

Machine Learning for Integrating Social Determinants in Cardiovascular Disease Prediction Models: A Systematic Review

Yuan Zhao¹, Erica P. Wood², Nicholas Mirin,² Rajesh Vedanthan,^{3,4} Stephanie H. Cook,^{2,5} Rumi Chunara^{5,6}

¹New York University, School of Global Public Health, Department of Epidemiology

²New York University, School of Global Public Health, Department of Social and Behavioral Sciences

³New York University Grossman School of Medicine, Department of Population Health

⁴New York University Grossman School of Medicine, Department of Medicine

⁵New York University, School of Global Public Health, Department of Biostatistics

⁶New York University Tandon School of Engineering, Department of Computer Science & Engineering

Summary

Background

Cardiovascular disease (CVD) is the number one cause of death worldwide, and CVD burden is increasing in low-resource settings and for lower socioeconomic groups worldwide. Machine learning (ML) algorithms are rapidly being developed and incorporated into clinical practice for CVD prediction and treatment decisions. Significant opportunities for reducing death and disability from cardiovascular disease worldwide lie with addressing the social determinants of cardiovascular outcomes. We sought to review how social determinants of health (SDoH) and variables along their causal pathway are being included in ML algorithms in order to develop best practices for development of future machine learning algorithms that include social determinants.

Methods

We conducted a systematic review using five databases (PubMed, Embase, Web of Science, IEEE Xplore and ACM Digital Library). We identified English language articles published from inception to April 10, 2020, which reported on the use of machine learning for cardiovascular disease prediction, that incorporated SDoH and related variables. We included studies that used data from any source or study type. Studies were excluded if they did not include the use of any machine learning algorithm, were developed for non-humans, the outcomes were bio-markers, mediators, surgery or medication of CVD, rehabilitation or mental health outcomes after CVD or cost-effective analysis of CVD, the manuscript was non-English, or was a review or meta-analysis. We also excluded articles presented at conferences as abstracts and the full texts were not obtainable. The study was registered with PROSPERO (CRD42020175466).

Findings

Of 2870 articles identified, 96 were eligible for inclusion. Most studies that compared ML and regression showed increased performance of ML, and most studies that compared performance with or without SDoH/related variables showed increased performance with them. The most frequently included SDoH variables were race/ethnicity, income, education and marital status. Studies were largely from North America, Europe and China, limiting the diversity of included populations and variance in social determinants.

Interpretation

Findings show that machine learning models, as well as SDoH and related variables, improve CVD prediction model performance. The limited variety of sources and data in studies emphasize that there is opportunity to include more SDoH variables, especially environmental ones, that are known CVD risk factors in machine learning CVD prediction models. Given their flexibility, ML may provide opportunity to incorporate and model the complex nature of social determinants. Such data should be recorded in electronic databases to enable their use.

Funding

We acknowledge funding from Blue Cross Blue Shield of Louisiana. The funder had no role in the decision to publish.

Introduction

An estimated 17.9 million people die each year from cardiovascular diseases (CVD), which represent 31% of all deaths worldwide and the number one cause of death.¹ Low-income and middle-income countries carry 75% of the burden of CVD deaths worldwide and in high-income countries, lower socioeconomic groups have a higher

56 incidence of CVD and higher mortality due to CVD.^{1,2} In high-income countries such as the United States, the
57 prevalence of CVD is expected to rise 10% between 2010 and 2030,³ not only in the aging population but also
58 notably via stark disparities among socioeconomic and racial groups.^{4,5} Direct causes for these shifts in CVD
59 burden have been well-studied, attributed to changes in diet (increased consumption of processed foods)⁶ and
60 physical activity (more sedentary lifestyles),⁷ resulting in a dramatic rise in conditions such as obesity,
61 hypertension, and diabetes mellitus. These changes are shaped by the “conditions in which people are born,
62 grow, live, work and age”, referred to by the World Health Organization as social determinants of health (SDoH).⁸

63 Multinational, prospective cohort studies as well as ecologic analyses have shown that SDoH contribute to
64 over 35% of the population attributable risk of various cardiovascular diseases,^{9,10} among which education,
65 income and occupation are particularly influential.¹¹ Research has also illuminated mechanisms of action; social
66 factors usually interact with each other through the mediation of or effect modification by psychological and
67 biological pathways, exerting a long-term effect on cardiovascular outcomes.^{5,12} Social determinants also result
68 in unequal sharing of the benefit of advances in CVD prevention and treatment.¹³ Given the critical importance of
69 social determinants with respect to disease risk, it is clear that better capturing the interaction and relative
70 influence of such factors in relation to traditional CVD risk factors of hypertension, diabetes and hyperlipidemia
71 provides the most significant opportunity to reduce CVD burden.^{5,11,12,14}

72 Meanwhile, artificial intelligence (AI) and machine learning (an application of AI for detecting patterns from
73 data)¹⁵ tools have started to be adopted in clinical research, prompted by recent progress in advanced computing
74 strategies as well as the proliferation of electronic medical record databases.¹⁶ Machine learning methods have
75 demonstrated improvement across multiple metrics for prediction of CVD risk, incidence and outcomes¹⁷⁻¹⁹ over
76 traditional risk scores such as those from the American College of Cardiology or American Heart Association.²⁰ As
77 a data-driven approach, machine learning provides more flexibility in modeling complex relationships between
78 predictors, which can be particularly advantageous in addressing the multi-level interactions between different
79 social determinants and CVD outcomes, as well as uncovering novel risk factors. Though the increased flexibility
80 of machine learning models is appealing, given the rapid rise of machine learning approaches including studies
81 which incorporate social determinants, we need to better understand best practices for such modelling
82 approaches for CVD risk prediction particularly in the context of those including SDoH.

83 Thus, we performed a systematic review to understand the current landscape of how social determinants are
84 being used in machine learning models for CVD prediction. Specifically, we sought to examine which types of
85 machine learning algorithms and types of social determinant variables are being used, and for which
86 populations. Indeed, understanding the manner in which SDoH are incorporated into such models is critical in
87 order to tease apart the distinct the biological and social influences, along with their interactions, that make
88 populations different and in need of a different standard of care. Findings from this review serve to inform the
89 design of future machine learning approaches and identify areas for methodological innovation in order to
90 improve early prediction of CVD and reduce its significant disease burden.^{21,22}

91

92 **Method**

93 **Search strategy and selection criteria**

94 First, YZ with the help of an expert librarian, did a comprehensive search of five databases: PubMed, Embase,
95 Web of Science, IEEE Xplore and ACM Digital Library on April 10th, 2020, to identify all relevant articles on
96 machine learning integrating social determinants in cardiovascular disease prediction models published in
97 English. IEEE Xplore and ACM Digital Library were included specifically to comprehensively capture computer

98 science articles related to our review. Papers from inception until the search date were included. To ensure the
99 quality of included papers, we only included peer-reviewed articles published on journals or accepted in
100 conferences and excluded non-peer reviewed grey literature or arXiv/medRxiv papers.

101 We identified the terms of social determinants of health (SDoH) using the broader definition from the World
102 Health Organization and Center for Disease Control and Prevention Healthy People 2020 initiative which delineates
103 SDoH in five key areas: economic stability, education, social and community context (e.g. “race/ethnicity”, “income”
104 and “education”), health and health care, neighborhood and built environment (e.g. “living environment”,
105 “pollution” and “residence characteristics”).²³ Figure 1 is a socio-ecological conceptual model adapted from Healthy
106 People 2020, the United States federal government’s national health agenda,²⁴ which illustrates the multifactorial
107 nature of social-ecological influences on health. The framework emphasizes the existence of proximate, or
108 “downstream,” health influences (e.g., smoking) that are shaped by distal, or “upstream,” factors (e.g., social norms
109 regarding smoking, tobacco regulations). Therefore, for a robust review, we also included prominent factors
110 including health-related behaviors along the causal pathway (e.g. “diet”, “smoking” and “physical activity”), as
111 although these are enacted at the individual level, they are shaped at social and economic levels.²⁵ This enables us to
112 understand comprehensively how social determinants and factors they directly shape are assessed in relation to
113 CVD. Age, gender and race are also in the causal pathway.²⁶ BMI was also included as it is influenced by social
114 factors and causes diabetes which directly affects CVD.²⁷

115 For search terms related to machine learning, we included all commonly used supervised machine learning
116 methods. Supervised machine learning algorithms are those that perform reasoning (i.e. prediction) from
117 observations of the features (e.g. clinical data, social determinants) based on externally supplied examples which
118 include the features linked to outcome “labels” (e.g. CVD outcomes). Thus supervised machine learning was a
119 focus as the types of tasks considered in the literature usually utilized labeled outcomes of CVD.²⁸ Commonly
120 used unsupervised machine learning algorithms captured by the search were also included in the abstract and
121 full text screening to ensure all types of possible studies were considered. We also added search terms to capture
122 deep learning and ensemble methods as they are widely used in current clinical research.²⁹

123 The search terms for CVD outcomes included cardiovascular ischemic outcomes, coronary heart disease and
124 cerebrovascular disease which are caused by atherosclerotic cardiovascular disease (ASCVD). These
125 cardiovascular diseases cause the highest mortality, and estimated years of lives lost attributed to these
126 conditions have increased in recent years.^{1,2,30} For each of the key areas of social determinants and included
127 variables, machine learning and CVD, we identified keywords by referencing previous review papers on social
128 determinants and cardiovascular diseases,^{11,31} related studies of different social determinants³¹⁻³⁵ or consulting
129 experts to include relevant concepts. Full search strategies are provided in the appendix.

