American Journal of
Preventive Medicine

REVIEW ARTICLE

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

Yuan Zhao, MPH,[1] Erica P. Wood, MPH,[2] Nicholas Mirin, BA,[2] Stephanie H. Cook, DrPH,[2,3] Rumi Chunara, SM, PhD[3,4]

Introduction: Cardiovascular disease is the leading cause of death worldwide, and cardiovascular disease burden is increasing in low-resource settings and for lower socioeconomic groups. Machine learning algorithms are being developed rapidly and incorporated into clinical practice for cardiovascular disease prediction and treatment decisions. Significant opportunities for reducing death and disability from cardiovascular disease worldwide lie with accounting for the social determinants of cardiovascular outcomes. This study reviews how social determinants of health are being included in machine learning algorithms to inform best practices for the development of algorithms that account for social determinants.

Methods: A systematic review using 5 databases was conducted in 2020. English language articles from any location published from inception to April 10, 2020, which reported on the use of machine learning for cardiovascular disease prediction that incorporated social determinants of health, were included.

Results: Most studies that compared machine learning algorithms and regression showed increased performance of machine learning, and most studies that compared performance with or without social determinants of health showed increased performance with them. The most frequently included social determinants of health variables were gender, race/ethnicity, marital status, occupation, and income. Studies were largely from North America, Europe, and China, limiting the diversity of the included populations and variance in social determinants of health.

Discussion: Given their flexibility, machine learning approaches may provide an opportunity to incorporate the complex nature of social determinants of health. The limited variety of sources and data in the reviewed studies emphasize that there is an opportunity to include more social determinants of health variables, especially environmental ones, that are known to impact cardiovascular disease risk and that recording such data in electronic databases will enable their use.

Am J Prev Med 2021;61(4):596–605.

INTRODUCTION

An estimated 17.9 million people die each year from cardiovascular disease (CVD), which represents 31% of all deaths worldwide.[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 incidence of disease and higher mortality.[1,2] In high-income countries such as the U.S., the prevalence of CVD is expected to rise by 10% between 2010 and

From the [1]Department of Epidemiology, NYU School of Global Public Health, New York University, New York, New York; [2]Department of Social and Behavioral Sciences, NYU School of Global Public Health, New York University, New York, New York; [3]Department of Biostatistics, NYU School of Global Public Health, New York University, New York, New York; and [4]Department of Computer Science and Engineering, NYU Tandon School of Engineering, New York University, Brooklyn, New York

Address correspondence to: Rumi Chunara, SM, PhD, Department of Computer Science & Engineering, Tandon School of Engineering, New York University, 370 Jay Street, 1106, Brooklyn NY 11201. E-mail: rumi.chunara@nyu.edu.

0749-3797/$36.00
https://doi.org/10.1016/j.amepre.2021.04.016

2030,[3] attributed not only to an aging population but also to shifts in societal and environmental conditions distributed unequally among groups. These shifts have led to changes in diet and physical activity resulting in a dramatic rise in conditions such as obesity, hypertension, diabetes mellitus, and physical inactivity. In the absence of large genetic changes (biologically infeasible over only a decade or 2), CVD risk has increased in China owing to an increase in cholesterol, blood pressure, smoking, and physical inactivity.[4] Other research shows increases in deaths due to CVD in immigrants to the U.S.,[5] Indians living in the United Kingdom,[6] and migrant twins,[7] illuminating the critical contribution of social and environmental factors in CVD risk. In summary, findings show that the most significant opportunities for reducing death and disability from CVD lie in addressing the social determinants of cardiovascular outcomes.[8,9]

