






Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations

Sarah Friedrich, Stefan Groß, Inke R. König, Sandy Engelhardt, Martin Bahls, Judith Heinz, Cynthia Huber, Lars Kaderali, Marcus Kelm, Andreas Leha, Jasmin Rühl, Jens Schaller, Clemens Scherer, Marcus Vollmer, Tim Seidler, Tim Friede

Angaben zur Veröffentlichung / Publication details:

Friedrich, Sarah, Stefan Groß, Inke R. König, Sandy Engelhardt, Martin Bahls, Judith Heinz, Cynthia Huber, et al. 2021. "Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations." *European Heart Journal - Digital Health* 2 (3): 424–36.
<https://doi.org/10.1093/ehjdh/ztab054>.

Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations

Sarah Friedrich ^{1*}, Stefan Groß ^{2,3}, Inke R. König^{4,5}, Sandy Engelhardt^{6,7,8}, Martin Bahls^{2,3}, Judith Heinz¹, Cynthia Huber¹, Lars Kaderali^{3,9}, Marcus Kelm ^{10,11,12,13}, Andreas Leha^{1,14}, Jasmin Rühl¹, Jens Schaller^{10,13}, Clemens Scherer ^{15,16}, Marcus Vollmer^{3,9}, Tim Seidler^{14,17}, and Tim Friede ^{1,14}

¹Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany; ²Department of Internal Medicine B, University Medicine Greifswald, Ferdinand-Sauerbruch-Straße, 17475 Greifswald, Germany; ³DZHK (German Centre for Cardiovascular Research), Partner Site Greifswald, Greifswald, Germany; ⁴Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany; ⁵DZHK (German Centre for Cardiovascular Research), Partner Site Hamburg/Kiel/Lübeck, Lübeck, Germany; ⁶Department of Internal Medicine III, University Hospital Heidelberg, Im Neuenheimer Feld 420, 69120 Heidelberg, Germany; ⁷DZHK (German Centre for Cardiovascular Research), Partner Site Mannheim/Heidelberg, Heidelberg, Germany; ⁸Informatics for Life, Heidelberg, Germany; ⁹Institute of Bioinformatics, University Medicine Greifswald, Felix-Hausdorff-Str. 8, 17475 Greifswald, Germany; ¹⁰Institute for Imaging Science and Computational Modelling in Cardiovascular Medicine, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; ¹¹Department of Congenital Heart Disease, Deutsches Herzzentrum Berlin (DHZB), Berlin, Germany; ¹²Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany; ¹³DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin, Germany; ¹⁴DZHK (German Centre for Cardiovascular Research), Partner Site Göttingen, Göttingen, Germany; ¹⁵Department of Medicine I, University Hospital, LMU Munich, Marchioninistr. 15, 81377 München, Germany; ¹⁶DZHK (German Centre for Cardiovascular Research), Partner Site Munich, Munich, Germany; and ¹⁷Clinic for Cardiology and Pneumology, University Medical Center Göttingen, Göttingen, Germany

Received 5 March 2021; revised 21 April 2021; editorial decision 7 June 2021; accepted 7 June 2021; online publish-ahead-of-print 8 June 2021

Aims

Artificial intelligence (AI) and machine learning (ML) promise vast advances in medicine. The current state of AI/ML applications in cardiovascular medicine is largely unknown. This systematic review aims to close this gap and provides recommendations for future applications.

Methods and results

Pubmed and EMBASE were searched for applied publications using AI/ML approaches in cardiovascular medicine without limitations regarding study design or study population. The PRISMA statement was followed in this review. A total of 215 studies were identified and included in the final analysis. The majority (87%) of methods applied belong to the context of supervised learning. Within this group, tree-based methods were most commonly used, followed by network and regression analyses as well as boosting approaches. Concerning the areas of application, the most common disease context was coronary artery disease followed by heart failure and heart rhythm disorders. Often, different input types such as electronic health records and images were combined in one AI/ML application. Only a minority of publications investigated reproducibility and generalizability or provided a clinical trial registration.

Conclusions

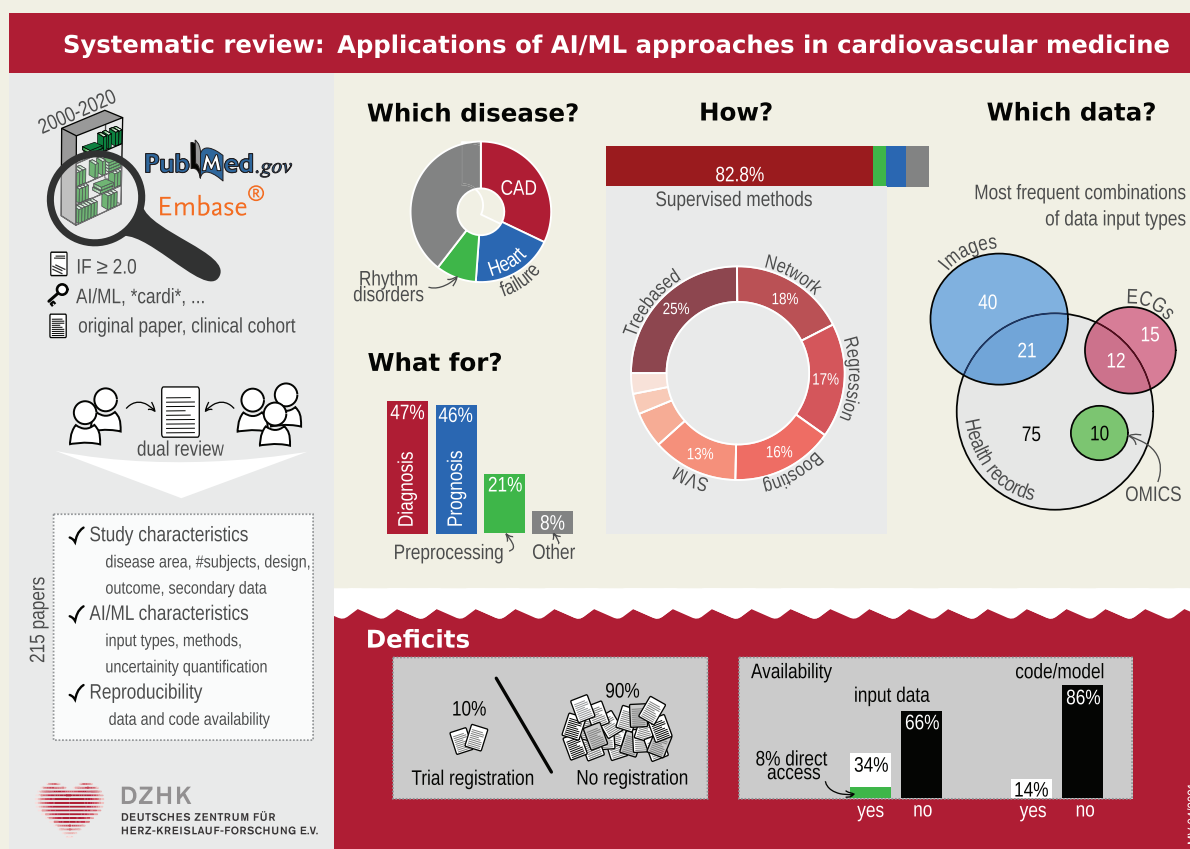
A major finding is that methodology may overlap even with similar data. Since we observed marked variation in quality, reporting of the evaluation and transparency of data and methods urgently need to be improved.

