

# Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging

---

**Sanchez-Martinez, Sergio; Camara, Oscar; Piella, Gemma; Čikes, Maja; González-Ballester, Miguel Ángel; Miron, Marius; Vellido, Alfredo; Gómez, Emilia; Fraser, Alan G.; Bijmens, Bart**

*Source / Izvornik:* **Frontiers in Cardiovascular Medicine, 2022, 8**

**Journal article, Published version**

**Rad u časopisu, Objavljena verzija rada (izdavačev PDF)**

<https://doi.org/10.3389/fcvm.2021.765693>

*Permanent link / Trajna poveznica:* <https://um.nsk.hr/um:nbn:hr:105:785880>

*Rights / Prava:* [Attribution 4.0 International](#)/[Imenovanje 4.0 međunarodna](#)

*Download date / Datum preuzimanja:* **2024-07-20**



*Repository / Repozitorij:*

[Dr Med - University of Zagreb School of Medicine Digital Repository](#)





# Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging

Sergio Sanchez-Martinez<sup>1</sup>, Oscar Camara<sup>2</sup>, Gemma Piella<sup>2</sup>, Maja Cikes<sup>3</sup>, Miguel Ángel González-Ballester<sup>2,4</sup>, Marius Miron<sup>5</sup>, Alfredo Vellido<sup>6</sup>, Emilia Gómez<sup>2,5\*</sup>, Alan G. Fraser<sup>7</sup> and Bart Bijnens<sup>1,4,8</sup>

<sup>1</sup> August Pi i Sunyer Biomedical Research Institute (IDIBAPS), Barcelona, Spain, <sup>2</sup> Department of Information and Communication Technologies, University Pompeu Fabra, Barcelona, Spain, <sup>3</sup> Department of Cardiovascular Diseases, University of Zagreb School of Medicine, University Hospital Centre Zagreb, Zagreb, Croatia, <sup>4</sup> ICREA, Barcelona, Spain, <sup>5</sup> Joint Research Centre, European Commission, Seville, Spain, <sup>6</sup> Computer Science Department, Intelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain, <sup>7</sup> School of Medicine, Cardiff University, Cardiff, United Kingdom, <sup>8</sup> Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

## OPEN ACCESS

### Edited by:

Pablo Lamata,  
King's College London,  
United Kingdom

### Reviewed by:

Jennifer Mancio,  
University of Porto, Portugal  
Karthik Seetharam,  
West Virginia State University,  
United States

### \*Correspondence:

Emilia Gómez  
emilia.gomez-gutierrez@ec.europa.eu

### Specialty section:

This article was submitted to  
Cardiovascular Biologics and  
Regenerative Medicine,  
a section of the journal  
Frontiers in Cardiovascular Medicine

**Received:** 27 August 2021

**Accepted:** 07 December 2021

**Published:** 04 January 2022

### Citation:

Sanchez-Martinez S, Camara O,  
Piella G, Cikes M, González-Ballester  
MÁ, Miron M, Vellido A, Gómez E,  
Fraser AG and Bijnens B (2022)  
Machine Learning for Clinical  
Decision-Making: Challenges and  
Opportunities in Cardiovascular  
Imaging.  
*Front. Cardiovasc. Med.* 8:765693.  
doi: 10.3389/fcvm.2021.765693

The use of machine learning (ML) approaches to target clinical problems is called to revolutionize clinical decision-making in cardiology. The success of these tools is dependent on the understanding of the intrinsic processes being used during the conventional pathway by which clinicians make decisions. In a parallelism with this pathway, ML can have an impact at four levels: for data acquisition, predominantly by extracting standardized, high-quality information with the smallest possible learning curve; for feature extraction, by discharging healthcare practitioners from performing tedious measurements on raw data; for interpretation, by digesting complex, heterogeneous data in order to augment the understanding of the patient status; and for decision support, by leveraging the previous steps to predict clinical outcomes, response to treatment or to recommend a specific intervention. This paper discusses the state-of-the-art, as well as the current clinical status and challenges associated with the two later tasks of interpretation and decision support, together with the challenges related to the learning process, the auditability/traceability, the system infrastructure and the integration within clinical processes in cardiovascular imaging.

**Keywords:** artificial intelligence, machine learning, deep learning, clinical decision making, cardiovascular imaging, diagnosis, prediction

## INTRODUCTION

Artificial intelligence (AI) systems are programmed to achieve complex tasks by perceiving their environment through data acquisition, interpreting the collected data and deciding the best action(s) to take to achieve a given goal. As a broad scientific discipline, AI includes several approaches and techniques, such as machine learning, machine reasoning, and robotics (1). Machine learning (ML) is the subfield of AI that focuses on the development of algorithms that allow computers to automatically discover patterns in the data and improve with experience, without being given a set of explicit instructions. Among ML techniques, Deep Learning (DL) is

the subfield concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. Unlike other ML techniques, DL bypasses the need of using hand-crafted features as input, automatically figuring out the data features that are important for solving complex problems. This is the main reason why DL stands out as the current state-of-the-art in virtually all medical imaging related tasks.

In the particular case of clinical decision-making in cardiology, ML methods would perceive an individual by collecting and interpreting his/her clinical data and would reason on them to suggest actions to maintain or improve that individual's cardiovascular health. This mimics the clinician's approach when examining and treating a sick patient, or when suggesting preventive actions to avoid illness. Therefore, in order to assess the challenges and opportunities of ML systems for clinical decision-making in cardiology, an in-depth understanding of this process, when performed by cardiologists, is paramount.

**Figure 1** summarises a typical paradigm for clinical decision-making. It starts by data acquisition, including the clinical history of the patient, demographics, physiological measurements, electrocardiogram, imaging and laboratory tests, and the relevant indices collected from these data. Next, clinicians construct and interpret the state of the patient by comparison with population-based information learned during their education or daily practice, or information derived from guidelines. This interpretation is based on reasoning on the data using the human innate capability of contextualizing information through pattern recognition. Furthermore, clinicians assess the uncertainty associated with measurements and the completeness of the available information to estimate how much they can rely on the data. Finally, they consider the knowledge from the (natural as well as treated) expected evolution of populations related to the patient's status to make decisions. The resulting actions can be to either collect more data to minimize the uncertainty associated with the decision, to make an intervention (drug/device therapy, surgery, etc.) to improve the patient's outcomes, or to send the patient home (whether or not with planned observation follow-up).

In the era of evidence-based, personalised medicine (2), millions of individuals are carefully examined, which results in a deluge of complex, heterogeneous data. The use of algorithmic approaches to digest these data and augment clinical decision-making is now feasible due to the ever-increasing computing power, and the latest advances in the ML field (3). Indeed, big data leveraged by ML can provide well-curated information to clinicians so they can make better informed diagnoses and treatment recommendations, while also estimating probabilities and costs for the possible outcomes. ML-augmented decisions made by clinicians have the potential to improve outcomes, lower costs of care, and increase patient and family satisfaction.

