical phase of evaluation requires prospective deployment of an AI-based algorithm into a clinical work flow and assessment that relevant clinical outcomes or work flow metrics are substantially improved. The goals of such evaluation are similar to those for a novel medical device. Such prospective clinical evaluation may be even more critical for ML software algorithms since, at least initially, they will likely still rely upon interaction with human clinicians to make a final diagnosis or plan of therapy. Methods that can help clinicians understand how the algorithm functions and makes its decisions, such as AI explainability or interpretability techniques, may not only help to build trust among clinicians who must integrate an algorithm's conclusions into their decision making but also provide context to clinicians as to when an algorithm may be making unreasonable predictions.

There has been a recent increase in AI-related cardiovascular clinical trials registered on [clinicaltrials.gov](https://clinicaltrials.gov), the US government-sponsored Web site on which clinical trials are required to be reported. At the time of writing, we identified 17 cardiology-related trials reported on [clinicaltrials.gov](https://clinicaltrials.gov): three used 12-lead ECGs, four used single-lead smartwatch ECGs, five used echos, three used angiograms, one used a proprietary take-home stethoscope, and one used free text from the electronic health record. Seven were reported as being completed; however, only one<sup>25</sup> had published results. As discussed above, the EAGLE trial<sup>25</sup> demonstrated that a DNN-enabled ECG analysis system for probability of low ejection fraction increased early detection of low ejection fraction compared with standard of care in primary

care clinics. Many future clinical trials examining the deployment of AI algorithms within clinical workflows and designed to evaluate clinically relevant questions—such as accuracy of diagnosis, speed to therapy or increased rate of early detection—will be needed to provide the evidence to support clinical adoption of AI algorithms. In some cases, it is expected that ML-based algorithms may either not outperform existing standard of care or may not provide the clinical outcomes to support adoption.

Governmental regulatory agencies, such as the United States Food and Drug Administration (FDA), govern the level of scrutiny required for an ML algorithm to obtain approval for clinical adoption. While clinical trials are not necessarily required for approval of an algorithm, the level of evidence required varies by governmental agency and generally correlates with the potential risk an algorithm poses to patients and the existence of previously approved similar algorithms.<sup>78</sup> FDA approval of ML algorithms has increased in recent years, with 130 ML-based algorithms approved between 1997 and 2020, most being after 2016.<sup>79</sup> Under current FDA rules, once these algorithms are approved, they must be locked, meaning that the algorithms should not be updated in real time as more data is obtained. While this may help to ensure algorithm consistency, it shortchanges one of the inherent benefits of ML algorithms, which is an ability to improve performance as more training data become available. Similar approaches can also be applied to allow AI algorithms to adapt their predictions as populations and data change over time, also known as input distribution drift.<sup>80–83</sup> In 2019, the FDA acknowledged this limitation and proposed a plan for implementing a control plan that would allow algorithms to be updated based on a predetermined process.<sup>84,85</sup> To date, this plan has not yet been codified but is a welcome sign that the inherent strengths of AI algorithms can be appropriately leveraged for medical applications. The coming years will likely mark a turning point in the adoption of AI algorithms in medicine, driven by algorithm validation in RCTs, increased algorithm approvals by the FDA, and improved FDA guidelines for retraining and monitoring deployed AI algorithms.

### LIMITATIONS AND RISKS OF ML IN CARDIOVASCULAR MEDICINE

The nature of ML systems presents several notable challenges to implementation in medical practice. These include opacity of algorithm decision making and potential for bias, which are related to the complexity of the algorithms and biases present in the training data. There is also potential for degradation of algorithm performance over time due to data distribution drift. Significant work is ongoing to address these issues not only in medicine but in the broader ML field.

