**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ April 20, 2022

Shawn Xavier Trimble Jr.

A Systematic Review of Machine Learning Prediction and Cardiovascular Risk Assessment

By

Shawn Xavier Trimble Jr.

MPH

Hubert Department of Global Health

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Mary Beth Weber, PhD, MPH

Committee Chair

A Systematic Review of Machine Learning Prediction and Cardiovascular Risk Assessment

By

Shawn Xavier Trimble Jr.

Pomona College

Bachelor of Arts, BA

2020

Thesis Committee Chair: Dr. Mary Beth Weber, PhD, MPH

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Global Health

2022

**Abstract**

A Systematic Review of Machine Learning Prediction and Cardiovascular Risk Assessment

By Shawn Xavier Trimble Jr.

Artificial intelligence (AI) techniques such as machine learning (ML) and deep learning (DL) promise considerable improvements in cardiovascular disease (CVD) prediction. This systematic review and meta-analysis aim to assess and compare the predictive ability of ML algorithms to conventional risk assessment methods in CVDs. Cochrane, Embase, Scopus and Web of Science databases were searched for studies published between January 1, 2012, and January 1, 2022. Studies including the predictive performance of ML models and conventional risk assessment methods were included. Studies without sufficient evaluation data and model validation were excluded. Diagnostic accuracy data was extracted and used to create contingency tables to derive performance metrics of interest: sensitivity, specificity, threshold limits, and areas under the curve (AUC). Studies were included in a meta-analysis, using a bivariate random effects model. The search identified 1688 studies of which 25 studies were included. For the prediction of CVD, ML models had a pooled AUC of 0.88 and conventional risk assessment had a pooled AUC of 0.74. For the prediction of coronary artery disease, ML models had a pooled AUC of 0.83. For the prediction of heart failure, ML models had a pooled AUC of 0.90. For the prediction of stroke, ML models had a pooled AUC of 0.83. Of the 25 studies, 9 studies provided conventional risk assessment comparators. ML models had a pooled AUC of 0.82 compared to conventional risk assessment who had a pooled AUC of 0.78. Insufficient samples sizes for ML models prevented within ML comparisons and analysis of cardiac arrythmias. The predictive ability of ML models is comparable to conventional risk assessment methods. However, heterogeneity among ML models and insufficient data reporting methods within the literature calls into question the clinical applicability of these findings. This review may assist clinicians in assessing the current state of AI and provide insights to how ML models can better translate into the clinical setting.

A Systematic Review of Machine Learning Prediction and Cardiovascular Risk Assessment

By

Shawn Xavier Trimble Jr.

Pomona College

Bachelor of Arts, BA

2020

Thesis Committee Chair: Dr. Mary Beth Weber, PhD, MPH

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Global Health

2022

**Acknowledgements**

Dr. Mary Beth Weber, PhD, MPH, Emory University

Dr. Michael Caudle, PhD, Emory University

**Table of Contents**

[Chapter 1: Introduction 1](#_Toc101361975)

[Section 1.1: Background 1](#_Toc101361976)

[Section 1.2: Statement of the Problem 3](#_Toc101361977)

[Section 1.3: Statement of the Purpose 3](#_Toc101361978)

[Section 1.4: Research Aims 3](#_Toc101361979)

[Section 1.5: Significance of the Study 3](#_Toc101361980)

[Chapter 2: Literature Review 4](#_Toc101361981)

[Section 2.1: Cardiovascular Disease 4](#_Toc101361982)

[Section 2.1.1: Current Burden 4](#_Toc101361983)

[Section 2.1.2: Epidemiological, Economic, and Social Impact 5](#_Toc101361984)

[Section 2.1.3: Risk Assessment 8](#_Toc101361985)

[Section 2.2: Artificial Intelligence, Machine Learning, Neural Networks, and Deep Learning 9](#_Toc101361986)

[Section 2.2.1: Artificial Intelligence 9](#_Toc101361987)

[Section 2.2.2: Machine Learning 10](#_Toc101361988)

[Section 2.2.3: Deep Learning 12](#_Toc101361989)

[Section 2.3: Artificial Intelligence Clinical Applications 13](#_Toc101361990)

[Section 2.3.1: Precision Medicine 13](#_Toc101361991)

[Section 2.3.2: Clinical Prediction 15](#_Toc101361992)

[Section 2.3.3: Diagnostic Imaging 16](#_Toc101361993)

[Section 2.4: AI Concerns and Limitations 17](#_Toc101361994)

[Section 2.4.1: Data Consent 17](#_Toc101361995)

[Section 2.4.2: Data Privacy 18](#_Toc101361996)

[Section 2.4.3: Data Transparency 20](#_Toc101361997)

[Chapter 3: Methods 21](#_Toc101361998)

[Section 3.1: Search Strategy 21](#_Toc101361999)

[Section 3.2: Study Selection 21](#_Toc101362000)

[Section 3.3: Data Extraction 22](#_Toc101362001)

[Section 3.4: Statistical Analysis 23](#_Toc101362002)

[Chapter 4: Results 24](#_Toc101362003)

[Section 4.1: Study Search 24](#_Toc101362004)

[Section 4.2: Study Characteristics 25](#_Toc101362005)

[Section 4.3: Study Results 25](#_Toc101362006)

[Section 4.3.1: Machine Learning Models and Prediction of Cardiovascular Disease 25](#_Toc101362007)

[Section 4.3.2: Machine Learning Models and Prediction of Cardiac Arrythmias 27](#_Toc101362008)

[Section 4.3.3: Machine Learning Models and Prediction of Coronary Artery Disease 27](#_Toc101362009)

[Section 4.3.4: Machine Learning Algorithms and Prediction of Heart Failure 28](#_Toc101362010)

[Section 4.3.5: Machine Learning Algorithms and Prediction of Stroke 29](#_Toc101362011)

[Section 4.3.6: Matched Prediction of Cardiovascular Disease 30](#_Toc101362012)

[Chapter 5: Discussion 32](#_Toc101362013)

[Section 5.1: Discussion of Key Findings 32](#_Toc101362014)

[Section 5.2: Strengths and Weaknesses 34](#_Toc101362015)

[Section 5.3: Recommendations 35](#_Toc101362016)

[Section 5.3.1: Quality Assessment Tools 35](#_Toc101362017)

[Section 5.3.2: Meta-Analysis of Test Performance Gold Standards 36](#_Toc101362018)

[Section 5.3.3: Clinical Level Contextualization 37](#_Toc101362019)

[Section 5.4: Conclusion 39](#_Toc101362020)

[References 41](#_Toc101362021)

[Appendix 58](#_Toc101362022)

[Section A-1: PRISMA Checklist 58](#_Toc101362023)

[Section A-2: Search Strategy 59](#_Toc101362024)

[Section A-3: Study Characteristics 62](#_Toc101362025)

# 

# **Chapter 1: Introduction**

## **Section 1.1: Background**

The rapid rise of CVD is attributed to the biological, behavioral, and psychosocial risk factors that are intrinsically linked with socioeconomic status (Schultz et al., 2018). Global health indices describe distinct patterns of variation regarding CVD incidence, prevalence, and mortality rate across lower- and middle-income countries (LMIC) and high-income countries (HIC) (Roth et al., 2020; Thomas et al., 2018). This variation reflects the difference in risk factors prevalent in each region, but also the availability of resources to prevent, predict, and treat cardiovascular outcomes. Epidemiologist have accentuated a link between the variation of CVD outcomes and the epidemiologic transition stage of a particular country (Omran 1971; Schultz et al., 2018, Yusuf, 2001). These epidemiologic transition stages are linked to the economic state of the country and the health policies implemented as a result (Mendoza & Miranda, 2017; Santosa et al., 2014). Many different root causes and barriers can be interpreted as the reason for a county’s current economic status, however as it relates to noncommunicable disease, CVD is the main contributor to economic burden in this category (GBD, 2017). All countries regardless of economic status require cost effective solutions to preserve their economic viability, and this is need is especially apparent for countries who have little resources to allocate to the issue.

Cardiovascular prevention medicine is a well-research topic within the field of cardiology. As a result, a number of behavioral, genetic, physiological, psychosocial risk factors have been identified and associated with different cardiovascular outcomes. Additionally, these findings have informed and led to the creation of a plethora of cardiovascular risk assessment tools such as the Framingham Risk Score (FRS), European Systemic Coronary Risk Evaluation Algorithm (SCORE), Prospective Cardiovascular Münster Model (PROCAM), and Reynolds Risk Score (RRS) (Khambhati et al., 2018). While these risk assessments have vastly improved clinicians’ ability to prevent, predict, and treat CVD, their guidelines have been scrutinized as being reductionist in their discriminative approach. CVD is complex and heterogenous in its pathophysiology and public health experts argue that it cannot be effectively predicted through a reductionist understanding of traditional and non-traditional risk factors (Leopold, Maron, & Loscalzo, 2020). Rather, a holistic understanding of its genotype-endophenotype-phenotype (GECP) relationship is imperative. At each level of the GECP relationship is heterogeneity leads to different manifestations of a phenotype that converges to a common end phenotype or clinical outcome. Due to large number of ways in which CVD can manifest as a result of its risk factors, the data needed to encapsulate this becomes too complex to assess by conventional means (Leopold et al., 2020; Shameer et al., 2018).

This issue is referred to as “big data”, where datasets are either too larger and/or too complex to be dealt with by conventional data processing methods. Big data coupled with more capable analytical approaches have been predicated to be the solution to both elucidating the heterogenetic pathways of CVD and the growing culmination of complex data across all disciplines within medicine (Leopold et al., 2020; Silverio et al., 2019; Weintraub 2019). For decades, studies have been employing AI approaches, ML models, to create more robust tools for medical research (Bohr & Memarzadeh, 2020; Briganti & Moine, 2020). Determining the clinical efficacy of these ML models is a research priority as it is believed that the use of complex datasets may be well positioned to resolve the issue of CVD heterogeneity through bridging the gap between current conventional risk assessment methods and the construction of the genotype-endophenotype-phenotype cardiovascular relationship.

## **Section 1.2: Statement of the Problem**

The literature on the applications of AI in medicine lacks substantive investigation into the performative ability of ML models. While there are many studies that have been conducted demonstrating efficacy in disease prediction, this has been in comparison to other ML models at the minimum and an insufficient number of clinicians at best. The literature requires a systematic assessment that may provide a more comprehensive review of DL models for CVD prediction and guide the focus of future research.

## **Section 1.3: Statement of the Purpose**

The primary purpose of this review is to compare the accuracy of DL models against conventional risk assessment methods applied in cardiovascular precision medicine, clinical prediction, and diagnostic imaging. The secondary purpose of this review is to assess the current state of AI in cardiovascular medicine, identifying the applications, implications, and limitations of its application. Conducting a systematic review of the literature with these underlying concerns can contribute to the discussions of how to establish ML models as trusted assessment tools in the clinical setting.

## **Section 1.4: Research Aims**

This review seeks to address the following research aims:

1. Compare predictive accuracy of ML models to conventional cardiovascular risk assessment methods.
2. Assess methodological approaches of AI applications in cardiovascular risk assessment.
3. Elucidate relevant concerns of the clinical community regarding the use of AI.

## **Section 1.5: Significance of the Study**

This review is the first to systematically compare the predictive accuracy of ML models against traditional risk assessment methods in the field of cardiology. A small number of studies evaluate the predictive accuracy of ML models in CVD prediction. An even smaller number makes direct comparisons between ML models and clinicians, let alone all forms of risk assessment. This review will provide an insight into the current state of ML based cardiovascular risk prediction. The review will also discuss, based on the prior literature and results, the appropriate use cases for each ML model.

# **Chapter 2: Literature Review**

## **Section 2.1: Cardiovascular Disease**

### **Section 2.1.1: Current Burden**

Noncommunicable diseases (NCD) are estimated to account for two-thirds of all deaths worldwide. These estimates are currently on the rise; NCDs accounted for 26.8 million deaths in 1990 and in 2017 this number rose to 41.1 million. NCDs are not isolated to HICs as 80% of premature deaths are attributed to NCDs in LMICs. CVD is the leading NCD, accounting for almost half of the world’s NCD related deaths and are a significant barrier to sustainable health development. CVD is preventable and linked to a range of modifiable health behaviors such as diet, tobacco use, alcohol use, and social stressors (Jagannathan et al., 2019). The question of public health importance that needs to be addressed is how cardiovascular behavioral risk factors affect the global burden of CVD. CVD is a class of diseases with a shared pathology involving heart and blood vessels. Common conditions include coronary artery disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease and congenital heart disease (Stewart, Manmathan & Wilkinson, 2017). The underlying mechanisms of CVD are theorized to be a combination of genetic, behavioral, physiological, and psychosocial factors that contribute both to its onset and development. The preponderance levels of CVD risk factors such as high such as high systolic blood pressure, high fasting plasma glucose, high lipoprotein cholesterol, poor diet, and low physical activity, has been found to manifest by difference of age, sex, and economic region (Jagannathan et al., 2019; Jousilahti et al., 1999; Roth et al., 2020).

Age and sex have been found to serve as a relevant factor for CVD development. Age has been associated with overall cardiovascular functioning and increased risk and prevalence of CVD in middle-aged and older adults. The American Heart Association (AHA) reports in the United States, that the incidence of CVD in men and women is 40% from 40–59 years, 75% from 60–79 years, and 86% in those above the age of 80. Global case studies suggest similar findings however describes a difference in the disability impact between men and women. When looking at global trends of disability adjusted life years, a measure of overall disease burden as the number of years lost due to ill health, total CVD disability adjusted life years were higher in men than women before age 80 to 84 years. After this age, the pattern reverses (Roth et al., 2020). When focusing on economic regions, HICs are observing an overall decline in age adjusted death rates of CVD with a number of exceptions. This overall decline has been largely driven by preventative interventions improving disease avoidance, treatment of acute conditions, and treatment of fully manifested CVD. In LMICs, CVD is largely the result of an increase in the prevalence of risk factors and a lack of access to prior mentioned interventions. This has resulted in increases in age adjusted death rates in these regions as well increases in CVD prevalence in those under the age of 60, specifically the working age population (Jameson et al., 2006; Gaziano et al., 2010).