ML analyses have, to date, demonstrated human-like performance in low-level tasks where pattern recognition or perception play a fundamental role. Some examples are **data acquisition**, standardization and classification (4, 5), and **feature extraction** (6, 7). For higher-level tasks involving reasoning, such as patient's **status interpretation** and **decision support**,

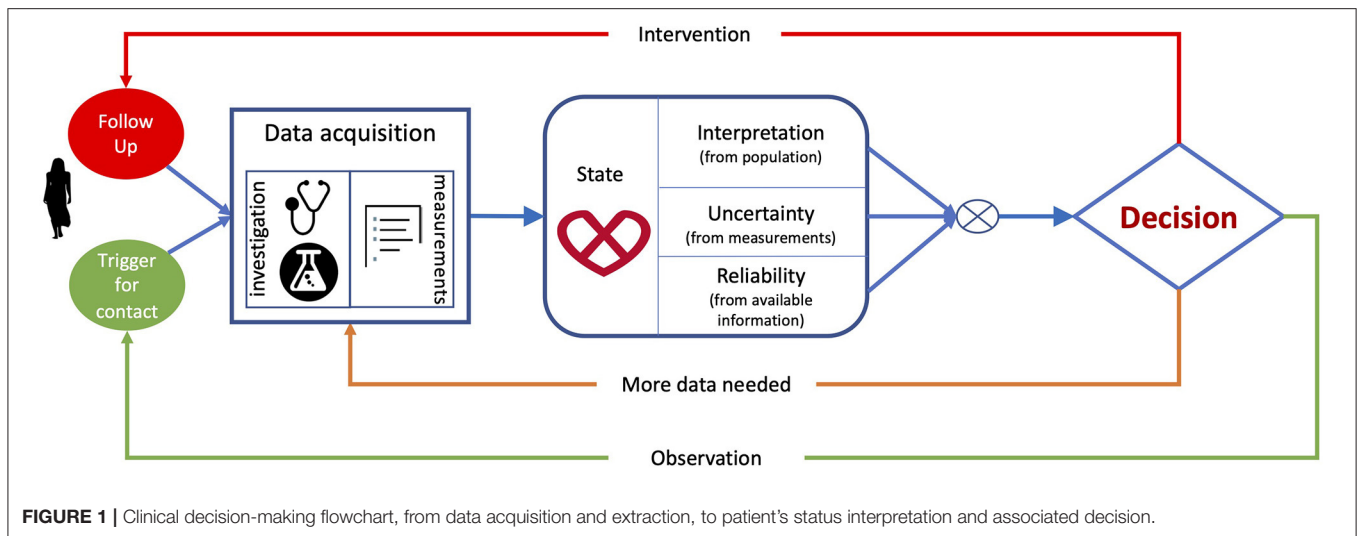
ML allows for the integration of complex, heterogenous data in the decision-making process, but these are still immature and need substantial validation (8). In parallel to **Figure 1**, which illustrated the process of making clinical decisions, **Figure 2** describes the tasks involved in this process according to how ML could contribute, and highlights how the risks to a patient from erroneous conclusions increase with each step.

There exist other review papers that cover the topic of AI in cardiovascular imaging from a broader perspective (9), or that highlight the synergy between machine learning and mechanistic models that enable the creation of a "digital twin" in pursuit of precision cardiology (10). Complementarily, this paper focuses on ML as a subfield of AI and on clinical decision-making as an essential part of cardiovascular medicine. Although cardiovascular imaging only constitutes a limited portion of the data spectrum in cardiology, we emphasize the imaging field in our literature review given that it is one of the areas to which ML has contributed the most (11). In the following, we discuss the higher-level tasks related to clinical decision-making that involve reasoning on clinical data, namely interpretation, and decision support. For each of these tasks, we review the ML state-of-the-art (indicating whether the implementation was based on DL or other ML technique), comment on the current penetration of ML tools into clinical practice (see *Clinical status* subsections), and elaborate on the current challenges that limit their implementation in clinical practice. Finally, we discuss the general challenges that may appear when tackling any clinical problem with ML approaches.

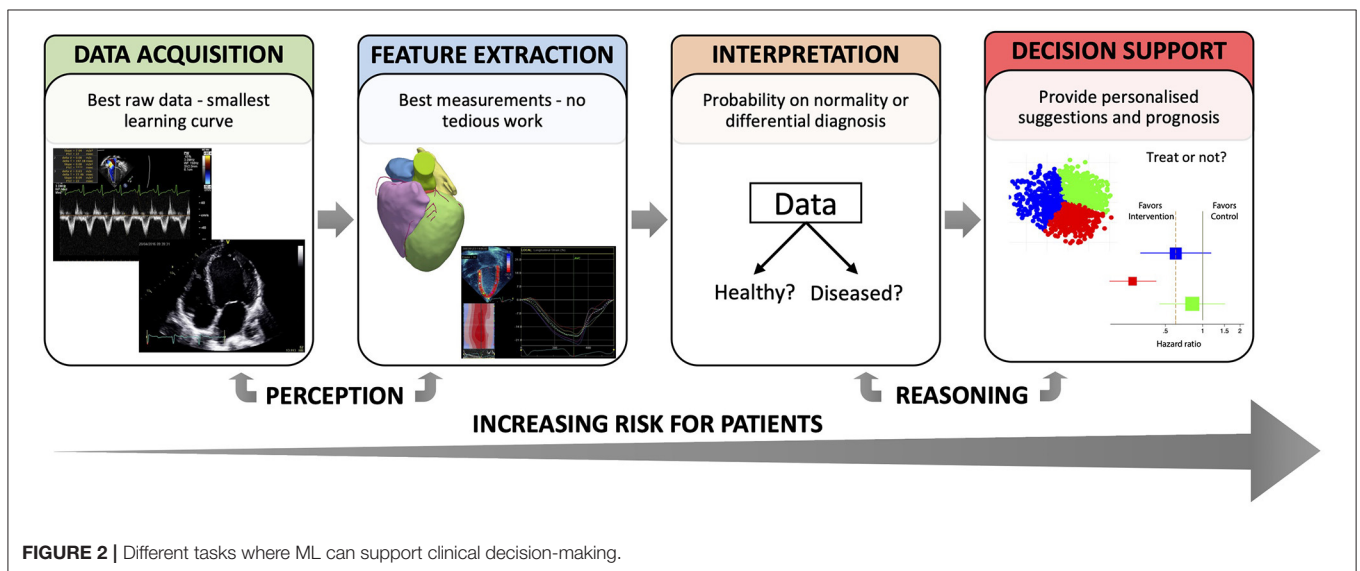
This paper addresses potential questions arising from data scientists, industrial partners and funding institutions, helping them understand clinical decision-making in cardiology and identify potential niches for their solutions to be helpful. At the same time, the paper aims at informing cardiologists about which ML tools could target their problems and what are their current limitations.

## STATUS INTERPRETATION—COMPARISON TO POPULATION

Let us assume that the clinical data of a patient have been properly acquired and relevant features are readily available. The next stage in the decision process consists in interpreting his/her status by comparison to populations. This comparison requires data normalization. When complex data are involved, such as cardiac images, the traditional approach to normalization is to build a statistical atlas—a reference model that captures the variability associated to a population (12). To build the atlas, the training data must be transformed into a common spatio(-temporal) framework, which can be achieved by registration. In this sense, registration appears as a crucial step for status interpretation toward diagnosis, and deep learning has emerged as a suitable tool to register 3D cardiac volumes (13), 3D pre-operative cardiac models to 2D intraoperative x-ray fluoroscopy to facilitate image-guided interventions (14), or cardiac MRI sequences (15).



**FIGURE 1** | Clinical decision-making flowchart, from data acquisition and extraction, to patient’s status interpretation and associated decision.