A common criticism of ML algorithms in general, and DNNs in particular, is the relative difficulty of interpreting how the algorithms make their decisions. These algorithms are extremely complex and derive their predictive power by allowing very high levels of interaction among the input data, which accordingly makes it nearly impossible to understand how the algorithms make a certain prediction simply by looking at the model parameters. In comparison, a simple generalized linear model

(such as linear regression) with a handful of predictors can be interpreted by looking at the magnitude and directionality of the trained weight associated with each predictor. Interpretability of algorithms carries particular value in medicine because the decisions they influence involve human health and the operational efficiency of health care systems. Therefore, understanding why and how an algorithm makes its prediction can provide valuable information that physicians can use to either incorporate the prediction into clinical care and/or learn what features the algorithm found important in the data, providing data-driven physiologic insights. In medicine, interpretability can help increase trust by physicians in the algorithm. AI explainability techniques have been developed, such as variable importance analysis, local interpretable model-agnostic explanations (LIME),<sup>86</sup> or gradient-weighted class activation mapping (Grad-CAM),<sup>87</sup> providing some insight into aspects of the input data that a given trained algorithm finds important to predict the target task. However, each of these methods provides only a limited view into algorithmic decisions, and none provide a comprehensive view. As shown in Ulloa Cerna et al.,<sup>50</sup> clinicians can have a difficult time interpreting the output of these methods on a case-by-case basis, due in part to the variability between cases and because some highlighted associations may represent artifact or noise. The ML field is actively working on developing novel approaches to improve ML interpretability.

It is important to recognize the potential for bias in any predictive model, and this phenomenon can be enhanced in some circumstances with ML. In part because ML models can so powerfully learn from the data on which they are trained, they are also positioned to learn biases present in these datasets.<sup>88</sup> Biases may be present in the proxies used to define the target task or labels used to train the model.<sup>89</sup> They may also arise if certain patient demographics are poorly represented in the training data,<sup>90</sup> or when training data are restricted to a limited geographic area.<sup>91</sup> The use of external validation datasets, and observing degradation of model performance therein, may help to illuminate presence of some biases, although this would only be true if the same biases do not also exist in the external validation dataset. Ideally algorithms should be trained and tested in demographically diverse target populations similar to the target population of interest.<sup>92</sup> Efforts should be made to identify potential sources of bias at every step of the research process from problem formulation to clinical validation.

Real-world clinical implementation of DL algorithms will require addressing data distribution shifts. Data distribution shift, or distribution drift, refers to changes that occur in either the input data, the output labels, or the relationship between them. For example, a model is trained to detect HCM. If HCM guidelines undergo revision and the criteria for HCM diagnosis change, then this previously trained algorithm may perform worse to detect HCM as defined by the new guidelines. Other examples can be more subtle. For example, if patient populations presenting to a given healthcare institution change, this may affect the disease prevalence in the population or other associated characteristics, which may degrade algorithm performance. Companies outside of medicine relying on ML models in production often retrain these models either at fixed intervals or have established monitoring systems to detect performance degradation to trigger model re-

training. Updated FDA guidance will help to inform how healthcare institutions should perform this kind of retraining while maintaining sufficient safety, oversight, and compliance.

### Conclusions

The field of modern AI, which can be defined as beginning with the development of DNNs in the past decade, has driven substantial progress in medical algorithms thus far. But in almost every respect this progress represents only the tip of the iceberg of the depth of impact ML could ultimately have on medicine. Much of the important work discussed above represents the early proof of concept showing how ML algorithms can be deployed in various ways related to cardiovascular disease. In many cases, similar ML approaches can be replicated for countless other diseases using similar algorithm architectures and data types. For every reported study in the literature there are likely dozens of related diseases, medical tasks, or applications for which similar ML algorithms can be trained. In many cases, access to sufficient multi-institutional labeled training data is the limiting factor in expanding the number of medical applications for AI approaches, highlighting an important area of future development for medical AI. Additional methodologic innovation in ML may expand the types of medical tasks that can be performed. The bulk of the effort, time, and capital that will be required to truly drive the clinical adoption of ML algorithms in medicine will almost certainly lie in the long-tailed process of iterative algorithm refinement and process improvements, including developing the infrastructure and methods that enable algorithms to be improved in real time based on human expert feedback, patient outcomes or increases in labeled training data. Therefore, the work still required by the broader medical community to achieve widespread clinical adoption of ML is also just at the tip of the iceberg.