* Corresponding author. Tel: +0049-551-3964064, Email: sarah.friedrich@med.uni-goettingen.de

© The Author(s) 2021. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Graphical Abstract



Keywords Artificial intelligence • Machine learning • Cardiology • Cardiovascular disease

Introduction

Despite significant improvements over the last decades, cardiovascular diseases remain the leading cause of morbidity and mortality in Europe and the USA.^{1,2} Due to complex disease pathways and heterogeneity, disease diagnostic and prognostic assessment remain a challenging task. On the other hand, modern technologies are constantly increasing the ability to collect large quantities of data, which require implementation of comprehensive automated analytical methods to improve the understanding of the underlying disease complexity and ultimately increase the quality of healthcare.

Artificial intelligence (AI) is an overarching term that describes the use of algorithms and software which demonstrate human-like cognition in analysing, interpreting, and understanding complicated medical and health data. An algorithm is simply a set of actions to be followed to get a solution. Algorithms are trained to learn how to process information. The term AI may also be applied to any machine that exhibits traits associated with a human mind, such as learning and problem-solving. When machines can extract information from data, improve their function or make predictions about future events, they are referred to as machine learning (ML), a subset of AI.³ The overall

objective of these approaches is to learn from samples and to generalize to new, yet unseen cases. Machine learning includes a range of advanced sub-branches, such as deep learning (DL) and neural networks.

AI/ML methods achieved remarkable progress, and their use has increased significantly over the last years in cardiovascular medicine, as indicated by recently published reviews.^{3–10}

Compared to other reviews such as Johnson *et al.*,⁸ we chose a different approach: our intention was to investigate what is currently published under the label 'AI/ML' in cardiovascular medicine as opposed to providing specific examples of AI/ML applications in a given disease context or comparing the predictive ability of various AI/ML methods in a meta-analysis.¹⁰ As Nagendran *et al.*⁵ note, there is a danger that the '[...] public and commercial appetite for health-care AI outpaces the development of a rigorous evidence base to support this comparatively young field'. Many authors also criticize the lack of details published on AI/ML methods, which hinder reproducibility and transparency.^{11,12} Lopez-Jimenez *et al.*⁹ provide a list of key aspects for evaluating AI literature. Moreover, there is still '[...] a scarcity of external validation studies and randomised trials [...] to evaluate the superiority of using these methods.¹³ Hence, the

CONSORT-AI guidelines were published very recently to improve transparency and completeness in reporting clinical trials for AI interventions.¹⁴ In parallel, the SPIRIT-AI extension¹⁵ was developed as a new reporting guideline for clinical trial protocols evaluating interventions with an AI component. A relevant issue concerning AI applications is the current overemphasis on the technical aspects, which sometimes leaves less attention to their interaction with the human users, see the DECIDE-AI statement for a discussion of this issue.¹⁶

In this systematic review, we provide an overview of the literature on applications of AI/ML methods in cardiovascular research. In the following, we describe the exact search strategy. We provide our results including descriptions of the specific methods applied in different research settings. Additionally, we evaluate whether the methods used were appropriately described and if code/data availability statements were provided. We conclude with some recommendations regarding the reporting and evaluation of methods as well as improving data and methods transparency. Our broad focus on all methods described by the respective authors as AI or ML methods without restriction with regards to specific disease areas or study designs allows for a clear view on the current state of AI/ML applications in cardiovascular medicine. As such, we aim to provide clear recommendations on how AI/ML studies should be conducted in contrast to describing criteria for the evaluation of AI/ML literature as provided by Vollmer et al.¹² or Lopez-Jimenez et al.⁹

Methods

In this systematic review, clinical studies applying AI/ML approaches in cardiovascular medicine without limitations regarding study design or study population were included. To specifically focus on clinical application, we excluded animal studies as well as publications reporting only the methodological aspect of an AI/ML approach without presentation of clinical data of the study population. The systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement¹⁷ and is registered in PROSPERO (registration number CRD42020196696).

Systematic literature search and study selection

We performed a literature search using PubMed and EMBASE to identify relevant publications. A first search in PubMed using the search terms '(artificial intelligence [Title/Abstract] OR machine learning [Title/Abstract]) AND (cardiac OR cardiology OR cardiosurgical OR cardiology[MeSH] OR heart OR heart[MeSH] OR *cardia* OR *cardio* OR *infarct* OR *valve*)' and considering publications from the year 2000 onwards resulted in 2410 abstracts, see the PRISMA flow chart in [Supplementary material online, Figure S1](#). In order to restrict our search to applied clinical studies and to exclude purely methodological publications, we limited our search to journals listed on the Web of Science in the categories 'CARDIAC & CARDIOVASCULAR SYSTEMS', 'MEDICINE, GENERAL & INTERNAL', 'MEDICINE, RESEARCH & EXPERIMENTAL', as well as 'MULTIDISCIPLINARY SCIENCES'. To furthermore restrict the analysis to articles with potential clinical impact, we only included journals with an impact factor of at least two in 2018 as given on the Web of Science (<https://jcr.clarivate.com/JCRJournalHomeAction.action>). This resulted in 228 distinct journals. Articles included in our analyses were published in 65 distinct journals, see [Supplementary material online, Table S1](#) for the complete list of

journals. Within the cardiovascular journals, the search terms used were '(artificial intelligence [Title/Abstract] OR machine learning [Title/Abstract])', while in the other journals we searched for '(artificial intelligence [Title/Abstract] OR machine learning [Title/Abstract]) AND (cardiac OR cardiology OR cardiosurgical OR cardiology [MeSH] OR heart OR heart [MeSH] OR *cardia* OR *cardio* OR *infarct* OR *valve*)'. The date of the last search was 5 March 2020.

Studies discussed in review papers and commentaries or editorials were also screened, but no additional studies fulfilling our inclusion criteria (see below) were found.

The list of abstracts was independently screened for inclusion by dual review of overall 16 reviewers using the following inclusion criteria: (i) an application of AI/ML methods, (ii) cardiovascular application, and (iii) the study has to present a clearly described clinical cohort. In particular, we excluded animal studies and review papers. Any disagreements were resolved in discussion or rescreened by a third reviewer (S.F., T.F., C.H., and J.R.).

Data extraction

We extracted data on study characteristics and study population (number of subjects included, study design, outcome scale, use of secondary data), characteristics of the applied AI/ML techniques such as uncertainty quantification and comparison to traditional statistical methods, reproducibility, disease area, and type of input data. Information was collected by at least two independent reviewers using a predefined data extraction form. Secondary data were specified as data originally collected for a different purpose. Studies were classified to quantify uncertainty of the AI/ML methods if they provided any measure of uncertainty such as confidence intervals for area under the curve estimates. AI/ML methods were extracted as defined by the authors of the corresponding study and the superiority to classical methods was defined based on the authors' claims.