130 Once papers were identified via the search terms, all study designs and all populations were included if the
131 article utilized any SDoH or health behaviors as features in the machine learning models (in addition to age and
132 gender, as we found that these were commonly included as standard practice and not specifically to represent their
133 contribution as social determinants) were deemed eligible. Eligibility was also considered if the outcomes were
134 CVD-related, including incidence, survival, mortality, hospital admission and readmission etc. We did not restrict
135 time of publication to enable capturing the trend of these types of papers over time. Studies were excluded if they
136 did not include the use of any machine learning algorithm, were developed for non-humans, the outcomes were
137 bio-markers, mediators, surgery or medication for CVD, rehabilitation or mental health outcomes after CVD
138 diagnosis or cost-effectiveness analysis of CVD treatment, the manuscript was non-English, or was a review or
139 meta-analysis. We also excluded articles presented at conferences as abstracts and the full texts were not
140 obtainable. This review was registered with PROSPERO (CRD42020175466) and conducted in accordance with the

141 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method. To supplement the
142 bibliographic database searches, we also used Google Scholar to scrutinize all keywords regarding their relevance
143 in articles as well as examine potential articles to identify if they were eligible. Duplicates were removed in the
144 process.

145 Three investigators (YZ, EPW, and NM) screened the title and abstract: each article retrieved was
146 independently assessed by two reviewers to determine its eligibility for full-text review. Conflicts were resolved
147 by discussion and validation from a third reviewer. After initial appraisal, we retrieved full texts of eligible
148 articles.

149 **Data analysis**

150 Data were extracted from individual articles independently by two reviewers (of YZ, EPW, and NM) and checked
151 by the third reviewer according to criteria in a standardized extraction form. All data extraction was cross-
152 checked, and disagreements were resolved by discussion or referral to the third reviewer. Information extracted
153 included year of publication, country, population, social determinants included in the machine learning
154 algorithm, machine learning algorithms, cardiovascular disease outcomes, data source and performance of the
155 algorithms. For each article, we defined several criteria to assess the quality of the study based on best practices
156 in machine learning³⁶ including (1) whether machine learning model performance was evaluated; (2) whether a
157 hyperparameter (a parameter whose value is used to control the learning process) tuning process was described;
158 (3) whether data-driven variable selection was performed; (4) whether methods were used to specifically
159 interpret the contribution of included variables in the prediction. Each item was scored as no (not present),
160 unclear, or yes (present), and then summarized alongside all items to get a study quality score.

162 **Role of the Funding Source**

163 The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing
164 of the report. The corresponding author had full access to all the data in the study and had final responsibility for
165 the decision to submit for publication.

167 **Results**

168 Our database search identified 2728 distinct articles; after a full-text review of 298 papers, 96 were included in
169 the systematic review (Figure 2). Among the included studies, one of the studies used data from a clinical trial,
170 while the others utilized observational data. Of the observational studies, data from cohort studies was the most
171 frequent (34 studies), followed by data from electronic medical records (32 studies), surveys (14 studies) and
172 data from open-access repositories of registry or national survey data (7 studies) (e.g. Scientific Registry of
173 Transplant Recipients Registry³⁷). Most of the observational data were structured data (clearly defined data
174 features), while 9 studies included unstructured data (e.g. electrocardiogram, image and heart sound). The
175 earliest year of publication was 1992 (artificial neural network algorithm)³⁸, and publications fulfilling our
176 inclusion criteria have been increasing over time (Figure 3). Figure 4 summarizes variables (4A), outcomes (4A),
177 author locations (4B) and types of venues where studies were published (4C). More details on the data sources
178 and populations included, along with all study details are in Table S1.

179

180 **Social determinant and variables in the causal pathway**

181 Included studies reported diverse variables across social determinants and variables considered, including
182 race/ethnicity, education, marital status, occupation/employment, individual or household income, medical
183 insurance, area of residence (e.g. urban versus rural or eastern vs. western USA) and other community-level factors
184 of deprivation, income and education and environmental pollutants as well as smoking, alcohol consumption,
185 physical activities, substance abuse and diet. In most studies, gender and age were included as standard variables
186 collected in the survey or EHR. A few studies assessed physical activities and diet as modifiable risk factors for early
187 prevention of CVD.^{14,39} Half of the studies reported feature importance of variables, in which age, gender, smoking
188 (e.g. current smoking/past smoking/non-smoking) and BMI were most frequently reported to contribute
189 significantly to the CVD outcome prediction. Other frequently reported determinants including race/ethnicity,
190 alcohol consumption (e.g. daily intake or alcoholism), and physical activity/exercise (e.g. weekly exercise time).
191 Besides age, gender, BMI and smoking which were frequently reported in all CVD outcomes, alcohol consumption
192 and physical activities were frequently associated with stroke while BMI was frequently associated with coronary
193 artery disease. The top ten variables considered in extracted papers, and their frequency, are illustrated in Figure
194 4A, which include marital status, education, income and race/ethnicity as the most common social determinants.
195 Four of the studies compared model performance with social determinants and without social determinants; three
196 showed social determinants significantly improved prediction, others showed improved prediction by addition of
197 age, gender and race.^{40,41} The study that showed decreased performance aimed to forecast the pattern of the
198 demand for hemorrhagic stroke healthcare services based on air quality; it is possible that the relationship between
199 specific variable tested and outcome have little direct relationship.⁴²

200

201 **Algorithms and model development**

202 The most common machine learning methods were neural network (NN, 36 studies), random forest (RF, 28
203 studies), and decision trees (DT, 21 studies). Three studies used unsupervised machine learning algorithms, such
204 as clustering to group CVD risk levels or principal component analysis (PCA) to extract features prior to
205 supervised machine learning classification.^{14,43,44} The most frequently used algorithms are described in Table 1.
206 Of the 35 studies using neural networks, 12 used one hidden layer, 23 used multiple hidden layers, including
207 most commonly three-layer perceptron, convolutional neural network and recurrent neural network. Here we
208 refer to these studies collectively as “neural networks” (NN) as deep learning typically refers to an neural
209 network with multiple layers.⁴⁵ Of the 42 studies including multiple machine learning algorithms, random forest
210 (9 studies) and neural network (9 studies) were most frequently reported as the best performing machine
211 learning algorithms. For most commonly studied CVD outcomes, random forest was frequently reported to have
212 the best prediction for stroke while support vector machine (SVM) performed best for coronary artery disease.

213 There were 24 studies that compared machine learning algorithms with standard linear regression, logistic
214 regression or survival analysis; among those 21 showed improved performance with machine learning. One
215 study of risk prediction for in-hospital mortality in women with ST-elevation myocardial infarction using data
216 from the National Inpatient Sample in the United States, found comparable performance using random forest and
217 logistic regression.⁴⁶ In another study, neural network models for prediction of acute coronary syndromes using
218 clinical data and NN showed similar performance to logistic regression in predicting acute coronary syndrome;
219 however, only 13 variables were considered.⁴⁷ A third study on predicting adverse cardiovascular events by
220 models integrating stress-related ventricular functional and angiographic data showed that while a logistic

221 model demonstrated better performance in this task and implementation, a Bayesian network model showed
222 good performance and also was highlighted as being better at defining causal relationships, and thus useful for
223 designing future models in which new variables can be incorporated in the prediction task.⁴⁸

224

225 **Model, validation and performance and study quality**

226 Most studies evaluated the performance of machine learning algorithm(s) developed. Area under the receiver
227 operating characteristic curve (AUC) was the most common evaluation metric used (45) studies, followed by
228 sensitivity (43 studies), specificity (32 studies) and accuracy (32 studies). At least three of the four metrics were
229 used in 31 studies. Other evaluation metrics used included accuracy, positive predictive value, negative
230 predictive value and F1-score, which is the harmonic mean of the precision and recall, commonly used to
231 evaluate machine learning methods via their balance of these metrics. External evaluation was performed in 11
232 studies, wherein the authors tested the machine learning models developed in one hospital on another hospital
233 or population. For example, one study specifically tested the generalizability of a recurrent neural network model
234 for predicting heart failure risk in a large dataset from 10 hospitals; evaluating the performance of a model
235 trained on each hospital's training (and validation) sets over the 10 hospitals' test sets. They also evaluated the
236 model that trained on all hospitals' training sets over the 10 hospitals' test sets.⁴¹ and another used data from one
237 hospital to train neural network models for diagnosis of acute coronary syndrome and tested the model on data
238 from two other hospitals.⁴⁷

239 Among those reported, most AUC were higher than 0.70 (Figure 5). As most studies were published in
240 biomedical and clinical journals, most studies explicitly interpreted the findings and their relevancy to clinical
241 applications. Almost half (40/96) of the studies compared more than one machine learning algorithm, of which
242 Random Forest was most commonly the best performing model. The mean score of included studies in the 4-item
243 quality assessment scale (based on evaluation of ML, data-driven selection of features, hyperparameter tuning
244 description, interpretation of the model) was 3.34. Half of the studies (49) had full scores and 30 studies missed
245 one of the four items. Commonly missed items were data-driven feature selection and details of hyperparameter
246 tuning (cross-validation or grid search strategies were utilized in 68 studies to tune hyperparameters; other
247 studies didn't give details about hyper-parameter tuning process). Half of all the studies utilized a data-driven
248 selection method to identify features before fitting machine learning models, which is defined as extracting a
249 subset of useful variables among the original variables and transforming data from a high- to a low-dimensional
250 space.⁴⁹ As deep learning models to extract features while training, those studies did not always include a feature
251 selection process.

252

253 **Discussion**

254 To our knowledge, this is the first systematic review to illustrate how machine learning is being used to
255 integrate SDoH in cardiovascular disease prediction models. This review distills which types of algorithms and
256 SDoH and related variables have been considered and resulting performance. We found that the flexibility of
257 machine learning models has proved useful in CVD prediction models, with them commonly performing better
258 than regression approaches. We find that models that consider SDoH and related variables also benefit from
259 flexible modeling approaches, with neural networks consistently outperforming regression across all CVD
260 outcomes.