The enormous potential of data-driven ML in medicine lies in the ability to employ algorithms that can help clinicians provide better care and enable physiologic discovery from large amounts of medical and biologic data. Although most modern medical data are accessible in digitized formats, they remain largely siloed within individual healthcare systems, walled off by administrative barriers that limit the large-scale data sharing that would likely drive the greatest ML-based medical advances. For all of the advances in AI technology over the past 10 years, perhaps the greatest medical AI innovations remain ahead of us: a combination of administrative and technical privacy-preserving approaches to allow ML algorithms to derive insights from diverse cross-institutional data, to more completely realize the fullest potential of AI in medicine.

### AUTHOR CONTRIBUTIONS

J.P.B. and G.H.T. contributed to manuscript conceptualization, literature review, writing, and revising.

### DECLARATION OF INTERESTS

G.H.T. has previously received research grants from General Electric, Janssen Pharmaceuticals, and MyoKardia, Inc., a subsidiary of Bristol Myers Squibb; and received consulting fees from MyoKardia. G.H.T. received funding from NHLBI-K23HL135274.



REFERENCES

1. Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.
2. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *N. Engl. J. Med.* 380, 1347–1358.
3. Ben-Assuli, O. (2015). Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. *Health Pol.* 119, 287–297. <https://doi.org/10.1016/j.healthpol.2014.11.014>.
4. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf.*, pp. 4171–4186.
5. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* 4, 86. <https://doi.org/10.1038/s41746-021-00455-y>.
6. Li, Y., Rao, S., Soares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. (2020). BEHRT: transformer for electronic health records. *Sci. Rep.* 10, 7155.
7. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., and Ng, A.Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* 25, 65–69. <https://doi.org/10.1038/s41591-018-0268-3>.
8. Attia, Z.I., Kapa, S., Lopez-Jimenez, F., McKie, P.M., Ladewig, D.J., Satam, G., Pellikka, P.A., Enriquez-Sarano, M., Noseworthy, P.A., Munger, T.M., et al. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat. Med.* 25, 70–74. <https://doi.org/10.1038/s41591-018-0240-2>.
9. Tison, G.H., Zhang, J., Delling, F.N., and Deo, R.C. (2019). Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. *Circ. Cardiovasc. Qual. Outcomes* 12, e005289.
10. Ko, W.Y., Siontis, K.C., Attia, Z.I., Carter, R.E., Kapa, S., Ommen, S.R., Demuth, S.J., Ackerman, M.J., Gersh, B.J., Arruda-Olson, A.M., et al. (2020). Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J. Am. Coll. Cardiol.* 75, 722–733.
11. Kwon, J.M., Kim, K.H., Medina-Inojosa, J., Jeon, K.H., Park, J., and Oh, B.H. (2020). Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. *J. Heart Lung Transplant.* 39, 805–814. <https://doi.org/10.1016/j.healun.2020.04.009>.
12. Goto, S., Mahara, K., Beussink-Nelson, L., Ikura, H., Katsumata, Y., Endo, J., Gaggin, H.K., Shah, S.J., Itabashi, Y., MacRae, C.A., and Deo, R.C. (2021). Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nat. Commun.* 12, 2726. <https://doi.org/10.1038/s41467-021-22877-8>.
13. Elias, P., Poterucha, T.J., Rajaram, V., Moller, L.M., Rodriguez, V., Bhavs, S., Hahn, R.T., Tison, G., Abreau, S.A., Barrios, J., et al. (2022). Deep learning electrocardiographic analysis for detection of left-sided valvular. *J. Am. Coll. Cardiol.* 80, 613–626.
14. Cohen-Shelly, M., Attia, Z.I., Friedman, P.A., Ito, S., Essayagh, B.A., Ko, W.Y., Murphree, D.H., Michelen, H.I., Enriquez-Sarano, M., Carter, R.E., et al. (2021). Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur. Heart J.* 42, 2885–2896.
15. Galloway, C.D., Valys, A.V., Shreibati, J.B., Treiman, D.L., Petterson, F.L., Gundotra, V.P., Albert, D.E., Attia, Z.I., Carter, R.E., Asirvatham, S.J., et al. (2019). Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiol.* 4, 428–436.