To investigate the increase of publications on AI/ML methods we also extracted the number of articles published per year in the journals considered in our literature search by counting the findings of the assigned journal IDs (using the ID of the bibliographic database of the National Library of Medicine: Nlmid) in PubMed as of 10 December 2020. Moreover, we extracted the number of citations of the reviewed papers (searched by its PMID in PubMed) as of 10 December 2020.

Since the investigations' aim was not always clearly stated in the articles, we applied a rather strict definition of prognostic and diagnostic approaches: we defined an approach as diagnostic when patients were classified or divided into subgroups without any time reference (e.g. publications 8 and 91 from [Supplementary material online, Table S1](#)), and as prognostic when there was a time reference (e.g. longitudinal outcomes in publications 50 and 167 from [Supplementary material online, Table S1](#)).

Data analysis

Descriptive summaries are used to describe study characteristics. Metric variables are characterized by median and interquartile range, while discrete variables are summarized by providing absolute and relative frequencies.

In order to enable comparisons and study interactions, the cardiovascular context is categorized into ten types, namely coronary artery disease (CAD), valvular heart disease, cardiomyopathies, heart rhythm disorders, peripheral vascular disease, hypertension, heart failure (HF), congenital heart disease, cardiometabolic, and other entities. These categories were chosen according to the disease context mentioned by the authors of the corresponding publication in title or abstract.

Similarly, the applied AI/ML methods are categorized into supervised, unsupervised and unspecified methods. The latter category refers to publications where the AI/ML approach was not mentioned or explained by

the authors, e.g. since a commercial software was used or due to statements such as 'we used a machine learning approach'. Additionally, the unsupervised methods are categorized into three and the supervised methods into eight sub-categories. A brief description of the methods along with some references is provided in [Supplementary material online, Table S2](#). [Supplementary material online, Table S3](#) shows an overview of the allocation of methods found in our search to the different sub-categories.

Interactions between categorical variables are presented as graphs displaying relative frequencies. All statistical analyses are performed in R 3.6.3 (R Foundation for statistical computing, Vienna, Austria).

Results

Included studies

The literature review identified 524 distinct publications that were screened for eligibility. A total of 215 studies were included in the final analysis, see also PRISMA flow chart in [Supplementary material online, Figure S1](#) as well as [Supplementary material online, Table S1](#) for the complete list of references. The study populations of the included publications are summarized in [Table 1](#), see [Supplementary material online, Table S4](#) for a stratified summary by disease area.

Time trends

Temporal trends in AI/ML applications in cardiovascular research were explored. We observed a relative and absolute increase in publications with AI/ML applications in the last years. [Figure 1A](#) shows the proportion of papers (in %) included in our study with respect to the total number of studies published in the considered journals. Note that the number for 2020 is based on the studies published until 5 March 2020 and is thus likely to be incomplete. A similar trend was observed for the articles we did not screen due to the restrictions of

our search strategy ([Supplementary material online, Figure S2](#)). [Figure 1B](#) displays the journals in which the identified articles were published most frequently, while [Figure 1C](#) shows the geographic distribution of the authors. With respect to the different AI/ML methods, no specific time trend could be observed ([Supplementary material online, Figure S3](#)).

Popular AI/ML methods and their areas of application

The majority (87%) of methods applied belonged to the context of supervised learning, see [Figure 2A](#). Within this group, tree-based methods were most commonly used, followed by network and regression analyses as well as boosting approaches. In 15 articles (7%), the authors did not describe the AI/ML approach in detail. In most of these cases, a commercial software was used. Among the unsupervised methods clustering was the most popular and included 67% of the unsupervised methods ([Figure 2A](#)). We also found that unsupervised and unspecified methods are more common when the AI/ML method is used for pre-processing than in other applications. In particular, supervised methods were applied in 60% of the studies that used AI/ML for pre-processing only as opposed to 92% in the other studies.

Concerning the areas of application, the most common disease context was CAD followed by HF and heart rhythm disorders as depicted in [Figure 2B](#). The most common input for the AI/ML methods were health records ([Figure 3](#)). Often, different types of input data were combined in one AI/ML-application. For example, 10 (5%) studies used omics data in combination with health records and 21 (10%) studies combined health records with images. [Figure 4](#) shows the distribution of the supervised methods applied in the most common disease areas (CAD, HF and all other diseases than CAD or HF) in more detail. In CAD, for example, boosting and regression methods are the most common methods of choice. In HF, on the other hand, tree-based methods are often used.

Comparison to non-AI/ML methods

Uncertainty of the AI/ML estimates was reported in 133 studies (62%). Results of AI/ML were compared to 'classical' methods (according to the authors' definition) in 111 (52%) studies. The majority of these, 94 (85%) decided in favour of the AI/ML approach, for example the work by Commandeur *et al.*¹⁸ (publication number 50 in [Supplementary material online, Table S1](#)) or the work by Leha *et al.*¹⁹ (publication number 91 in [Supplementary material online, Table S1](#)). Sample sizes reported in these studies are displayed in [Supplementary material online, Figure S4](#).

A positive example concerning the investigation of reproducibility and generalizability is the work from Bhuva *et al.*²⁰ (publication number 122 in [Supplementary material online, Table S1](#)) which describes a multicentre, scan-rescan cardiac magnetic resonance study to test generalizability for imaging biomarkers.

Trial registration and reproducible research

Only 21 studies (10%) provided a clinical trial registration; these were mostly randomized clinical trials. Only 3 (6%) of the prospective cohort studies and 11 (8%) of the retrospective cohort studies

Table 1 Study characteristics

Variable	Level	Total
Subjects	Median (IQR)	1083.0 (213.5–10 757.0)
Subject categories	<100	31 (14.4)
	100–1000	73 (34.0)
	1000–10 000	53 (24.7)
	10 000–100 000	45 (20.9)
	100 000–1 000 000	11 (5.1)
	>1 000 000	2 (0.9)
Design	Prospective cohort study	48 (22.3)
	Retrospective cohort study	138 (64.2)
	Case-control study	20 (9.3)
	RCT	9 (4.2)
Outcome	Binary	153 (71.2)
	Categorical	13 (6.0)
	Continuous	27 (12.6)
	Time to event	22 (10.2)
Secondary data	No	64 (29.8)
	Yes	151 (70.2)

Values are n (%) unless otherwise stated.