**FIGURE 2** | Different tasks where ML can support clinical decision-making.

The ML interpretation of the state of the patient can augment the diagnosis made by clinicians. Indeed, a recently published meta-analysis highlighted the promising potential of ML and DL models to predict conditions such as coronary artery disease, heart failure, stroke, and cardiac arrhythmias using data derived from routinely used imaging techniques and ECG (16). Based on imaging, ensemble ML models, which group the prediction of different weak learners, have demonstrated higher accuracy than expert readers for the diagnosis of obstructive coronary artery disease (17); a DL model automated the diagnosis of acute ischemic infarction using CT studies (18); and another DL model achieved 92.3% accuracy for left ventricular hypertrophy classification analysing echocardiographic images (19). Different ML models have also operated on electronic health records (EHR) for triaging of low-risk vs. high-risk cardiovascular patients, grading findings as requiring non-urgent, urgent or critical attention, as a strategy to improve efficiency and

allocation of the finite resources available in the emergency department (20). Lastly, a ML ensemble model combined clinical data, quantitative stenosis, and plaque metrics from CT angiography to effectively detect lesion-specific ischemia (21).

Another data-driven example of status interpretation is unsupervised machine learning for dimensionality reduction; a label-agnostic approach that orders individuals according to their similarity, i.e., those with a similar clinical presentation are grouped together, whereas those showing distinct pathophysiological features are positioned far apart (22). This allows identifying different levels of abnormality, or assessing the effect of therapies and interventions, as these are aimed to restore an individual toward increased “normality.” An implementation of unsupervised dimensionality reduction provided useful insight into treatment response in large patient populations (23), and quantified patient changes after an intervention using temporally dynamic data (24).

## Clinical Status

ML approaches for interpreting a patient's status enhance discovery in massive databases by offering the possibility to identify similar cases, build normality statistics, and spot outliers. Whether these approaches are intended for diagnosis or risk assessment, they could contribute to deliver better healthcare.

Unfortunately, many current ML applications for interpreting clinical data present a technically sound contribution, but do not address real clinical needs, and they often focus on binary classification of normal vs. abnormal (19), which strongly limits their use in routine clinical practice. Furthermore, studies showing impact on hard clinical endpoints rather than on surrogate measures are still needed. The way forward is through further integration of technical and clinical contributions, and through the elaboration of consensus recommendations on how to tackle a clinical necessity using ML.

## DECISION-MAKING (PREDICTION)

Based on the interpretation of the patient's status, clinicians should decide on whether: (1) observe the patient and wait until an event triggers the need for a decision; (2) collect more data to improve the odds of making the right decision; or (3) perform an intervention and monitor the outcome (see **Figure 1**). Machine learning methods can help clinicians to decide which pathway to follow (25), in a way that is cost-effective (26).

Several studies have assessed the predictive power of ML techniques based on imaging. An echocardiography-based DL model was shown to improve the prediction of in-hospital mortality among coronary heart disease and heart failure patients as compared to traditionally used prediction models (27). An ensemble ML approach interrogating SPECT myocardial perfusion studies demonstrated superior performance at predicting early revascularization in patients with suspected coronary artery disease as compared to an experienced reader (28), or in combination with clinical and ECG data outperformed the reading physicians at predicting the occurrence of major adverse cardiovascular events (29). Lastly, a DL implementation fed with CT scans from asymptomatic as well as stable and acute chest pain cohorts demonstrated the added clinical value of automated systems to predict cardiovascular events (30). Leaving imaging aside, deep learning based on clinical, laboratory and demographic data, ECG parameters, and cardiopulmonary exercise testing estimated prognosis and guided therapy in a large population of adults with congenital heart disease (31).

The interplay between different a priori non-related imaging tests has recently been discovered by DL through the identification of previously unnoticed associations. For example, breast arterial calcifications and the likelihood of patients at a high cardiovascular risk was sorted out using a DL model that operated with mammograms (32). Similarly, the power of ML in combination with the non-invasiveness of retinal scanning has been used to predict abnormalities in the macrovasculature based on the microvascular features of the eye. One such example is the DL model that predicted cardiovascular

risk factors using retinal fundus photographs, thus allowing for an easier and cost-effective cardiovascular risk stratification (33), or the DL implementation that inferred coronary artery calcium (CAC) scores from retinal photographs, which turned out to be as accurate as CT scan-measured CAC in predicting cardiovascular events (34).

## Clinical Status

The few examples of FDA-cleared cutting-edge ML applications to cardiovascular imaging that are thus suitable for routine use, focus on the low-level tasks of data acquisition and feature extraction, both in cardiac MRI (35) and echocardiography (36, 37), although the latter contribution did actually prove useful to predict a poor prognosis in acute COVID-19 patients based on DL-enabled automated quantification of echocardiographic images. However, the use of these ML applications for prediction and decision-making is still in its early days, as most models are still incapable of making predictions at the individual level (8, 38). More effort is needed toward integration in a clinical environment, interpretability, and validation if we want to see these models embedded in routine patient care.

## CHALLENGES COMMON TO STATUS INTERPRETATION & DECISION-MAKING

Applications concerning patient's status interpretation and decision-making, which entails learning what is the risk associated with each possible clinical decision, imply a much higher risk as compared to the low-level tasks of data acquisition and feature extraction, since decisions derived from them could harm patients. Accordingly, ML outcomes need to be intuitively interpretable by the cardiologist and validated in a much more exhaustive way (as required by medical device regulators; e.g., class IIa or IIb routes to commercialization), ultimately with the launch of randomized prospective trials.

One of the main challenges for ML approaches to status interpretation lies in the extraction of meaningful concepts from raw data. This challenge entails many others, related to the data themselves. The first concerns the reliability and representativeness of training and outcome data. If representative, ML models need to find a reliable metric to compare heterogeneous data, which is not trivial. Furthermore, for a successful interpretation, data collection protocols should be designed to cover gender-, ethnicity- and age-related changes, and capture the rare outliers (39). On top of this, ML systems should be designed to consider longitudinal data, as to assess a patient over time, e.g., during a stress protocol (40) or disease progression (41). Finally, ML models are trained on three different kinds of data; ranging from higher to lower quality and completeness: (1) randomized clinical trials, (2) cohorts, and (3) clinical routine real-world data. The exchange of knowledge throughout these collections of data is challenging, since what was learned from highly curated data (e.g., randomized clinical trials) may not generalize to routinely collected data.

Another crucial problem associated to currently available data is bias, i.e., when the training sample is not representative

of the population of interest (see section “General challenges” for more details). Accordingly, caution is needed when testing a trained model in new clinical centres. As ML users can attest, there will always be a trade-off between improving the system performance locally and having systems that generalize well (42). Automation bias, defined as the human tendency to accept a computer-generated solution without searching for contradictory information (43), may also affect clinical interpretation and decision-making. As shown by Goddard et al. (44), when the ML solution is reliable it augments human performance, but when the solution is incorrect human errors increase. Thus, who is to blame if a diagnostic algorithm fails? The further we move along the clinical decision-making flowchart (**Figure 2**), the more ethical and legal barriers the ML practitioner/company faces. To mitigate some of these issues, the training data should be accessible, and the learning systems equipped with tools that allow reconstructing the reasoning behind the decision taken.

**Table 1** organizes the ideas discussed for status interpretation and decision-making in the form of a SWOT analysis.

## GENERAL CHALLENGES

We have previously described the specific challenges that may arise when ML models are given the tasks of interpreting the patient’s status or making predictions to guide the clinical decision. In the following, we discuss the general challenges that may appear when tackling any clinical problem with ML approaches. These are divided into different sections, depending on whether they relate to the learning itself, the auditability/traceability aspects, the system/infrastructure, or the integration within clinical processes.