### **Section 2.1.2: Epidemiological, Economic, and Social Impact**

The global rise in CVD is the result of a transformations in the causes of morbidity and mortality over the time course of the 20th century. These transformations are termed as the “epidemiologic transition” and are demographic shifts driven by industrialization, urbanization, and other socioeconomic changes that take place in a population, region, and or country. The epidemiologic transition is divided into five stages: pestilence and famine, receding pandemics, degenerative diseases, delayed degenerative diseases, epidemic inactivity (Omran, 1971; Yusuf et al., 2001; Santosa et al., 2014) (Figure 1). CVD embodies different clinical outcomes depending on the stage that a country and/or region is experiencing. Countries with epidemiologic profiles in the first four stages manifest CVD outcomes and their associated risk factors differently. Sub-Saharan Africa, rural India, and low-income regions in South America are localized to the first stage as they tend to suffer disproportionately from rheumatic heart disease and cardiomyopathies due to infectious disease. Other regions are dispersed across the next three stages for their differences in coronary heart disease and stroke prevalence. Many HICs have proceeded through the first four stages of the epidemiologic transition and describe coronary heart disease to be the dominant form of CVD, with rates tending to be between two and five folds higher than stroke rates (Gaziano & Gaizano, 2014). The high burdens of CVD in LMICs are posited to be attributable to higher incidences of atherosclerosis due to urbanization and higher risk factor levels (obesity, diabetes, hypertension, etc.) that manifest at an early age. Currently across all countries, CVD has accounted for 6.2 million deaths between the ages of 30 and 70 (Roth et al., 2020).

Linked to the epidemiologic transitions of LMICs and HICs are the economic and social impacts of CVD. The economic burden of CVD is increasing worldwide with the estimated burden cost equating to $863 billion USD in 2010. It is predicted to rise to $1.044 billion USD by 2030, a 22% increase in cost (Kuehn, 2013). Additionally, although the disease burden and the societal and healthcare cost are well established, resources to devote towards healthcare can be scarce. The gross national income per capita of HICs ($12,476) is nearly 12-fold that of a LICs ($1,025) (World Bank, 2021). Additionally, out-of-pocket spending per capita in HICs amounts to $565 whereas it is $18 in LICs (World Health Organization, 2021). Given the limited resources available, only interventions that can lead to large reductions in CVD burden at relatively low cost are likely to be sustainable. The burden of CVD mortality and morbidity has an enormous impact on not only healthcare systems but their patients’ quality of life, productivity and on informal caregivers (Song et al., 2015). Most studies on the burden of CVD analyze on the direct cost, however indirect cost, those arising from productivity losses are much less studied despite representing a major negative impact on individual CVD patients, their families, and society overall (Moreno et al., 2017). In a study conducted by the Health Economic Working Group of the Stent For Life program assessing the socioeconomic impact and clinical benefits of timely interventions for myocardial infarction, the costs of the program were outweighed by the reduction in indirect cost obtained from the number of lives saved demonstrating a positive socioeconomic impact (Wein et al., 2019).

|  |  |  |  |
| --- | --- | --- | --- |
| Stage | Description | CVD-Related Deaths, % | CVD Outcome |
| Stage 1:  Pestilence and Famine | Predominance of malnutrition and infectious diseases; High rates of infant and child mortality; Low mean life expectancy | <10 | Rheumatic heart disease and cardiomyopathies |
| Stage 2:  Receding Pandemic | Improvements in nutrition and public health leads to decreases in malnutrition and infectious disease-related deaths; Declines in infant and child mortality rates | 10-35 | Rhuematic valvular disease, hypertension, CHD, and stroke (hemorrhagic) |
| Stage 3:  Degenerative Diseases | Increases in fat and caloric intake and decreases in physical activity leads to higher rates of hypertension and atherosclerosis; Chronic and noncommunicable disease mortality exceeds malnutrition and infectious disease mortality | 35-65 | CHD and stroke (ischemic and hemorrhagic) |
| Stage 4:  Delayed Degenerative Diseases | CVD and cancer are the major causes of morbidity and mortality; prevention and treatment delay disease primary events; age-adjusted CVD mortality rate declines; CVD affects older adults | 40-50 | CHD, stroke, and congestive heart failure |
| Stage 5:  Endemic Inactivity | Population obesity levels increase; diabetes and hypertension increase; declines in smoking and tobacco use; decreases in the level of the population meeting physical activity recommendation standards | Age-adjusted decline in mortality | CHD, stroke, congestive heart failure, and peripheral vascular disease |

***Abbreviations:***CHE, coronary heart failure; CVD, cardiovascular disease.  
***Source:*** Gaziano, T. A., & Gaziano, J. M. (2014).

**Figure 1.** **Five Stages of Epidemiologic Transition.** A table describing the epidemiologic stages of cardiovascular health and related outcomes.

### **Section 2.1.3: Risk Assessment**

CVD risk assessment begins with an assessment of traditional risk factors (i.e., age, family history, hypertension, dyslipidemia, diabetes, and obesity). Population-based risk calculators (i.e., FRS, SCORE, PROCAM, RRS) are used to estimate both absolute and lifetime CVD risk. Based on this estimation, a risk profile is produced and informs what evidence-based guidelines should be considered when determining preventative treatments. It is common practice to also consider non-traditional risk factors (i.e., metabolic syndrome, autoimmune disease, inflammatory processes, comorbidities, and social determinants of health) before finalizing this risk profile (Khambhati et al., 2018). The risk profile determines form of prevention that is needed by a patient. Preventative care separated into three categories: primordial, primary, and secondary. Primordial prevention is focused on risk factor prevention, through targeting underlying social conditions that prevent disease onset. An example includes improving access to an urban neighborhood to safe sidewalks to promote physical activity, in turn, decreases risk factors for CVD. Primary prevention consists of measures aimed at a susceptible individual or population and institutes activities that limit risk further exposure such as prescribing medication for high blood pressure. Secondary prevention emphasizes early disease detection and commonly takes the form of diagnostic screenings. Tertiary and quaternary prevention are often cited in cardiovascular literature but are concerned with the effects of a disease following its onset (Kisling & Das, 2021).

Epidemiological studies have demonstrated the efficacy of population-based risk assessments in managing CVD burden. The North Karelia Project, a community wide health intervention program, was launched to reduce the toll of heart disease in North Karelia, Finland. The program focused on the reduction of risk factors assessed through population surveys over the course of 15 years. This intervention led to major declines were seen in cholesterol, blood pressure and smoking levels. Coronary-related mortalities were reduced in middle aged adults by 84% from 1972 to 2014 and about two-thirds of the mortality rate decline was explained by risk factor changes. One-third of improvement by new treatments developed since 1980s (Vartiainen, 2018). Conventional risk assessment methods have demonstrated to be effective in improve cardiovascular health outcomes, however, tend to have low discriminative ability when attempting to classify patients who fall between low and high-risk profiles. This creates issues for clinicians regarding how treatments are decided. Additional tools such as biomarkers and noninvasive techniques to better classify are generally used to improve CVD outcomes and treatment for patients (Khambhati et al., 2018).

## **Section 2.2: Artificial Intelligence, Machine Learning, Neural Networks, and Deep Learning**

### **Section 2.2.1: Artificial Intelligence**

AI is a field of computer science concerned with building machines capable of performing task that traditionally require human intelligence (McCarthy & Wright, 2004; Turing, 1950). The criterion of an intelligent machine is its ability to demonstrate forms of intelligence characterized by the cognitive function domains: perception, language, memory, learning, and decision-making (Jiang et al., 2017; Kiely, 2014). The term machine serves as a broad definition for any form of technology that is able to manifest and perform the abilities previously mentioned. Machines in this instance take the form of algorithms and models. Algorithms are the procedures implemented in code and are performed on datasets. Models are the output of algorithms are outputs produced by algorithms and represent what has been learned from the algorithm. A model can be saved and act as a program, storing the functionality of an algorithm to make predictions and conduct analyses on similar datasets (Sidney-Gibbons & Sidney-Gibbons, 2019). In the field of medicine, the primary use of algorithms and models are to interpret and understand health data and make clinical predictions. (Friedrich et al., 2021; Leopold et al., 2020). As a whole, AI contains subfields that categorize the modes in which machines learn and adapt: ML, DL, Neural Networks, Natural Language Processing, and Computer Vision. This review will focus on the ML, neural networks, and the DL subfields, as they represent the majority of clinical use cases of AI (Amisha et al., 2019; Basu et al., 2020; Friedrich et al., 2021) (Figure 2).Timeline

Description automatically generated

**Figure 2.** **Visualization of Artificial Intelligence, Machine Learning, and Deep Learning.** Definitions and visualization of the relationships between AI, ML, and DL

### **Section 2.2.2: Machine Learning**

ML, neural networks, and DL account for the majority of AI applications in medicine because they are subfields of each other. That is, DL is a subfield of neural networks and neural networks is a subfield of ML. ML is defined as the study of computer algorithms and models that can improve through the experience and use of datasets (Mitchell, 1997). Algorithms used in ML fall into three model categories: supervised, unsupervised, and reinforcement learning. The primary distinctions between these three categories are the types of data that is input into the model and how the models use this data to inform their predictions (Krittanawong et al, 2020; IBM Cloud Education, 2020) (Figure 3). Supervised learning models utilize structured data to inform its predictions for a desired outcome. The model incorporates a feedback loop to indicate when a prediction is incorrect and determine that the algorithm parameters need to be altered. If a prediction is incorrect, the algorithm modifies its parameters using the structured data as a reference in an attempt to make a correct prediction. Structured data have specific features that are defined for the model to organize into classes. For data to become structured, manual pre-processing of data must be done in-person. Unsupervised learning utilizes unstructured data and does not incorporate a feedback loop to indicate prediction correctness. Unstructured data does not have to be pre-processed. The algorithm categorizes the data based on similar characteristics or “features” and organizes the data into segments or “classes”. Each class contains a portion of the data set with common features and the model determines its prediction based on patterns and trends within the data. Reinforcement learning utilizes the concept of “states”, “actions” and “rewards”. States serve as input data. States refer to a current situation and actions refer to a finite number of options for what a model can do in a given state. When a model takes an action in a state, it receives a reward. Reward is a concept that describes feedback from the environment. Feedback can be positive or negative (i.e, reward, or punishment). The model utilizes this feedback to inform the actions it selects to reach a desired outcome (IBM Cloud Education, 2020; Mathur et al., 2020).

Diagram

Description automatically generated with medium confidence**Figure 3. Machine Learning Models.** (A) supervised learning (B) unsupervised learning and (C) reinforcement learning model.

### **Section 2.2.3: Deep Learning**

Neural networks, commonly referred to as artificial neural networks (ANN), are a specific form of supervised learning that are inspired by the human brain and mimic the process by which neurons signal to one another. Neural networks are comprised of “nodes” which culminate together to form “node layers”. Neural networks care comprised of node layers consisting of an input layer, hidden layer(s), and an output layer. Nodes share connections with other nodes of varying strengths that represented by “weight”. Weight determines the influence of input data and whether it should impact the prediction of a model. Similarly, to a neuron, other features are permeated in this model such as activation thresholds, forward propagation, and back propagation. These all contribute to a neural network’s ability to make robust predictions with complex data (Figure 4). DL is an extension of neural networks. The “deep” in DL refers to the depth of layers in a neural network. A DL algorithm is a neural network consisting of four or more layers, including the input and output layers. The distinct difference between ML and DL is that ML is more reliant on human intervention to learn. DL can leverage structured and unstructured data and automatically determine features for which to create classes from (Haleem et al., 2021; IBM Cloud Education, 2020; Mathur et al., 2020).

Diagram

Description automatically generated

**Figure 4. Artificial Neural Network.** (A) an individual model of an individual input – node – output connection (i.e., McCulloch–Pitts neuron model) . (B) a 4-layer ANN architecture.

## **Section 2.3: Artificial Intelligence Clinical Applications**

### **Section 2.3.1: Precision Medicine**

Precision medicine is an integrative approach to healthcare via optimizing treatments based on a patient’s biological, psychological, and social environments. Precision medicine applications in cardiology is fairly novel however attempts to directly address the gaps in knowledge surrounding the GECP cardiovascular relationship (Leopold & Loscalzo, 2018; Yan et al., 2019). Medical reductivism assumes patients that exhibit similar symptomology share a common pathophenotype, should be provided similar treatment. Current literature, in contrast, suggest that CVDs and their similar symptomology can originate from different genotypic, endophenotypic, and phenotypic origins (Leopold & Loscalzo, 2018; Leopold et al., 2020; Shameer et al., 2018). Currently, precision medicine approaches have been applied to disease classification, diagnosis, risk prediction, and patient management most commonly through the use of supervised learning algorithms (logistic regression (LR), decision tree (DF), random forest (RF), and support vector machine(SVM)) and DL models (convolutional neural networks and recurrent neural networks) (Sevakula et al., 2020). AI has demonstrated its greatest utility as a clinical tool through classifying clinical health data and panomic data for more accurate patient phenotype predictions.