The WHO defines *social determinants of health (SDH)* as "the conditions in which people are born, grow, live, work and age,"[10] which are shaped by the distribution of resources at global, national, and local levels. Figure 1 (adapted from Lund et al.) shows that the theoretic framework of SDH from different domains can work multidimensionally on health outcomes. In addition to limiting access to CVD care and treatment, substantial recent literature has also shown how social factors also exert independent influence over cardiovascular health.[11] Multinational, prospective cohort studies as well as ecologic analyses have shown that SDH contribute to >35% of the population-attributable risk of various CVDs,[12,13] among which education, income, and occupation are particularly influential.[11] Given the foundation of research undergirding the critical importance of SDH as a driver of differential disease risk, it is clear that modeling methods that incorporate such factors, including capturing the interaction and relative influence of such factors in relation to other physiologic CVD risk factors, are needed.[14]

Artificial intelligence and machine learning (an application of artificial intelligence for detecting patterns from data)[15] tools are seeing rapid adoption in clinical research, particularly given the proliferation of electronic health records and advanced computing strategies. These approaches have been shown to improve the prediction of CVD risk, incidence, and outcomes[16–18] over traditional risk scores such as those from the American College of Cardiology or the American Heart Association.[19] As a data-driven approach, machine learning may make fewer assumptions and provide more flexibility.[20] This is particularly advantageous when considering SDH, given the feedback mechanisms, complex mediation, and interactions involved in these variables and their action on CVD. However, machine learning

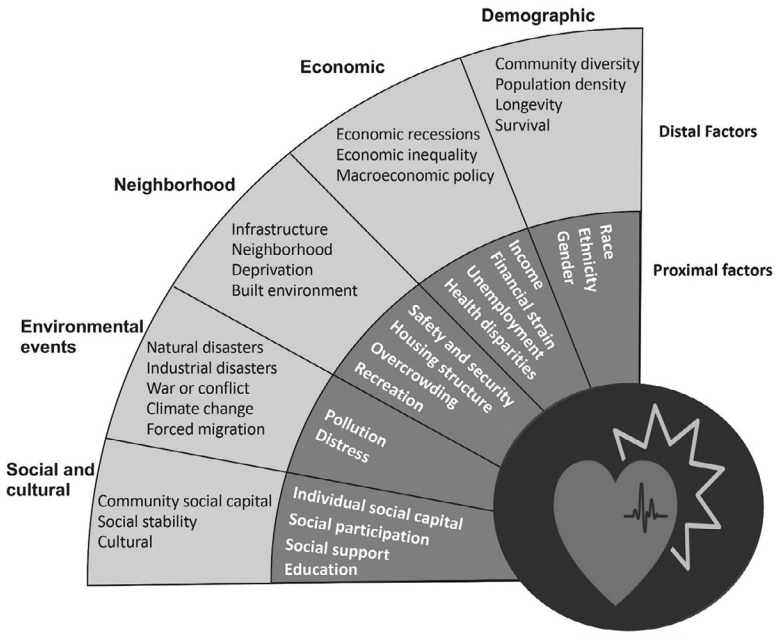

Figure 1. Conceptual framework of SDH, based on a socioecologic framework approach to outline the multidimensional factors that shape health outcomes.

SDH, social determinants of health.

Source: Figure adapted from Lund et al.[59]

models deployed in clinical settings have not typically accounted for SDH. Thus, although the increased flexibility and precision of machine learning models are appealing, given the rapid rise of machine learning approaches and the few studies that do incorporate SDH in machine learning models, a better understanding of how they should be incorporated into algorithms is needed to improve early prediction of CVD and reduce their significant disease burden.[21,22]

Given the direct relevance of SDH in CVD and the significant promise of machine learning for better CVD risk prediction, a systematic review to understand the machine learning methods by which SDH have been incorporated in CVD prediction models, which machine learning algorithms are being used in model development, and for which populations was performed. This review serves to inform the best practices for the design of machine learning approaches and to identify spaces for methodologic innovation in incorporating all relevant SDH into machine learning models for CVD.