261 Broadly, we found several limitations in the content covered by included papers. First, the studies were highly
262 skewed to originate from USA, Europe and China, with lower-income locations not being well represented.
263 Moreover, we found that the race and ethnicity distribution in some studies was also not very representative of
264 underlying populations. This is particularly striking given the high and increasing CVD burden and changing
265 socio-environmental circumstances in lower-income countries and regions and disparities in CVD burden. The
266 variance of social determinants incorporated into models, and thus the performance and applicability of those
267 models across contexts, will be decreased with less diversity in the study sample.⁵⁰ SDoH variables themselves
268 were also not very frequently included, with only marital status, education, income and race/ethnicity in the top
269 ten. Environmental attributes that have been shown as important modifiable components of CVD risk such as
270 green spaces and stress^{32,51,52} were very few, and even then were very broad (e.g. region of country).⁴² If such
271 area-based variables are included, machine learning may also prove useful to unweave the strands of
272 environmental influences but also integrate the effects of the various components of the environment into a
273 comprehensive model.³² We also find that models did not take into account social processes associated with
274 socioeconomic conditions across the life course. Socioeconomic position, psychosocial factors and behaviors
275 during adolescence and youth are important likely to be important in the development of CVD and precursors
276 (dyslipidemia, hypertension, and smoking).⁵³ Finally, studies generally included gender interchangeably with sex,
277 which precludes consideration of the socially-determined aspect of gender.⁵⁴

278 Despite these limitations, our results largely found social determinants and variables considered to improve
279 model performance. In terms of algorithms, several types of machine learning algorithms were evaluated, with
280 results showing that when compared within studies, the most flexible models such as neural networks and
281 random forest models were best performing. Neural networks also most commonly outperformed regression
282 models. This is understood to be the case because neural networks include hidden layers which can take into
283 account more complex relations in the data, and therefore this may be another possible explanation for the
284 improved performance.^{55,56} Moreover, recent studies uncovering network and spillover effects (social
285 environment) and shared decision-making⁵⁷ involved in physical activity,^{58,59} diet⁶⁰ and smoking⁶¹ indicate that
286 the pathways that inform these behaviors are intricate. However, this may illuminate an opportunity for machine
287 learning, which based on flexibility, can help capture such complex interactions.

288 The constraints on included data are likely due to difficulties in capturing certain SDoH variables and linking
289 them with individual records in databases used in many of the included studies. Studies have largely used social
290 variables from available data sources; commonly those in the electronic health record. The use of flexible,
291 machine learning models also bring concerns regarding interpretability and potential over-fitting to data,⁶²
292 though this was not a common discussion topic across all papers. This is likely because most models selected
293 variables based on prior clinical significance, thus prediction performance would be based on such factors which
294 are known to be relevant to CVD even if the specific importance of each variable was not measured. Furthermore,
295 most papers (66) papers used methods such as automatic relevance determination⁵⁵ or feature selection⁶³ to
296 examine and/or rank the importance of variables in machine learning models. This was the case even as articles
297 were published in a variety of venues (Figure 4D).

298 While this is the first review that gives findings related to the use of machine learning and social determinants
299 for CVD prediction, there are individual studies that support components of the findings of this study. First,
300 machine learning in general has shown promise with respect to cardiovascular disease prediction.⁶⁴⁻⁶⁶ Compared
301 to the established American College of Cardiology/American Heart Association risk calculator to predict
302 incidence and prognosis of ASCVD,²⁰ previous work has shown that machine-learning algorithms (especially
303 random forest, gradient boosting machines and neural networks) were better at identifying individuals who will

304 develop CVD and those who will not.⁷⁻⁹ These studies have attributed this to the fact that standard CVD risk
305 assessment models make an implicit assumption that each risk factor is related in a linear fashion to CVD
306 outcomes and such models may thus oversimplify complex relationships which include large numbers of risk
307 factors with non-linear interactions. The role of social determinants in cardiovascular disease (not specifically
308 machine learning-related) has been studied through several papers and systematic reviews. While full
309 summaries of this work have been performed elsewhere,¹¹ we note that there have been several studies of
310 various proximal and distal social determinants and cardiovascular disease. In general, studies indicate that the
311 changing burden of disease due to societal and environmental conditions, as well as increasing advances in
312 treatment and prevention have not been shared equally across economic, racial and ethnic groups, compelling
313 the need for broad range consideration of social determinants in CVD prediction.^{11,31} Finally, the models that have
314 incorporated social determinants and machine learning for CVD prediction also reflect limitations of many
315 machine learning algorithms that have been highlighted recently, which are based on homogenous populations,
316 particular with respect to race (captured through the limited geographic diversity in Figure 4C).¹⁷

317 Our review was limited in several aspects. First, the included studies evaluated different types of
318 cardiovascular outcomes, and heterogeneity of outcome metrics makes it difficult to compare machine learning
319 performance across studies. The population considered also includes samples from different data sources,
320 hospitals and countries which taken together make the comparison across studies not standardized. Third, most
321 studies did not evaluate external validity, leaving the applicability of the algorithms to other populations or
322 healthcare settings inconclusive. Fourth, the review was also limited to studies published in English, which might
323 have created some bias in the articles that were ultimately retained for the analysis.

324 Findings emphasize the need to comprehensively capture both proximal and distal social determinant
325 variables in models. Where mechanisms are not well understood, machine learning can also be used to
326 understand relationships between social and biological variables comprehensively. For example, race is often
327 conceptualized as a proxy for variables for socioeconomic position or cultural factors and better ways to capture
328 as well as understand relationships between these factors and their impact on CVD risk should be investigated.
329 Indeed, identification of potential mediating and moderating factors in these pathways of social determinants
330 will inform public health interventions. Improved constructs will also help in incorporation of environmental and
331 behavioral variables such as diet and physical activity which were not well represented in current studies. Our
332 findings support bodies of work that promote inclusion of such information in the electronic health record^{67,68}
333 and we add reasoning that this would also enable study of social determinants in machine learning in large
334 enough sample sizes to reduce overfitting of models. Finally, results emphasize the need for studies that include
335 more diverse populations with varied environmental and social influences, which would represent and ensure
336 validity of prediction models across these diverse interactions⁵⁰ to improve cardiovascular disease prediction in
337 diverse settings, in particular those where disease risk is increasing.

338

339 **Research in context**

340 **Evidence before this study**

341 While there are no reviews that specifically address social determinants, machine learning and cardiovascular
342 disease (CVD), the latest research on cardiovascular disease indicates the imperative relevance of social
343 determinants. Societal and environmental conditions distributed unequally among groups are driving a
344 significant and increasing global burden of cardiovascular diseases particularly to low and middle-income
345 countries as well as lower-socioeconomic groups in high-income countries. At the same time, research shows

346 that machine learning offers the potential for capturing flexible relationships compared to the linear
347 relationships assumed in typical CVD risk scores, which is of particular relevance for the consideration of social
348 variables. Select models have examined the performance of certain social variables in CVD prediction, including
349 those that use machine learning, but given these rapidly advancing research areas, a systematic examination into
350 which social determinants have been modelled and how different methods have performed, is needed.

351 **Added value of this study**

352 Through a rigorous and comprehensive systematic review, we assessed the state-of-the art methods prediction
353 of the two types of CVD with the highest recent mortality (ischemic heart disease and stroke), that use machine
354 learning and incorporate social determinants and related variables in their causal pathway. We show that
355 environmental and area-based determinants are lacking from most models. Machine learning, especially flexible
356 models such as neural networks, show good performance in relation to regression models. We accounted for
357 model and information gain differences across by examining within study performance of best algorithm
358 compared to regression. We assessed the quality of their implementations via best-practices from the machine
359 learning literature, finding that quality was generally rigorous. Finally, we show that the origin of studies is
360 highly skewed to USA and middle/high-income countries in Europe and Asia, which indicates that knowledge
361 regarding the diversity of social determinants and their impact is limited.

362 **Implications of all the available evidence**

363 With the significant burden of CVD and large burden in low- and middle-income countries, this work directly
364 informs how we can augment prediction models, using state of the art machine learning methods, while also
365 taking into account growing social, environmental risk factors that shape CVD risk. According to the findings of
366 this review, strategies to capture social variables, especially environmental determinants are needed in the
367 electronic health record databases from which machine learning methods are commonly developed. Finally,
368 studies to-date represent a narrow set of locations; we need to support studies in low- and middle-income
369 countries to identify and tailor our understanding to the specific social determinants in these populations.

370

371 **Acknowledgements**

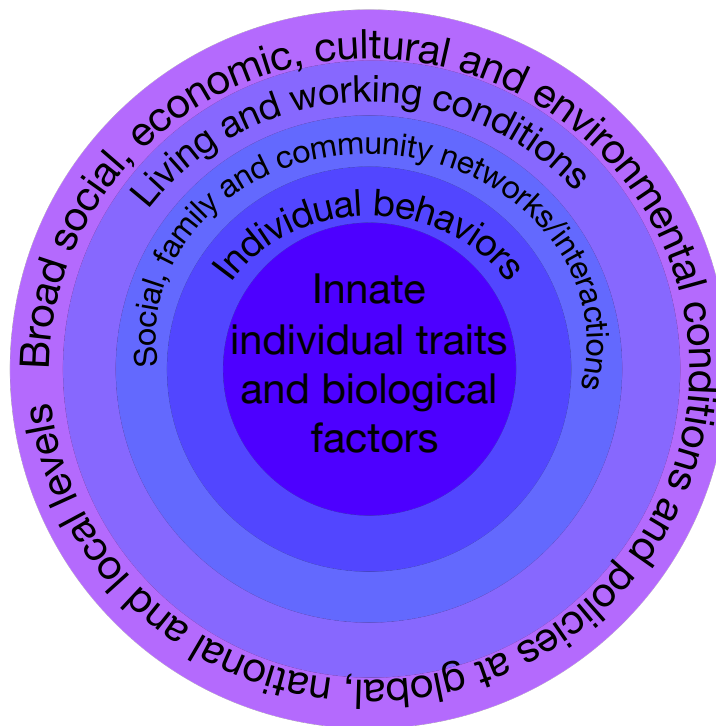
372 We thank Dorice Vieira for valuable help with the search process.