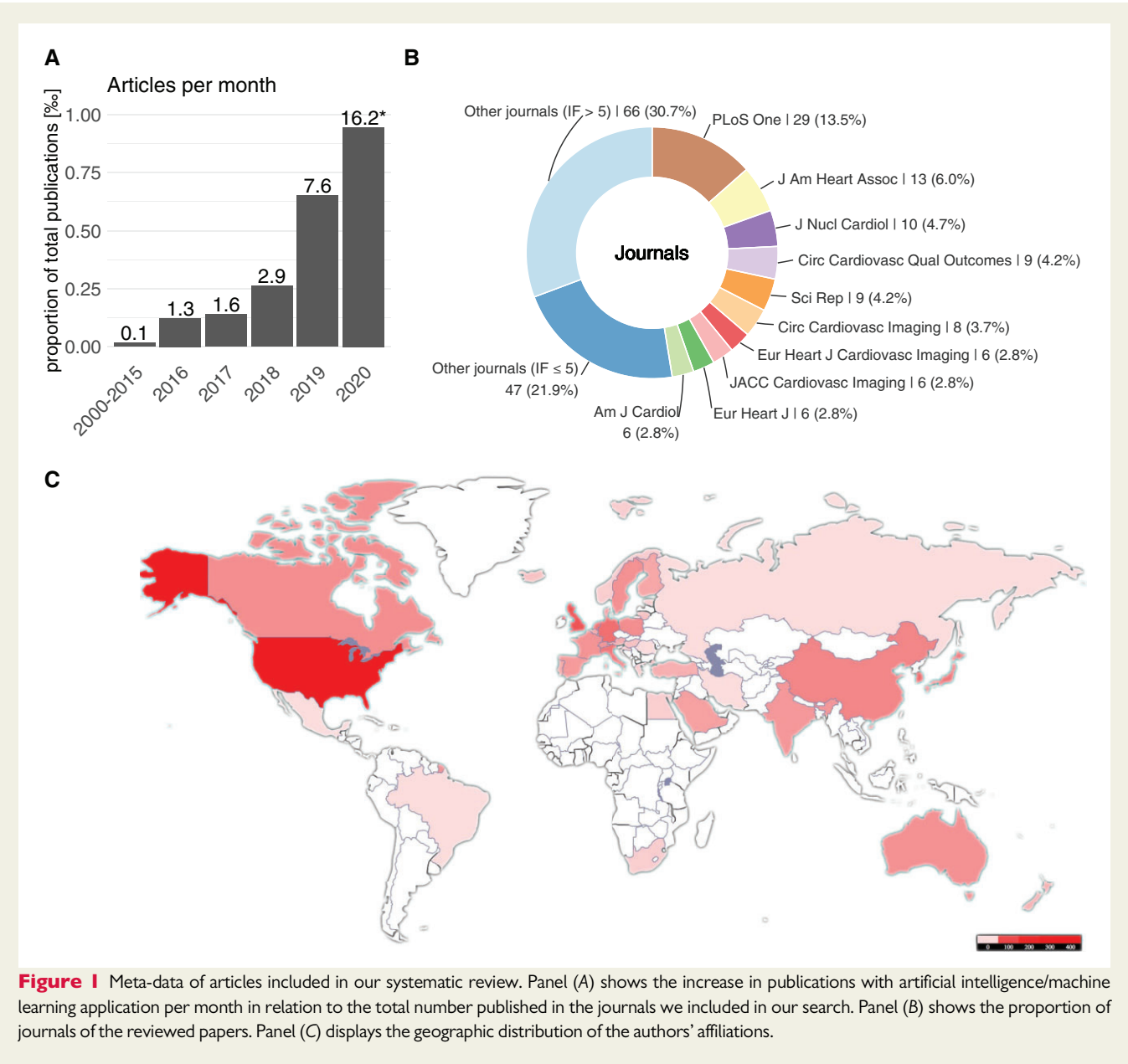


Figure 1 Meta-data of articles included in our systematic review. Panel (A) shows the increase in publications with artificial intelligence/machine learning application per month in relation to the total number published in the journals we included in our search. Panel (B) shows the proportion of journals of the reviewed papers. Panel (C) displays the geographic distribution of the authors' affiliations.

were registered. Of the case-control studies, 2 (10%) were registered.

Of the studies analysed, 73 (34%) stated that the data used for the analysis was available. However, only 17 studies (8%) provided direct access to the data. With respect to code, only 31 studies (14%) had made their code publicly available.

Prognostic vs. diagnostic analyses

According to the definition described above, the aim of the AI/ML approach was diagnostic in 91 articles (42%). Another 93 articles (43%) used the AI/ML approach to build prognostic models. Both diagnostic and prognostic aims were considered in only 4 (2%) of the studies. A total of 27 (13%) articles applied AI/ML algorithms for neither diagnostic nor prognostic models, but rather as part of the pre-processing, e.g. to extract features from images.

We found that methods such as clustering, k-nearest neighbour and network analyses were mainly used for diagnostic purposes, whereas prognostic models rather used tree-based approaches and regression models (data not shown).

Typical cases of AI/ML algorithms

Given the larger number of different methods and their applications in cardiovascular medicine, inspection of specific examples is helpful to understand how the use of AI/ML algorithms might have a potential benefit for clinical practice. Therefore, Table 2 lists typical cases, which were selected based on their high number of overall citations. Interestingly, the number of trial subjects differed widely in the top-cited publications. Furthermore, there was no general preference for one AI/ML method. Of note, none of the highly cited examples were randomized controlled trials.