### Learning

#### (Non-standardized) Data

Medical data are normally kept in many separate systems, which hampers accessibility and makes comparisons at a population level nearly impossible. Electronic health records mostly contain unstructured data, and so they are underutilized by care providers and clinical researchers. Machine learning systems can help organize and standardize information, or can be designed to directly integrate unstructured complex data for high-throughput phenotyping to identify patient cohorts (45).

#### Bias and Confounding

As discussed in the previous section, bias is another risk that arises with the use of ML. Indeed, a recent review of cardiovascular risk prediction models revealed potential problems in the generalizability of multicentre studies that often show a wide variation in reporting, and thus these models may be biased toward the methods of care routinely used in the interrogated centres (46). For example, in the case of cardiac MRI, protocols are not standardised, varying by institution and machine vendor (47). This bias may amplify the gap in health outcomes between the dominant social group, whose data are used to train algorithms, and the minorities (8). Luckily enough, there are studies that make sure that all minorities are represented

in the training data (12, 39), but this should be the rule, not the exception. Another challenge when applying ML models that are designed to recognise patterns lies in the tendency of these to overfit the dataset because they fail at distinguishing a true contributing factor to the clinical outcome from noise (48).

ML solutions can similarly inherit human-like biases (49), such as the model whose recommendation of care after a heart attack was unequal among sex groups (50), as a consequence of a biased training dataset. This bias may also appear in ML-powered echocardiography, where studies are dependent on both the operator performing the study and the interpreter analysing it (47). This bias occurs because we ask ML solutions to predict which decisions the humans profiled in the training data would have made, thus we should not expect the ML method to be fair. The effect of human cognitive biases in ML algorithms have already been addressed and different “debiasing” techniques have been proposed (51), but this is a topic that certainly needs more attention.

Similar to bias, the learning process can be undermined by confounding, i.e., the finding of a spurious association between the input data and the outcome under study. Such is the case of the deep learning model that attempted at predicting ischaemia by looking at ECG records, but rather learnt at detecting the background electrical activity noise present in the ischaemic ECG examples, which was not present in the control cases (52). Unsupervised learning can be beneficial to avoid confounding, as it does not force the output of the model to match a given label but rather finds natural associations within the data (53). Similarly, randomization of experiments is highly recommended to avoid confounding (54).

### Validation and Continuous Improvement

Even if an algorithm proves to outperform humans in prediction tasks, systematic debugging, audit and extensive validation should be mandatory. For ML algorithms to be deployed in hospitals, they must improve patient as well as financial outcomes (8). Validation should be through multi-centre randomized prospective trials, to assess whether models trained at one site can be applied elsewhere. Examples of prospective ML trials assessed in a “real world” clinical environment are scarce—only 6% of 516 surveyed studies performed external validation, according to (55). Among these rare examples, finding prospective validation studies to prove the suitability of ML-enabled applications in cardiovascular imaging is even rarer. In (56), a prospective study concluded that a ML model that integrates clinical and quantitative imaging-based features outperforms the prediction of myocardial infarction and death as compared to standard clinical risk assessment. Attia et al. (57) conducted a prospective study to validate a DL algorithm that detected left ventricular systolic dysfunction. Another pivotal prospective multicentre trial was launched to demonstrate the feasibility of a ML-powered image guided acquisition software that helps novices to perform transthoracic echocardiography studies (58). Lastly, a validation study was performed to prove the feasibility of using DL to automatically segment and quantify the ventricular volumes in cardiac MRI (35).

**TABLE 1** | SWOT analysis—status interpretation and decision-making.

Strengths	Weaknesses
<ul style="list-style-type: none"> <li>• Allow objective and thorough comparison to populations</li> <li>• Allow the integration of complex, heterogenous features</li> <li>• May enhance the prediction of clinical outcomes, or the prediction of response to a given treatment or intervention.</li> </ul>	<ul style="list-style-type: none"> <li>• Need well-curated, representative databases for training</li> <li>• Affected by data reliability, representativeness, and bias</li> <li>• Need to extract meaningful, interpretable concepts</li> <li>• Need thorough validation—prospective trials</li> <li>• Need to integrate longitudinal data</li> <li>• Ensure transference of knowledge across populations</li> <li>• Need to prove clinical benefit</li> <li>• Need to be integrated within clinical systems</li> <li>• Need to prove cost-effectiveness</li> </ul>
Opportunities	Threats
<ul style="list-style-type: none"> <li>• Stimulate the man/machine collaboration</li> <li>• Reach diagnosis in a shorter time</li> <li>• Separate ambiguous cases that deserve more attention from clear cases—triaging</li> <li>• Help in the organization of healthcare—diagnosis, risk assessment and urgency assessment</li> <li>• Lower cost of healthcare by suggesting cost-effective decisions</li> </ul>	<ul style="list-style-type: none"> <li>• Harm patients if wrong decisions are taken—high-risk</li> <li>• Disappoint users, especially after all the striking news on ML failures</li> <li>• Affect human decisions in a negative way—automation bias</li> <li>• Make decisions for the average patient, not at the individual level</li> </ul>

One of the greatest benefits of ML models resides in their ability to improve their performance as more data become available. However, this might be challenging particularly for neural networks, which are prone to “catastrophic forgetting”—to abruptly forget previously learned information upon learning from new data. Furthermore, re-training on the whole database is time and resource consuming. To solve these problems, federated learning, a novel de-centralized computational architecture where machines run models locally to improve them with a single user’s data (59), could be helpful. Given the evolving nature of ML models, medical device providers are obliged to periodically monitor the performance of their programs, using a continuous validation paradigm (60).

## Auditability/Traceability Interpretability vs. Explainability

Interpretability is understood as the ability to explain or to present in understandable terms to a human (61). In a strictly-regulated field such as cardiovascular medicine, the lack of interpretability of ML models is one of the main limitations hindering adoption (62). Indeed, from the example of predicting ischaemia by looking at ECG records discussed above (52), it is evident that the non-intelligible use of ML outputs can lead to controversial results and therefore translation to clinical practice should be done cautiously. Unfortunately, many ML implementations available do not comply with the European General Data Protection Regulation (GDPR), which compels ML providers to reveal the information and logic involved in each decision (63).

When reasoning on the data to make decisions, the human brain can follow two approaches (64): the *fast/intuitive* (Type 1) vs. the *slow/reasoned* (Type 2) one. Type 1 is almost instantaneous and based on the human ability to apply heuristics to identify patterns from raw information. However, it is prone to error and bias, as it can lead to an incomplete *perception* of the patient due to low quality or lack of relevant data (65, 66). In contrast, Type 2 is deductive, deliberate, and demands a greater

intellectual, time and cost investment, but often turns out to be more accurate.

The above is highly relevant for both “traditional” and ML-based clinical decision-making, as ML systems ultimately mimic different aspects of human reasoning and can lead to the same errors. For the sake of explaining ML decisions, researchers provide attention maps (67), reveal which data the model “looked at” for each individual decision (68), provide estimates of feature importance (69), or explain the local behaviour of complex models by changing the input and evaluating the impact on the prediction (70). However, caution is needed with this entire research trend, as explainability is not a synonym of interpretability (71). Explainable models tend to reach conclusions by fast/intuitive black-box reasoning (Type 1, see also causal vs. predictive learning in the following subsection), while interpretable models demand a slow/reasoned (Type 2) approach throughout the entire learning path. In this sense, explainable models that follow a fast/intuitive reasoning might incur more diagnostic errors than interpretable models, which follow an analytical reasoning (64). The research field that focuses on enhancing models’ interpretability is still in its infancy. Generative synthesis (69), which uses ML to generate a simplified version of a neural network, or mathematical attempts to explain the inner working of neural networks (72) could provide key insight into why and how a network behaves the way it does, thus unravelling the black-box enigma (73).