16. Raghunath, S., Ulloa Cerna, A.E., Jing, L., vanMaanen, D.P., Stough, J., Hartzel, D.N., Leader, J.B., Kirchner, H.L., Stumpe, M.C., Hafez, A., et al. (2020). Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat. Med.* 26, 886–891. <https://doi.org/10.1038/s41591-020-0870-z>.
17. Jentzer, J.C., Kashou, A.H., Lopez-Jimenez, F., Attia, Z.I., Kapa, S., Friedman, P.A., and Noseworthy, P.A. (2021). Mortality risk stratification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients. *Eur. Heart J. Acute Cardiovasc. Care* 10, 532–541.
18. Hughes, J.W., Olgin, J.E., Avram, R., Abreau, S.A., Sittler, T., Radia, K., Hsia, H., Walters, T., Lee, B., Gonzalez, J.E., and Tison, G.H. (2021). Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA Cardiol.* 6, 1285–1295.
19. Tison, G.H., Siontis, K.C., Abreau, S., Attia, Z., Agarwal, P., Balasubramanyam, A., Li, Y., Sehnert, A.J., Edelberg, J.M., Friedman, P.A., et al. (2022). Assessment of disease status and treatment response with artificial Intelligence-Enhanced electrocardiography in obstructive hypertrophic cardiomyopathy. *J. Am. Coll. Cardiol.* 79, 1032–1034.
20. Ulloa-Cerna, A.E., Jing, L., Pfeifer, J.M., Raghunath, S., Ruhl, J.A., Rocha, D.B., Leader, J.B., Zimmerman, N., Lee, G., Steinhubl, S.R., et al. (2022). RECHOMmend: an ECG-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography. *Circulation* 146, 36–47.
21. Attia, Z.I., Noseworthy, P.A., Lopez-Jimenez, F., Asirvatham, S.J., Deshmukh, A.J., Gersh, B.J., Carter, R.E., Yao, X., Rabinstein, A.A., Erickson, B.J., et al. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 394, 861–867. [https://doi.org/10.1016/S0140-6736\(19\)31721-0](https://doi.org/10.1016/S0140-6736(19)31721-0).
22. Raghunath, S., Pfeifer, J.M., Ulloa-Cerna, A.E., Nemani, A., Carbonati, T., Jing, L., vanMaanen, D.P., Hartzel, D.N., Ruhl, J.A., Lagerman, B.F., et al. (2021). Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related stroke. *Circulation* 143, 1287–1298.
23. Kwon, J.M., Cho, Y., Jeon, K.H., Cho, S., Kim, K.H., Baek, S.D., Jeung, S., Park, J., and Oh, B.H. (2020). A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. *Lancet. Digit. Health* 2, e358–e367.
24. Bos, J.M., Attia, Z.I., Albert, D.E., Noseworthy, P.A., Friedman, P.A., and Ackerman, M.J. (2021). Use of artificial intelligence and deep neural networks in evaluation of patients with electrocardiographically concealed long QT syndrome from the surface 12-lead electrocardiogram. *JAMA Cardiol.* 6, 532–538.
25. Yao, X., Rushlow, D.R., Inselman, J.W., McCoy, R.G., Thacher, T.D., Behnken, E.M., Bernard, M.E., Rosas, S.L., Akfaly, A., Misra, A., et al. (2021). Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat. Med.* 27, 815–819. <https://doi.org/10.1038/s41591-021-01335-4>.
26. Khamis, H., Zurakhov, G., Azar, V., Raz, A., Friedman, Z., and Adam, D. (2017). Automatic apical view classification of echocardiograms using a discriminative learning dictionary. *Med. Image Anal.* 36, 15–21. <https://doi.org/10.1016/j.media.2016.10.007>.
27. Agarwal, D., Shriram, K.S., and Subramanian, N. (2013). Automatic view classification of echocardiograms using Histogram of Oriented Gradients. *Proc - Int Symp Biomed Imaging*, 1368–1371.
28. Ebadollahi, S., Chang, S.F., and Wu, H. (2004). Automatic view recognition in echocardiogram videos using parts-based representation. In *In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
29. Balaji, G.N., Subashini, T.S., and Chidambaram, N. (2015). Automatic classification of cardiac views in echocardiogram using histogram and statistical features. *Procedia Comput. Sci.* 46, 1569–1576. <https://doi.org/10.1016/j.procs.2015.02.084>.
30. Gao, X., Li, W., Loomes, M., and Wang, L. (2017). A fused deep learning architecture for viewpoint classification of echocardiography. *Inf. Fusion* 36, 103–113. <https://doi.org/10.1016/j.inffus.2016.11.007>.