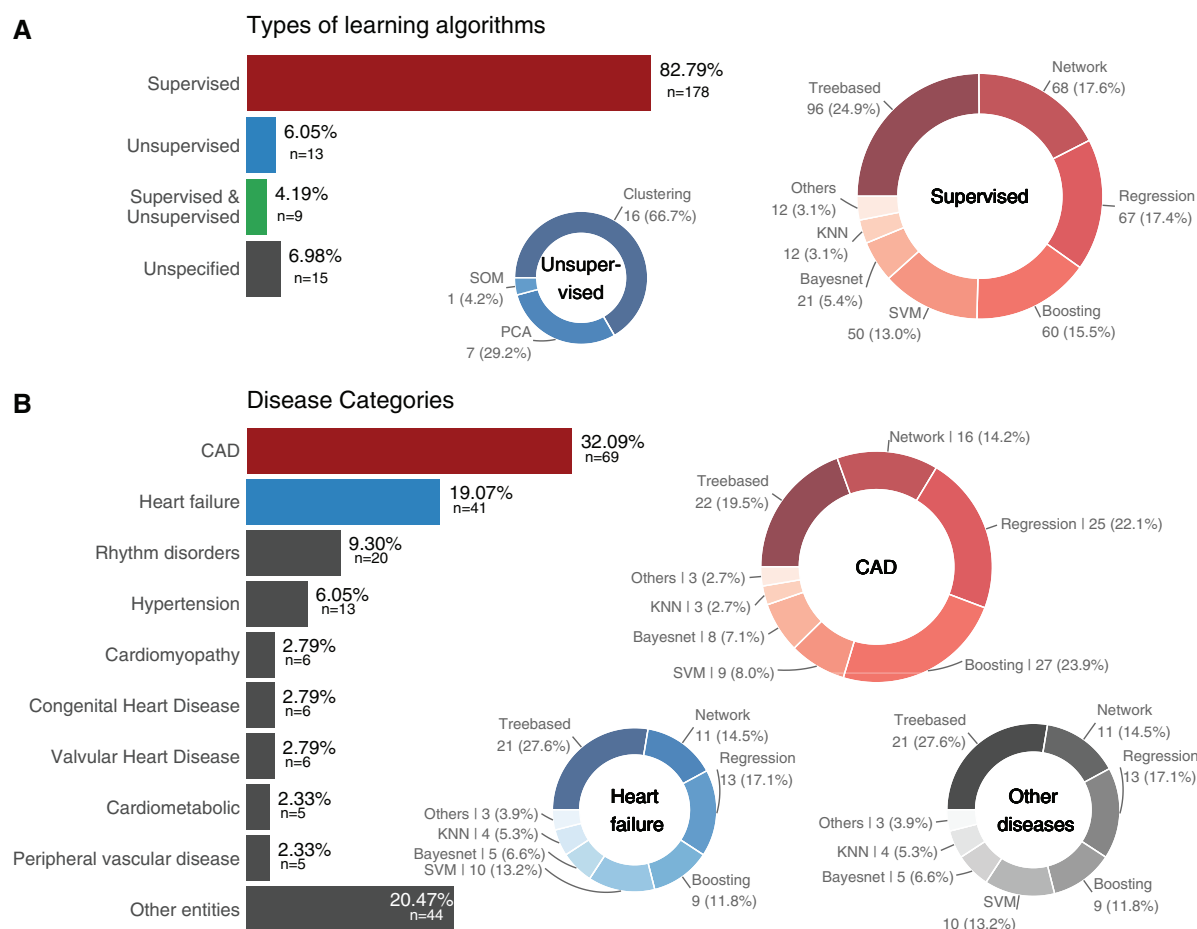


Figure 2 Overview of the methods and disease areas presented in the articles. Panel (A) shows the types of artificial intelligence/machine learning algorithms applied. Panel (B) displays the distribution of disease areas as well as which supervised methods are most commonly applied in which disease area.

Recommendations

The choice of a specific AI/ML method is complex and depends on various parameters that are specific to the individual problem to be solved. From this extensive review, a few broad recommendations on this choice can be derived. In feature selection with, e.g. tabular data such as health records, tree-based, regression or boosting methods are most commonly applied. The application of DL to tabular data is generally possible and might perform similarly or even outperform other methods especially on larger data sets,²¹ but specific adaptations of DL to tabular data are still an area of active research.^{22,23} For image (and similarly electrocardiogram signal or omics) data, network-based DL methods are the most commonly used method. Besides the excellent performance of DL on image data, the possibility to do transfer-learning easily with DL methods is one of the key components that make DL the go-to method for image data, if the number of patients is sufficiently large.

We explicitly restricted our search to clinical applications of AI/ML methods, excluding methodological publications from our literature

search. However, some recent reviews provide nice overviews for specific methodologies. For example, Chen *et al.*²⁴ reviewed the use of DL in image segmentation, while Bizopoulos and Koutsouris⁴ provide an overview of DL applications in structured data, signal and imaging modalities. In line with Chen *et al.*,²⁴ we recommend that future research explicitly targets the deployment of novel methodology such as DL in real-world clinical applications.

Furthermore, we recommend that some points are considered with regard to the (i) evaluation of an ML algorithm in itself and in comparison to alternative algorithms; (ii) reporting of the evaluation; and (iii) transparency of data and methods. The recommendations are summarized in Figure 5.

The evaluation of an ML algorithm requires that a model has been developed and validated in a carefully designed study. This includes, among other aspects, that predictor variables were assessed independently from outcome variables, and that sample sizes were sufficient for stable model building and a precise estimation. Vollmer *et al.*¹² provide a list of critical questions to assess the quality of AI/ML applications in medical applications. Moreover, tools to assess data

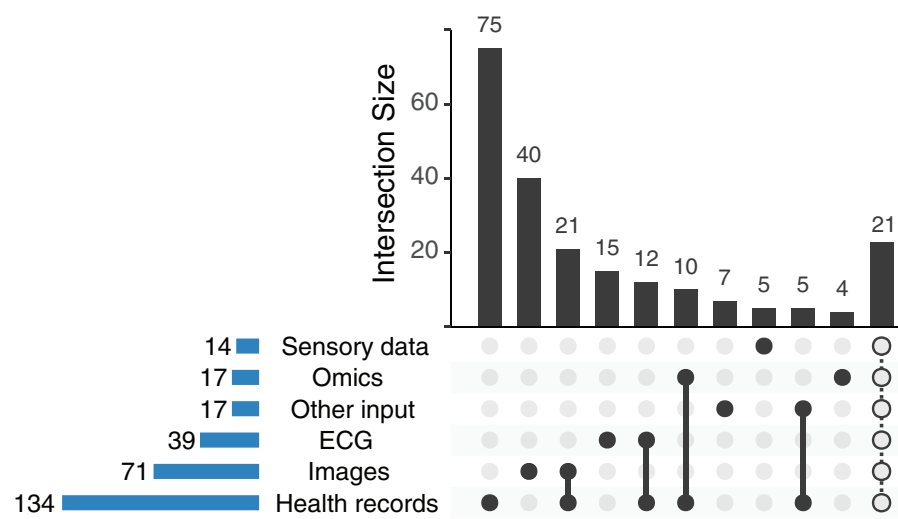


Figure 3 Input type used for the artificial intelligence/machine learning algorithms. Displayed are the absolute number a respective input type was used (lower left bars) and the most common combinations of input types (upper bar plot). The last bar summarizes all other combinations that occurred less than four times.

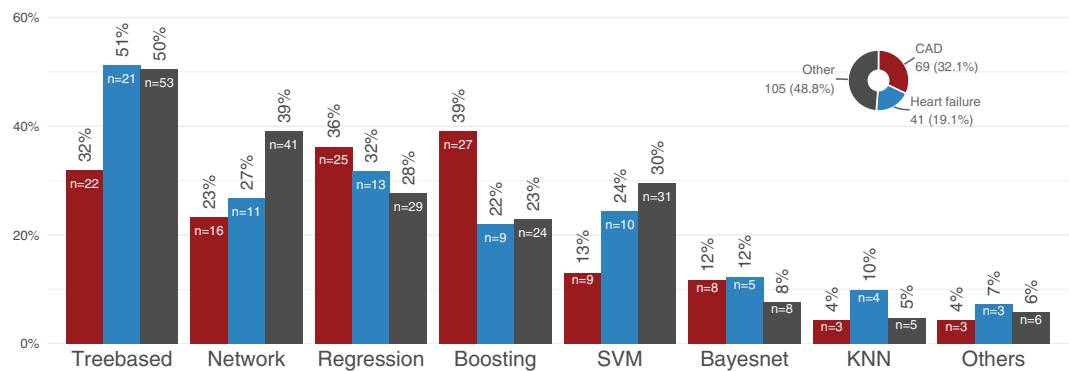


Figure 4 Overview of the distribution of the applied supervised methods stratified by disease area (coronary artery disease, heart failure, and all other diseases) in more detail. The bars will not sum up to 100% because multiple methods were used in the reviewed papers.

quality in observational studies have recently been proposed and implemented.²⁵ The metrics used to assess algorithms should generally cover aspects of calibration as well as discrimination and specifically need to match the aim of the study. For instance, the focus could be on an overall performance in all study participants versus in specific subgroups (sensitivity or specificity); the absolute performance of the algorithm could be of interest versus its incremental value above established methods; the focus could be on the statistical performance or the clinical utility of applying the algorithm in practice. Comparisons to classical regression methods or established predictive models should be 'fair': model parameters for classical regression methods should be tuned to the best performance possible in the

sample under investigation for AI/ML. Established predictive models have to be re-calibrated to the sample used for AI/ML to have the best possible performance. Ideally, method comparison studies should be conducted by independent research groups.^{26,27} Where possible, the reporting of methods' comparisons and applications should adhere to available guidelines. Oftentimes, AI/ML is used to train predictive models (for prognostic or diagnostic settings) which are usually multivariable models. A first important step in this context is to highlight the aim of the AI/ML approach, i.e. whether a longitudinal prognostic (incidence, risk for future event) or cross-sectional diagnostic/classification outcome is of interest. With the TRIPOD statement²⁸ well established guidelines on the reporting of