In a study by Juhola et al., the authors were able to classify different genetic CVD on the basis of Ca2+ transient profiles using ML. The diseases of interest were catecholaminergic polymorphic ventricular tachycardia, long QT syndrome, and hypertrophic cardiomyopathy. Utilizing k-nearest neighbors (KNN), RF, and SVM, classification accuracy of up to 87% was obtained and determined Ca2+ transient was disease specific. In another study, Kera et al. examined the relationship between familial hypercholesterolemia and polygenic risk score pathways for early-onset myocardial infection. Using LR models adjusted for race, the authors found that the presence of familial hypercholesterolemia genomic mutations and high polygenic scores are associated with a >3-fold increased odds of early-onset myocardial infarction. Additionally, that a high polygenic score has a 10-fold higher prevalence among patients with early onset-myocardial infarction. Several large genome association studies suggest that the advancement of precision medicine is contingent on the ability of AI to create endotype classifications based on its current genotypic and phenotypic abilities (Kalinin et al., 2018; Li et al., 2015).

### **Section 2.3.2: Clinical Prediction**

The implementation of AI in clinical prediction is through the use of clinical decision support systems. Clinical decision support systems are computer-based programs that analyze electronic health records to assist clinicians in providing evidence-based clinical recommendations at the point of care (Centers for Disease Control and Prevention, 2017). The distinction between AI clinical decision support systems and clinical decision support systems is the ability to learn and adapt based on feedback, synonymous with the definitions of supervised and reinforcement learning (Choi et al., 2020). In cardiovascular medicine, the aims of AI clinical decision support systems are to administer support directly to patients through personal health records or other systems, augment the extraction and visualization of medical images and laboratory test, and to support with the adherence of clinical guidelines (Sutton et al., 2020). AI clinical decision support systems have seen use in both in-patient and out-patient settings but are argued to have the best utilization in intensive car unit and hospitalization settings (Sevakula et al., 2020).Outside of clinical settings, applications in mobile and wearable technology have additionally, provided opportunities for early diagnosis and prevention of CVDs (Piette et al., 2015).

Ong et al. conducted a study utilizing support vector machine algorithms to predict the likelihood of cardiac arrest within 72 hours in hospitalized patients and compared it to an existing department early warning system. Features such as age, heart rate and vital signs were used. Patients were assigned a risk score of either low risk, intermediate risk, or high risk. The patients were then assigned to one of two categories, patients predicted to experience cardiac arrest or death as outcomes and, patients predicted to experience neither outcome. According to the study, ML model was more accurate in predicting cardiac arrest. Additionally, the author suggest that risk stratification based on cardiac arrest prediction could be conducted with further study. Similar studies by Dawes et al. and Motwani et al. created custom ML models to evaluate survival probabilities in patients with pulmonary hypertension and coronary heart disease respectively. Dawes et al., found that patient survival rates comparable to clinicians could be found based on a combination of electronic health records, diagnostic blood test, and cardiac MRIs. Motwani et al. found that with the use of coronary computed tomographic angiographs, the ML model was able to surpass clinician prediction.

### **Section 2.3.3: Diagnostic Imaging**

Current literature indicates that in the short term, AI has the capacity to reduce human error and save time in the clinical workflow through automated classification of anatomical structures. In the long term, AI may expand the clinical value of diagnostic images either based on images alone or in combination with other clinical data (Lim, Tison, & Delling, 2020). AI permeates the field of diagnostic imaging including echocardiography, nuclear imaging, computed topography, and magnetic resonance (Yan et al., 2019). This permeation is a result of the inability to conduct traditional analytic methods to process and model imaging data. This inability is best described as a product the three main characteristics of big data: volume, velocity, and volume. According to other definitions, big data is also characterized by veracity, the concern of quality, authenticity, and trustworthiness of the data (Lee & Yoon et al., 2017; Pastorino et al., 2019). While all of these factors contribute to the complexity of medical data overall, diagnostic imaging data is argued to be disproportionately affected. It is estimated that 20% of Medicare patients undergo echocardiography examinations, accounting for approximately 7.07 million echocardiography examinations. Given that a single examination outputs 2 GB of data, this means approximately 14 PB worth of results are collected annually (Shameer et al., 2018). However, due to AI and its ability to handle data with large inputs, it can make each step of imaging less workflow intensive from image acquisition to reporting while improving overall optimal patient care.

In a study conducted by Madani et al., the authors trained a convolutional neural network to classify 15 echocardiograph views from labeled still images and videos using a dataset of over 200,000 images from 240 studies to train and validate the model. The authors also included clinicians to serve as a comparison. The model achieved an accuracy of 91.7% in comparison clinician accuracy of 84%. Additional studies such as Samad et al. and Tabassaian et al. similarly demonstrated that DL models can predict survival likelihoods with higher accuracy than traditional assessments with the use of echocardiograph training datasets. A large number of studies unlike the ones previously mentioned, use a supervised approach to training their DL models. This approach requires a significant amount of expertise and time to preprocess data, however there are methods that exist to hasten the process. While AI has been demonstrated to contribute to the issue of big data, current methods must be improved upon to further contribute to the field of cardiology and medicine as a whole (Chen et al., 2017).

## **Section 2.4: AI Concerns and Limitations**

### **Section 2.4.1: Data Consent**

AI applications are currently in the process of transforming the patient clinician relationship. However, this transition to AI assisted care delivery prompts the question of how AI assist in the principles of patient informed consent will. Informed consent is one of the most immediate challenges of integrating AI into clinical. Additionally, there is the conundrum of informed consent should be used to train AI (Cohen et al., 2014; Gerke, Minssen, & Cohen, 2020). Questions surrounding to what extent clinicians are responsible for educating patients about AI, the potential risk of bias and misprediction, and whether patients should be notified of the use of AI at all. This becomes more complicated when ML algorithms use techniques that are noninterpretable to clinician and patient alike. This lack of knowledge then poses the question of to what extent does a clinician need to know to disclose that they cannot fully interpret or understand a treatment recommendation.

Online health applications and chat bots are also being used more, as clinicians move to online platforms. These applications raise the question about user agreements and their relationship to informed consent. In-person health assessments require a user agreement that an individual agrees to through person-to-person dialog. In contrast, online applications do not require this, and it can be argued that most people do not read or comprehend user agreements. In addition, regular software updates make it difficult for user to know how their terms of service has been changed (Cohen et al., 2014; Sevakula et al., 2020). Research must be conducted to attempt to address the question what an ethical user agreement looks like in this telehealth context and how can we better educate consumers of the future use of online applications that are conditional on accepting changes of terms of service.

### **Section 2.4.2: Data Privacy**

A significant concern regarding the use of AI are the systems in place designed for the use and oversight of patient data. A majority of existing technology that implements ML is proprietary and owned by a small number of tech corporations. Google, Amazon, Apple, Microsoft, and IBM and others are large proprietors of technological implementations within the healthcare sector (Powles & Hodson, 2017). The industry standard of privacy protection comes in the form of information sharing agreements. These agreements grant private entities access to patient health information, however either fail to provide adequate protection of patient data either through poor implementation or contract clause loopholes. DeepMind, a Google subsidiary, entered a partnership with the Royal Free London National Health Service Foundation Trust in 2016 to implement ML to assist acute kidney management initiatives (Cuttler, 2019; Murdoch 2021). Critics of the partnership pointed to a lack of patient agency over the use of their health information, furthermore, the privacy agreements were never adequately discussed. Additionally, Google requisitioned United Kingdom sectors of DeepMind, effectively transferring patient data to the United States (Murdoch 2021; Vincent, 2018). The ability mass accumulate data through means of corporate annexation is a reality of commercial healthcare and can even extend to the power dynamics of partnerships between public and private institutions.

Another concern is the risk of privacy breaches patient data from the ML models and algorithms themselves. Studies have been conducted demonstrating how computational strategies utilizing existing models can identify patients in health data repositories managed by private and public institutions. This is true even after information is anonymized and redacted of all identifiers. A study by Na et al., found that an algorithm could be used to re-identify 85.6% of adults and 69.8% of children in a physical activity cohort study, despite removal of protected health information. Furthermore, a 2019 study successfully used a “linkage attack framework”, an algorithm aimed at re-identifying anonymous health information, which could link online health data to real world people, demonstrating the vulnerabilities of existing online health data (Ji et al., 2019). The use of AI and the ease of access to health data, even in circumstances where anonymity occurs, raises questions of liability and insurability for both private and public institutions. Considering the direction of AI in healthcare, agreements delineating the rights and obligations of all involved parties must be made apparent. Concurrently, solutions must explore to address the inevitable actions that harmful parties will take to breach patient health data privacy.

### **Section 2.4.3: Data Transparency**

AI not only has the capability to improve healthcare but democratize expertise and provide capacity to low resource areas (Wahl et al., 2018). However, ML models and algorithms are only as effective as the data it was trained with. AI can also inherit bias and lead to discrimination or misrepresentation of study populations. Therefore, it is important that clinicians and programmers are aware of the impact of their inherent bias when designing and interacting with AI at every stage of development (Mittelstadt et al., 2016; Char, Shah & Magnus, 2018). Considering the appropriate ML model and datasets to use to train algorithms are the most effective ways to consider the risk of bias. Several studies have demonstrated that algorithms can exhibit bias based on age, race, gender, ethnicity, and disability. (Gerke et al., 2020). These studies have found that unrepresentative datasets tend to be the primary cause. If this is representative cause of algorithmic bias, then bias may be resolved due to increased data availability and better attempts to collect data from underrepresented populations.

Another concern of AI is the tradeoff between ease of use and data transparency. ML models are adept in analyzing and compiling data, but as models become more convoluted, it becomes difficult for clinicians to derive explanations for how predictions and health recommendations are being made. These more complex models such as neural networks are referred to as “black box” models and algorithms (Rowe et al., 2019). The black box issue compiled with the issue of ML models focusing on the strength of correlations between variables rather than their causality, many clinicians demonstrate concerns surrounding the ethicality of their use within a clinical setting. Proponents for AI argue that what matters is not how the model or algorithm reaches its decision, but that it is accurate (London, 2019). The safety and effectiveness of algorithms and models that are black boxes could, be demonstrated, similar to the handling of drugs, by positive results of randomized clinical trials. Regardless, clinical judgement and rationale must be paramount in determining when and how to use AI in a transparent and ethical manner.

# **Chapter 3: Methods**

The objective of this review was to systematically identify, collate, and evaluate published literature that details DL model prediction in CVD and compare the predictive accuracy between DL models and in comparison, to conventional risk assessment methods.

This review is reported in accordance with the Preferred Reporting Information for Systematic Reviews and Meta-Analysis (PRISMA) recommendations. See Appendix A-1 for the full the PRISMA checklist. Ethical approval was not required for this study.

## **Section 3.1: Search Strategy**

A comprehensive search strategy was designed and conducted in electronic databases and published referenced list to identify articles employing one or more DL models for CVD between January 2012 and January 2022. Cochrane, Embase, Scopus, and Web of Science databases were searched (January 2022 - March 2022). A manual search and completeness reviews of articles, bibliographies, and citations were conducted to identify additional relevant studies. The author designed and conducted the search strategy. See Appendix A-2 for the full search strategy.

## **Section 3.2: Study Selection**

Search results were exported from all databases and imported into Covidence, an online systematic review tool. Duplicates were identified, screened, and removed using Covidence’s de-duplication functionality. Two eligibility assessments were conducted to identify studies that met the pre-specified eligibility criteria. The initial screening examined the titles and abstracts of studies retrieved from the Covidence portal. Abstracts with sufficient data regarding evaluation data, methodology, and outcome definitions were included. The secondary screening examined the full content of initially screened studies. Studies with sufficient data regarding contingency tables, evaluation data, methodology, and outcome definitions were included. Studies without external or internal DL model validation methods and/or sufficient performance metrics were excluded. All reviews, editorials, letters, and non-human studies were excluded. No limits were placed on the target population, the disease outcome of interest, or intended use case of the DL models. The author conducted the pre-specified initial and secondary screening.

## **Section 3.3: Data Extraction**

The following characteristics were extracted from each study: First Author, Model, Year of Publication, Sample Size, Clinical Indication, Diagnostic Imaging, and Performance Measures. The performance measures were intended to consist of reported sensitivity, specificity, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In cases where studies did not report positive and negative cases, performance measures were reverse calculated with the use of the positive predictive value (PPV), negative predictive value (NPV), sample size and threshold limit granted they were provided (Figure 5). See Appendix A-3 & A-4 for study characteristics and performance metrics.

Table

Description automatically generated

**Figure 5. Confusion Matrix / Contingency Table.** A 2x2 cross-tabulation (bolded 2x2 square) representing the instances of an actual class and predicted class. Performance metrics listed in this cross tabulation are reported by standard in epidemiological and machine learning literature.

## **Section 3.4: Statistical Analysis**

Bivariate summary receiver operating characteristic (SROC) models were used to jointly estimate sensitivity, specificity and the arear under the receiver operating characteristics curve (AUC). Subgroup analysis was stratified by CVD (cardiac arrythmia, coronary artery disease, heart failure, and stroke) and DL model when applicable. All statistical analyses were performed using R version 4.1.2 (Mada and Meta packages) (Doebler & Holling, 2020; R Core Team, 2021).

# **Chapter 4: Results**

## **Section 4.1: Study Search**

The database search yielded 1688 results published between January 2012 and 2022. Duplicates (n = 257) were removed by Covidence. Following the screening process, 252 articles were selected for full-text review. After full-text and supplementary review, 227 studies were excluded due to insufficient data, inaccessibility, and no specified disease target among other exclusion criteria, resulting in the extraction of 25 studies. The disposition of excluded studies following the full-text review (Figure 6).

Text

Description automatically generated

**Figure 6. Study Design.** A flow chart illustrating the selection process for published studies.

## **Section 4.2: Study Characteristics**

Appendix A-3 shows the basic characteristics of the studies included in the meta-analysis. The 25 patient studies were characterized by cardiac arrhythmia (1 study), coronary artery disease (14 studies), heart failure (4 studies), and stroke (6 studies). Reference standards were varied in line with the target condition and the modality of diagnostic prediction being used, with some studies using multiple methods to diagnose the condition of interest. Conventional risk assessment methods consisted of clinicians, risk assessment tools (FRS), imaging indices (THRIVE, HIAT, and SPAN101), and diagnostic analysis (CT-FFR and HRV).