31. Madani, A., Arnaout, R., Mofrad, M., and Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit. Med.* **1**, 6–8. <https://doi.org/10.1038/s41746-017-0013-1>.
32. Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E., Aras, M.A., Jordan, C., et al. (2018). Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* **138**, 1623–1635.
33. Barbosa, D., Heyde, B., Dietenbeck, T., Houle, H., Friboulet, D., Bernard, O., and D'hooge, J. (2013). Quantification of left ventricular volume and global function using a fast automated segmentation tool: validation in a clinical setting. *Int. J. Cardiovasc. Imaging* **29**, 309–316.
34. Yang, L., Georgescu, B., Zheng, Y., Foran, D.J., and Comaniciu, D. (2008). A fast and accurate tracking algorithm of left ventricles in 3D echocardiography. In 2008 5th IEEE Int Symp Biomed Imaging From Nano to Macro, Proceedings, ISBI., pp. 221–224.
35. Tsang, W., Salgo, I.S., Medvedofsky, D., Takeuchi, M., Prater, D., Weinert, L., Yamat, M., Mor-Avi, V., Patel, A.R., and Lang, R.M. (2016). Transthoracic 3D echocardiographic left heart chamber quantification using an automated adaptive analytics algorithm. *JACC. Cardiovasc. Imaging* **9**, 769–782. <https://doi.org/10.1016/j.jcmg.2015.12.020>.
36. Costa, E., Martins, N., Sultan, M.S., Veiga, D., Ferreira, M., Mattos, S., et al. (2019). Mitral valve leaflets segmentation in echocardiography using convolutional neural networks. In 6th IEEE Port Meet Bioeng ENBENG 2019 - Proc., pp. 1–4.
37. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., and Zou, J.Y. (2020). Deep learning interpretation of echocardiograms. *NPJ Digit. Med.* **3**, 10. <https://doi.org/10.1038/s41746-019-0216-8>.
38. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., and Zou, J.Y. (2020). Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256. <https://doi.org/10.1038/s41586-020-2145-8>.
39. Narula, S., Shameer, K., Salem Omar, A.M., Dudley, J.T., and Sengupta, P.P. (2016). Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. *J. Am. Coll. Cardiol.* **68**, 2287–2295.
40. Upton, R., Mumith, A., Beqiri, A., Parker, A., Hawkes, W., Gao, S., Porumb, M., Sarwar, R., Marques, P., Markham, D., et al. (2022). Automated echocardiographic detection of severe coronary artery disease using artificial intelligence. *JACC. Cardiovasc. Imaging* **15**, 715–727.
41. Kusunose, K., Abe, T., Haga, A., Fukuda, D., Yamada, H., Harada, M., and Sata, M. (2020). A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *JACC. Cardiovasc. Imaging* **13**, 374–381.
42. Duffy, G., Cheng, P.P., Yuan, N., He, B., Kwan, A.C., Shun-Shin, M.J., Alexander, K.M., Ebinger, J., Lungren, M.P., Rader, F., et al. (2022). High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiol.* **7**, 386–395.
43. Huang, M.S., Wang, C.S., Chiang, J.H., Liu, P.Y., and Tsai, W.C. (2020). Automated recognition of regional wall motion abnormalities through deep neural network interpretation of transthoracic echocardiography. *Circulation* **142**, 1510–1520.
44. Arnaout, R., Curran, L., Zhao, Y., Levine, J.C., Chinn, E., and Moon-Grady, A.J. (2021). An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat. Med.* **27**, 882–891. <https://doi.org/10.1038/s41591-021-01342-5>.
45. Sanchez-Martinez, S., Duchateau, N., Erdei, T., Kunszt, G., Aakhus, S., Degiovanni, A., Marino, P., Carluccio, E., Piella, G., Fraser, A.G., and Bjnens, B.H. (2018). Machine learning analysis of left ventricular function to characterize heart failure with preserved ejection fraction. *Circ. Cardiovasc. Imaging* **11**, e007138. <http://ahajournals.org>.
46. Tabassian, M., Sunderji, I., Erdei, T., Sanchez-Martinez, S., Degiovanni, A., Marino, P., Fraser, A.G., and D'hooge, J. (2018). Diagnosis of heart failure with preserved ejection fraction: machine learning of spatiotemporal variations in left ventricular deformation. *J. Am. Soc. Echocardiogr.* **31**, 1272–1284, e9. <https://doi.org/10.1016/j.echo.2018.07.013>.