METHODS

Search Strategy and Selection Criteria

With the help of an expert librarian, YZ performed a comprehensive search of 5 databases: PubMed, Embase, Web of Science, IEEE Xplore, and ACM Digital Library on April 10, 2020 to identify all relevant articles on machine learning integrating SDH in CVD prediction models published in English from inception to the search date. IEEE Xplore and ACM Digital Library were included to comprehensively capture computer science articles related to this review. Only peer-reviewed articles published in journals or accepted in conferences were included, and nonpeer-reviewed gray literature or *arXiv/medRxiv* papers were excluded.

Terms representing SDH were derived using the broader definition from the WHO and Centers for Disease Control and Prevention *Healthy People 2020* initiative, which delineates SDH in 5 areas: economic stability, education, social and community context, health and health care, and neighborhood and built environment. For each area, keywords were identified by referencing previous review papers on SDH and CVD[11,23] and related studies of different SDH[23–27] or by consulting experts. Full search strategies are described in the Appendix (available online).

For search terms related to machine learning, all commonly used supervised machine learning methods were included. Supervised machine learning algorithms are those that perform reasoning (i.e., prediction) from observations of the features based on externally supplied examples that include the features linked to outcome labels (e.g., CVD outcomes). Thus, supervised machine learning was a focus because the types of tasks considered usually utilized labeled outcomes of CVD.[28] Commonly used unsupervised machine learning algorithms captured in the search were also included in the abstract and full-text screening to ensure that all types of possible studies were included. Search terms to capture Deep Learning and Ensemble methods, as they are widely used in current clinical research, were added.[29] For prediction outcomes, out of ischemic, cerebrovascular, carditis, and rheumatic CVD outcomes, the focus was narrowed to cardiovascular ischemic outcomes, coronary heart disease, and cerebrovascular disease, which are caused by *atherosclerotic CVD*, defined as plaque buildup in arterial walls, because these are the highest causes of mortality and because estimated years of lives lost attributed to these have increased recently.[30]

All study designs and populations were included if the article utilized any SDH as features in the machine learning models (in addition to age or gender, which are commonly included as standard practice and not specifically to represent their contribution as SDH) and if the outcomes were CVD related, including incidence, survival, mortality, and hospital admission and readmission. Time of publication was not restricted. Studies were excluded if they did not use any machine learning algorithm; if they were developed for nonhumans; if outcomes were biomarkers, mediators, surgery or medication of CVD, rehabilitation or mental health outcomes after CVD, or cost-effectiveness analysis of CVD; or if the manuscript was non-English or was a review or meta-analysis. Articles presented at conferences as abstracts and for which the full text was not obtainable were excluded. This review, conducted in 2020, was registered with the International Prospective Register of Systematic Reviews (CRD42020175466), and the PRISMA method was followed. To supplement the bibliographic database searches, Google Scholar was used to scrutinize all keywords regarding their relevance in articles as well as examine potential article eligibility. Duplicates were removed in the process.

A total of 3 investigators (YZ, NM, and EPW) screened the title and abstract; each retrieved article was independently assessed by 2 reviewers to determine whether it was an eligible full-text article. Conflicts were resolved by discussion and validation from a third reviewer. After initial appraisal, full texts and information of eligible articles were retrieved.

Data Analysis

Data were extracted from individual articles independently by 2 reviewers (among YZ, NM, and EPW) and checked by the third reviewer according to the standardized extraction form. All data extraction was cross-checked, and disagreements were resolved by discussion or referral to the third reviewer. Extracted information included the year of publication, country, population, SDH, machine learning algorithms, CVD outcomes, data source, and performance of the algorithms. Several criteria to assess the quality of the study based on best practices in machine learning[31] were defined, including (1) whether machine learning model performance was evaluated, (2) whether a hyperparameter (parameters to control the learning process; for example, number of hidden layers in a Deep Neural Network) tuning process was described, (3) whether data-driven variable selection was performed, and (4) whether model results were interpreted. Each item was scored as no (not present), unclear, or yes (present), and all items were summarized in a quality score. Bias was investigated for each study by assessing whether an external validity evaluation was performed. Use of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis checklist was assessed.