373

374

375 **Figures and Tables**

376



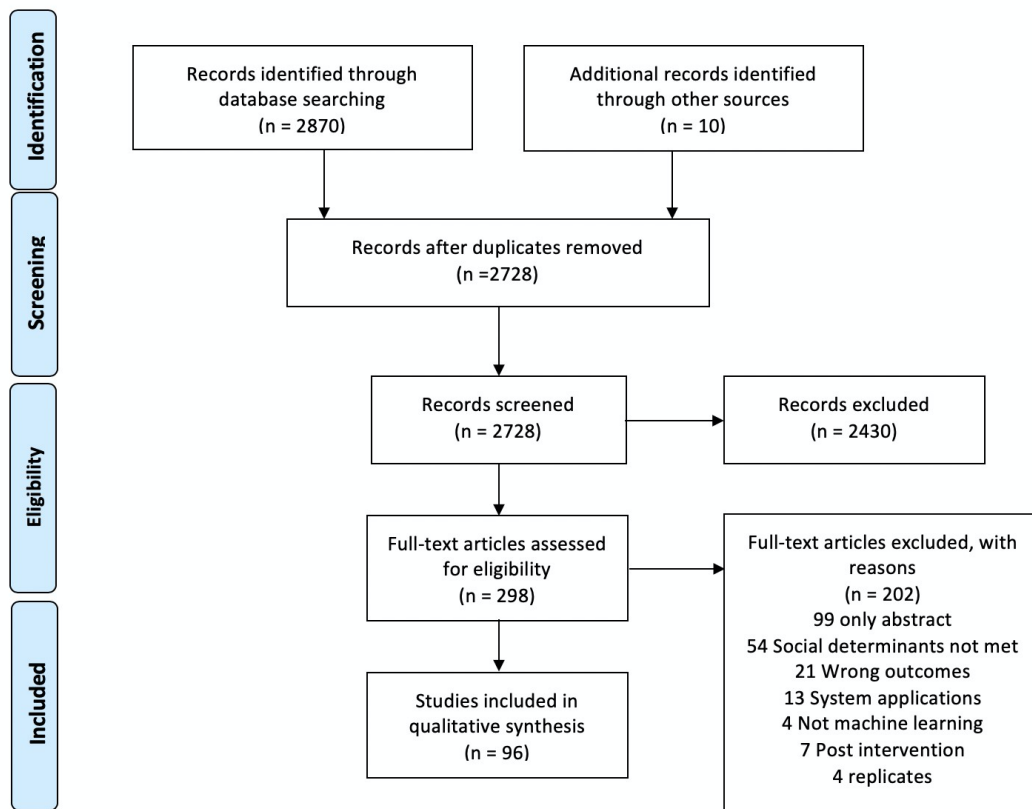
377

378

379

Figure 1: Socio-ecological framework of health; conceptual model used in the study, adapted from Healthy People 2020, the United States federal government’s national health agenda

380

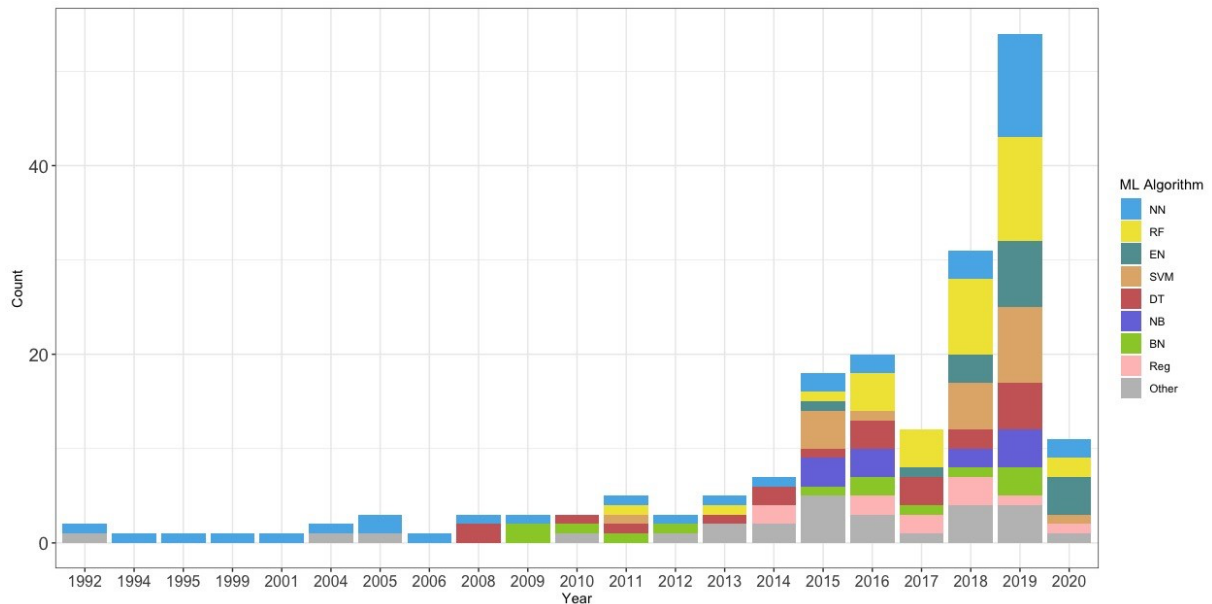


381

382

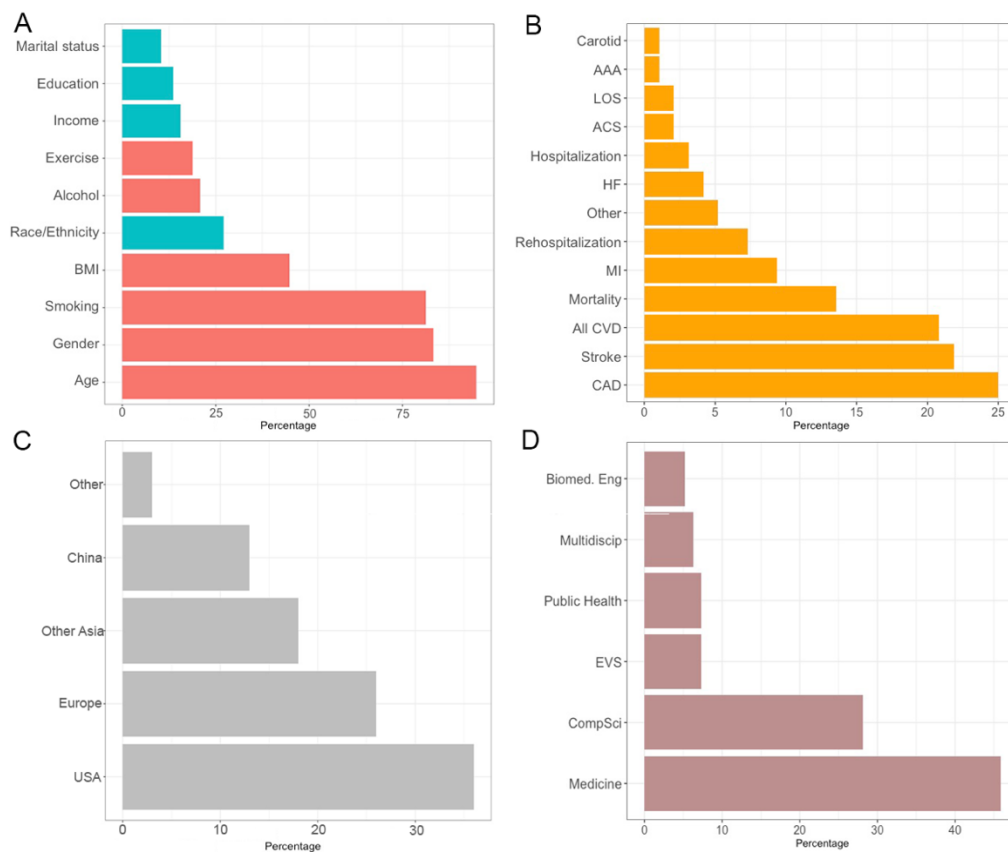
Figure 2: PRISMA flowchart of study review process and exclusion of papers

383



384

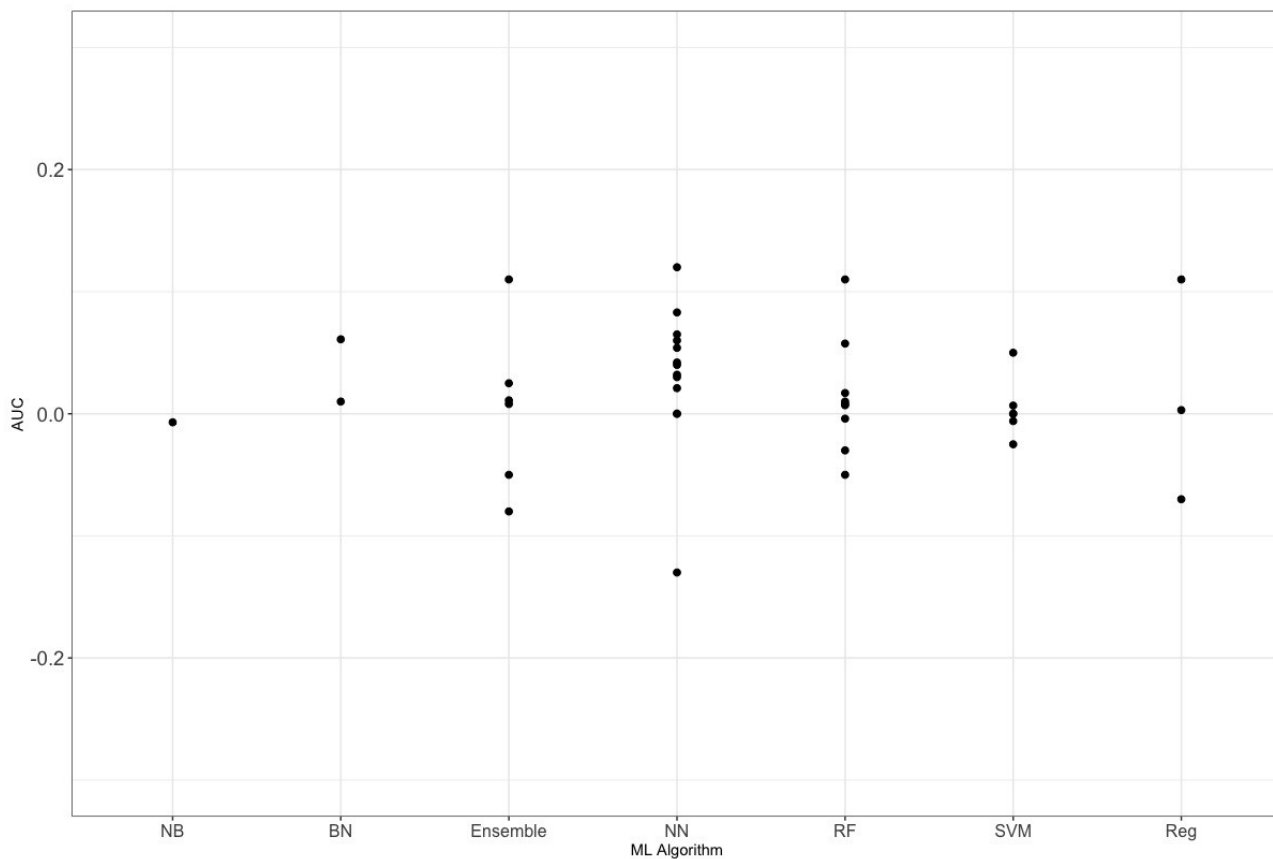
385 Figure 3: Number of ML algorithms used in publications by year and type (NN: neural network, RF: random forest,
 386 EN: ensemble methods (e.g. Adaboost, gradient boosting, bagged decision tree), SVM: support vector machine, DT:
 387 decision tree, NB: naïve Bayes, BN: Bayesian network, Reg: regularization methods (ridge/lasso regression), Other:
 388 multilayer perceptron, maximum entropy, adversarial network, linear discriminant analysis, k-nearest neighbors,
 389 recursive partitioning, clustering, quadratic discriminant, radial basis function kernel)
 390