For ML models to be applied in clinical decision-making, they cannot merely be interpretable, but they also must be credible. A credible model is an interpretable model that: (1) provides arguments for its predictions that are, at least in part, in-line with domain knowledge; (2) is at least as good as previous standards in terms of predictive performance (74). For ML models to achieve credibility, the medical expert must be included in the interpretation loop (75).

Together with models, we should also develop strategies to objectively evaluate their interpretability. This can be done by assessing fairness; privacy of data; generalizability; causality, to

prevent spurious correlations; or trust, to make sure the model is right for the right reasons (61). Depending on the application, the interpretability needs might be different. For cardiology applications in particular, ML for status interpretation and decision-making should be equipped with the most-advanced interpretability tools.

### Causal ML Rather Than Predictive ML

Predictive ML based on correlations of input data and outcomes may not be enough to truly impact the healthcare system. Indeed, this form of learning can be misleading if important causal variables are not analysed. For auditability reasons, we should probably shift toward finding the root causes of *why* that decision was made, and interpreting the process followed by the algorithm to reach that (diagnostic) decision, i.e., *how* the diagnosis was made. These two questions are addressed by causal ML, a powerful type of analysis aimed at inferring the mechanisms of the system producing the output data. In practice, causal models provide detailed maps of interaction between variables, so the users can simulate cause and effect of future actions (76).

### System-Related Security

Machine learning raises a handful of data security and privacy issues, as DL models require enormous datasets for training purposes. The most secure way to transfer data between healthcare organizations is still unclear, and stakeholders no longer underestimate the hazards of a high-profile data leakage. Hacking is even more harmful, as hackers could manipulate a decision-making model to damage people at a large scale.

The European GDPR compels to adopt security measures against hacking and data breaches. A potential solution that has been largely discussed is Blockchain, a technology that enables data exchange systems that are cryptographically secured and irrevocable, by providing a public and immutable log of transactions and “smart contracts” to regulate data access. The downsides of Blockchain’s technology are that it is slow, costly to maintain, and hard to scale (77). As an alternative, federated learning could guarantee the security of patient data (see “Validation and continuous improvement subsection”), as this model-training paradigm allows updating a learning model locally without sharing individual information with a central system (78).

### Regulatory

The use of ML for clinical decision-making unavoidably brings legal challenges regarding medical negligence derived from learning failures. When such negligence arises, the legal system needs to provide guidance on what entity holds liability, for which recommendations have been developed (79). Furthermore, the evolving nature of ML models poses a unique challenge to regulatory agencies, and the best way to evaluate updates remains unclear (60). Policymakers should generate specific criteria for demonstrating non-inferiority of algorithms compared to existing standards, specially emphasizing the quality of the training data and the validation process (80). Regulatory bodies must also ensure that algorithms are used properly and for

people’s welfare. In summary, for ML technology to be adopted by cardiology departments within the next years many legal aspects still need to be addressed, and decision- and policy-makers should join efforts toward this end.

### Integration (Man/Machine Coexistence)

The scenario of ML tools replacing humans in clinical medicine is highly unlikely (81). Besides the formidable challenges for ML solutions discussed above (8), cardiologists will still be needed to interact with the patients and perform physical examinations, navigate diagnostic procedures, integrate and adapt ML solutions according to the changing stages of disease or patient’s preferences, inform the patient’s family about therapy options, or console them if the disease stage is very advanced.

Accordingly, instead of a human-machine competition, we should rather think of a cooperation paradigm, where ML is used to augment human intelligence—targeting repetitive sub-tasks to assist physicians to reach a more informed decision, more efficiently. Indeed, ML and humans possess complementary skills: ML stands out at pattern recognition on massive amounts of data, whereas people are far better at understanding the context, abstracting knowledge from their experience, and transferring it across domains. Human-in-the-loop approaches facilitate cooperation by enabling users to interact with ML models without requiring in-depth technical knowledge. However, understanding where ML models can be used and at which level is crucial to avoid preventable errors attributed to automation bias (43). Examples of human-machine collaboration already exist. Indeed, a ML algorithm cleared by the FDA improved the diagnosis of wrist fractures when clinicians used it, as compared to clinicians alone (82). In diabetic retinopathy diagnosis (83), model assistance increased the accuracy of retina specialists above that of the unassisted reader or model alone. In cardiovascular imaging, most examples of human-machine collaboration thus far focus on segmentation, and detection-classification of imaging planes (84).

In light of this, the current clinical workflow could be rethought: the ML system would propose a diagnosis, the human operator then revises the data on which the conclusions are drawn, informing the system of potential measurement errors or confounders, and finally accepts or rejects the diagnosis. Thus, the human operator preserves the overall control, while machines perform measurements and integrate and compare data at request (75). Ultimately, this human-machine symbiosis will be beneficial to release physicians from low-level tasks such as cardiac measurements, data preparation, and standardization, to give them more time on higher-level tasks such as patient care and clinical decision-making (85).

### ML Applied to Real Clinical Data

In human decision-making, a clinician would explore all available data and compare them to patients they have seen before or were trained to recognize. Once an individual is put into context with regards to expected normality and typical cases, previous knowledge on treatment effect is used to manage this individual patient. This ‘eminence-based’ approach is only within reach of very experienced clinicians. For standardization, many



professional organizations provide diagnostic guidelines based on data from large cohorts or clinical trials (86–88). Although guidelines have significantly contributed to improve medical care, they do not consider the full data available. In this sense, the use of ML seems amply justified.

Most ML models are trained with data collected following strict input criteria and well-defined protocols used in randomized clinical trials (89, 90), but routinely collected data is often much noisier, heterogeneous and incomplete. ML techniques need to deal with incompleteness, either by performing imputation or by adopting formulations that explicitly consider that the data can be incomplete. Furthermore, patients often lie outside the narrow selection criteria of cardiology trials (including co-morbidities, ethnicity, gender, age, lifestyle, etc.), may have been differently treated before the investigation, may present at a different stage of disease, and most importantly, may undergo different decision pathways during the study. On top of this, obtaining a hard outcome to train an algorithm is often difficult, e.g., to register death, the study would need to be conducted until everybody dies, which is unfeasible both for time and economic constraints. Even if registered, often the outcome is scarce, and when appearing, the reason for experiencing it may be different among patients (91).

All these aspects make it extremely challenging to associate input descriptors to outcomes using supervised predictive ML/DL techniques, which may fail to understand the context from which data have been drawn, and thus yield unwanted results that might harm patients. A more promising approach could be based on unsupervised dimensionality reduction, a label-agnostic approach where input descriptors are used to position individuals according to their similarity and combined with previous knowledge this similarity can be used to infer diagnosis or to predict treatment response (23).

## FUTURE PERSPECTIVES

The foreseeable application of ML in the short to mid-term is to perform specific and well-defined tasks relating to data acquisition, predominantly by extracting standardized, high-quality information with the smallest possible learning curve. In this sense, DL solutions already help extracting information with minimal or even without the need of human intervention (8, 92), or aid selecting the images that are good enough for subsequent clinical interrogation (93). Another evident application of ML that will soon be ubiquitous in clinical practice is that of image analysis, which will discharge cardiologists from monotonous activities related to feature extraction from images (94), thus freeing them up to dedicate more time to higher-level tasks involving interpretation, patient care, and decision-making.