RESULTS

Database search identified 1,655 distinct articles; after a full-text review of 178 articles, 48 were included in the review (Figure 2). All included studies used data

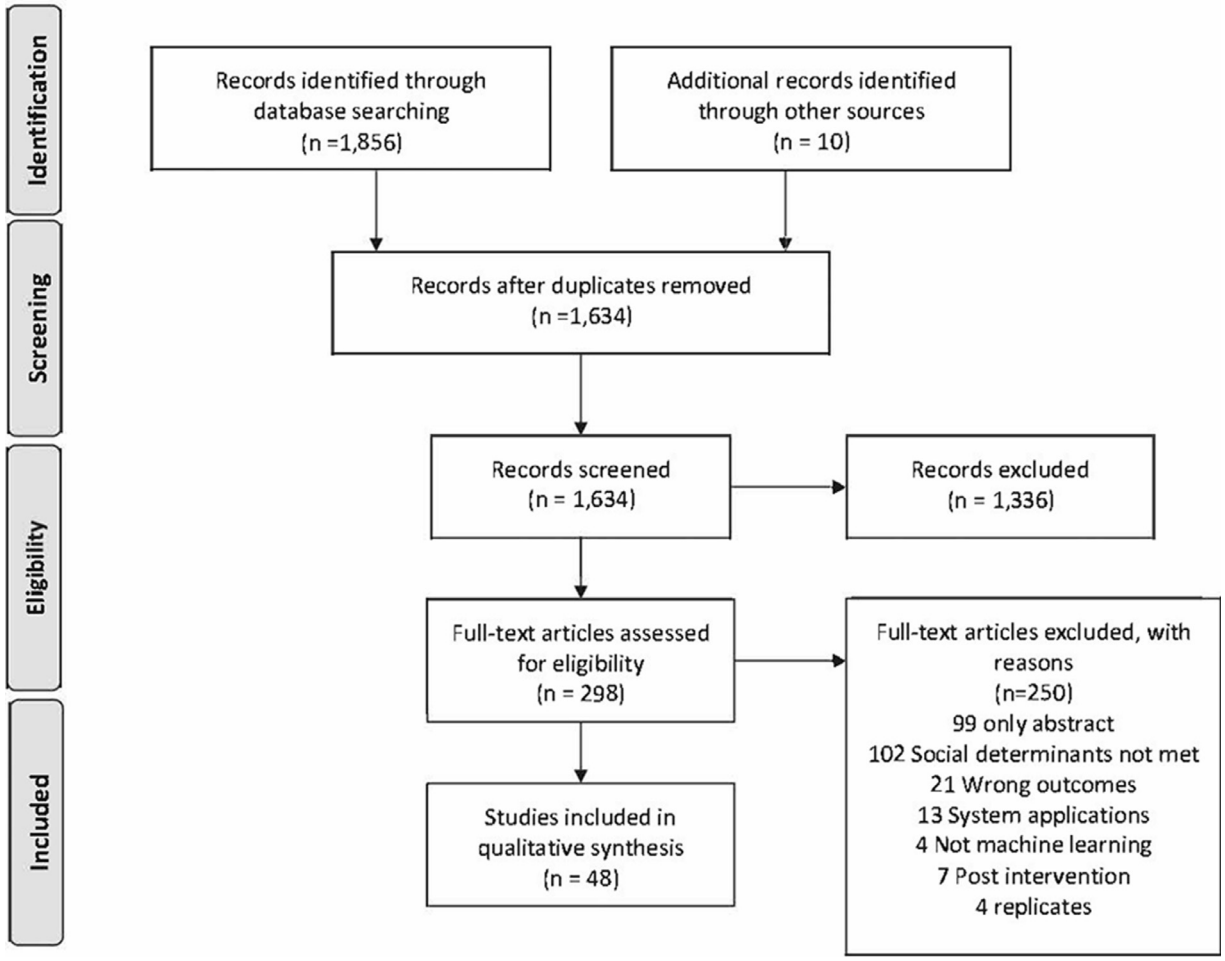

Figure 2. Flowchart of the study review process that included 48 studies from 1,887 articles in the initial database search (April 2020). Each abstract and full-text article was screened by 2 reviewers independently, and when consensus could not be reached, a third reviewer assessed these articles and decided whether they should be included or excluded.