391

392 Figure 4: (A) Top ten social determinant and related variables included based on study inclusion criteria (social
 393 determinants in blue, other variables in red), (B) most frequently reported CVD outcomes (AAA: abdominal aortic
 394 aneurysm, LOS: length of stay , ACS: acute coronary syndrome, HF: heart failure, MI: myocardial infarction, CAD:
 395 coronary artery disease) (C) countries of corresponding authors and (D) journal types of publication reported in
 396 systematic review papers (EVS: environmental sciences) all with respect to the percentage of included papers they
 397 appear in

398



399

400

Figure 5: Difference of Area under the ROC curve between ML and LR by ML algorithm type

401

Table 1: Summary of machine learning algorithms, best performing and sample sizes used in the studies

Algorithm	Number (%) of papers**	Num as best algorithm when multiple algorithms	Sample Size			
			<100	100-1000	1000-10,000	>10,000
Neural Net	35 (36.5%)	9	5	8	13	9
Random Forest	32 (33.3%)	9	0	3	14	15
Decision Tree	21 (21.9%)	2	2	6	8	5
Support Vector Machine	20 (20.8%)	7	1	3	11	5
Ensemble	17 (17.7%)	5	0	2	7	8
Bayesian Network	13 (13.5%)	1	1	4	2	6
Naïve Bayes	12 (12.5%)	1	2	2	5	3
Regularization*	11 (11.5%)	0	0	2	6	3
Other	28 (29.2%)	1	4	6	12	6

402

*regularization included Lasso, Ridge and Elastic net

403

**Note: each paper could include multiple versions or multiple algorithms

404

405

406

407

408 **Appendix**

409 **Table S1:** Summary of all included papers (attached at end).

410 **Search terms and search strategies PubMed:**

411 (Social determinants of health[Mesh] OR demography[Mesh] OR demographic*[tw] OR race [tw] OR racism[Mesh]
412 OR "ethnicity"[tw] OR gender identity[Mesh] OR gender[tw] OR social[tw] OR social support[Mesh] OR
413 income[Mesh] OR education[Mesh] OR employment[Mesh] OR marital status[Mesh] OR occupation[tw] OR
414 "health insurance"[tw] OR health literacy[Mesh] OR marriage[tw] OR insurance[tw] OR housing[tw] OR
415 home[tw] OR religion[tw] OR socioeconomic factors[Mesh] OR social class[Mesh] OR "social status"[tw] OR
416 "access healthcare"[tw] OR healthcare disparities[Mesh] OR "financial difficulties"[tw] OR poverty[Mesh] OR
417 "social disparity"[tw] OR unemployment[Mesh] OR social condition[Mesh] OR "social inequality"[tw] OR
418 vulnerable population[Mesh] OR "social environment"[tw] OR sociodemographic*[tw] OR sociological
419 factors[Mesh] OR body mass index[Mesh] OR physical activity[Mesh] OR diet[Mesh] OR smoking[Mesh] OR
420 "alcohol consumption"[tw] OR tobacco[Mesh] OR "substance use"[tw] OR "physical inactivity"[tw] OR "substance
421 abuse"[tw] OR health* behavi*r*[tw] OR health* service[tw] OR environment[Mesh] OR "living environment" OR
422 "birthplace"[tw] OR "pollution"[tw] OR residence characteristics[Mesh] OR "geographic locations"[tw] OR
423 "rural"[tw] OR "urban health"[tw] OR neighborhood[tw] OR cultur*[tw]) AND (machine learning[Mesh] OR
424 supervised machine learning[Mesh] OR decision trees[Mesh] OR neural networks[Mesh] OR "Naive Bayes"[tw]
425 OR "kNN"[tw] OR support vector machine[Mesh] OR perceptron[tw] OR "radial basis function"[tw] OR "Bayesian
426 Network"[tw] OR "random forest"[tw] OR "classification tree"[tw] OR "elastic net"[tw] OR "multilayer
427 perceptron"[tw] OR lasso[tw] OR ridge[tw] OR "nearest neighbor"[tw] OR deep learning[Mesh] OR boosting[tw]
428 OR bagging[tw] OR ensemble[tw]) AND ("atherosclerotic cardiovascular disease"[tw] OR cardiovascular
429 abnormalities[Mesh] OR heart disease*[tw] OR heart arrest[Mesh] OR myocardial ischemia[Mesh] OR arterial
430 occlusive diseases[Mesh] OR cerebrovascular disorders[Mesh] OR peripheral vascular diseases[Mesh]) **Embase:**
431 (exp "social determinants of health"/ or exp "demography"/ or demographic* or *race"/ or "racism" or
432 *ethnicity"/ or exp "gender identity"/ or "gender" or "social" or exp "social support"/ or exp "education"/ or exp
433 "employment"/ or "income" or "marital status" or exp "occupation"/ or exp "health insurance"/ or "health
434 literacy"/ or exp "marriage"/ or "insurance" or exp "housing"/ or "home" or "religion" or "socioeconomic factors"
435 or exp "socioeconomics"/ or "social class" or exp "healthcare access"/ or exp "health care disparities"/ or
436 "financial difficulties" or exp "poverty"/ or "social disparity" or exp "unemployment"/ or exp "social status"/ or
437 "social inequality" or exp "vulnerable population"/ or *social environment/ or sociodemographic* or *body
438 mass/ or *physical activity/ or *diet/ or exp "smoking"/ or exp "alcohol consumption"/ or "tobacco use" or exp
439 "substance use"/ or exp "physical inactivity"/ or exp "substance abuse"/ or *environment/ or *birthplace/ or exp
440 "pollution"/ or "residence characteristics" or *geography/ or "neighborhood" or cultur* or exp "rural health"/ or
441 exp "urban health"/) and (exp "machine learning"/ or "supervised machine learning" or exp "decision trees"/ or
442 "neural networks" or exp "artificial neural network"/ or "Naive Bayes" or exp "Bayesian learning"/ or exp "k
443 nearest neighbor"/ or "knn" or exp "support vector machine"/ or "SVM" or exp "perceptron"/ or exp "radial
444 based function"/ or "Bayesian Network" or exp "random forest"/ or "classification tree" or "elastic net" or
445 "multilayer perceptron" or "lasso" or "ridge" or exp "deep learning"/ or "boosting" or "ensemble") and
446 ("atherosclerotic cardiovascular disease" or *coronary artery atherosclerosis/ or "cardiovascular abnormalities"
447 or exp "cardiovascular malformation"/ or *heart disease/ or exp "heart arrest"/ or exp "myocardial ischemia"/ or
448 "arterial occlusive diseases" or exp "cerebrovascular disorders"/ or exp "peripheral vascular diseases"/)

449 **Web of Science:**

450 TS=((("Social determinants of health" OR demography OR demographic* OR race OR ethnicity OR "gender
451 identity" OR gender OR social OR "social support" OR income OR education OR employment OR "marital status"
452 OR occupation OR "health insurance" OR marriage OR insurance OR housing OR religion OR "socioeconomic
453 factors" OR "social class" OR "access healthcare" OR "healthcare disparities" OR "financial difficult" OR poverty