47. Shah, S.J., Katz, D.H., Selvaraj, S., Burke, M.A., Yancy, C.W., Gheorghide, M., Bonow, R.O., Huang, C.C., and Deo, R.C. (2015). Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* **131**, 269–279. <https://doi.org/10.1161/CIRCULATIONAHA.114.010637>.
48. Chiou, Y.A., Hung, C.L., and Lin, S.F. (2021). AI-assisted echocardiographic prescreening of heart failure with preserved ejection fraction on the basis of intrabeat dynamics. *JACC. Cardiovasc. Imaging* **14**, 2091–2104.
49. Laumer, F., Di Vece, D., Cammann, V.L., Würdinger, M., Petkova, V., Schönberger, M., Schönberger, A., Mercier, J.C., Niederseer, D., Seifert, B., et al. (2022). Assessment of artificial intelligence in echocardiography diagnostics in differentiating takotsubo syndrome from myocardial infarction. *JAMA Cardiol.* **7**, 494–503.
50. Ulloa Cerna, A.E., Jing, L., Good, C.W., vanMaanen, D.P., Raghunath, S., Suever, J.D., Nevius, C.D., Wehner, G.J., Hartzel, D.N., Leader, J.B., et al. (2021). Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality. *Nat. Biomed. Eng.* **5**, 546–554. <https://doi.org/10.1038/s41551-020-00667-9>.
51. Hughes, J.W., Yuan, N., He, B., Ouyang, J., Ebinger, J., Botting, P., Lee, J., Theurer, J., Tooley, J.E., Nieman, K., and Lungren, M.P. (2021). Deep learning evaluation of biomarkers from echocardiogram videos. *EBioMedicine* **73**, 103613. <https://doi.org/10.1016/j.ebiom.2021.103613>.
52. Narang, A., Bae, R., Hong, H., Thomas, Y., Surette, S., Cadieu, C., Chaudhry, A., Martin, R.P., McCarthy, P.M., Rubenson, D.S., et al. (2021). Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. *JAMA Cardiol.* **6**, 624–632.
53. Simes, P.A., Myreng, Y., Molstad, P., and Golf, S. (1996). Reproducibility of quantitative coronary analysis. *Int. J. Card. Imaging* **12**, 197–203.
54. Leape, L.L., Park, R.E., Bashore, T.M., Harrison, J.K., Davidson, C.J., and Brook, R.H. (2000). Effect of variability in the interpretation of coronary angiograms on the appropriateness of use of coronary revascularization procedures. *Am. Heart J.* **139**, 106–113.
55. Zir, L.M., Miller, S.W., Dinsmore, R.E., Gilbert, J.P., and Harthorne, J.W. (1976). Interobserver variability in coronary angiography. *Circulation* **53**, 627–632.
56. Avram, R., Olgin, J.E., Wan, A., Ahmed, Z., Verreault, L., Abreau, S., Wan, D., Gonzalez, J.E., So, D.Y., Soni, K., and Tison, G.H. CathAI: fully automated interpretation of coronary angiograms using neural networks. Preprint at arXiv. Available from: <https://arxiv.org/pdf/2106.07708.pdf>
57. Zhou, C., Dinh, T.V., Kong, H., Yap, J., Yeo, K.K., Lee, H.K., and Liang, K. (2021). Automated deep learning analysis of angiography video sequences for coronary artery disease. Preprint at arXiv. 1–10. <http://arxiv.org/abs/2101.12505>.
58. Wolterink, J.M., Leiner, T., de Vos, B.D., van Hamersvelt, R.W., Viergever, M.A., and Išgum, I. (2016). Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med. Image Anal.* **34**, 123–136. <https://doi.org/10.1016/j.media.2016.04.004>.
59. Lessmann, N., Van Ginneken, B., Zreik, M., De Jong, P.A., De Vos, B.D., Viergever, M.A., and Išgum, I. (2018). Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. *IEEE Trans. Med. Imaging* **37**, 615–625.
60. Commandeur, F., Goeller, M., Betancur, J., Cadet, S., Doris, M., Chen, X., Berman, D.S., Slomka, P.J., Tamarappoo, B.K., and Dey, D. (2018). Deep learning for quantification of epicardial and thoracic adipose tissue from non-contrast CT. *IEEE Trans. Med. Imaging* **37**, 1835–1846.
61. Zreik, M., Lessmann, N., van Hamersvelt, R.W., Wolterink, J.M., Voskuil, M., Viergever, M.A., Leiner, T., and Išgum, I. (2018). Deep learning analysis of the myocardium in coronary CT angiography for identification of