collected in an observational manner instead of data from an experiment in which treatments or interventions were randomized. The most frequent study design was cohort (20 studies), followed by data extracted from electronic medical records (17 studies), cross-sectional studies, or surveys (11 studies). Most data used were structured, although 9 studies included unstructured data (e.g., electrocardiogram, image, and heart sound). The earliest year of publication was 1995 (artificial neural network algorithm) (Appendix Figure 1, available online). Almost half (23 studies) of the articles were published by authors from the U.S., with others mostly from Europe (11 studies) and China (5 studies) (Figure 3A).

The 10 most common algorithms were Bayesian Network; Decision Tree; AdaBoost; Gradient Boosting and other Ensemble methods; Naive Bayes; Neural Net, including Artificial Neural Networks; Convolutional Neural Networks; Recurrent Neural Networks and Deep Learning; Ridge/Lasso/Elastic Net Regularization; Random Forest; and Support Vector Machine. A total of 3 studies used unsupervised machine learning algorithms such as clustering to group CVD risk levels or principal component analysis to extract features before supervised machine learning classification.[14,32,33] Median sample size was 2,510; more than two thirds of the studies had sample size >1,000, and 13 of those had >10,000. There was a sample size <100 in 5 studies, which mostly used a Bayesian Network method. Neural Net, Random Forest, and Decision Tree studies often have larger sample sizes than Support Vector Machine and Naive Bayes studies (Table 1). The overall study characteristics are summarized in Table 1 and Appendix Figure 1 (available online) and discussed further in the following paragraph.

Most studies included demographic variables from routine clinical practice or survey questions, whereas 10 specifically studied the association of SDH with CVD outcomes or CVD risk factors as potential intervention