Table 2 Examples of AI/ML applications to cardiovascular medicine

Authors/title/doi	Disease area	AI/ML method	Trial design	No. of subjects	Aim	Citations	Abstract no.
Weng <i>et al.</i> Can machine-learning improve cardiovascular risk prediction using routine clinical data? 10.1371/journal.pone.0174944	Other entities	Random forest, logistic regression, gradient boosting machines, neural networks	Prospective cohort study	378 256	Cardiovascular risk prediction	125	427
Zahid <i>et al.</i> Patient-derived models link re-entrant driver localization in atrial fibrillation to fibrosis spatial pattern. 10.1093/cvr/cww073	Rhythm disorders	Supervised machine learning algorithm	Prospective cohort study	20	Testing the hypothesis that atrial fibrillation re-entrant drivers persist only in regions with specific fibrosis patterns	80	464
Motwani <i>et al.</i> Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. 10.1093/eurheartj/ehw188	CAD	Gradient boosting	Retrospective cohort study	10 030	Prediction of all-cause mortality in patients with suspected coronary artery disease	78	460
McConnell <i>et al.</i> Feasibility of obtaining measures of lifestyle from a smartphone app: the MyHeart Counts Cardiovascular Health Study. 10.1001/jamacardio.2016.4395	Other entities	Unsupervised	Prospective cohort study	40 017	Assessing the feasibility of obtaining measures of lifestyle from smartphones	53	443
Ambale-Venkatesh <i>et al.</i> Cardiovascular event prediction by machine learning: the Multi-Ethnic Study of Atherosclerosis. 10.1161/CIRCRESAHA.117.311312	Other entities	Random survival forests	Retrospective cohort study	6814	Prediction of six cardiovascular outcomes in comparison to standard cardiovascular risk scores	50	410
Narula <i>et al.</i> Machine-learning algorithms to automate morphological and functional assessments in 2D	Cardiomyopathy	Support vector machines, random forests, and artificial neural networks	Retrospective cohort study	139	Automated discrimination of hypertrophic cardiomyopathy from physiological hypertrophy seen in athletes	49	446

Continued

Table 2 Continued

Authors/title/doi	Disease area	AI/ML method	Trial design	No. of subjects	Aim	Citations	Abstract no.
echocardiography. 10.1016/j.jacc.2016.08.062							
Attia et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram.	Heart failure	Convolutional neural network	Retrospective cohort study	44 959/52 870	Identification of cardiac contractile dysfunction by ECG	48	156
10.1038/s41591-018-0240-2							
Frizzell et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches.	Heart failure	Tree-augmented naive Bayesian network, random forest, gradient-boosted, logistic regression, least absolute shrinkage, and selection operator models	Retrospective cohort study	56 477	Prediction of 30-day readmission rate in patients discharged following hospitalization for heart failure	48	447
10.1001/jamacardio.2016.3956							
Mortazavi et al. Analysis of machine learning techniques for heart failure readmissions.	Heart failure	Random forests, boosting, random forests, support vector machines, logistic regression, Poisson regression	Retrospective cohort study	1004	Prediction of readmission after hospitalization for heart failure	46	436
10.1161/CIRCOUTCOMES.116.003039							
Attia et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction.	Rhythm disorders	Convolutional neural network	Retrospective cohort study	180 922	Identification of patients with atrial fibrillation during sinus rhythm by ECG	37	156
10.1016/S0140-6736(19)31721-0							

The examples were chosen according to the overall number of citations. The table depicts the disease area, the ML method(s) considered in the paper, the trial design, the sample size, and the aim of the study. Abstract number refers to [Supplementary material online, Table S1](#).

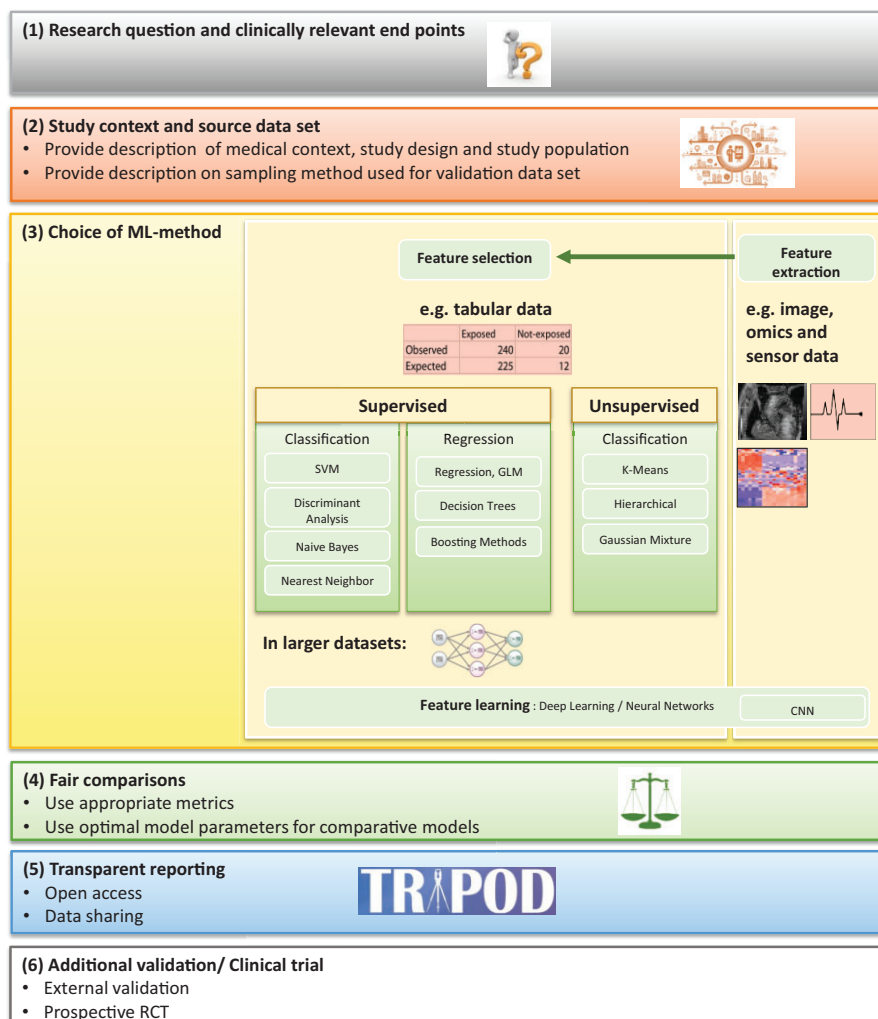


Figure 5 Recommended steps to be taken into account when using artificial intelligence/machine learning methods in cardiovascular research. Feature selection (selecting the most relevant subset of features, e.g. a biomarker, age or sex of a patient or image information), feature extraction (finding a minimalistic representation of a larger data set, e.g. an image), and feature learning (the algorithm chooses/learns relevant features from the data).

such models are available. While an explicit adaptation to AI/ML methods is still under development,²⁹ the statement is overall well applicable to predictive AI/ML models. The 22 items on the checklist relate to all parts of a typical academic report from the title to the supplement and cover areas as the data source (items 4 and 5), the outcome (item 6), the model building (item 10), and the clinical implications (item 20).