## **Section 4.3: Study Results**

### **Section 4.3.1: Machine Learning Models and Prediction of Cardiovascular Disease**

The 25 studies provided sufficient information to enable the calculation of contingency tables and test parameters, with a total of 71 tables across these studies. In line with the aims of the review, all studies were included regardless of the disease target. Averaged across studies, the pooled performance metrics for machine learning models (n = 53 contingency tables) reported an AUC = 0.88, sensitivity = 0.77, 95% CI [0.64, 0.84], and specificity = 0.88, 95% CI [0.82, 0.91] (Figure 7). Conventional methods (n = 18 contingency tables) reported an AUC = 0.74, sensitivity = 0.73, 95% CI [0.64, 0.88], and specificity = 0.67, 95% CI [0.59, 0.74] for all cardiovascular diseases (Figure 8).

Diagram

Description automatically generated

**Figure 7. Machine Learning Bivariate SROC Curve (Cardiovascular Disease)**. Bivariate SROC curve of all machine learning contingency tables (n = 53). The prediction accuracy is characterized by an AUC = 0.88, sensitivity = 0.77, and specificity = 0.88.

Diagram

Description automatically generated

**Figure 7. Machine Learning Bivariate SROC Curve (Cardiovascular Disease)**. Bivariate SROC curve of all conventional method contingency tables (n = 18). The prediction accuracy is characterized by an AUC = 0.74, sensitivity = 0.73, and specificity = 0.67.

### **Section 4.3.2: Machine Learning Models and Prediction of Cardiac Arrythmias**

For cardiac arrythmias, 1 study used a total of 3039 individuals. This study used a CNN model (Poh et al., 2018). Conventional methods, pooled algorithmic and algorithmic specific statistical analysis could not be performed due to an insufficient sample size (≤ 5 contingency tables per model).

### **Section 4.3.3: Machine Learning Models and Prediction of Coronary Artery Disease**

For coronary artery disease, 13 studies included a total of 36278 individuals. Two study used ANN models, (Joloudari et al., 2022; Miao et al., 2018), 1 study used a boosting algorithm (Han et al., 2018), 1 study used a CNN model (Muscogiuri et al., 2019), 7 studies use custom built algorithms (Baumann et al., 2019; Coenen et al., 2018; Dutta et al., 2020; Eisenberg et al., 2021; Freiman et al., 2017; Xiuhua et al., 2018; Zellweger et al., 2018) and 2 studies used SVM models (Araki et al., 2016; Song et al., 2014). The pooled performance for machine learning algorithms (n = 18 contingency tables) reported an AUC = 0.83, sensitivity = 0.80, 95% CI [0.74, 0.85], and specificity = 0.79, 95% CI [0.72, 0.85] (Figure 9). Conventional methods and algorithmic specific statistical analysis could not be performed due to an insufficient sample size (≤ 5 contingency tables per model).

Diagram

Description automatically generated

**Figure 9. Machine Learning Bivariate SROC Curve (Coronary Heart Disease)**. Bivariate SROC curve of all reported machine learning model contingency tables for coronary heart disease (n = 18). The prediction accuracy is characterized by an AUC = 0.83, sensitivity = 0.80, and specificity = 0.79.

### **Section 4.3.4: Machine Learning Algorithms and Prediction of Heart Failure**

For heart failure, 4 studies used a total of 173734 individuals. Two study used multiple algorithms for comparison (Blecker et al., 2018; Weng et al., 2017). Two studies used ANN algorithms (Attia et al., 2020; Weng et al., 2017), 1 study used a boosting algorithm (Weng et al., 2017), 1 study used custom built algorithms (Blecker et al., 2018), 1 study used a LR model (Blecker et al., 2018), 1 study used a RF model (Weng et al., 2017), and 1 study used a SVM model (Rossing et al., 2016). The pooled performance for machine learning algorithms (n = 9 contingency tables) reported an AUC = 0.90, sensitivity = 0.89, 95% CI [0.76, 0.95], and specificity = 0.83, 95% CI [0.75, 0.88] (Figure 10). Conventional methods and algorithmic specific statistical analysis could not be performed due to an insufficient sample size (≤ 5 contingency tables per model).

Diagram

Description automatically generated

**Figure 10. Machine Learning Bivariate SROC Curve (Heart Failure)**. Bivariate SROC curve of all reported machine learning model contingency tables for heart failure (n = 7). The prediction accuracy is characterized by an AUC = 0.90, sensitivity = 0.89, and specificity = 0.83.

### **Section 4.3.5: Machine Learning Algorithms and Prediction of Stroke**

For stroke, 6 studies used a total of 885 individuals. Multiple studies used multiple algorithms for comparison (Arslan et al., 2016; Bacchi et al., 2020; Dharmasaroja et al., 2013; Ho et al., 2018; Matsuo et al 2021). Three studies used ANN algorithm (Bacchi et al., 2020; Dharmasaroja et al., 2013; Matsuo et al., 2021), 3 studies used boosting algorithm (Arslan et al., 2016, Ho et al., 2018; Matsuo et al., 20201), 3 studies used CNN algorithm (Bacchi et al., 2020; Beecy et al., 2017; Dharmasaroja et al., 2013), 1 study used a LI model (Ho et al., 2018), 1 study used a LR model (Matsuo et al., 2021), 1 study used s RF (Ho et al., 2018), and 3 studies used SVM models (Dharmasaroja et al., 2013; Ho et al., 2018; Matsuo et al., 2021). The pooled performance for machine learning models (n = 24 contingency tables) reported and AUC = 0.83, sensitivity = 0.59, 95% CI [0.42, 0.75], and specificity = 0.91, 95% CI [0.83, 0.96] (Figure 11). Conventional methods and algorithmic specific statistical analysis could not be performed due to an insufficient sample size (≤ 5 contingency tables per model).

Diagram

Description automatically generated

**Figure 11. Machine Learning Bivariate SROC Curve (Stroke)**. Bivariate SROC curve of all reported machine learning model contingency tables for stroke (n = 24). The prediction accuracy is characterized by an AUC = 0.83, sensitivity = 0.59, and specificity = 0.91.

### **Section 4.3.6: Matched Prediction of Cardiovascular Disease**

Of the 8 studies that provided reference standards for comparison, a single contingency table reporting the highest accuracy for both models and conventional methods were used (Bacchi et al., 2020; Baumann et al., 2019; Coenen et al., 2018; Han et al., 2018; Song et al., 2014; Weng et al., 2017; Xiuhua et al., 2018; Zellweger et al., 2018). The pooled performance metrics for machine learning models (n = 8 contingency tables) reported an AUC = 0.82, sensitivity = 0.72, 95% CI [0.63, 0.80], and specificity = 0.81, 95% CI [0.75, 0.86] (Figure 12). Conventional methods (n = 9 contingency tables) reported an AUC = 0.78, sensitivity = 0.75, 95% CI [0.66, 0.83], and specificity = 0.77, 95% CI [0.73, 0.81] (Figure 13).

Diagram

Description automatically generated

**Figure 12. Machine Learning Bivariate SROC Curve (Matched)**. Bivariate SROC curve of comparator matched machine learning model contingency tables for CVD (n = 9). The prediction accuracy is characterized by an AUC = 0.82, sensitivity = 0.72, and specificity = 0.81.

Diagram

Description automatically generated

**Figure 13. Machine Learning Bivariate SROC Curve (Matched)**. Bivariate SROC curve of comparator matched reported conventional method contingency tables for CVD (n = 9). The prediction accuracy is characterized by an AUC = 0.78, sensitivity = 0.75, and specificity = 0.77.

# **Chapter 5: Discussion**

## **Section 5.1: Discussion of Key Findings**

Findings suggest that ML models share similar accuracy, sensitivity, and specificity values to conventional risk assessment methods. The pooled analyses suggest that ML models are accurate in overall CVD prediction, coronary artery disease prediction, heart failure and stroke prediction. Statistical could not be performed for cardiac arrythmias due to insufficient data. Although these estimates support the claim that ML models are accurate in clinical prediction and can match, albeit surpass conventional risk assessment predictive accuracy; several methodological issues were common across many included studies must be taken considered. Additionally, significant limitations within the meta-analysis itself are to be addressed to provide an informed interpretation of the results.

Most studies in the review assessed ML model performance either in isolation or in comparison to another ML model(s). This form of study design is prevalent within the domain of AI. To date, only two meta-analyses have been conducted with similar underlying concepts and interest similar to this review (Krittanawong et al., 2020; Liu et al., 2019). Both reviews have also found that a majority of studies screened do not report comparisons with clinicians, conventional risk scores, or diagnostic imaging techniques. When comparison groups are reported, they are typically small. This study had a median of two comparators. There can be wide intra- and inter- case/cohort variation between comparators; especially when it comes to clinicians. Additionally, the inclusion of non-expert clinicians who interpret risk scores and/or diagnostic imaging techniques may mitigate human performance (Nagendran et al., 2020).

The studies reported a wide range of performance metrics to evaluate the performance of the employed ML models and comparators. The minimum data needed for a meta-analysis to be conducted is a contingency table (i.e., the frequency of true positives, false positives, true negatives, and false negatives), the sample size, and the specific threshold for the receiver operator characteristics curve. In this review, a number of studies provided incomplete contingency tables which required reverse calculations based on positive and negative predictive values if present. Additionally, no studies reported the threshold at which sensitivity and specificity were reported. It is a convention in ML training and development to set the threshold at an arbitrary value of 0.5. Conducting risk assessment test at a prevalence of 50% proves to be of little use for clinical practice (Nagendran et al., 2020). Performance metrics that were commonly provided were accuracy, precision, and the F1 score, metrics that are reported as common practice in DL development literature (Uddin et al., 2019). A standard for data reporting should be adopted, with at minimum reporting contingency tables and sample sizes.

There was inconsistent use in the term “validation”. In ML, there are three datasets that are commonly used in different stages of model creation: training, validation, and testing sets. The training set trains the model and allows for it to learn hidden features and patterns in the data. The validation set is data, independent from the training set which is used to tune a ML model’s performance. The test set is another separate set of data used to test the ML model after completing the training set. Validation refers to the process of evaluating a model with the testing set. The studies in this meta-analysis used the term interchangeable with some authors using it appropriately for testing the ML models and others describing the tuning of a ML model (IBM Cloud Computing, 2020; Uddin et al., 2019). This inconsistent use has made it difficult to verify with certainty the independence of the test set, but also the types of validation if not outright reported. The reversal of the meaning of validation and test sets is not novel to this meta-analysis and the included studies but pervades AI research. Understanding this, it may be of benefit to adopt a standard of distinguishing the datasets involved for the training, validation, and testing sets (Altman & Royston, 2000; Krittanawong et al., 2020).

## **Section 5.2: Strengths and Weaknesses**

To date this is one of few reviews with the intent of evaluating AI literature for both efficiency and accuracy outside of isolation. Additionally, this is one of few reviews to acknowledge the lack of reporting standards within the field of AI. Regardless, this study provides encouraging findings albeit its limitations when comparing prediction accuracy between ML models and conventional risk assessment methods. Such findings point to strong clinical applicability in CVD assessment, especially coronary heart disease and stroke. This brings to attention the need for more research and ML assessments of cardiac arrythmias and heart failure, CVDs that were unable to be analyzed in this review. There are limitations that occurred within the methodology and design of the study that should be placed under scrutiny when evaluating its findings. This meta-analysis was subject to single screening rather than conventional double screening. Studies suggest that single-viewer abstract screenings do not fulfill the comprehensive standards that decision makers expect from systematic reviews. It can be argued that while this may be a more viable option for rapid reviews, this is at the cost of methodologic rigor, regardless of expertise (Higgins et al., 2022; Gartlehner et al., 2020). In addition, a systematic quality assessment for the transparency of reporting in this review was not conducted. This decision was based in the understanding that there was insufficient guidelines and literature on how to apply AI assessments to clinical prediction analysis. Quality assessment standards in AI predictive accuracy are fairly new and meta-analyses are currently being researched looking at the applicability of general predictive accuracy assessment tools such as the Quality Assessment of Diagnostic Accuracy Studies (QADAS) (Sounderajah et al., 2020).

## **Section 5.3: Recommendations**

This review demonstrates the need for quality assessment standards in ML predictive accuracy studies and reviews and highlights variations in ML specific methodological aspects and reporting across different factors of study design. These factors include study generalizability, reporting guidelines, terminology, and the development of training, validation, and testing datasets. When evaluating study quality and applicability of ML predictive accuracy, it is important to consider these factors. While there are tools that exist to aid in evaluating ML models, it is left to the author to modify and signal the criteria as needed to evaluate predictive accuracy and translate their findings into a clinical setting. Based on the results and limitations of this review, the following recommendations are proposed for future studies.

### **Section 5.3.1: Quality Assessment Tools**

The perceived limitations of current quality assessment tools highlight the need for a ML specific guideline to evaluate predictive accuracy, this review included. Currently, a number of quality assessment tools are either being created or modified to counteract the limitations previously mentioned. Standards for Reporting of Diagnostic Accuracy Studies (STARD) and Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) are currently in the process of being extended to develop guidelines that evaluate ML diagnostic accuracy studies. As of current, Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and Consolidated Standards of Reporting Trials (CONSORT) have been published (Collins & Moons, 2019; Liu et al., 2020; Sounderajah et al., 2020). Each assessment tool shares differences in their evaluation approach and their suitability may vary depending on the underlying focuses of one’s review. Per example, TRIPOD focuses on regression-based ML models and provides insufficient guidance on how to apply its checklist to models outside of the regression-oriented framework. With this in mind, employing multiple assessment tools in parallel may balance the evaluation of predictive accuracy studies and maximize the benefits of ML for future clinical diagnostics and care (Sounderajah et al., 2020).