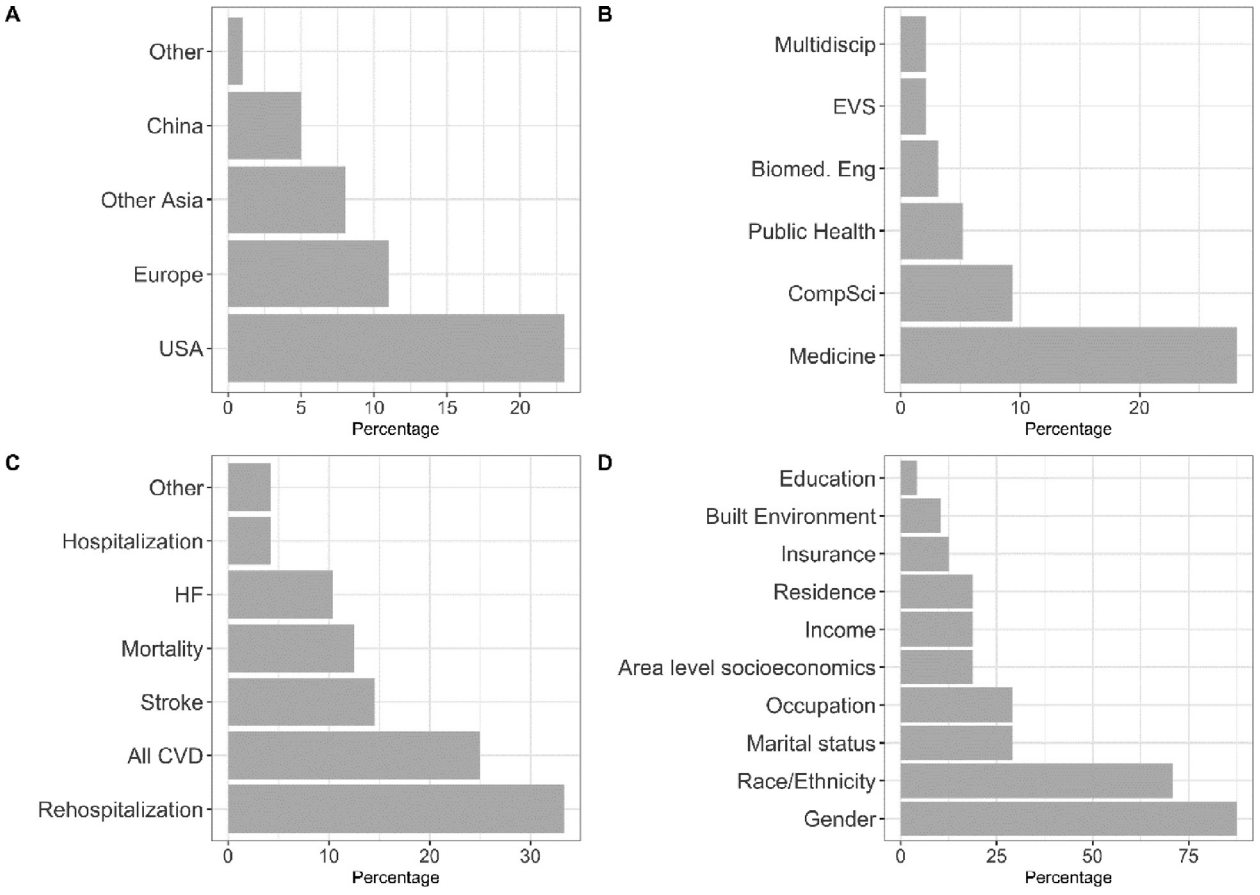

Figure 3. Descriptive summaries of included papers. (A) Countries of corresponding authors, (B) journal types of the publications reported in systematic review papers (EVS) with respect to the percentage of included papers they appear in, (C) most frequently reported cardiovascular outcomes (HF), and (D) the top 10 social determinants of health. Area-level socioeconomics is typically defined by an index to represent SES in a certain area, for example, ZIP code level or census tract level, and *built environment* refers to measures of human-made surroundings, such as parks or green space.

Biomed. Eng, biomedical engineering; CompSci, computer sciences; CVD, cardiovascular disease; EVS, environmental sciences; HF, heart failure; Multidiscip, multidisciplinary.

Table 1. Summary of Machine Learning Algorithms, Best Performing Algorithms and Sample Sizes Used in the Studies

Algorithm	Number of papers[a] (%)	Number as best algorithm when multiple algorithms are used	Sample size			
			<100	100–1,000	1,000–10,000	>10,000
Neural Net	14 (29.2)	3	1	2	9	2
Random Forest	14 (29.2)	6	0	0	8	6
Support Vector Machine	12 (25.0)	4	1	1	7	3
Decision tree	8 (18.8)	1	1	3	3	1
Ensemble	8 (18.8)	2	0	1	3	4
Regularization[b]	7 (14.6)	1	0	1	4	2
Bayesian Network	5 (10.4)	1	1	2	1	1
Naive Bayes	5 (10.4)	0	2	0	2	1
Other	14 (29.2)	2	1	2	7	4

[a]Note that each paper could include multiple versions or multiple algorithms.
[b]Regularization included Lasso, Ridge, and Elastic Net.

targets. More than half of the studies used data from CVD-related patient visits (25 studies), and around a third used information from general populations (14 studies); the rest used information from hospital visits unrelated to CVD issues (9 studies). Studies were published on a diverse range of journal types, including medicine, computer science, environmental science, and public health journals (Figure 3B). Studies described outcomes, including hospitalization (16 studies), stroke (7 studies), CVD-related mortality (6 studies), heart failure (5 studies), as well as others (e.g., transient ischemic attack) (Figure 3C).