454 OR "social disparity" OR unemployment OR "social condition" OR "social inequality" OR "vulnerable population"
455 OR "social environment" OR sociodemographic* OR "body mass index" OR "physical activity" OR diet OR smoking
456 OR "alcohol consumption" OR tobacco OR "substance use" OR "physical inactivity" OR "substance abuse" OR
457 environment OR birthplace OR pollution OR "residence characteristics" OR "geographic locations" OR "rural" OR
458 "urban health") AND ("machine learning" OR "supervised machine learning" OR "decision trees" OR "neural
459 networks" OR "Naive Bayes" OR kNN OR "support vector machine" OR "perceptron" OR "radial basis function" OR
460 "Bayesian Network" OR "random forest" OR "classification tree" OR "elastic net" OR "multilayer perceptron" OR
461 "lasso" OR "ridge" OR "nearest neighbor" OR "deep learning" OR "boosting" OR "ensemble") AND
462 ("atherosclerotic cardiovascular disease" OR "cardiovascular abnormalities" OR heart disease* OR "heart arrest"
463 OR "myocardial ischemia" OR "arterial occlusive diseases" OR "cerebrovascular disorders" OR "peripheral
464 vascular diseases"))
465 **IEEE:**
466 ("Social determinants of health" OR demography OR demographic* OR race OR ethnicity OR "gender identity" OR
467 gender OR social OR "social support" OR income OR education OR employment OR "marital status" OR
468 occupation OR "health insurance" OR marriage OR insurance OR housing OR religion OR "socioeconomic factors"
469 OR "social class" OR "access healthcare" OR "healthcare disparities" OR "financial difficult" OR poverty OR "social
470 disparity" OR unemployment OR "social condition" OR "social inequality" OR "vulnerable population" OR "social
471 environment" OR sociodemographic* OR "body mass index" OR "physical activity" OR diet OR smoking OR
472 "alcohol consumption" OR tobacco OR "substance use" OR "physical inactivity" OR "substance abuse" OR
473 environment OR birthplace OR pollution OR "residence characteristics" OR "geographic locations" OR "rural" OR
474 "urban health") AND ("machine learning" OR "supervised machine learning" OR "decision trees" OR "neural
475 networks" OR "Naive Bayes" OR kNN OR "support vector machine" OR "perceptron" OR "radial basis function" OR
476 "Bayesian Network" OR "random forest" OR "classification tree" OR "elastic net" OR "multilayer perceptron" OR
477 "lasso" OR "ridge" OR "nearest neighbor" OR "deep learning" OR "boosting" OR "ensemble") AND
478 ("atherosclerotic cardiovascular disease" OR "cardiovascular abnormalities" OR heart disease* OR "heart arrest"
479 OR "myocardial ischemia" OR "arterial occlusive diseases" OR "cerebrovascular disorders" OR "peripheral
480 vascular diseases"))

481 References

- 482 1 World Health Organization. Cardiovascular diseases (CVDs) fact sheet. *World Health Organization*
483 (2017).
- 484 2 Deaton, C. *et al.* The global burden of cardiovascular disease. *European Journal of Cardiovascular Nursing*
485 **10**, S5-S13 (2011).
- 486 3 Heidenreich, P. A. *et al.* Forecasting the impact of heart failure in the United States: a policy statement
487 from the American Heart Association. *Circulation: Heart Failure* **6**, 606-619 (2013).
- 488 4 Kuzawa, C. W. & Sweet, E. Epigenetics and the embodiment of race: developmental origins of US racial
489 disparities in cardiovascular health. *American Journal of Human Biology: The Official Journal of the*
490 *Human Biology Association* **21**, 2-15 (2009).
- 491 5 Carnethon, M. R. *et al.* Cardiovascular health in African Americans: a scientific statement from the
492 American Heart Association. *Circulation* **136**, e393-e423 (2017).
- 493 6 Stuckler, D., McKee, M., Ebrahim, S. & Basu, S. Manufacturing epidemics: the role of global producers in
494 increased consumption of unhealthy commodities including processed foods, alcohol, and tobacco. *PLoS*
495 *Med* **9**, e1001235 (2012).
- 496 7 Lakka, T. A. *et al.* Sedentary lifestyle, poor cardiorespiratory fitness, and the metabolic syndrome.
497 *Medicine & Science in Sports & Exercise* (2003).
- 498 8 Health, W. C. o. S. D. o. & Organization, W. H. *Closing the gap in a generation: health equity through action*
499 *on the social determinants of health: Commission on Social Determinants of Health final report.* (World
500 Health Organization, 2008).
- 501 9 Joseph, P. *et al.* Reducing the global burden of cardiovascular disease, part 1: the epidemiology and risk
502 factors. *Circulation research* **121**, 677-694 (2017).
- 503 10 Tillmann, T. *et al.* Psychosocial and socioeconomic determinants of cardiovascular mortality in Eastern
504 Europe: A multicentre prospective cohort study. *PLoS medicine* **14**, e1002459 (2017).

- 505 11 Havranek, E. P. *et al.* Social determinants of risk and outcomes for cardiovascular disease: a scientific
506 statement from the American Heart Association. *Circulation* **132**, 873-898 (2015).
- 507 12 Theodore, R. F. *et al.* Childhood to early-midlife systolic blood pressure trajectories: early-life predictors,
508 effect modifiers, and adult cardiovascular outcomes. *Hypertension* **66**, 1108-1115 (2015).
- 509 13 Cooper, R. *et al.* Trends and disparities in coronary heart disease, stroke, and other cardiovascular
510 diseases in the United States: findings of the national conference on cardiovascular disease prevention.
511 *Circulation* **102**, 3137-3147 (2000).
- 512 14 He, X., Matam, B. R., Bellary, S., Ghosh, G. & Chattopadhyay, A. K. CHD Risk Minimization through Lifestyle
513 control: Machine Learning Gateway. *Scientific reports* **10**, 1-10 (2020).
- 514 15 Watson, D. S. *et al.* Clinical applications of machine learning algorithms: beyond the black box. *Bmj* **364**
515 (2019).
- 516 16 Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *New England Journal of Medicine* **380**,
517 1347-1358 (2019).
- 518 17 Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. & van der Schaar, M. Cardiovascular disease risk
519 prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants.
520 *PloS one* **14**, e0213653 (2019).
- 521 18 Dimopoulos, A. C. *et al.* Machine learning methodologies versus cardiovascular risk scores, in predicting
522 disease risk. *BMC Medical Research Methodology* **18**, 179 (2018).
- 523 19 Kakadiaris, I. A. *et al.* Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *Journal of*
524 *the American Heart Association* **7**, e009476 (2018).
- 525 20 Cook, N. R. & Ridker, P. M. Further insight into the cardiovascular risk calculator: the roles of statins,
526 revascularizations, and underascertainment in the Women's Health Study. *JAMA internal medicine* **174**,
527 1964-1971 (2014).
- 528 21 Caballero, F. F. *et al.* Advanced analytical methodologies for measuring healthy ageing and its
529 determinants, using factor analysis and machine learning techniques: the ATHLOS project. *Scientific*
530 *Reports* **7**, 43955 (2017).
- 531 22 Seligman, B., Tuljapurkar, S. & Rehkopf, D. Machine learning approaches to the social determinants of
532 health in the health and retirement study. *SSM-population health* **4**, 95-99 (2018).
- 533 23 People, C. o. L. H. I. f. H., Health, B. o. P., Practice, P. H. & Medicine, I. o. *Leading health indicators for*
534 *healthy people 2020: letter report*. (National Academies Press, 2011).
- 535 24 Council, N. R. & Population, C. o. *US health in international perspective: Shorter lives, poorer health*.
536 (National Academies Press, 2013).
- 537 25 Short, S. E. & Mollborn, S. Social determinants and health behaviors: conceptual frames and empirical
538 advances. *Current opinion in psychology* **5**, 78-84 (2015).
- 539 26 Shiroma, E. J. & Lee, I.-M. Physical activity and cardiovascular health: lessons learned from
540 epidemiological studies across age, gender, and race/ethnicity. *Circulation* **122**, 743-752 (2010).
- 541 27 Nuttall, F. Q. Body mass index: obesity, BMI, and health: a critical review. *Nutrition today* **50**, 117 (2015).
- 542 28 Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. Supervised machine learning: A review of classification
543 techniques. *Emerging artificial intelligence applications in computer engineering* **160**, 3-24 (2007).
- 544 29 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436-444 (2015).
- 545 30 Roth, G. A. *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195
546 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017.
547 *The Lancet* **392**, 1736-1788 (2018).
- 548 31 Kreatsoulas, C. & Anand, S. S. The impact of social determinants on cardiovascular disease. *Canadian*
549 *Journal of Cardiology* **26**, 8C-13C (2010).
- 550 32 Bhatnagar, A. Environmental determinants of cardiovascular disease. *Circulation research* **121**, 162-180
551 (2017).
- 552 33 Cheng, I., Ho, W. E., Woo, B. K. & Tsiang, J. T. Correlations between health insurance status and risk
553 factors for cardiovascular disease in the elderly Asian American population. *Cureus* **10** (2018).
- 554 34 Fang, J. *et al.* Association of birthplace and coronary heart disease and stroke among US adults: National
555 Health Interview Survey, 2006 to 2014. *Journal of the American Heart Association* **7**, e008153 (2018).
- 556 35 Lapane, K. L., Lasater, T. M., Allan, C. & Carleton, R. A. Religion and cardiovascular disease risk. *Journal of*
557 *Religion and Health* **36**, 155-164 (1997).
- 558 36 Chen, P.-H. C., Liu, Y. & Peng, L. How to develop machine learning models for healthcare. *Nature*
559 *materials* **18**, 410 (2019).
- 560 37 Hsich, E. M. *et al.* Variables of importance in the Scientific Registry of Transplant Recipients database
561 predictive of heart transplant waitlist mortality. *American Journal of Transplantation* **19**, 2067-2076
562 (2019).
- 563 38 Akay, M. Noninvasive diagnosis of coronary artery disease using a neural network algorithm. *Biological*
564 *cybernetics* **67**, 361-367 (1992).
- 565 39 Shao, Z., Chen, C., Li, W., Ren, H. & Chen, W. Assessment of the risk factors in the daily life of stroke
566 patients based on an optimized decision tree. *Technology and Health Care* **27**, 317-329 (2019).
- 567 40 McGeachie, M. *et al.* An integrative predictive model of coronary artery calcification in arteriosclerosis.
568 *Circulation* **120**, 2448 (2009).