### **Section 5.3.2: Meta-Analysis of Test Performance Gold Standards**

The meta-analysis of predictive test accuracy studies differs from that of interventional studies in that they are required to simultaneously analyze two outcome measures rather than one. In this meta-analysis, these measures are sensitivity and specificity; Sensitivity and specificity are generally inversely correlated and can be affected by a threshold effect. The threshold effect describes an effect in outcome that does not occur until a certain level, or threshold, is reached in a predictor (Lee et al., 2015). For example, a drug may have no effect until a certain dosage level is reached. Additionally, diagnostic test studies generally have larger between-study heterogeneity and can vary depending on a study’s design, study population, and reference standards. It is standard practice to utilize either SROC or hierarchal summary receiver operating characteristics (HSROC) curve to account for both correlation between sensitivity and specificity, and within- and between-study heterogeneity (Shim, Kim, & Lee, 2019). In this study, SROC was chosen due to its simplicity. In future studies where multiple thresholds are reported, HSROC should be used. In this study, thresholds were reverse calculated to compensate for lack of study transparency (Rutjes et al., 2007). The importance of choosing between SROC and HSROC lies in the distinction between how the predictors are used to explore between study heterogeneity. For example, “spectrum bias,” where the subjects included in a study are not representative of the patients who will receive the test in practice,might impact test accuracy rather than the threshold, and is advised to be investigated using the HSROC. SROC would be more appropriate if there was a need for direct evaluation of predictors’ impact on sensitivity, specificity, or both (Trikalinos et al., 2012).

### **Section 5.3.3: Clinical Level Contextualization**

Despite the underlying focus of the review comparing predictive accuracy between ML models and conventional risk assessment methods, the aspects of translation to a reliable clinical design are overlooked. The feasibility of conducting a review that provides both quantitative and qualitative assessments of predictive accuracy within a specific field of medicine may be low, though it should be entertained. Qualitative interviews of clinicians have led researchers to determine that a well-designed ML model design should address (Tonekaboni et al., 2019): Feature Importance, Instance Level Explanations, Uncertainty, and Design Transparency

Feature importance describes the techniques that determine what features or predictors are relevant for determining an outcome (Saarela & Jauhiainen, 2021).Clinicians describe that knowing the subset of features that are used to help a model reach its outcome, is imperative. This knowledge allows for them to compare their clinical judgement to the model decision. Clinicians also necessitated a distinction between the use of population-level features and patient-level features. Patient-level features are far less explored in ML medicine applications (Tonekaboni et al., 2019).

Instance level explanations are a set of features that are considered most responsible for the prediction of an outcome; where feature importance identifies what is relevant, instance level explanations assign them by level of importance (Murdoch et al., 2019). Clinicians have described these explanations as helpful only when the population used for training of the model is similar to their current patients. However, despite similar outcomes in patients with these instance level features, the clinical trajectories that led these patients to their outcomes may differ significantly (Dabek & Caban, 2016). It may be of interest to clinicians and also future studies to provide information if accessible, surrounding previous clinical interventions that were taken in certain cases and the associated outcomes of these interventions (Tonekaboni et al., 2019).

Research suggests that overall ML predictive accuracy is not enough to influence trust and sustained use, but rather its alignment with clinical judgement. Clinicians anticipate a clinical change in a patient's health status that aligns with the prediction of the respective DL models (Guidi et al., 2015; Tonekaboni et al., 2019; Umscheid et al., 2015) . This understanding is important because it provides insight into the expectations of clinicians. There is an expectation that DL models are to trigger alerts or produce predictions that would be in line with the occurrence or change in health status of a patient. More so, that the sustained use and trust in the dep learning model is based on this form of performance. The use of a certainty score is not an uncommon solution that has been presented to augment clinician decisions but is rarely something that is implemented in ML predictive accuracy studies (Abdair et al., 2021; Valen et al., 2022) . At minimum, mentioning model uncertainty and data uncertainty can help impact DL model mistrust.

A need for ML models to mirror the process of evidence-based decision making is a very important feature in the opinion of healthcare clinicians. A ML model should share emphasis on identifying and recognizing the clinical features that drive prediction. Having a transparent design is useful for healthcare clinicians who may want to be able to understand and rationalize the ins and outs of the ML model’s decision-making process (Cutillo et al., 2020; Mourby, Cathaoir, & Collin 2021; Nagendran et al., 2020). Using rule-based methods, regularization-based methods, and model distillation are all attempts at enhancing interpretability of ML algorithm results. The current understanding is that there is a trade-off between simplicity and performance of ML models and if this is the case as of current, no evaluations as of late have explored the acceptability of the potential trade-offs (Tonekaboni et al, 2019).

## **Section 5.4: Conclusion**

Although there are a number of methodological issues and review limitations to overcome to be able to implement ML algorithms in clinical practice, overall results suggest future promising results. Prediction of CVD outcomes are of similar accuracy to that of conventional l risk assessment methods. The limitations of this are intrinsically linked to the need for ML predictive accuracy study standards that necessitate adherence to appropriate guidelines. Prospective studies comparing conventional risk assessments and clinicians, reporting of all contingency table information, and explicit labeling of datasets are needed for the proper interpretation of review findings for a clinical context. Only then could it be conceivable that ML models could be applied to clinical practice and support clinical decision making.

# **References**

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model?. *Statistics in medicine*, *19*(4), 453–473. [https://doi.org/10.1002/(sici)1097-0258(20000229)19:4<453::aid-sim350>3.0.co;2-5](https://doi.org/10.1002/(sici)1097-0258(20000229)19:4%3c453::aid-sim350%3e3.0.co;2-5)

Amisha, Malik, P., Pathania, M., & Rathaur, V. K. (2019). Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*, *8*(7), 2328–2331. <https://doi.org/10.4103/jfmpc.jfmpc_440_19>

Basu, K., Sinha, R., Ong, A., & Basu, T. (2020). Artificial Intelligence: How is It Changing Medical Sciences and Its Future?. *Indian journal of dermatology*, *65*(5), 365–370. <https://doi.org/10.4103/ijd.IJD_421_20>

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, 25–60. <https://doi.org/10.1016/B978-0-12-818438-7.00002->2

Briganti, G., & Le Moine, O. (2020). Artificial Intelligence in Medicine: Today and Tomorrow. *Frontiers in medicine*, *7*, 27. <https://doi.org/10.3389/fmed.2020.00027>

Centers for Disease Control and Prevention. (2021, July 22). *Implementing clinical decision support systems*. Centers for Disease Control and Prevention. Retrieved April 15, 2022, from https://www.cdc.gov/dhdsp/pubs/guides/best-practices/clinical-decision-support.htm

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *The New England journal of medicine*, *378*(11), 981–983. <https://doi.org/10.1056/NEJMp1714229>

Choi, D.J., Park, J.J., Ali, T., & Lee, S. (2020). Artificial intelligence for the diagnosis of heart failure. *NPJ Digital Medicine, 3*. <http://doi.org/0.1038/s41746-020-0261-3>

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health affairs (Project Hope)*, *33*(7), 1139–1147. <https://doi.org/10.1377/hlthaff.2014.0048>

Collins, G. S., & Moons, K. (2019). Reporting of artificial intelligence prediction models. *Lancet (London, England)*, *393*(10181), 1577–1579. <https://doi.org/10.1016/S0140-6736(19)30037-6>

Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., & Mandl, K. D. (2020). Machine intelligence in healthcare—Perspectives on trustworthiness, explainability, usability, and transparency. *Npj Digital Medicine*, *3*(1), 1–5. <https://doi.org/10.1038/s41746-020-0254-2>

Cuttler, M. Transforming health care: how artificial intelligence is reshaping the medical landscape. CBC News. 2019. [https://www.cbc.ca/news/health/artificial-intelligence-health-care-1.5110892. Accessed 15 Apr 2022](https://www.cbc.ca/news/health/artificial-intelligence-health-care-1.5110892.%20Accessed%2015%20Apr%202022).

Dabek, F. J., & Caban, J. J. (2017). A K-Reversible Approach to Model Clinical Trajectories. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2016*, 460–469.

Dawes, T., de Marvao, A., Shi, W., Fletcher, T., Watson, G., Wharton, J., Rhodes, C. J., Howard, L., Gibbs, J., Rueckert, D., Cook, S. A., Wilkins, M. R., & O'Regan, D. P. (2017). Machine Learning of Three-dimensional Right Ventricular Motion Enables Outcome Prediction in Pulmonary Hypertension: A Cardiac MR Imaging Study. *Radiology*, *283*(2), 381–390. <https://doi.org/10.1148/radiol.2016161315>

Doebler, P. & Holling, H. (2015). Meta-Analysis of Diagnostic Accuracy with mada.

Friedrich, S., Groß, S., König, I. R., Engelhardt, S., Bahls, M., Heinz, J., Huber, C., Kaderali, L., Kelm, M., Leha, A., Rühl, J., Schaller, J., Scherer, C., Vollmer, M., Seidler, T., & Friede, T.. (2021). Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations. *European Heart Journal - Digital Health*, *2*(3), 424–436. <https://doi.org/10.1093/ehjdh/ztab054>

Gartlehner, G., Affengruber, L., Titscher, V., Noel-Storr, A., Dooley, G., Ballarini, N., & König, F. (2020). Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *Journal of clinical epidemiology*, *121*, 20–28. <https://doi.org/10.1016/j.jclinepi.2020.01.005>

Gaziano T. A. (2007). Reducing the growing burden of cardiovascular disease in the developing world. *Health affairs (Project Hope)*, *26*(1), 13–24. <https://doi.org/10.1377/hlthaff.26.1.13>

Gaziano, T. A., & Gaziano, J. M. (2014). Epidemiology of Cardiovascular Disease. In D. Kasper, A. Fauci, S. Hauser, D. Longo, J. L. Jameson, & J. Loscalzo (Eds.), *Harrison’s Principles of Internal Medicine* (19th ed.). McGraw-Hill Education. [accessmedicine.mhmedical.com/content.aspx?aid=1120804383](https://doi.org/accessmedicine.mhmedical.com/content.aspx?aid=1120804383)

GBD 2016 Disease and Injury Incidence and Prevalence Collaborators (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet (London, England)*, *390*(10100), 1211–1259. <https://doi.org/10.1016/S0140-6736(17)32154-2>

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, 295–336. <https://doi.org/10.1016/B978-0-12-818438-7.00012-5>

Guidi, J. L., Clark, K., Upton, M. T., Faust, H., Umscheid, C. A., Lane-Fall, M. B., Mikkelsen, M. E., Schweickert, W. D., Vanzandbergen, C. A., Betesh, J., et al. (2015). Clinician perception of the effectiveness of an automated early warning and response system for sepsis in an academic medical center. Annals of the American Thoracic Society, 12.

Haddad, M., Walters, P., Phillips, R., Tsakok, J., Williams, P., Mann, A., & Tylee, A. (2013). Detecting depression in patients with coronary heart disease: a diagnostic evaluation of the PHQ-9 and HADS-D in primary care, findings from the UPBEAT-UK study. *PloS one*, *8*(10), e78493. <https://doi.org/10.1371/journal.pone.0078493>

Haleem, A., Javaid, M., Singh, R. P., & Suman, R.. (2021). Applications of Artificial Intelligence (AI) for cardiology during COVID-19 pandemic. *Sustainable Operations and Computers*, *2*, 71–78. <https://doi.org/10.1016/j.susoc.2021.04.003>

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane, 2022. Available from [www.training.cochrane.org/handbook](http://www.training.cochrane.org/handbook).

IBM Cloud Education. (2020, July 15). *What is machine learning?* IBM. Retrieved April 15, 2022, from <https://www.ibm.com/cloud/learn/machine-learning>

Jagannathan, R., Patel, S.A., Ali, M.K. *et al.* Global Updates on Cardiovascular Disease Mortality Trends and Attribution of Traditional Risk Factors. *Curr Diab Rep* **19,**44 (2019). <https://doi.org/10.1007/s11892-019-1161-2>

Jamison, D. T., Breman, J. G., Measham, A. R., Alleyne, G., Claeson, M., Evans, D. B., ... & Musgrove, P. (Eds.). (2006). Disease control priorities in developing countries.