Studies included diverse SDH variables such as gender; race/ethnicity; education; marital status; occupation/employment; individual or household income; medical insurance; area of residence (e.g., urban versus rural or eastern versus western U.S.); and other community-level factors of deprivation, income, and education and environmental pollutants. Figure 3D illustrates the top 10 SDH considered in the extracted papers and their frequency. In most studies, demographic information such as gender and race were included as standard variables collected in a survey or electronic health record, and the terms gender and sex were used interchangeably, therefore inhibiting the capturing of the biological and social aspects of sex and gender and their interaction with other SDH separately. More than half of the studies reported feature importance of SDH (25 studies), in which gender, ethnicity, and environmental pollutants were most frequently reported to contribute significantly to CVD outcome prediction (Figure 3D). Other frequently reported determinants were social isolation and health insurance. A total of 3 studies compared model performance with and without social determinants, all of which showed that SDH significantly improved prediction. A total of 2 studies showed improved prediction by adding gender and race.[34,35] The study that showed decreased performance aimed to forecast the pattern of the demand for hemorrhagic stroke healthcare services on the basis of air quality; it is possible that the relationship between the specific variable tested and the outcome has little direct relationship.[36]

In terms of algorithm development, of the 14 studies using Neural Networks (the most common algorithm), 8 used multiple hidden layers, including most commonly 3-layer perceptron, Convolutional Neural Network, and Recurrent Neural Network. The authors refer to these studies collectively as Deep Learning/Neural Networks.[37] The publication of machine learning algorithms for CVD prediction has been increasing quickly since 2015, with the wide application of Neural Networks and Random Forest (Appendix Figure 1, available online), likely owing to the availability of software for ease of implementation and the availability of computing power resources for these algorithms that may otherwise take long compute times. Of the 21 studies including multiple algorithms, Random Forest (6 studies) and Support Vector Machine (4 studies) were most frequently reported as the best performing algorithms. Of the 16 studies that compared machine learning algorithms with standard linear regression, logistic regression, or survival analysis, 15 showed improved performance with machine learning (Appendix Figure 2, available online). Only 13 variables were considered in the 1 study that showed that Deep Learning/Neural Networks had similar performance to logistic regression in predicting acute coronary syndrome.[38]

Most studies evaluated the performance of machine learning algorithm(s) developed. Area under the receiver operating characteristic curve was the most common evaluation metric (23 studies), followed by sensitivity (20 studies), specificity (15 studies), and accuracy (13 studies). At least 3 of the 4 metrics were used in 15 studies. Other evaluation metrics used included accuracy, positive predictive value, negative predictive value, and F1-score (the harmonic mean of precision and recall, commonly used to evaluate machine learning methods through a balance of these metrics). External evaluation, in which the machine learning models developed in 1 hospital was tested on another hospital or population, was performed in 5 studies.

Among those reported, most areas under the receiver operating characteristic curve were >0.70. Because most studies were published in biomedical and clinical journals, most studies explicitly interpreted the findings and their relevance to clinical applications. The mean score of included studies in the quality assessment scale (based on evaluation of machine learning, data-driven selection of features, hyperparameter tuning description, and interpretation of the model) was 3.4. One of the articles reported using the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis checklist.[39] Half of the studies (26) had full scores, and 14 studies missed 1 of the 4 items. The commonly missed items were data-driven feature selection and details of hyperparameter tuning (cross-validation or grid search strategies were utilized in 35 studies to tune hyperparameters; other studies did not give details about the hyperparameter tuning process). Half of all the studies utilized a data-driven selection method to identify features before fitting machine learning models (defined as extracting a subset of useful variables among the original variables or transforming data from a high- to a low-dimensional space).[40] Because Deep Learning learns to extract features while training, those studies sometimes do not provide feature selection details.