- patients with functionally significant coronary artery stenosis. *Med. Image Anal.* **44**, 72–85. <https://doi.org/10.1016/j.media.2017.11.008>.
62. Ito, L., Rapaka, S., Passerini, T., Georgescu, B., Schwemmer, C., Schoebinger, M., Flohr, T., Sharma, P., and Comaniciu, D. (2016). A machine-learning approach for computation of fractional flow reserve from coronary computed tomography. *J. Appl. Physiol.* **121**, 42–52.
63. Schlemper, J., Caballero, J., Hajnal, J.V., Price, A.N., and Rueckert, D. (2018). A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Trans. Med. Imaging* **37**, 491–503.
64. Blansit, K., Retson, T., Masutani, E., Bahrami, N., and Hsiao, A. (2019). Deep learning-based prescription of cardiac MRI planes. *Radiol. Artif. Intell.* **1**, e180069.
65. Romaguera, L.V., Romero, F.P., Fernandes Costa Filho, C.F., and Fernandes Costa, M.G. (2018). Myocardial segmentation in cardiac magnetic resonance images using fully convolutional neural networks. *Biomed. Signal Process Control* **44**, 48–57. <https://doi.org/10.1016/j.bspc.2018.04.008>.
66. Bernard, O., Lalonde, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans. Med. Imaging* **37**, 2514–2525.
67. Wang, S., Chauhan, D., Patel, H., Amir-Khalili, A., da Silva, I.F., Sojoudi, A., Friedrich, S., Singh, A., Landers, L., Miller, T., et al. (2022). Assessment of right ventricular size and function from cardiovascular magnetic resonance images using artificial intelligence. *J. Cardiovasc. Magn. Reson.* **24**, 27–12. <https://doi.org/10.1186/s12968-022-00861-5>.
68. Maicas, G. (2019). End-to-end diagnosis and segmentation learning from cardiac magnetic resonance imaging aiml (School of Computer Science, The University of Adelaide, Australia Imaging Physics), pp. 802–805. Faculty of Applied Sciences, Delft University of Technology.
69. Wolterink, J.M., Leiner, T., Viergever, M.A., and Išgum, I. (2018). Automatic segmentation and disease classification using cardiac cine MR images. *Lect. Notes Comput. Sci.* **10663**, 101–110.
70. Khened, M., Alex, V., and Krishnamurthi, G. (2018). Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest. *Lect. Notes Comput. Sci.* **10663**, 140–151.
71. Ammar, A., Bouattane, O., and Youssfi, M. (2021). Automatic cardiac cine MRI segmentation and heart disease classification. *Comput Med Imaging Graph* **88**, 101864. <https://doi.org/10.1016/j.compmedimag.2021.101864>.
72. Moon, S., Liu, S., Scott, C.G., Samudrala, S., Abidjan, M.M., Geske, J.B., Noseworthy, P.A., Shellum, J.L., Chaudhry, R., Ommen, S.R., et al. (2019). Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. *Int. J. Med. Inform.* **128**, 32–38. <https://doi.org/10.1016/j.ijmedinf.2019.05.008>.
73. Blecker, S., Katz, S.D., Horwitz, L.I., Kuperman, G., Park, H., Gold, A., and Sontag, D. (2016). Comparison of approaches for heart failure case identification from electronic health record data. *JAMA Cardiol.* **1**, 1014–1020.
74. Oguz, C., Sen, S.K., Davis, A.R., Fu, Y.P., O'Donnell, C.J., and Gibbons, G.H. (2017). Genotype-driven identification of a molecular network predictive of advanced coronary calcium in ClinSeq® and Framingham Heart Study cohorts. *BMC Syst. Biol.* **11**, 1–14.