Most challenging to address is item 15, which asks to present the full model and to explain how the trained model is applied and how to interpret the results. Complex AI/ML models involving many variables (e.g. random forest, deep neural network) are not easily shared in printed form and other means to make the trained model available have to be found in these cases, see the Discussion section for possibilities.

Finally, we recommend a *high level of transparency* throughout the process of model development and validation with regard to the

utilized data, the specific final model(s) built, and the program code that was used. While performing the extensive literature review, we often recognized a lack of reporting in the exact methods used, which limits comparability and reproducibility of the proposed approaches. To overcome this shortcoming, we recommend that source code should be openly available, e.g. in a web-based version control system such as git/github (<https://github.com/>). The respective URL should be included in the manuscript.

It is of crucial importance to further encourage data sharing. Safe and trusted open data initiatives such as Zenodo (<https://zenodo.org/>) are recommended for sharing data. The platform provides a DOI to each upload to make data citable and traceable. It also offers a sophisticated data access model to restrict data access only to certain groups. Google Cloud provides a suite of tools as part of their AI platform offering (<https://cloud.google.com/ai-platform>) to build, validate, and explain models. As for proprietary data sharing, Triple

Blind (<https://tripleblind.ai/ai-in-healthcare/>) is an example of a platform to reproduce results using the same datasets and models while maintaining privacy.

If data sharing is not possible due to legal issues, it is recommended to make the trained model openly available such that other research groups can re-use model weights or estimated model coefficients. This is a very common approach in traditional computer vision, where network models like VGG16 are shared with the public for re-usage. To allow for an easy integration of such models into novel applications, such pre-trained networks are now even built-in common libraries, such as Keras (<https://keras.io/api/applications/>). However, possible issues of model inversion³⁰ need to be taken into account, i.e. Zhu et al.³⁰ could show that it is possible to recover the (private) training data from the publicly shared models.

Discussion

In this article, we have reviewed the current state of AI/ML applications in cardiovascular medicine. We provide a comprehensive overview of the spectrum of the various different AI/ML methods and illustrate the context in which these were applied to address questions in a variety of cardiovascular diagnostic applications and diseases. Since a major finding is that methodology may overlap even in similar data and since we observed marked variation in quality we also provide some recommendations with respect to applying AI/ML methods in practice. This methodological overlap may be explained by the fact that to date no consensus exists as to which method should best be applied in which disease context. Therefore, many publications included in our review investigated and compared several methods simultaneously. Indeed, the choice of a specific AI/ML method is complex as are the various parameters and pitfalls that determine appropriate use. We found that AI/ML-based work frequently lacks aspects of quality such as transparency regarding methodology and data as well as validation of the methods. Other important aspects of AI/ML research include data partition and cross-validation. Krittanawong et al.¹⁰ found a large heterogeneity with respect to these aspects in their meta-analysis. Therefore, after a period of rather intense AI/ML research, which we document herein, we advocate a more vigorous approach to scientific standards, which should be a prerequisite for clinical application.

Our review is limited by our literature search, where we explicitly required the search terms 'artificial intelligence' or 'machine learning' mentioned in title or abstract. Moreover, we have focused on clinical applications and thus not considered publications in methodological journals, since we specifically wanted to focus on AI/ML applications to real-world clinical data. In contrast, methodological papers often demonstrate the usefulness or applicability of new methods on freely available benchmark data sets. Thus, they play an important role with respect to proof of concept and feasibility of newly developed methods and can be seen as an important intermediate step between method development and widespread clinical use.

A potential source of bias in our study is the exclusion of journals with an impact factor less than 2. The rationale behind this approach was to limit our analyses to articles with potential clinical impact. The threshold of 2 was chosen since it lies between the median impact factor in cardiology (median IF 2.3) and general internal medicine (median

IF 1.6) according to the Web of Science.³¹ In total, 228 journals were searched and articles from 65 distinct journals were included in our analyses. The effect of this restriction is also displayed in the PRISMA flow chart, see [Supplementary material online, Figure S1](#).

A further limitation results from the categorization of AI/ML approaches and disease categories. Here we used the terms primarily used by the authors of the articles included, but a potential overlap, e.g. between HF and cardiomyopathies, cannot be ruled out completely.

The broad scope of our publication limits in-depth discussion with respect to specific disease areas or data types. However, we deliberately chose this approach in order to give a broad overview. In this, our systematic review complements previously published work that focused on particular applications. Finally, we provide only descriptive analyses with respect to superiority of the AI/ML methods as compared to 'classical' methods, therein relying solely on the authors' definition of superiority. As already mentioned in the recommendations above, however, these comparisons are often not conducted fairly.³² From a clinical perspective, there is still a lack of randomized controlled trials as the mainstay of evidence-based medicine in the cardiovascular field of AI/ML. Comparison of AI/ML-incorporated algorithms to standard of care by means of clinically relevant endpoints and validation in prospective studies are prerequisites for further integration and acceptance. Only the minority of publications investigated reproducibility and generalizability. However, such studies are necessary to foster large-scale clinical implementation of novel AI/ML approaches. But even if prospective validation is not implemented at this stage, now is the time for advancing quality of AI/ML-based work, given an increasing body of practical recommendations. For example, the essential TRIPOD guidelines have been extended by additional important work, such as recommendations for proper reporting of AI prediction models.²⁹ Likewise, standards for avoiding bias and fostering reproducibility have been communicated and should be demanded, ultimately, to avoid harm to patients.^{14,15} Our review suggests that most of the time, standards were set too low. On the other hand, demanding that any data used to train AI/ML models must be (come) open source, while certainly ideal, might significantly preclude important hypothesis generating work. Given the shortage of open-source training data, work on closed source data, such as some registries, is indeed important for hypothesis generation and, provided it is labelled as such, deserves attention. However, at a stage where routine clinical decision making takes place, we consider external validation essential.

Our extensive review also showed that some promising AI/ML methods are currently underutilized in clinical practice. To encourage wider use of potentially superior AI/ML methods and to push such research on urgent, clinically relevant problems, one promising approach is to conduct medical challenges, as has been done frequently in various research areas. Linked to this is the definition of the task and appropriate metrics to evaluate the incoming results. Participants, mostly volunteers, can register and are asked to upload their code and/or results before the predefined deadline. Platforms like Grand Challenge (<https://grand-challenge.org/challenges/>) or Kaggle (<https://www.kaggle.com/>) provide options for data upload, participant registration and leaderboard visualization. The main benefit is that the developed methods are directly comparable, because

they were, unlike in many other works, trained and tested on the same data sets. Moreover, a spill of training data into validation data-sets, a problem that is hard to control for in several AI/ML settings, is excluded by design. To allow for such challenges, grants supporting the purpose of data acquisition, including an incentive to provide open or closed source data to such a challenge should be promoted.