For the topics covered in this paper, i.e., the higher-level tasks involving reasoning, such as patient's status interpretation and decision support, ML applications are still immature and need substantial validation. A modest number of ongoing clinical trials have been conceived to tackle these drawbacks. One such example is the current investigation aiming at validating a DL model that diagnoses different arrhythmias (AF, supraventricular tachycardia, AV-block, asystole, ventricular tachycardia and ventricular fibrillation) on 12-lead ECGs and

single-lead Holter monitoring registered in 25,458 participants (95). Another example is the clinical study that will interrogate stress echocardiography scans with ML models to discriminate normal hearts from those at risk of a heart attack in a prospective cohort of 1,250 participants (96). Considering the above, we do not expect to see a vast penetration of ML-enabled applications for patient's status interpretation and decision support in clinical practice in the foreseeable future.

Whatever the application, the penetration of ML models into routine practice will be subject to their seamless integration into the clinical decision pathway used by cardiologists. Furthermore, we consider that the upcoming policies for ML research in healthcare should address the challenges described in the previous section, which can only be achieved by multidisciplinary teams. On the algorithmic side, more attention should be dedicated to dealing with longitudinal data, and how to relate the ML conclusions with pathophysiological knowledge. Data integration and what is the best approach for dealing with incomplete data and outliers should be also surveyed. On the validation side, generalization performance should be systematically reported, and uncertainty quantification methods should be developed to establish trust in the (predictive) models. Finally, the practical considerations that will affect adoption of the ML technology, such as how ML software should be integrated with the archiving and communication system of the hospital or how it would be paid for by facilities, should be explored. For these, a clear demonstration of the cost-effectiveness of ML technology in healthcare systems and its positive impact on patients' outcomes is needed.

## CONCLUSION

ML algorithms allow computers to automatically discover patterns in the data and improve with experience. Together with the enormous computational capacities of modern servers and the overwhelming amount of data resulting from the digitalization of healthcare systems, these algorithms open the door for a paradigm shift in clinical decision-making in cardiology. However, their seamless integration is dependent on the understanding of the intrinsic processes being used during the conventional pathway by which clinicians make decisions, which in turn helps identifying the areas where certain types of ML models can be most beneficial. If the obstacles and pitfalls that have been covered in this paper can be addressed satisfactorily, then ML might indeed revolutionize many aspects of healthcare, including cardiovascular medicine. For the promise to be fulfilled, engineers and clinicians will need to engage jointly in intensive development and validation of specific ML-enabled clinical applications.

## AUTHOR CONTRIBUTIONS

SS-M participated in the conception and design of the review, drafted the work, approved the final version and agreed on the accuracy and integrity of the work. OC, GP, MC, MG-B, MM, AV, EG, and AF helped drafting the review, revised it critically

for important intellectual content, approved the final version, and agreed on the accuracy and integrity of the work. BB participated in the conception and design of the review, drafted the work, approved the final version, and agreed on the accuracy and integrity of the work. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was supported by the Spanish Ministry of Economy and Competitiveness (María de Maeztu Programme for R&D

[MDM-2015-0502], Madrid, Spain) and by the Fundació La Marató de TV3 (n°20154031, Barcelona, Spain). The work of SS-M was supported by IDIBAPS and by the HUMAINT project of the Joint Research Centre of the European Commission. AV's contribution is funded by Spanish research project PID2019-104551RB-I00.

## ACKNOWLEDGMENTS

The author thanks to Nicolas Duchateau, who carefully read this manuscript and provided valuable feedback.

## REFERENCES

- Smuha N, AI, HLEG. High-level expert group on artificial intelligence. In: *Ethics Guidelines for Trustworthy AI*. (2019).
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. (2001) 409:860–921.
- Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. (2016) 316:2402. doi: 10.1001/jama.2016.17216
- Nie D, Cao X, Gao Y, Wang L, Shen D. “Estimating CT image from MRI data using 3d fully convolutional networks,” In: *International Workshop on Deep Learning in Medical Image Analysis*. (Cham: Springer), 170–178.
- Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *npj Digit Med*. (2018) 1:6. doi: 10.1038/s41746-017-0013-1
- Desai AS, Jhund PS, Yancy CW, Lopatin M, Stevenson LW, Marco T De, et al. After TOPCAT: what to do now in heart failure with preserved ejection fraction. *Eur Heart J*. (2016) 47:1510–8. doi: 10.1093/eurheartj/ehw114
- Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. (2017) 36:61–78. doi: 10.1016/j.media.2016.10.004
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. (2019) 25:44–56. doi: 10.1038/s41591-018-0300-7
- Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, et al. Artificial intelligence in cardiovascular imaging: jacc state-of-the-art review. *J Am Coll Cardiol*. (2019) 73:1317–35. doi: 10.1016/j.jacc.2018.12.054
- Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, et al. The “digital twin” to enable the vision of precision cardiology state of the art review. *Eur Heart J*. (2020) 41:4556–64. doi: 10.1093/eurheartj/ehaa159
- Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digit Health*. (2019) 1:e271–97. doi: 10.1016/S2589-7500(19)30123-2
- Fonseca CG, Backhaus M, Bluemke DA, Britten RD, Chung J, Cowan BR, et al. The Cardiac Atlas Project—an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics*. (2011) 27:2288–95. doi: 10.1093/bioinformatics/btr360
- Rohé M-M, Datar M, Heimann T, Sermesant M, Pennec X. SVF-Net: learning deformable image registration using shape matching. *Lecture Notes Comput Sci*. (2017) 10433:266–74. doi: 10.1007/978-3-319-66182-7\_31
- Toth D, Miao S, Kurzendorfer T, Rinaldi CA, Liao R, Mansi T, et al. 3D/2D model-to-image registration by imitation learning for cardiac procedures. *Int J Comput Assist Radiol Surg*. (2018) 13:1141. doi: 10.1007/s11548-018-1774-y
- Krebs J, Delingette H, Mailhe B, Ayache N, Mansi T. Learning a probabilistic model for diffeomorphic registration. *IEEE Trans Med Imaging*. (2019) 38:2165–76. doi: 10.1109/TMI.2019.2897112
- Krittawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Rep*. (2020) 10:1–11. doi: 10.1038/s41598-020-72685-1
- Arsanjani R, Xu Y, Dey D, Vahistha V, Shalev A, Nakanishi R, et al. Improved accuracy of myocardial perfusion SPECT for detection of coronary artery disease by machine learning in a large population. *J Nucl Cardiol*. (2013) 20:553–62. doi: 10.1007/s12350-013-9706-2
- Beecy AN, Chang Q, Anchouche K, Baskaran L, Elmore K, Kolli K, et al. A novel deep learning approach for automated diagnosis of acute ischemic infarction on computed tomography. *JACC: Cardiovascular Imaging*. (2018) 11:1723–5. doi: 10.1016/j.jcmg.2018.03.012
- Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Med*. (2018) 1:59. doi: 10.1038/s41746-018-0065-x
- Jiang H, Mao H, Lu H, Lin P, Garry W, Lu H, et al. Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *Int J Med Inform*. (2021) 145:104326. doi: 10.1016/j.ijmedinf.2020.104326
- Dey D, Gaur S, Ovrehus KA, Slomka PJ, Betancur J, Goeller M, et al. Integrated prediction of lesion-specific ischaemia from quantitative coronary CT angiography using machine learning: a multicentre study. *Euro Radiol*. (2018) 28:2655–2664. doi: 10.1007/s00330-017-5223-z
- Loncaric F, Marti Castellote P-M, Sanchez-Martinez S, Fabijanovic D, Nunno L, Mimbreno M, et al. Automated pattern recognition in whole-cardiac cycle echocardiographic data: capturing functional phenotypes with machine learning. *J Am Soc Echocardiograph*. (2021) 6:14. doi: 10.1016/j.echo.2021.06.014
- Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenotyping in heart failure to identify responders to cardiac resynchronization therapy. *Eur J Heart Fail*. (2019) 21:74–85. doi: 10.1002/ejhf.1333
- Nogueira M, De Craene M, Sanchez-Martinez S, Chowdhury D, Bijmens B, Piella G. Analysis of nonstandardized stress echocardiography sequences using multiview dimensionality reduction. *Med Image Anal*. (2020) 60:101594. doi: 10.1016/j.media.2019.101594
- Funkner AA, Yakovlev AN, Kovalchuk S V. Data-driven modeling of clinical pathways using electronic health records. *Procedia Comput Sci*. (2017) 121:835–42. doi: 10.1016/j.procs.2017.11.108
- Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. *AMIA Sympo Proceed*. (2017) 2017:1312–21.
- Kwon J, Kim K-H, Jeon K-H, Park J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography*. (2019) 36:213–8. doi: 10.1111/echo.14220
- Arsanjani R, Dey D, Khachatryan T, Shalev A, Hayes SW, Fish M, et al. Prediction of revascularization after myocardial perfusion spect by machine learning in a large population. *J Nuclear Cardiol Off Publ Am Soc Nucl Cardiol*. (2015) 22:877. doi: 10.1007/s12350-014-0027-x
- Betancur J, Otaki Y, Motwani M, Fish MB, Lemley M, Dey D, et al. Prognostic value of combined clinical and myocardial perfusion imaging