Ji, S., Gu, Q., Weng, H., Liu, Q., Zhou, P., Chen, J., Li, Z., Beyah, R., & Wang, T. (2020). De-Health: All your online health information are belong to us. In *Proceedings - 2020 IEEE 36th International Conference on Data Engineering, ICDE 2020*(pp. 1609-1620). [9101713] (Proceedings - International Conference on Data Engineering; Vol. 2020-April). IEEE Computer Society. <https://doi.org/10.1109/ICDE48307.2020.00143>

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, *2*(4), 230–243. <https://doi.org/10.1136/svn-2017-000101>

Jousilahti, P., Vartiainen, E., Tuomilehto, J., & Puska, P. (1999). Sex, Age, Cardiovascular Risk Factors, and Coronary Heart Disease. *Circulation*, *99*(9), 1165–1172. <https://doi.org/10.1161/01.CIR.99.9.1165>

Juhola, M., Joutsijoki, H., Penttinen, K., & Aalto-Setälä, K. (2018). Detection of genetic cardiac diseases by Ca2+ transient profiles using machine learning methods. *Scientific reports*, *8*(1), 9355. <https://doi.org/10.1038/s41598-018-27695-5>

Kalinin, A. A., Higgins, G. A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I. D., Najarian, K., & Athey, B. D. (2018). Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics*, *19*(7), 629–650. <https://doi.org/10.2217/pgs-2018-0008>

Kalinin, A. A., Higgins, G. A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I. D., Najarian, K., & Athey, B. D. (2018). Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics*, *19*(7), 629–650. <https://doi.org/10.2217/pgs-2018-0008>

Khambhati, J., Allard-Ratick, M., Dhindsa, D., Lee, S., Chen, J., Sandesara, P. B., O'Neal, W., Quyyumi, A. A., Wong, N. D., Blumenthal, R. S., & Sperling, L. S. (2018). The art of cardiovascular risk assessment. *Clinical cardiology*, *41*(5), 677–684. <https://doi.org/10.1002/clc.22930>

Khera, A. V., Chaffin, M., Zekavat, S. M., Collins, R. L., Roselli, C., Natarajan, P., Lichtman, J. H., D'Onofrio, G., Mattera, J., Dreyer, R., Spertus, J. A., Taylor, K. D., Psaty, B. M., Rich, S. S., Post, W., Gupta, N., Gabriel, S., Lander, E., Ida Chen, Y. D., Talkowski, M. E., … Kathiresan, S. (2019). Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation*, *139*(13), 1593–1602. <https://doi.org/10.1161/CIRCULATIONAHA.118.035658>

Kiely, K.M. (2014) Cognitive Function. In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. <https://doi.org/10.1007/978-94-007-0753-5_426>

Kisling, L. A., & M Das, J. (2022). Prevention Strategies. In *StatPearls*. StatPearls Publishing. <http://www.ncbi.nlm.nih.gov/books/NBK537222/>

Krittanawong, C., Virk, H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific reports*, *10*(1), 16057. <https://doi.org/10.1038/s41598-020-72685-1>

Kuehn B. M. (2013). Costs of cardiac care likely to increase, despite advances in prevention, care. *JAMA*, *310*(19), 2029. https://doi.org/10.1001/jama.2013.282805

Larsen, B. A., & Christenfeld, N. J. (2009). Cardiovascular disease and psychiatric comorbidity: the potential role of perseverative cognition. *Cardiovascular psychiatry and neurology*, *2009*, 791017. <https://doi.org/10.1155/2009/791017>

Lee, C. H., & Yoon, H. J. (2017). Medical big data: promise and challenges. *Kidney research and clinical practice*, *36*(1), 3–11. <https://doi.org/10.23876/j.krcp.2017.36.1.3>

Lee, J., Kim, K. W., Choi, S. H., Huh, J., & Park, S. H. (2015). Systematic Review and Meta-Analysis of Studies Evaluating Diagnostic Test Accuracy: A Practical Review for Clinical Researchers-Part II. Statistical Methods of Meta-Analysis. *Korean journal of radiology*, *16*(6), 1188–1196. <https://doi.org/10.3348/kjr.2015.16.6.1188>

Leeder, S., Raymond, S., Greenberg, H., Liu, H., & Esson, K. (2004). A race against time: the challenge of cardiovascular disease in developing economies. *New York: Columbia University*.

Leopold, J. A., & Loscalzo, J. (2018). Emerging Role of Precision Medicine in Cardiovascular Disease. *Circulation research*, *122*(9), 1302–1315. <https://doi.org/10.1161/CIRCRESAHA.117.310782>

Leopold, J. A., Maron, B. A., & Loscalzo, J. (2020). The application of big data to cardiovascular disease: paths to precision medicine. *The Journal of clinical investigation*, *130*(1), 29–38. <https://doi.org/10.1172/JCI129203>

Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), 311ra174. <https://doi.org/10.1126/scitranslmed.aaa9364>

Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), 311ra174. <https://doi.org/10.1126/scitranslmed.aaa9364>

Lim, L. J., Tison, G. H., & Delling, F. N. (2020). Artificial Intelligence in Cardiovascular Imaging. *Methodist DeBakey cardiovascular journal*, *16*(2), 138–145. <https://doi.org/10.14797/mdcj-16-2-138>

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature medicine*, *26*(9), 1364–1374. <https://doi.org/10.1038/s41591-020-1034-x>

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet. Digital health*, *1*(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2

London A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *The Hastings Center report*, *49*(1), 15–21. <https://doi.org/10.1002/hast.973>

M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

Madani, A., Arnaout, R., Mofrad, M., & Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine*, *1*, 6. <https://doi.org/10.1038/s41746-017-0013-1>

Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L.. (2020). Artificial Intelligence, Machine Learning, and Cardiovascular Disease. *Clinical Medicine Insights: Cardiology*, *14*, 117954682092740. <https://doi.org/10.1177/1179546820927404>

McCarthy, J., & Wright, P. (2004). Technology as experience. *Interactions*, *11*(5), 42–43. <https://doi.org/10.1145/1015530.1015549>

Mendoza, W., & Miranda, J. J. (2017). Global Shifts in Cardiovascular Disease, the Epidemiologic Transition, and Other Contributing Factors: Toward a New Practice of Global Health Cardiology. *Cardiology clinics*, *35*(1), 1–12. <https://doi.org/10.1016/j.ccl.2016.08.004>

Mitchell, T. M. (1997). *Machine Learning*.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L.. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. <https://doi.org/10.1177/2053951716679679>

Motwani, M., Dey, D., Berman, D. S., Germano, G., Achenbach, S., Al-Mallah, M. H., Andreini, D., Budoff, M. J., Cademartiri, F., Callister, T. Q., Chang, H. J., Chinnaiyan, K., Chow, B. J., Cury, R. C., Delago, A., Gomez, M., Gransar, H., Hadamitzky, M., Hausleiter, J., Hindoyan, N., … Slomka, P. J. (2017). Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *European heart journal*, *38*(7), 500–507. <https://doi.org/10.1093/eurheartj/ehw188>

Mourby, M., Ó Cathaoir, K., & Collin, C. B. (2021). Transparency of machine-learning in healthcare: The GDPR & European health law. *Computer Law & Security Review*, *43*, 105611. <https://doi.org/10.1016/j.clsr.2021.105611>

Murdoch B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC medical ethics*, *22*(1), 122. <https://doi.org/10.1186/s12910-021-00687-3>

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. <https://doi.org/10.1073/pnas.1900654116>

Murray, C. J., & Lopez, A. D. (1996). Evidence-based health policy--lessons from the Global Burden of Disease Study. *Science (New York, N.Y.)*, *274*(5288), 740–743. <https://doi.org/10.1126/science.274.5288.740>

Nagendran M, Chen Y, Lovejoy C A, Gordon A C, Komorowski M, Harvey H et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies *BMJ*2020; 368 :m689 doi:10.1136/bmj.m689

*New World Bank country classifications by income level: 2021-2022*. (n.d.). Retrieved April 16, 2022, from <https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2021-2022>

Omran A. R. (1971). The epidemiologic transition. A theory of the epidemiology of population change. *The Milbank Memorial Fund quarterly*, *49*(4), 509–538.

Ong, M. E., Lee Ng, C. H., Goh, K., Liu, N., Koh, Z. X., Shahidah, N., Zhang, T. T., Fook-Chong, S., & Lin, Z. (2012). Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. *Critical care (London, England)*, *16*(3), R108. <https://doi.org/10.1186/cc11396>

Pastorino, R., De Vito, C., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W., & Boccia, S. (2019). Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. *European journal of public health*, *29*(Supplement\_3), 23–27. <https://doi.org/10.1093/eurpub/ckz168>

Piette, J. D., Striplin, D., Marinec, N., Chen, J., Trivedi, R. B., Aron, D. C., Fisher, L., & Aikens, J. E. (2015). A Mobile Health Intervention Supporting Heart Failure Patients and Their Informal Caregivers: A Randomized Comparative Effectiveness Trial. *Journal of medical Internet research*, *17*(6), e142. <https://doi.org/10.2196/jmir.4550>

Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. *Health and technology*, *7*(4), 351–367. <https://doi.org/10.1007/s12553-017-0179-1>

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reddy, K. S., & Yusuf, S. (1998). Emerging epidemic of cardiovascular disease in developing countries. *Circulation*, *97*(6), 596–601. <https://doi.org/10.1161/01.cir.97.6.596>

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., Criqui, M., … GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group (2020). Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology*, *76*(25), 2982–3021. <https://doi.org/10.1016/j.jacc.2020.11.010>

Rowe M. (2019). An Introduction to Machine Learning for Clinicians. *Academic medicine : journal of the Association of American Medical Colleges*, *94*(10), 1433–1436. https://doi.org/10.1097/ACM.0000000000002792

Rutjes, A. W., Reitsma, J. B., Coomarasamy, A., Khan, K. S., & Bossuyt, P. M. (2007). Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health technology assessment (Winchester, England)*, *11*(50), iii–51. <https://doi.org/10.3310/hta11500>

Saarela, M., Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. *SN Appl. Sci.* **3,**272 (2021). <https://doi.org/10.1007/s42452-021-04148-9>

Santosa, A., Wall, S., Fottrell, E., Högberg, U., & Byass, P. (2014). The development and experience of epidemiological transition theory over four decades: a systematic review. *Global health action*, *7*, 23574. <https://doi.org/10.3402/gha.v7.23574>

Schultz, W. M., Kelli, H. M., Lisko, J. C., Varghese, T., Shen, J., Sandesara, P., Quyyumi, A. A., Taylor, H. A., Gulati, M., Harold, J. G., Mieres, J. H., Ferdinand, K. C., Mensah, G. A., & Sperling, L. S. (2018). Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. *Circulation*, *137*(20), 2166–2178. <https://doi.org/10.1161/CIRCULATIONAHA.117.029652>

Sevakula, R. K., Au-Yeung, W. M., Singh, J. P., Heist, E. K., Isselbacher, E. M., & Armoundas, A. A. (2020). State-of-the-Art Machine Learning Techniques Aiming to Improve Patient Outcomes Pertaining to the Cardiovascular System. *Journal of the American Heart Association*, *9*(4), e013924. <https://doi.org/10.1161/JAHA.119.013924>

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: are we there yet?. *Heart (British Cardiac Society)*, *104*(14), 1156–1164. <https://doi.org/10.1136/heartjnl-2017-311198>

Shim, S. R., Kim, S. J., & Lee, J. (2019). Diagnostic test accuracy: application and practice using R software. *Epidemiology and health*, *41*, e2019007. <https://doi.org/10.4178/epih.e2019007>

Sidey-Gibbons, J., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC medical research methodology*, *19*(1), 64. <https://doi.org/10.1186/s12874-019-0681-4>

Silverio, A., Cavallo, P., De Rosa, R., & Galasso, G. (2019). Big Health Data and Cardiovascular Diseases: A Challenge for Research, an Opportunity for Clinical Care. *Frontiers in medicine*, *6*, 36. <https://doi.org/10.3389/fmed.2019.00036>

Song X, Quek RG, Gandra SR, et al. Productivity loss and indirect costs associated with cardiovascular events and related clinical procedures. BMC Health Serv Res. 2015;15:245

Sounderajah, V., Ashrafian, H., Rose, S., Shah, N. H., Ghassemi, M., Golub, R., Kahn, C. E., Jr, Esteva, A., Karthikesalingam, A., Mateen, B., Webster, D., Milea, D., Ting, D., Treanor, D., Cushnan, D., King, D., McPherson, D., Glocker, B., Greaves, F., Harling, L., … Darzi, A. (2021). A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nature medicine*, *27*(10), 1663–1665. <https://doi.org/10.1038/s41591-021-01517-0>

Stewart, J., Manmathan, G., & Wilkinson, P. (2017). Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. *JRSM cardiovascular disease*, *6*, 2048004016687211. <https://doi.org/10.1177/2048004016687211>

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, *3*, 17. <https://doi.org/10.1038/s41746-020-0221-y>

Tabassian, M., Sunderji, I., Erdei, T., Sanchez-Martinez, S., Degiovanni, A., Marino, P., Fraser, A. G., & D'hooge, J. (2018). Diagnosis of Heart Failure With Preserved Ejection Fraction: Machine Learning of Spatiotemporal Variations in Left Ventricular Deformation. *Journal of the American Society of Echocardiography : official publication of the American Society of Echocardiography*, *31*(12), 1272–1284.e9. <https://doi.org/10.1016/j.echo.2018.07.013>

Thomas, H., Diamond, J., Vieco, A., Chaudhuri, S., Shinnar, E., Cromer, S., Perel, P., Mensah, G. A., Narula, J., Johnson, C. O., Roth, G. A., & Moran, A. E. (2018). Global Atlas of Cardiovascular Disease 2000-2016: The Path to Prevention and Control. *Global heart*, *13*(3), 143–163. <https://doi.org/10.1016/j.gheart.2018.09.511>

Tonekaboni, S., Joshi, S., McCradden, M.D. &amp; Goldenberg, A.. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. <i>Proceedings of the 4th Machine Learning for Healthcare Conference</i>, in <i>Proceedings of Machine Learning Research</i> 106:359-380 Available from <https://proceedings.mlr.press/v106/tonekaboni19a.html>.

Trikalinos, T. A., Balion, C. M., Coleman, C. I., Griffith, L., Santaguida, P. L., Vandermeer, B., & Fu, R. (2012). Chapter 8: meta-analysis of test performance when there is a "gold standard". *Journal of general internal medicine*, *27 Suppl 1*(Suppl 1), S56–S66. https://doi.org/10.1007/s11606-012-2029-1

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>

Uddin, S., Khan, A., Hossain, M. *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* **19,**281 (2019). https://doi.org/10.1186/s12911-019-1004-8

Umscheid, C. A., Betesh, J., VanZandbergen, C., Hanish, A., Tait, G., Mikkelsen, M. E., French, B., and Fuchs, B. D. (2015). Development, implementation, and impact of an automated early warning and response system for sepsis. Journal of hospital medicine

Valen, J., Balki, I., Mendez, M., Qu, W., Levman, J., Bilbily, A., & Tyrrell, P. N. (2022). Quantifying uncertainty in machine learning classifiers for medical imaging. *International Journal of Computer Assisted Radiology and Surgery*, *17*(4), 711–718. <https://doi.org/10.1007/s11548-022-02578-3>

Vartiainen E. (2018). The North Karelia Project: Cardiovascular disease prevention in Finland. *Global cardiology science & practice*, *2018*(2), 13. <https://doi.org/10.21542/gcsp.2018.13>

Vincent, J. Privacy advocates sound the alarm after Google grabs DeepMind UK health app. The Verge. 2018. <https://www.theverge.com/2018/11/14/18094874/google-deepmind-health-app-privacy-concerns-uk-nhs-medical-data>. Accessed 15 Apr 2022.