Another possible path for future directions entails the use of federated learning. Federated learning means 'to let the algorithms travel and not the data'. Rieke *et al.*³³ propose to use federated learning to avoid the complexity of data sharing that is associated from a legal point of view. This may be realized by linking the data infrastructure of the hospital to an in-house computational node that trains models and sends the trained model weights to a central node outside the hospital, where the models are aggregated to create a novel powerful approach, which better accounts for more variants of data.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health* online.

Acknowledgements

The authors would like to thank the members of the DZHK project group AI/ML.

Funding

M.K. is a fellow in the BIH Charité Digital Clinician Scientist Program funded by DFG. We acknowledge the support of the German Center for Cardiovascular Research funded by the Bundesministerium für Bildung und Forschung, grant 81Z1700102 (to I.R.K.), 81Z0300108 (to T.F.), 81Y0400120 (to S.G.), 81X3400108 (to S.G.), 81Z0400101 (to M.B.), 81X2400123 (to M.B.), and 81X2400143 (to M.B.).

Conflict of interest: T.F. reports personal fees from Novartis, personal fees from Bayer, personal fees from Janssen, personal fees from Roche, personal fees from Boehringer Ingelheim, personal fees from Daiichi-Sankyo, personal fees from Galapagos, personal fees from Penumbra, personal fees from Parexel, personal fees from Vifor, personal fees from BiosenseWebster, personal fees from CSL Behring, personal fees from Fresenius Kabi, personal fees from Cohere Medical, personal fees from LivaNova, personal fees from Minoryx, outside the submitted work. I.R.K. reports grants from German Center for Cardiovascular Research, grants from German Center for Lung Research, grants from Deutsche Forschungsgemeinschaft, grants from Deutsche Krebshilfe, grants from Federal Ministry of Education and Research, outside the submitted work. J.S. reports grants from DZHK e.V., during the conduct of the study. All other authors have declared no conflict of interest.

Data availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. Timmis A, Townsend N, Gale CP, Torbica A, Lettino M, Petersen SE, Mossialos EA, Maggioni AP, Kazakiewicz D, May HT, De Smedt D, Flather M, Zuhke L, Beltrame JF, Huculeci R, Tavazzi L, Hindricks G, Bax J, Casadei B, Achenbach S, Wright L, Vardas P. European Society of Cardiology: cardiovascular disease statistics 2019. *Eur Heart J* 2020;**41**:12–85.
2. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Shay CM, Spartano NL, Stokes A, Tirschwell DL, VanWagner LB, Tsao CW; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2020 update: a report from the American Heart Association. *Circulation* 2020;**141**: e139–e596.
3. Kilic A. Artificial intelligence and machine learning in cardiovascular health care. *Ann Thorac Surg* 2020;**109**:1323–1329.
4. Bizopoulos P, Koutsouris D. Deep learning in cardiology. *IEEE Rev Biomed Eng* 2018;**12**:168–193.
5. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;**368**:m689.
6. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018;**104**: 1156–1164.
7. Seetharam K, Shrestha S, Sengupta PP. Artificial intelligence in cardiovascular medicine. *Curr Treat Options Cardiovasc Med* 2019;**21**:1–14.
8. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT. Artificial intelligence in cardiology. *J Am Coll Cardiol* 2018;**71**: 2668–2679.
9. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, Kapa S, Lerman A, Luong C, Medina-Inojosa JR, Noseworthy PA, Pellikka PA, Redfield MM, Roger VL, Sandhu GS, Senecal C, Friedman PA. Artificial intelligence in cardiology: present and future. *Mayo Clin Proc* 2020;**95**: 1015–1039.
10. Krittawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, Zhang HJ, Kaplan S, Narasimhan B, Kitai T, Baber U, Halperin JL, Tang, WHW. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* 2020;**10**:16057.
11. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F; Massive Analysis Quality Control (MAQC) Society Board of Directors, L Waldron, B Wang, C McIntosh, A Goldenberg, A Kundaje, CS Greene, T Broderick, MM Hoffman, JT Leek, K Korthauer, W Huber, A Brazma, J Pineau, R Tibshirani, T Hastie, JPA Ioannidis, J Quackenbush, HJWL. Aerts Transparency and reproducibility in artificial intelligence. *Nature* 2020;**586**:E14–E16.
12. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KSL, Myles P, Granger D, Birse M, Branson R, Moons KGM, Collins GS, Ioannidis JPA, Holmes C, Hemingway H. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;**368**:i6927.
13. Wynants L, Smits LJM, Van Calster B. Demystifying AI in healthcare. *BMJ* 2020;**370**:m3505.
14. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health* 2020;**2**:e537–e548.
15. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;**2**:e549–e560.
16. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;**27**:186–187.
17. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;**6**:e1000100.
18. Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, McElhinney P, Gransar H, Cantu S, Miller RJH, Rozanski A, Achenbach S, Tamarappoo BK, Berman DS, Dey D. 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**:2216–2225.
19. Leha A, Hellenkamp K, Unsöld B, Mushemi-Blake S, Shah AM, Hasenfuß G, Seidler T. A machine learning approach for the prediction of pulmonary hypertension. *PLoS One* 2019;**14**:e0224453.

20. Bhuvana AN, Bai W, Lau C, Davies RH, Ye Y, Bulluck H, McAlindon E, Culotta V, Swoboda PP, Captur G, Treibel TA, Augusto JB, Knott KD, Seraphim A, Cole GD, Petersen SE, Edwards NC, Greenwood JP, Bucciarelli-Ducci C, Hughes AD, Rueckert D, Moon JC, Manisty CH. A multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis. *Circ Cardiovasc Imaging* 2019;**12**:e009214.
21. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. arXiv preprint arXiv:1706.02515; 2017.
22. Arik SO, Pfister T. Tabnet: attentive interpretable tabular learning. arXiv preprint arXiv:1908.07442; 2019.
23. Popov S, Morozov S, Babenko A. Neural oblivious decision ensembles for deep learning on tabular data. arXiv preprint arXiv:1909.06312; 2019.
24. Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W & Rueckert D. Deep learning for cardiac image segmentation: a review. *Front Cardiovasc Med* 2020;**7**:25.
25. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, Huebner M, Schmidt B, Sauerbrei W, Richter A. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Med Res Methodol* 2021;**21**:1–15.
26. Boulesteix AL, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Med Res Methodol* 2017;**17**:1–12.
27. Weber LM, Saelens W, Cannoodt R, Soneson C, Hapfelmeier A, Gardner PP, Boulesteix AL, Saeys Y, Robinson MD. Essential guidelines for computational method benchmarking. *Genome Biol* 2019;**20**:1–12.
28. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Circulation* 2015;**131**:211–219.
29. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;**393**:1577–1579.
30. Zhu L., Han S. Deep leakage from gradients. In Q Yang, L Fan, H Yu, eds. *Federated Learning. Lecture Notes in Computer Science*, vol. **12500**. Cham: Springer, 2020, pp. 17–31. 10.1007/978-3-030-63076-8_2.
31. Journal Impact Factor 2020, Journal Citation Reports Science Edition, Clarivate Analytics. <https://jcr.clarivate.com/JCRHomePageAction.action?>
32. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;**110**:12–22.
33. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M, Cardoso MJ. The future of digital health with federated learning. *NPJ Digit Med* 2020;**3**:119.