75. Bauer, Y., de Bernard, S., Hickey, P., Ballard, K., Cruz, J., Cornelisse, P., Chadha-Boreham, H., Distler, O., Rosenberg, D., Doelberg, M., et al. (2021). Identifying early pulmonary arterial hypertension biomarkers in systemic sclerosis: machine learning on proteomics from the DETECT cohort. *Eur. Respir. J.* **57**, 2002591. <https://doi.org/10.1183/13993003.02591-2020>.
76. Luo, R., Sedlazeck, F.J., Lam, T.W., and Schatz, M.C. (2019). A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat. Commun.* **10**, 998. <https://doi.org/10.1038/s41467-019-09025-z>.
77. Schatz, M.C., Luo, R., and Lam, T.W. (2020). Skyhawk: an artificial neural network-based discriminator for reviewing clinically significant genomic variants. *Int. J. Comput. Biol. Drug Des.* **13**, 431–437.
78. IMDRF Software as a Medical Device (SaMD) Working Group (2014). Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations (IMDRF), pp. 1–30. <http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf>.
79. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., and Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584. <https://www.nature.com/articles/s41591-021-01312-x>.
80. Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004). Learning with drift detection. In *Advances in Artificial Intelligence – SBIA 2004*, A.L.C. Bazzan and S. Labidi, eds. (Springer Berlin Heidelberg), pp. 286–295.
81. Widmer, G., and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Mach. Learn.* **23**, 69–101.
82. Webb, G.I., Lee, L.K., Petitjean, F., and Goethals, B. (2017). Understanding concept drift. Preprint at arXiv. <http://arxiv.org/abs/1704.00362>.
83. Feng, J., Emerson, S., and Simon, N. (2021). Approval policies for modifications to machine learning-based software as a medical device: a study of bio-creep. *Biometrics* **77**, 31–44.
84. US FDA (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML) -based software as a medical device (SaMD) - discussion paper and request for feedback. *US Food Drug Adm*, 1–20.
85. US FDA (2021). How Is the FDA Considering Regulation of Artificial Intelligence and Machine Learning Medical Devices?. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>.
86. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *NAACL-HLT 2016 - 2016 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Proc Demonstr Sess*, pp. 97–101.
87. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359.
88. Liu, Y., Chen, P.H.C., Krause, J., and Peng, L. (2019). How to read articles that use machine learning: users' guides to the medical literature. *JAMA, J. Am. Med. Assoc.* **322**, 1806–1816.
89. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453.
90. Adamson, A.S., and Smith, A. (2018). Machine learning and health care disparities in dermatology. *JAMA Dermatology* **154**, 1247–1248.
91. Kaushal, A., Altman, R., and Langlotz, C. (2020). Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA, J. Am. Med. Assoc.* **324**, 1212–1213.
92. Noseworthy, P.A., Attia, Z.I., Brewer, L.P.C., Hayes, S.N., Yao, X., Kapa, S., Friedman, P.A., Lopez-Jimenez, F., et al. (2020). Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythmia Electrophysiol* **13**, 208–214.