Wahl B, Cossy-Gantner A, Germann S, et alArtificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Global Health 2018;3:e000798.

Wein B, Bashkireva A, Au-Yeung A, et al. Systematic investment in the delivery of guideline-coherent therapy reduces mortality and overall costs in patients with ST-elevation myocardial infarction: results from the Stent for Life economic model for Romania, Portugal, Basque Country and Kemerovo region [published online ahead of print, 2019 Sep 26]. Eur Heart J Acute Cardiovasc Care. 2019.

Weintraub W. S. (2019). Role of Big Data in Cardiovascular Research. *Journal of the American Heart Association*, *8*(14), e012791. <https://doi.org/10.1161/JAHA.119.012791>

World Health Organization. (2019). Countries are spending more on health, but people are still paying too much out of their own pockets. (n.d.). Retrieved April 16, 2022, from <https://www.who.int/news/item/20-02-2019-countries-are-spending-more-on-health-but-people-are-still-paying-too-much-out-of-their-own-pockets>

Yan, Y., Zhang, J. W., Zang, G. Y., & Pu, J. (2019). The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine?. *Journal of geriatric cardiology : JGC*, *16*(8), 585–591. <https://doi.org/10.11909/j.issn.1671-5411.2019.08.010>

Yusuf, S., Reddy, S., Ounpuu, S., & Anand, S. (2001). Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation*, *104*(22), 2746–2753. <https://doi.org/10.1161/hc4601.099487>

# **Appendix**

## **Section A-1: PRISMA Checklist**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Item #** | **Checklist Item** | **Page #** |
| **TITLE** | | | |
| Title | 1 | Identify the report as a systematic review. | **-** |
| **ABSTRACT** | | | |
| Abstract | 2 | See the PRISMA 2020 for Abstracts checklist. | **-** |
| **INTRODUCTION** | | | |
| Rationale | 3 | Describe the rationale for the review in the context of existing knowledge | 3 |
| Objectives | 4 | Provide an explicit statement of the objective(s) or question(s) the review addresses. | 3 |
| **METHODS** | | | |
| Eligibility Criteria | 5 | Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses. | 21 |
| Information Sources | 6 | Specify all databases, registers, websites, organizations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted. | 21 |
| Search Strategy | 7 | Present the full search strategies for all databases, registers and websites, including any filters and limits used. | 21 |
| Selection Process | 8 | Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process. | 21 |
| Data Collection Process | 9 | Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process. | 22 |
| Data Items | 10a | List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g., for all measures, time points, analyses), and if not, the methods used to decide which results to collect. | 22 |
| 10b | List and define all other variables for which data were sought (e.g., participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information. | 22 |
| Study Risk of Bias Assessment | 11 | Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process. | 21 |
| Effect Measures | 12 | Specify for each outcome the effect measure(s) (e.g., risk ratio, mean difference) used in the synthesis or presentation of results. | 22 |
| Synthesis Methods | 13a | Describe the processes used to decide which studies were eligible for each synthesis (e.g., tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)). | 22 |
| 13b | Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions. | 22 |
| 13c | Describe any methods used to tabulate or visually display results of individual studies and syntheses. | 21 |
| 13d | Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used. | 23 |
| 13e | Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression). | - |
| 13f | Describe any sensitivity analyses conducted to assess robustness of the synthesized results. | - |
| Reporting Bias Assessment | 14 | Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases). | - |
| Certainty Assessment | 15 | Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome. | - |
| **RESULTS** | | | |
| Study Selection | 16a | Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram. | 24 |
| 16b | Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded. | 24 |
| Study Characteristics | 17 | Cite each included study and present its characteristics. | 25 |
| Risk of Bias in Study | 18 | Present assessments of risk of bias for each included study. | - |
| Results of Individual Studies | 19 | For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimates and its precision (e.g., confidence/credible interval), ideally using structured tables or plots. | - |
| Results of Syntheses | 20a | For each synthesis, briefly summarize the characteristics and risk of bias among contributing studies. | 25 |
| 20b | Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g., confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect. | - |
| 20c | Present results of all investigations of possible causes of heterogeneity among study results. | - |
| 20d | Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results. | - |
| Reporting Bias | 21 | Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed. | - |
| Certainty of Evidence | 22 | Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed. | - |
| **DISCUSSION** | | | |
| Discussion | 23a | Provide a general interpretation of the results in the context of other evidence. | 32 |
| 23b | Discuss any limitations of the evidence included in the review. | 34 |
| 23c | Discuss any limitations of the review processes used. | 34 |
| 23d | Discuss implications of the results for practice, policy, and future research. | 35 |
| **OTHER INFORMATION** | | | |
| Registration and Protocol | 24a | Provide registration information for the review, including register name and registration number, or state that the review was not registered. | - |
| 24b | Indicate where the review protocol can be accessed, or state that a protocol was not prepared. | - |
| 24c | Describe and explain any amendments to information provided at registration or in the protocol. | - |
| Support | 25 | Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review. | - |
| Competing Interest | 26 | Declare any competing interests of review authors. | - |
| Availability of Data, Code and Other Materials | 27 | Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review. | - |

***Source:*** Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

## **Section A-2: Search Strategy**

COCHRANE

(( "artificial intelligence" OR "machine learning" )):ti,ab,kw AND (( "linear regression" OR "logistic regression" OR "support vector machine" OR "decision tree" OR "random forest" OR "naive bayes" OR "k nearest neighbor" OR "k nearest neighbour" OR "neural network" )):ti,ab,kw AND (( ( classification OR probability OR regression ) NEAR/3 tree\* ) OR ( ( artificial OR bayesian OR convolutional OR neural ) NEAR/3 network\* ) OR ( fuzzy NEAR/3 ( logit OR logic OR logistic ) )):ti,ab,kw AND (( diagnosis OR imaging OR prediction OR risk) OR ( disease NEAR/2 ( prediction OR risk ) ) OR ( aided NEAR/2 ( diagnos\* OR decision\* ) ) OR ( clinician OR healthcare OR practitioner) OR ( healthcare NEAR/2 ( clinician OR practitioner ) )):ti,ab,kw AND (( apoplexy OR arteriosclerosis OR “blood pressure” OR "cardiovascular disease" OR "cerebrovascular disease" OR hypercholesterol OR hyperlipidemia OR hypertension OR “peripheral artery disease” OR angina\* OR arrythmi\* OR arterio\* OR brain\* OR cardio\* OR cardia\* OR cerebrovasc\* OR coronary\* OR endocardi\* OR infarct\* OR ischem\* OR lacunar\* OR myocard\* OR pericard\* OR stroke\* )):ti,ab,kw

EMBASE

('artificial intelligence':ti,ab,kw OR 'machine learning':ti,ab,kw) AND ('linear regression':ti,ab,kw OR 'logistic regression':ti,ab,kw OR 'support vector machine':ti,ab,kw OR 'decision tree':ti,ab,kw OR 'random forest':ti,ab,kw OR 'naive bayes':ti,ab,kw OR 'k nearest neighbor':ti,ab,kw OR 'k nearest neighbour':ti,ab,kw OR 'neural network':ti,ab,kw) AND ((((classification OR probability OR regression) NEAR/3 tree\*):ti,ab,kw) OR (((artificial OR bayesian OR convolutional OR neural) NEAR/3 network\*):ti,ab,kw) OR ((fuzzy NEAR/3 (logit OR logic OR logistic)):ti,ab,kw)) AND (diagnosis:ti,ab,kw OR imaging:ti,ab,kw OR prediction:ti,ab,kw OR risk:ti,ab,kw OR ((disease NEAR/2 (prediction OR risk)):ti,ab,kw) OR ((aided NEAR/2 (diagnos\* OR decision\*)):ti,ab,kw)) AND (apoplexy:ti,ab,kw OR arteriosclerosis:ti,ab,kw OR 'blood pressure':ti,ab,kw OR 'cardiovascular disease':ti,ab,kw OR 'cerebrovascular disease':ti,ab,kw OR hypercholesterol:ti,ab,kw OR hyperlipidemia:ti,ab,kw OR hypertension:ti,ab,kw OR 'peripheral artery disease':ti,ab,kw OR angina\*:ti,ab,kw OR arrythmi\*:ti,ab,kw OR arterio\*:ti,ab,kw OR brain\*:ti,ab,kw OR cardio\*:ti,ab,kw OR cardia\*:ti,ab,kw OR cerebrovasc\*:ti,ab,kw OR coronary\*:ti,ab,kw OR endocardi\*:ti,ab,kw OR infarct\*:ti,ab,kw OR ischem\*:ti,ab,kw OR lacunar\*:ti,ab,kw OR myocard\*:ti,ab,kw OR pericard\*:ti,ab,kw OR stroke\*:ti,ab,kw)

SCOPUS

( ( TITLE-ABS-KEY ( "artificial intelligence" OR "machine learning" ) ) AND ( TITLE-ABS-KEY ( "linear regression" OR "logistic regression" OR "support vector machine" OR "decision tree" OR "random forest" OR "naive bayes" OR "k nearest neighbor" OR "k nearest neighbour" OR "neural network" ) ) AND ( TITLE-ABS-KEY ( ( classification OR probability OR regression ) W/3 tree\* ) OR ( ( artificial OR bayesian OR convolutional OR neural ) W/3 network\* ) OR ( fuzzy W/3 ( logit OR logic OR logistic ) ) ) AND ( TITLE-ABS-KEY ( diagnosis OR imaging OR prediction OR risk) OR ( disease W/2 ( prediction OR risk ) ) OR ( aided W/2 ( diagnos\* OR decision\* ) ) ) AND ( TITLE-ABS-KEY ( clinician OR healthcare OR practitioner) OR ( healthcare W/2 ( clinician OR practitioner ) ) ) AND ( TITLE-ABS-KEY ( apoplexy OR arteriosclerosis OR “blood pressure” OR "cardiovascular disease" OR "cerebrovascular disease" OR hypercholesterol OR hyperlipidemia OR hypertension OR “peripheral artery disease” OR angina\* OR arrythmi\* OR arterio\* OR brain\* OR cardio\* OR cardia\* OR cerebrovasc\* OR coronary\* OR endocardi\* OR infarct\* OR ischem\* OR lacunar\* OR myocard\* OR pericard\* OR stroke\* ) ) )

WEB OF SCIENCE

#6 AND #5 AND #4 AND #3 AND #2 AND #1  
*Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC  
Timespan=2012-01-01 to 2022-03-01*

1. TS = (( apoplexy OR arteriosclerosis OR “blood pressure” OR "cardiovascular disease" OR "cerebrovascular disease" OR hypercholesterol OR hyperlipidemia OR hypertension OR “peripheral artery disease” OR angina\* OR arrythmi\* OR arterio\* OR brain\* OR cardio\* OR cardia\* OR cerebrovasc\* OR coronary\* OR endocardi\* OR infarct\* OR ischem\* OR lacunar\* OR myocard\* OR pericard\* OR stroke\* ) )

*Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC  
Timespan=2012-01-01 to 2022-03-01*

1. TS = (( clinician OR healthcare OR practitioner) OR ( healthcare NEAR/2 ( clinician OR practitioner ) ) )

*Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC  
Timespan=2012-01-01 to 2022-03-01*

1. TS = (( diagnosis OR imaging OR prediction OR risk) OR ( disease W/2 ( prediction OR risk ) ) OR ( aided W/2 ( diagnos\* OR decision\* ) ) )

*Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC  
Timespan=2012-01-01 to 2022-03-01*

1. TS = ( ( ( classification OR probability OR regression ) NEAR/3 tree\* ) OR ( ( artificial OR bayesian OR convolutional OR neural ) NEAR/3 network\* ) OR ( fuzzy NEAR/3 ( logit OR logic OR logistic ) ) )

*Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC  
Timespan=2012-01-01 to 2022-03-01*

1. TS = (( "linear regression" OR "logistic regression" OR "support vector machine" OR "decision tree" OR "random forest" OR "naive bayes" OR "k nearest neighbor" OR "k nearest neighbour" OR "neural network" ) )  
   *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC  
   Timespan=2012-01-01 to 2022-03-01*
2. TS = ( "artificial intelligence" OR "machine learning")

*Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC  
Timespan=2012-01-01 to 2022-03-01*