- 569 41 Rasmy, L. *et al.* A study of generalizability of recurrent neural network-based predictive models for heart
570 failure onset risk using a large and heterogeneous EHR data set. *Journal of biomedical informatics* **84**, 11-
571 16 (2018).
- 572 42 Chen, J. *et al.* Machine Learning-Based Forecast of Hemorrhagic Stroke Healthcare Service Demand
573 considering Air Pollution. *Journal of healthcare engineering* **2019** (2019).
- 574 43 Cheon, S., Kim, J. & Lim, J. The use of deep learning to predict stroke patient mortality. *International*
575 *journal of environmental research and public health* **16**, 1876 (2019).
- 576 44 Jabbar, M., Deekshatulu, B. & Chndra, P. in *International Conference on Circuits, Communication, Control*
577 *and Computing*. 322-328 (IEEE).
- 578 45 Illing, B., Gerstner, W. & Brea, J. Biologically plausible deep learning—But how far can we go with
579 shallow networks? *Neural Networks* **118**, 90-101 (2019).
- 580 46 Mansoor, H., Elgendy, I. Y., Segal, R., Bavry, A. A. & Bian, J. Risk prediction model for in-hospital mortality
581 in women with ST-elevation myocardial infarction: A machine learning approach. *Heart & Lung* **46**, 405-
582 411 (2017).
- 583 47 Harrison, R. F. & Kennedy, R. L. Artificial neural network models for prediction of acute coronary
584 syndromes using clinical data from the time of presentation. *Annals of emergency medicine* **46**, 431-439
585 (2005).
- 586 48 Berchiulla, P., Foltran, F., Bigi, R. & Gregori, D. Integrating stress-related ventricular functional and
587 angiographic data in preventive cardiology: a unified approach implementing a Bayesian network.
588 *Journal of Evaluation in Clinical Practice* **18**, 637-643 (2012).
- 589 49 Chu, C. *et al.* Does feature selection improve classification accuracy? Impact of sample size and feature
590 selection on classification using anatomical magnetic resonance images. *Neuroimage* **60**, 59-70 (2012).
- 591 50 Harper, S., Lynch, J. & Smith, G. D. Social determinants and the decline of cardiovascular diseases:
592 understanding the links. *Annual review of public health* **32**, 39-69 (2011).
- 593 51 Richardson, E. A. & Mitchell, R. Gender differences in relationships between urban green space and
594 health in the United Kingdom. *Social science & medicine* **71**, 568-575 (2010).
- 595 52 Steptoe, A. & Kivimäki, M. Stress and cardiovascular disease. *Nature Reviews Cardiology* **9**, 360-370
596 (2012).
- 597 53 Pollitt, R. A., Rose, K. M. & Kaufman, J. S. Evaluating the evidence for models of life course socioeconomic
598 factors and cardiovascular outcomes: a systematic review. *BMC public health* **5**, 7 (2005).
- 599 54 Phillips, S. P. Defining and measuring gender: a social determinant of health whose time has come.
600 *International Journal for Equity in Health* **4**, 1-4 (2005).
- 601 55 Bishop, C. M. Bayesian methods for neural networks. (1995).
- 602 56 Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a
603 methodology review. *Journal of biomedical informatics* **35**, 352-359 (2002).
- 604 57 Smith, K. P. & Christakis, N. A. Social networks and health. *Annu. Rev. Sociol* **34**, 405-429 (2008).
- 605 58 Suglia, S. F. *et al.* Why the neighborhood social environment is critical in obesity prevention. *Journal of*
606 *Urban Health* **93**, 206-212 (2016).
- 607 59 Bahr, D. B., Browning, R. C., Wyatt, H. R. & Hill, J. O. Exploiting social networks to mitigate the obesity
608 epidemic. *Obesity* **17**, 723-728 (2009).
- 609 60 Pachucki, M. A., Jacques, P. F. & Christakis, N. A. Social network concordance in food choice among
610 spouses, friends, and siblings. *American journal of public health* **101**, 2170-2177 (2011).
- 611 61 Årnes, A. P. & Krokstrand, T. T. *The incidence and prevalence of Chronic Fatigue Syndrome, Back Pain of*
612 *unknown origin, Fibromyalgia, and Myalgia in Norwegian women, and their association to physical*
613 *activity. A prospective cohort study of material from the Norwegian Women and Cancer (NOWAC) study,*
614 *UiT Norges arktiske universitet, (2014).*
- 615 62 Ahmad, M. A., Eckert, C. & Teredesai, A. in *Proceedings of the 2018 ACM international conference on*
616 *bioinformatics, computational biology, and health informatics*. 559-560.
- 617 63 Ni, Y. *et al.* Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to
618 detect stroke diagnosis. *PloS one* **13**, e0192586 (2018).
- 619 64 Ambale-Venkatesh, B. *et al.* Cardiovascular event prediction by machine learning: the multi-ethnic study
620 of atherosclerosis. *Circulation research* **121**, 1092-1101 (2017).
- 621 65 Sitar-tăut, A., Zdrenghea, D., Pop, D. & Sitar-tăut, D. Using machine learning algorithms in cardiovascular
622 disease risk evaluation. *Age* **1**, 4 (2009).
- 623 66 Weng, S. F., Reys, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular
624 risk prediction using routine clinical data? *PloS one* **12**, e0174944 (2017).
- 625 67 Bazemore, A. W. *et al.* “Community vital signs”: incorporating geocoded social determinants into
626 electronic records to promote patient and population health. *Journal of the American Medical Informatics*
627 *Association* **23**, 407-412 (2016).
- 628 68 Cantor, M. N. & Thorpe, L. Integrating data on social determinants of health into electronic health
629 records. *Health Affairs* **37**, 585-590 (2018).

Study	Country	Study Design	SDoH and Related Variables Included	CVD Outcomes	Algorithms
Alizadehsani, Roohallah, et al. (2018)	Australia	Unclear	Age, BMI, obesity, sex, smoking	Coronary artery disease	DT, SVM, NB
Tamal, Maruf Ahmed, et al. (2019)	Bangladesh	Observational (survey)	Age, physical exercise, smoking	Heart disease	DT, SVM, NB, QDA, RF, LR
Al'Aref, Subhi J., et al. (2020)	Canada, Germany, Italy, Korea, Switzerland, US	Prospective observational	Age, BMI, ethnicity, sex, smoking	Coronary artery disease	Gradient boosting
Chan, Ka Lung, et al. (2018)	China	Cohort	Age, sex, smoking	Stroke	NN, SVM, NB
Chen, Jian, et al. (2019)	China	Observational (EHR)	Environmental pollutants	Stroke	RF, DT, XGB, SVM, LR, KNN
Gan, Xiu-min, et al. (2011)	China	Case control	Age, alcohol intake, BMI, education, other body metrics, physical activities, diet, smoking	Stroke	DT
Hsiao, Han CW, et al. (2016)	China	Observational (EHR)	Age, gender, area level social determinants	All types of CVD	DL
Hu, Danqing, et al. (2016)	China	Observational (EHR))	Age, smoking	Other CVD	RF, SVM, NB, lasso, other*
Huang, Zhengxing, et al. (2015)	China	Observational (EHR)	Age, gender, smoking	Coronary artery disease	SVM, other
Huang, Zhengxing, et al. (2019)	China	Observational (EHR)	Age, gender, smoking	Other CVD	NN, DL, other
Shao, Zeguo, et al. (2019)	China	Observational (survey)	Age, alcohol intake, BMI, diet, physical activities, smoking	Stroke	RF, DT
Wan, Eric Yuk Fai, et al. (2017)	China	Retrospective cohort	Age, BMI, gender, smoking	All types of CVD	DT
Xu, Yuan, et al. (2019)	China	Retrospective cohort	Age, alcohol intake, gender, smoking	Rehospitalization	Gradient boosting
Karaolis, Minas A., et al. (2010)	Cyprus	Observational (EHR)	Age, gender, smoking	Myocardial infarction	DT

Anselmino, Matteo, et al. (2009)	Europe	Cross-sectional	Age, BMI, gender, smoking, other body metrics	Stroke, myocardial infarction	NN
Deguen, Séverine, et al. (2010)	France	Cross-sectional	Age, education, gender, income, occupation, residence, area-level social determinants	Coronary artery disease	Other
Baars, Theodor, et al. (2020)	Germany	Cohort	Age, BMI, gender, smoking	Mortality	Other ensemble**
Exarchos, Konstantinos P., et al. (2015)	Greece	Observational (EHR)	Age, gender, smoking	Other	NB
Tsipouras, Markos G., et al. (2008)	Greece	Observational (EHR)	Age, BMI, gender, smoking, other body metrics	Coronary artery disease	NN, DT
Jabbar, M. A, et al. (2014)	India	Observational (EHR)	Age, gender, residence	All types of CVD	DT, PCA
Naushad, Shaik Mohammad, et al. (2018)	India	Case control	Age, alcohol intake, BMI, diet gender, smoking	Coronary artery disease	Other ensemble, other
Amin, Syed Umar, et al. (2013)	India	Observational (survey)	Age, alcohol intake, diet, gender, physical activities, smoking	All types of CVD	NN, other
Afarideh, Mohsen, et al. (2016)	Iran	Open cohort	Age, BMI, gender, smoking, other body metrics,	All types of CVD	NN
Amini, Leila, et al. (2013)	Iran	Observational (survey)	Age, alcohol intake, BMI, gender, physical activities, smoking	Stroke	DT, other
Ayatollahi, Haleh, et al. (2019)	Iran	Observational (EHR)	Age, gender, marital status, occupation, residence, smoking	Coronary artery disease	NN, SVM
Parizadeh, Donna, et al. (2017)	Iran	Cohort	Age, gender, smoking, other body metrics,	Stroke	DT
Shakerkhatibi, M., et al. (2015)	Iran	Case–crossover design	Age, gender, area-level social determinants	Hospital admission	NN
Berchiolla, Paola, et al. (2012)	Italy	Cohort	Age, smoking	Myocardial infarction, mortality	RF, NN, SVM, BN