- data using machine learning. *JACC: Cardiovascul Imaging*. (2018) 11:1000–9. doi: 10.1016/j.jcmg.2017.07.024
30. Zeleznik R, Foldyna B, Eslami P, Weiss J, Alexander I, Taron J, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nature Commun*. (2021) 12:1–9. doi: 10.1038/s41467-021-20966-2
  31. Diller G-P, Kempny A, Babu-Narayan S V, Henrichs M, Brida M, Uebing A, et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *Eur Heart J*. (2019) 40:1069–77. doi: 10.1093/eurheartj/ehy915
  32. Wang J, Ding H, Bidgoli FA, Zhou B, Iribarren C, Molloy S, et al. Detecting cardiovascular disease from mammograms with deep learning. *IEEE Trans Med Imaging*. (2017) 36:1172–81. doi: 10.1109/TMI.2017.2655486
  33. Poplin R, Varadarajan A V., Blumer K, Liu Y, McConnell M V, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. (2018) 2:158–64. doi: 10.1038/s41551-018-0195-0
  34. Rim TH, Lee CJ, Tham Y-C, Cheung N, Yu M, Lee G, et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *Lancet Digit Health*. (2021) 3:e306–16. doi: 10.1016/S2589-7500(21)00043-1
  35. Retson TA, Masutani EM, Golden D, Hsiao A. Clinical performance and role of expert supervision of deep learning for cardiac ventricular volumetry: a validation study. *Radiol Artif Intell*. (2020) 2:e190064. doi: 10.1148/ryai.2020190064
  36. Karagodin I, Singulane CC, Woodward GM, Xie M, Tucay ES, Rodrigues ACT, et al. Echocardiographic correlates of in-hospital death in patients with acute Covid-19 infection: the world alliance societies of echocardiography (WASE-COVID) study. *J Am Soc Echocardiograph*. (2021) 34:819–30. doi: 10.1016/j.echo.2021.05.010
  37. Asch FM, Mor-Avi V, Rubenson D, Goldstein S, Saric M, Mikati I, et al. Deep learning-based automated echocardiographic quantification of left ventricular ejection fraction: a point-of-care solution. *Circul Cardiovascul Imaging*. (2021) 21:528–537. doi: 10.1161/CIRCIMAGING.120.012293
  38. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*. (2018) 363:4245. doi: 10.1136/bmj.k4245
  39. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux A V, Folsom AR, et al. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol*. (2002) 156:871–81. doi: 10.1093/aje/kwf113
  40. Nogueira M, Piella G, Sanchez-Martinez S, Langet H, Saloux E, Bijmens B, et al. “Characterizing patterns of response during mild stress-testing in continuous echocardiography recordings using a multiview dimensionality reduction technique,” In *Functional Imaging and Modelling of the Heart—Conference Proceedings*.
  41. Lee G, Kang B, Nho K, Sohn K-A, Kim D. MildInt: deep learning-based multimodal longitudinal data integration framework. *Front Genet*. (2019) 10:617. doi: 10.3389/fgene.2019.00617
  42. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*. (2020) 2:e489–92. doi: 10.1016/S2589-7500(20)30186-2
  43. Cummings ML. “Automation bias in intelligent time critical decision support systems,” In: *AIAA 3rd Intelligent Systems conference*. (2004), p. 6313.
  44. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc : JAMIA*. (2012) 19:121–7. doi: 10.1136/amiainjnl-2011-000089
  45. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DS, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARpN consortium. *J Am Med Inform Assoc*. (2013) 20:e341–8. doi: 10.1136/amiainjnl-2013-001939
  46. Wynants L, Kent DM, Timmerman D, Lundquist CM, Calster B. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. *Diagnostic Prognostic Res*. (2019) 3:9. doi: 10.1186/s41512-019-0046-9
  47. Tat E, Bhatt DL, Rabbat MG. Addressing bias: artificial intelligence in cardiovascular medicine. *The Lancet Digital Health*. (2020) 2:e635–6. doi: 10.1016/S2589-7500(20)30249-1
  48. Kagiya N, Shrestha S, Farjo PD, Sengupta PP. Artificial intelligence: practical primer for clinical research in cardiovascular disease. *J Am Heart Assoc*. (2019) 8:12788. doi: 10.1161/JAHA.119.012788
  49. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. (2017) 356:183–6. doi: 10.1126/science.aal4230
  50. Wilkinson C, Bebb O, Dondo TB, Munyombwe T, Casadei B, Clarke S, et al. Sex differences in quality indicator attainment for myocardial infarction: a nationwide cohort study. *Heart*. (2019) 105:516–23. doi: 10.1136/heartjnl-2018-313959
  51. Klieger T, Stěpán Bahník, Fůrnkranz J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *AI*. (2021) 103458. doi: 10.1016/j.artint.2021.103458
  52. Brisk R, Bond R, Finlay D, McLaughlin J, Peadro A, Leslie SJ, et al. The effect of confounding data features on a deep learning algorithm to predict complete coronary occlusion in a retrospective observational setting. *Euro Heart J Digit Health*. (2021) 2:127–34. doi: 10.1093/ehjdh/ztab002
  53. Hastie T, Tibshirani R, Friedman J. *The elements of Statistical Learning (Vol.1)*. Springer, Berlin: Springer series in statistics (2001).
  54. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol bed to bench*. (2012) 5:79–83.
  55. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol*. (2019) 20:405. doi: 10.3348/kjr.2019.0025
  56. Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, et al. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. *Cardiovasc Res*. (2020) 116:2216–25. doi: 10.1093/cvr/cvz321
  57. Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pelikka PA, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol*. (2019) 30:668–74. doi: 10.1111/jce.13889
  58. Surette S, Narang A, Bae R, Hong H, Thomas Y, Cadieu C, et al. Artificial intelligence-guided image acquisition on patients with implanted electrophysiological devices: results from a pivotal prospective multi-center clinical trial. *Euro Heart J*. (2020) 41:6. doi: 10.1093/ehjci/ehaa946.0006
  59. Konečný J, McMahan HB, Yu FX, Richtarik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. *arXiv [preprint]*. (2016)
  60. FDA. 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. (2019).
  61. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv [preprint]*. (2017).
  62. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. (2017) 318:517. doi: 10.1001/jama.2017.7797
  63. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a “right to explanation.” *AI Magazine*. (2017) 38:2741. doi: 10.1609/aimag.v38i3.2741
  64. Croskerry P. A. universal model of diagnostic reasoning. *Acad Med*. (2009) 84:1022–8. doi: 10.1097/ACM.0b013e3181ace703
  65. Benavidez OJ, Gauvreau K, Geva T. Diagnostic errors in congenital echocardiography: importance of study conditions. *J Am Soc Echocardiograph*. (2014) 27:616–23. doi: 10.1016/j.echo.2014.03.001
  66. Balogh EP, Miller BT, Ball JR. *Improving diagnosis in health care and M National Academies of Sciences, Engineering*. Washington, DC: National Academies Press. (2015).
  67. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng*. (2018) 1:9. doi: 10.1038/s41551-018-0324-9
  68. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Med*. (2018) 1:18. doi: 10.1038/s41746-018-0029-1

69. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng.* (2018) 2:749–60. doi: 10.1038/s41551-018-0304-0
70. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (2016). p. 1135–44. doi: 10.1145/2939672.2939778
71. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x
72. Mallat S. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* (2016) 374:20150203. doi: 10.1098/rsta.2015.0203
73. Kumar D, Taylor GW, Wong A. Discovery radiomics with CLEAR-DR: interpretable computer aided diagnosis of diabetic retinopathy. *IEEE Access.* (2017) 7:25891–6. doi: 10.1109/ACCESS.2019.2893635
74. Wang J, Oh J, Wang H, Wiens J. Learning credible models. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* (2018) 2417–26. doi: 10.1145/3219819.3220070
75. D’hooge J, Fraser AG. Learning about machine learning to create a self-driving echocardiographic laboratory. *Circulation.* (2018) 138:1636–8. doi: 10.1161/CIRCULATIONAHA.118.037094
76. Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause And Effect.* London: Basic Books, Inc. (2018).
77. Song J. Why blockchain is hard. *Medium.* (2018)
78. Silva S, Gutman B, Romero E, Thompson PM, Altmann A, Lorenzi M. Federated learning in distributed medical databases: meta-analysis of large-scale subcortical brain data. In: *2019 IEEE 16th International Symposium on Biomedical Imaging.* IEEE (2018). p. 270–4. doi: 10.1109/ISBI.2019.8759317
79. Expert Group on Liability and New Technologies. *Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies - European Commission.* (2019).
80. Yu K, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng.* (2018) 2:719–31. doi: 10.1038/s41551-018-0305-z
81. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *NE J Med.* (2016) 375:1216–9. doi: 10.1056/NEJMp1606181
82. Voelker R. Diagnosing fractures with AI. *JAMA.* (2018) 320:23. doi: 10.1001/jama.2018.8565
83. Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology.* (2018) 11:16. doi: 10.1016/j.ophtha.2018.11.016
84. Soares De Siqueira V, De Castro Rodrigues D, Nunes Dourado C, Marcos Borges M, Gomes Furtado R, Pereira Delfino H, et al. “Machine learning applied to support medical decision in transthoracic echocardiogram exams: a systematic review.” In: *Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020.* (2020), p. 400–407.
85. Sengupta PP, Adjeroh DA. Will artificial intelligence replace the human echocardiographer? *Circulation.* (2018) 138:1639–42. doi: 10.1161/CIRCULATIONAHA.118.037095
86. Ponikowski P, Voors AA, Anker SD, Bueno H, González-Juanatey JR, Harjola V-P, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J.* (2016) 37:2129–200. doi: 10.1093/eurheartj/ehw128
87. Brignole M, Auricchio A, Baron-Esquivias G, Bordachar P, Boriani G, Breithardt OA, et al. ESC Guidelines on cardiac pacing and cardiac resynchronization therapy. *Eur Heart J.* (2013) 15:1070–118. doi: 10.15829/1560-4071-2014-4-5-63
88. Epstein AE, Dimarco JP, Ellenbogen KA, Estes NAM, Freedman RA, Gettes LS, et al. ACC/AHA/HRS 2008 Guidelines of cardiac rhythm abnormalities. A report of the American College of Cardiology/American Heart Association task force on practice guidelines (writing committee to revise the ACC/AHA/NASPE 2002 Guideline update for implantation. *Circulation.* (2008) 117:350–408.
89. Kalscheur MM, Kipp RT, Tattersall MC, Mei C, Buhr KA, DeMets DL, et al. Machine Learning Algorithm Predicts Cardiac Resynchronization Therapy Outcomes. *Circulation: Arrhythm Electrophysiol.* (2018) 11:499. doi: 10.1161/CIRCEP.117.005499
90. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Gregory Hundley W, McClelland R, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ Res.* (2017) 121:1092–101. doi: 10.1161/CIRCRESAHA.117.311312
91. Oladapo OT, Souza JP, Bohren MA, Tunçalp Ö, Vogel JP, Fawole B, et al. Better Outcomes in Labour Difficulty (BOLD) project: innovating to improve quality of care around the time of childbirth. *Reprod Health.* (2015) 12:48. doi: 10.1186/s12978-015-0027-6
92. Housden J, Wang S, Noh Y, Singh D, Singh A, Back J, et al. “Control strategy for a new extra-corporeal robotic ultrasound system,” In: *MPEC* (2017).
93. Wu L, Cheng J-Z, Li S, Lei B, Wang T, Ni D, et al. fetal ultrasound image quality assessment with deep convolutional networks. *IEEE Trans Cybern.* (2017) 47:1336–49. doi: 10.1109/TCYB.2017.2671898
94. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation.* (2018) 138:1623–35. doi: 10.1161/CIRCULATIONAHA.118.034338
95. NCT03662802. *Development of a Novel Convolution Neural Network for Arrhythmia Classification.* *ClinicalTrials.gov.* Available online at: [https://clinicaltrials.gov/ct2/show/NCT03662802?term=Machine\\$+\\$Learning%2C\\$+\\$AI&draw=3&rank=29](https://clinicaltrials.gov/ct2/show/NCT03662802?term=Machine$+$Learning%2C$+$AI&draw=3&rank=29)
96. NCT04193475. *Machine Learning in Quantitative Stress Echocardiography.* *ClinicalTrials.gov.* Available online at: [https://clinicaltrials.gov/ct2/show/NCT04193475?term=machine\\$+\\$learning%2C\\$+\\$echocardiography&draw=2&rank=1](https://clinicaltrials.gov/ct2/show/NCT04193475?term=machine$+$learning%2C$+$echocardiography&draw=2&rank=1)

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Sanchez-Martinez, Camara, Piella, Cikes, González-Ballester, Miron, Vellido, Gómez, Fraser and Bijmens. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.