## **Section A-3: Study Characteristics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Author | Algorithm | Sample | Indication | Imaging | Comparison | Database |
| Cardiac Arrythmias | | | | | | |
| Poh et al. (2018) | CNN | 3039 | Arterial Fibrillation | Electrocardiographs | - | PPG-RHYTHM |
| Coronary Artery Disease | | | | | | |
| Araki et al. (2016) | SVM | 19 | Coronary Risk Assessment | Intravascular Ultrasound | - | Single Center |
| Araki et al. (2016) | SVM | 15 | Plaque Risk Assessment | Intravascular Ultrasound | - | Single Center |
| Baumann et al. (2019) | Custom | 398 | Coronary Lesion Detection | Computer Topography Angiography | Invasive Fractional Flow Reserve | MACHINE Registry |
| Coenen et al. (2018) | Custom | 351 | Computational Flow Dynamics Based Invasive Fractional Flow Reserve | Computer Topography Angiography | Computational Flow Dynamics Based Invasive Fractional Flow Reserve | MACHINE Registry |
| Dutta et al. (2020) | Custom | 31779 | Coronary Artery Disease Prediction | Clinical Data | - | NHANES |
| Eisenberg et al. (2021) | BA | 2709 | Myocardial Perfusion Imaging | Single Photon Emission Computerized Tomography | Clinician Image Analysis | REFINE SPECT Registry |
| Freiman et al. (2017) | Custom | 132 | Coronary Artery Stenosis Detection | Computer Topography Angiography | Cardiac Image Analysis | MICCAI 2012 Registry |
| Han et al. (2018) | BA | 252 | Coronary Artery Disease Prediction | Computer Topography Angiography / Fractional Flow Reserve | - | DeFACTO Study |
| Joloudari et al. (2022) | ANN | 30 | Coronary Artery Disease Prediction | Magnetic Resonance Imaging | - | Single Center |
| Miao et al. (2018) | ANN | 147 | Coronary Artery Disease Prediction | Clinical Data | - | Single Center |
| Muscogiuri et al. (2019) | CNN | 284 | Coronary Artery Disease Prediction | Computer Topography Angiography | Algorithm Comparison | Single Center |
| Song et al. (2014) | SVM | 66 | Acute Coronary Syndrome Risk Assessment | - | - | Single Center |
| Xiuhua et al. (2018) | Custom | 105 | Coronary Lesion Detection | Computer Topography Angiography / Fractional Flow Reserve | Computer Topography Angiography / Fractional Flow Reserve | Single Center |
| Zellweger et al. (2018) | Custom | 262 | Coronary Artery Disease Prediction | - | Framingham Scores | LURIC |
| Heart Failure | | | | | | |
| Attia et al. (2020) | ANN | 52870 | Asymptomatic Left Ventricular Dysfunction | Electrocardiographs | - | Single Center |
| Blecker et al. (2018) | LGR | 37229 | Acute Decompensated Heart Failure | Clincial Data | Algorithm Comparison | Single Center |
| Rossing et al. (2016) | SVM | 646 | Heart Failure with Preserved Ejection Fraction | Urine Proteome | - | Single Center |
| Weng et al. (2017) | ANN + BA + LGR + RF | 82989 | Heart Failure Event | Electronic Health Records | American College of Cardiology Guidelines | UK CRPD |
| Stroke | | | | | | |
| Arslan et al. (2016) | BA + LGR + SVM | 190 | Ischemic Stroke | Electronic Medical Records | - | Single Center |
| Bacchi et al. (2020) | ANN + CNN | 204 | Intravenous Thrombolysis | Computer Topography Head | THRIVE/HIAT/SPAN-100 Scores | Single Center |
| Beecy et al. (2017) | CNN | 22 | Stroke | Computer Topography Head | Clinician Consensus | Single Center |
| Dharmasaroja et al. (2013) | ANN + CNN + SVM | 194 | Intracranial Hemorrhage | Computer Topography Head | Post Stroke Thrombolysis | Single Center |
| Ho et al. (2018) | BA + LNR + RF + SVM | 105 | Acute Ischemic Stroke | Magnetic Resonance Imaging | Algorithm Comparison | Single Center |
| Matsuo et al. (2021) | ANN + BA + LGR + RF + SVM | 170 | Carotid Artery Stenting | Clinical Data | Clinician Comparison | Single Center |

***Abbreviations:***ANN, artificial neural network; BA, boosting algorithm; CNN, convolutional neural network; LGR, logistic regression; LNR, linear regression; RF, random forest; SVM, support vector machine.

**Section A-4: Meta Analysis Contingency Tables**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Author | Model | Sample | TP | FP | FN | TN | Sensitivity | Specificity | AUC |
| Araki et al. (2016) | SVM | 19 | 10 | 1 | 0 | 8 | 0.979 | 0.989 | 0.980 |
| Araki et al. (2016) | SVM | 15 | 7 | 0 | 1 | 7 | 0.929 | 0.966 | 0.950 |
| Arslan et al. (2016) | BA | 190 | 78 | 1 | 4 | 107 | 0.951 | 0.991 | 0.976 |
| Arslan et al. (2016) | LGR | 190 | 71 | 8 | 12 | 99 | 0.855 | 0.925 | 0.895 |
| Arslan et al. (2016) | SVM | 190 | 77 | 2 | 2 | 109 | 0.975 | 0.982 | 0.978 |
| Attia et al. (2020) | ANN | 52870 | 3406 | 6847 | 554 | 42063 | 0.860 | 0.860 | 0.932 |
| Bacchi et al. (2020) | CNN | 204 | 60 | 7 | 48 | 89 | 0.560 | 0.930 | 0.750 |
| Bacchi et al. (2020) | ANN | 204 | 85 | 52 | 20 | 47 | 0.810 | 0.470 | 0.610 |
| Bacchi et al. (2020) | CNN | 204 | 93 | 59 | 13 | 39 | 0.880 | 0.400 | 0.540 |
| Bacchi et al. (2020) | Reference | 204 | 85 | 52 | 20 | 47 | 0.810 | 0.470 | 0.690 |
| Bacchi et al. (2020) | Reference | 204 | 58 | 33 | 46 | 67 | 0.560 | 0.670 | 0.630 |
| Bacchi et al. (2020) | Reference | 204 | 107 | 65 | 0 | 32 | 1.000 | 0.330 | - |
| Bacchi et al. (2020) | CNN | 204 | 86 | 53 | 6 | 59 | 0.930 | 0.530 | 0.700 |
| Bacchi et al. (2020) | ANN | 204 | 40 | 13 | 53 | 98 | 0.430 | 0.880 | 0.680 |
| Bacchi et al. (2020) | CNN | 204 | 66 | 39 | 27 | 72 | 0.710 | 0.650 | 0.630 |
| Bacchi et al. (2020) | Reference | 204 | 40 | 27 | 53 | 84 | 0.430 | 0.760 | 0.630 |
| Bacchi et al. (2020) | Reference | 204 | 66 | 45 | 27 | 66 | 0.710 | 0.590 | 0.670 |
| Bacchi et al. (2020) | Reference | 204 | 20 | 13 | 74 | 97 | 0.210 | 0.880 | - |
| Baumann et al. (2019) | Reference | 398 | 157 | 84 | 19 | 138 | 0.890 | 0.620 | 0.740 |
| Baumann et al. (2019) | Reference | 127 | 30 | 33 | 6 | 58 | 0.830 | 0.640 | 0.760 |
| Baumann et al. (2019) | Custom | 398 | 142 | 48 | 40 | 168 | 0.780 | 0.780 | 0.830 |
| Baumann et al. (2019) | Custom | 127 | 27 | 17 | 9 | 74 | 0.750 | 0.810 | 0.803 |
| Beecy et al. (2017) | CNN | 22 | 7 | 1 | 1 | 13 | 0.930 | 0.920 | 0.973 |
| Beecy et al. (2017) | CNN | 22 | 2 | 2 | 1 | 17 | 0.650 | 0.910 | 0.910 |
| Blecker et al. (2018) | LGR | 37229 | 1324 | 7499 | 27 | 28379 | 0.980 | 0.791 | 0.960 |
| Blecker et al. (2018) | Custom | 37229 | 1306 | 3051 | 27 | 32845 | 0.980 | 0.915 | 0.990 |
| Blecker et al. (2018) | Custom | 37229 | 1295 | 2514 | 26 | 33394 | 0.980 | 0.930 | 0.990 |
| Coenen et al. (2018) | Reference | 351 | 125 | 130 | 17 | 79 | 0.880 | 0.380 | 0.690 |
| Coenen et al. (2018) | Reference | 351 | 61 | 38 | 81 | 171 | 0.430 | 0.820 | 0.690 |
| Coenen et al. (2018) | Reference | 351 | 116 | 50 | 26 | 159 | 0.820 | 0.760 | 0.840 |
| Coenen et al. (2018) | Custom | 351 | 116 | 50 | 27 | 158 | 0.810 | 0.760 | 0.840 |
| Dharmasaroja et al. (2013) | ANN | 194 | 5 | 18 | 26 | 145 | 0.177 | 0.888 | 0.687 |
| Dharmasaroja et al. (2013) | ANN | 194 | 4 | 2 | 27 | 161 | 0.118 | 0.989 | 0.638 |
| Dharmasaroja et al. (2013) | CNN | 194 | 7 | 0 | 13 | 174 | 0.353 | 1.000 | 0.788 |
| Dharmasaroja et al. (2013) | SVM | 194 | 0 | 0 | 19 | 175 | 0.000 | 1.000 | 0.416 |
| Dutta et al. (2020) | Custom | 31779 | 161 | 5743 | 47 | 25828 | 0.773 | 0.818 | 0.768 |
| Eisenberg et al. (2021) | Custom | 2709 | 1330 | 640 | 148 | 591 | 0.900 | 0.480 | 0.840 |
| Freiman et al. (2017) | Custom | 132 | 40 | 27 | 8 | 57 | 0.830 | 0.680 | 0.800 |
| Freiman et al. (2017) | Custom | 132 | 39 | 34 | 8 | 51 | 0.830 | 0.600 | 0.760 |
| Han et al. (2018) | Reference | 252 | 91 | 48 | 38 | 75 | 0.705 | 0.610 | 0.680 |
| Han et al. (2018) | BA | 252 | 66 | 18 | 61 | 107 | 0.519 | 0.854 | 0.750 |
| Ho et al. (2018) | BA | 105 | 47 | 7 | 37 | 14 | 0.560 | 0.667 | 0.632 |
| Ho et al. (2018) | LNR | 105 | 57 | 9 | 26 | 13 | 0.687 | 0.591 | 0.683 |
| Ho et al. (2018) | RF | 105 | 63 | 7 | 21 | 14 | 0.750 | 0.667 | 0.651 |
| Ho et al. (2018) | SVM | 105 | 73 | 14 | 10 | 8 | 0.880 | 0.364 | 0.640 |
| Joloudari et al. (2022) | ANN | 30 | 27 | 0 | 1 | 2 | 0.964 | 0.878 | 0.933 |
| Joloudari et al. (2022) | ANN | 30 | 30 | 0 | 0 | 0 | 0.997 | 0.997 | 0.999 |
| Joloudari et al. (2022) | ANN | 30 | 30 | 0 | 0 | 0 | 0.998 | 0.999 | 1.000 |
| Matsuo et al. (2021) | BA | 170 | 4 | 4 | 26 | 136 | 0.146 | 0.974 | 0.719 |
| Matsuo et al. (2021) | ANN | 170 | 15 | 6 | 40 | 109 | 0.265 | 0.949 | 0.702 |
| Matsuo et al. (2021) | RF | 170 | 10 | 6 | 24 | 130 | 0.290 | 0.955 | 0.692 |
| Matsuo et al. (2021) | SVM | 170 | 22 | 6 | 80 | 62 | 0.220 | 0.912 | 0.683 |
| Matsuo et al. (2021) | LGR | 170 | 3 | 3 | 5 | 159 | 0.362 | 0.983 | 0.680 |
| Miao et al. (2018) | ANN | 147 | 51 | 5 | 19 | 72 | 0.935 | 0.853 | 0.892 |
| Muscogiuri et al. (2019) | CNN | 284 | 35 | 21 | 18 | 210 | 0.660 | 0.910 | 0.890 |
| Muscogiuri et al. (2019) | CNN | 284 | 125 | 55 | 27 | 77 | 0.820 | 0.580 | 0.780 |
| Poh et al. (2018) | CNN | 3039 | 79 | 29 | 4 | 2927 | 0.952 | 0.990 | 0.997 |
| Poh et al. (2018) | CNN | 3039 | 83 | 12 | 0 | 2944 | 1.000 | 0.996 | 0.997 |
| Rossing et al. (2016) | SVM | 646 | 88 | 39 | 6 | 513 | 0.936 | 0.929 | 0.972 |
| Song et al. (2014) | Reference | 66 | 4 | 13 | 0 | 49 | 0.917 | 0.791 | 0.890 |
| Song et al. (2014) | SVM | 66 | 4 | 13 | 0 | 49 | 0.917 | 0.786 | 0.888 |
| Song et al. (2014) | SVM | 66 | 3 | 13 | 1 | 49 | 0.833 | 0.791 | 0.858 |
| Weng et al. (2017) | LGR | 82989 | 4967 | 22155 | 2437 | 53430 | 0.671 | 0.707 | 0.760 |
| Weng et al. (2017) | RF | 82989 | 4834 | 22288 | 2570 | 53297 | 0.653 | 0.705 | 0.745 |
| Weng et al. (2017) | ANN | 82989 | 4998 | 22124 | 2406 | 53461 | 0.675 | 0.707 | 0.764 |
| Weng et al. (2017) | BA | 81989 | 4997 | 22127 | 2407 | 52458 | 0.675 | 0.707 | 0.761 |
| Weng et al. (2017) | Reference | 82989 | 4643 | 22479 | 2761 | 53106 | 0.627 | 0.703 | 0.728 |
| Xiuhua et al. (2018) | Reference | 105 | 20 | 18 | 11 | 56 | 0.690 | 0.780 | 0.775 |
| Xiuhua et al. (2018) | Reference | 105 | 24 | 16 | 8 | 57 | 0.780 | 0.800 | 0.846 |
| Xiuhua et al. (2018) | Reference | 105 | 21 | 36 | 11 | 37 | 0.690 | 0.560 | 0.609 |
| Xiuhua et al. (2018) | Custom | 105 | 20 | 7 | 14 | 64 | 0.610 | 0.910 | 0.864 |
| Zellweger et al. (2018) | Reference | 262 | 206 | 13 | 2 | 41 | 0.412 | 0.877 | 0.690 |
| Zellweger et al. (2018) | Custom | 215 | 140 | 19 | 19 | 37 | 0.606 | 0.916 | 0.870 |

***Abbreviations:***ANN, artificial neural network; AUC, area under the curve; BA, boosting algorithm; CNN, convolutional neural network; FN, false negative; FP, false positive; LGR, logistic regression; LNR, linear regression; RF, random forest; SVM, support vector machine; TN, true negative; TP, true positive.