Bigi, Riccardo, et al. (2005)	Italy	Cohort	Age, gender, smoking	Myocardial infarction, mortality	NN, other
Foltran, Francesca, et al. (2011)	Italy	Observational (EHR)	Age, BMI, gender, smoking, other body metrics	Coronary artery disease	BN
Pasanisi, Stefania, et al. (2018)	Italy	Cohort	Age, gender, smoking	All types of CVD	NN
Cho, In Jeong et al. (2020)	Korea	Cohort	BMI, physical activities, smoking	All types of CVD	DL
Hae, Hyeonyong, et al. (2018)	Korea	Retrospective Cohort	Age, BMI, gender, smoking	Coronary artery disease	RF, DT, GB, SVM, NB, Ridge, other
Kwon, Joon-myung, et al. (2019)	Korea	Retrospective Cohort	Age, BMI, gender, smoking	Mortality	RF, DL
Juarez-Orozco, Luis Eduardo, et al. (2020)	Netherlands	Observational (EHR)	Gender, BMI, smoking	Coronary artery disease, other	ensemble
Tay, Darwin, et al. (2015)	Singapore	Cohort	Age, diet, gender, physical activities, built environment	All types of CVD	NN, SVM
Fuster-Parra, Pilar, et al. (2016)	Spain	Observational (survey)	BMI, gender, physical activities, smoking, other body metrics	All types of CVD	DT, BN, NB, other
Green, Michael, et al. (2006)	Sweden	Observational (EHR)	Age, gender, smoking	Acute coronary syndrome	NN
Marshall, Adele H., et al. (2010)	UK	Cohort	BMI, Smoking	Coronary artery disease, mortality	BN
Alaa, Ahmed M., et al. (2019)	UK	Cohort	BMI, diet, physical activities, residence, smoking	All types of CVD	RF, NN, ensemble, GB, Adaboost
Harrison, Robert F., et al. (2005)	UK	Observational (EHR)	Gender, smoking	Acute coronary syndrome	NN
He, Xi, et al. (2020)	UK	Observational (survey)	Gender, smoking	Coronary artery disease	PCA, other
Ayala Solares, Jose Roberto, et al. (2019)	UK	Observational (EHR)	Gender, income, smoking, area-level social determinants	All types of CVD	BN
Yang, Hui, et al. (2015)	UK	Observational (EHR)	BMI, smoking	Coronary artery disease	NB, other

Ahmad, Tariq, et al. (2018)	USA	Registry	Age, alcohol intake, BMI, education, gender, income, marital status	Heart failure	RF, other
Akay, Metin., et al. (1992)	USA	Cross-sectional	Age, BMI, gender, smoking	Coronary artery disease	NN, other
Ambale-Venkatesh, Bharath, et al. (2017)	USA	Cohort	Age, alcohol intake, BMI, education, gender, income, race, smoking, other body metrics	Stroke, all types of CVD, heart failure, coronary artery disease, mortality	RF, lasso
Basu, Sanjay, et al. (2017)	USA	Cohort	Age, gender, race, smoking	Stroke, heart failure, myocardial infarction, mortality	Lasso
Dinh, An, et al. (2019)	USA	Cross-sectional	Age, alcohol intake, BMI, gender, income, physical activities, race	All types of CVD	RF, GB, ensemble, SVM
Dogan, Meeshanthini V., et al. (2018)	USA	Cohort	Age, alcohol intake, BMI, gender, physical activities, smoking	Stroke	RF, ensemble
Edwards, Dorothy F., et al. (1999)	USA	Observational (EHR)	Age, gender, race	Mortality	NN
Golas, Sara Bersche, et al. (2018)	USA	Observational (EHR)	Education, gender, marital status, occupation, race	Rehospitalization	Deep unified networks, GB
Gonzales, Tina K., et al. (2017)	USA	Cohort	Alcohol intake, BMI, gender, income, physical activities, smoking	Myocardial infarction	RF, other
Hsich, Eileen M., et al. (2019)	USA	Observational (survey)	Age, BMI, medical insurance, race, smoking	Mortality	RF
Hu, Danqing, et al. (2016)	USA	Clinical trial	Age, BMI, gender, income, race, residence,	Carotid atherosclerosis	RF, NB, other
Imran, Tasnim F., et al. (2018)	USA	Observational (EHR)	Age, BMI, gender, race, smoking	Stroke	Lasso
Kerut, Edmund Kenneth, et al. (2019)	USA	Observational (survey)	Gender, race, smoking	Abdominal aortic aneurysm	NN
Kogan, Emily, et al. (2020)	USA	Observational (EHR)	Gender, residence, area-level social determinants	Stroke	RF, NN, GB

Leach, Heather J., et al. (2016)	USA	Cohort	BMI, diet, income, physical activities, smoking, built environment	All types of CVD	DT
Akay, Metin, et al. (1994)	USA	Observational (EHR)	Gender, smoking	Coronary artery disease	NN
Mansoor, Hend, et al. (2017)	USA	Cohort	Alcohol intake, income, medical insurance, race, smoking, substance abuse, area-level social determinants	Mortality	RF
McGeachie, Michael, et al. (2009)	USA	Cohort	Age, BMI, education, gender, smoking	Coronary artery disease	BN
Mobley, Bert A, et al. (1995)	USA	Observational (EHR)	Gender, medical insurance, race	Length of stay in hospital	NN
Motwani, Manish, et al. (2017)	USA	Cohort	Age, BMI, race, smoking	Mortality	Other ensemble
Ni, Yizhao, et al. (2018)	USA	Cohort	Alcohol intake, gender, marital status, occupation, race, smoking, substance abuse	Stroke	RF, NN, SVM
Ottenbacher, Kenneth J., et al. (2001)	USA	Retrospective Cohort	Age, gender, marital status, medical insurance, occupation, residence	Rehospitalization	NN
Rasmy, Laila, et al. (2018)	USA	Observational (EHR)	Age, gender, race	Heart failure	DL, ridge, lasso
Baldassarre, Damiano, et al. (2004)	Italy	Cross-sectional	Age, BMI, gender, smoking	All types of CVD	NN, other
Bandyopadhyay, Sunayan, et al. (2015)	USA	Observational (EHR)	Age, BMI, gender, smoking	All types of CVD	BN
Beunza, Juan-Jose, et al. (2019)	Spain	Cohort	Age, BMI, education, gender, smoking	Coronary artery disease	RF, NN, DT, AdaBoost, SVM
Biesbroek, Sander, et al. (2015)	Netherlands	Cohort	Age, alcohol intake, diet, education, gender, physical activities, , smoking	Stroke, Coronary artery disease	RF, DT, PCA, other
Brisimi, Theodora S., et al. (2018)	USA	Observational (EHR)	Age, gender, race, smoking, area-level social determinants	Hospitalization	RF, SVM

Çelik, Güner, et al. (2014)	Turkey	Observational (EHR)	Age, gender, smoking	Stroke	NN, other
Cheon, Songhee, et al. (2019)	Korea	Observational (survey)	Age, gender, medical insurance, area-level social determinants	Stroke	RF, Adaboost, SVM, DL, PCA, NB, other
Corsetti, James P., et al. (2016)	USA	Observational (EHR)	BMI, race	All types of CVD	BN
Cox Jr, Louis Anthony Tony. (2017)	USA	Observational (survey)	Age, education, gender, income, marital status, smoking, area-level social determinants	Stroke	RF, DT, BN
Cox Jr, Louis Anthony Tony. (2018)	USA	Observational (survey)	Age, education, income, gender, marital status, smoking, area-level social determinants	All CVD	BN
Daghistani, Tahani A., et al. (2019)	Saudi Arabia	Observational (EHR)	Age, gender, medical insurance, smoking	Length of stay	RF, NN, SVM, BN
Dai, Wuyang, et al. (2015)	USA	Observational (EHR)	Age, gender, race, smoking, area-level social determinants	Hospitalization	AdaBoost, SVM, NB, other
Dogan, Meeshanthini V., et al. 2018	USA	Cohort	Age, gender, smoking	Coronary artery disease	RF
Li, Yan, et al. (2019)	USA	Observational (survey)	Age, alcohol intake, education, gender, income, medical insurance, physical activities, race, smoking, area-level social determinants	Stroke, coronary artery disease	RF
Li, Xuemeng, et al. (2019)	China	Observational (survey)	Age, alcohol intake, BMI, gender, physical activities, smoking	Stroke	RF, NN, DT, other ensemble, BN, NB
Karaolis, M., et al. (2008)	Cyprus	Observational (EHR)	Age, gender, smoking	Coronary artery disease	DT
Raihan, M., et al. (2019)	Bangladesh	Observational (EHR)	Age, gender, physical activities, smoking, substance abuse,	Coronary artery disease	NN
Martínez-García, M., et al. (2018)	Mexico	Retrospective Cohort	Age, alcohol intake, BMI, education, gender, income, marital status, medical insurance, smoking	Myocardial infarction, rehospitalization	SVM
Miller, C. S., et al. (2014)	USA	Cross-sectional case-controlled	Age, BMI, gender, race, smoking	Myocardial infarction	DT

Eswaran, Chikkannan, et al. (2012)	Malaysia	Cross-sectional	Age, BMI, gender, smoking	Myocardial infarction	NN, BN, other
Ross, Elsie Gyang, et al. (2016)	USA	Prospective observational	Age, alcohol intake, BMI, education, income, marital status, occupation, physical activities, race, residence, smoking	Mortality, other	RF, ridge
Saito, Hiroshi, et al. (2019)	Japan	Observational (survey)	Age, gender, occupation, residence, smoking	Rehospitalization	Lasso
Van Loo, Hanna M., et al. (2014)	Netherlands	Observational study	Age, BMI, gender, smoking	Mortality	Lasso
Vedomské, Michael A. et al. (2013)	USA	Observational (EHR)	Age, gender, medical insurance, race	Rehospitalization	RF
Ayyagari, Rajeev, et al. (2014)	USA	Retrospective Cohort	Age, gender, race, smoking	Stroke	Lasso
Vistisen, Dorte, et al. (2016)	Denmark	Cohort	Age, alcohol intake, BMI, gender, physical activities, smoking	Stroke	RF, DT
Wu, Yafei, et al. (2020)	China	Prospective observational	Age, alcohol intake, gender, smoking	Stroke	RF, SVM, lasso
Yu, Shipeng, et al. (2015)	USA	Retrospective Cohort	Age, gender, marital status, race	Rehospitalization	SVM, lasso
Zhuang, Xiaodong, et al. (2018)	China	Observational (survey)	Age, BMI, gender, income, race, smoking, area-level social determinants	All types of CVD	RF

Abbreviations used: NN: neural network, RF: random forest, SVM: support vector machine, DT: decision tree, NB: naïve Bayes, BN: Bayesian network, Lasso: lasso regression, Ridge: ridge regression, QDA: quadratic discriminant analysis

*“Other” algorithms used include: multilayer perceptron, maximum entropy, adversarial network, k-nearest neighbors, recursive partitioning, clustering, quadratic discriminant, RBF

**Other ensemble: methods other than: Adaboost, Gradient boosting