DISCUSSION

This systematic review provides unique insight into the use of SDH in machine learning CVD prediction models. The flexibility of machine learning models has proved useful in CVD prediction models, with their improved performance being over that of regression approaches. To date, models largely have not been constructed to explicitly and broadly examine and include SDH. Instead, studies have most frequently examined the contribution of a specific set of SDH variables, commonly those available in the electronic health record. These most frequently examined variables were individual-level instead of neighborhood- or community-level attributes (Figure 3D). Acknowledging this limited diversity of evaluated SDH, in a wide array of CVD outcomes (e.g., rehospitalization, stroke, mortality, heart failure), SDH were largely found to improve model performance. In terms of algorithms, several types of machine learning algorithms were evaluated. When multiple algorithms were compared within studies, the most flexible models such as Neural Networks and Random Forest models were the best performing. Neural Networks also most commonly outperformed regression models. This is understood because Neural Networks include hidden layers that can take into account more complex relations in the data.[41,42]

Although this is the first review that gives findings related to the growing opportunity of machine learning and SDH for CVD prediction, there are individual studies that support the components of the findings of this study. First, machine learning, in general, has shown promise with respect to CVD prediction.[43–45] Previous work has shown that machine learning algorithms (Random Forest, Logistic Regression, Gradient Boosting Machines, and Neural Networks) were better at identifying individuals who will develop CVD and those who will not, and especially for those in minority groups, than the established American College of Cardiology/American Heart Association risk calculator to predict the incidence and prognosis of atherosclerotic CVD.[19] These studies have attributed this to the fact that standard CVD risk assessment models make an implicit assumption that each risk factor is related in a linear fashion to CVD outcomes and that such models may thus oversimplify complex relationships that include large numbers of risk factors with nonlinear interactions. This review finds that prediction models that consider social variables also benefit from flexible modeling approaches.

Very flexible models also bring concerns regarding interpretability and potential overfitting to data.[46] The authors found that few included papers had an in-depth discussion on this issue. First, most models selected variables on the basis of previous clinical significance; thus, prediction performance would be based on such factors that are known to be relevant to CVD even if the specific importance of each variable was not measured. Second, papers did use methods such as automatic relevance determination[41] to examine the importance of variables in Neural Networks.

The role of SDH in CVD (not specifically machine learning related) has been examined by several studies and systematic reviews. Although full summaries of this work have been performed elsewhere,[11] the authors note that there have been several studies of various proximal and distal SDH and CVD (Figure 1). In general, studies indicate that the changing burden of disease due to societal and environmental conditions as well as the increasing advances in treatment and prevention have not been shared equally across economic, racial, and ethnic groups, imploring the need for considering a broad range of SDH in CVD prediction.[11,23] Of the studies considered in this review, the one that included the widest array of SDH (using data from the UK Biobank) showed that the most predictive SDH were gender, race/ethnicity, income, Townsend index (a measure of material deprivation of a population), and parents' ages at death.[16] In this systematic review, the most common SDH considered in machine learning models for CVD prediction, besides gender, were race/ethnicity and marital status. The mechanisms of action of SDH have been well studied; societal and environmental conditions affect diet and physical activity, which in turn have consequences for obesity (commonly quantified through body measures such as BMI). These environments can also influence behaviors such as smoking, which along with obesity are known to increase the risk of CVD through hemodynamics caused by blood clotting induced through cigarette smoke or body composition changes.[47,48] A recent review on environmental determinants of CVD studied built and social environmental factors, concluding with the need to not only systematically unweave the strands of environmental influences but also integrate the effects of the various components of the environment into a comprehensive model.[24] In this study, it is noted that the considered community-level attributes were very few and even then were not very precisely localized (e.g., hospital region). Only 5 papers considered environmental factors from the built environment such as walkability and recreation and the availability of health-promoting resources (e.g., some grocery stores and playgrounds). This is likely due to the added difficulties in capturing these data to be linked with existing large databases used in studies included in this review. Fortunately, there is a growing emphasis in recent years on using electronic health records to collect

SDH data and several screening tools that have been developed to capture this information in appropriate constructs. Numerous screening tools have been developed to capture individual-level SDH such as employment, education, food, and social support in a routine process.[49,50] This development also provides an opportunity to make more diverse SDH information available and integrated into risk prediction models.

Simultaneously, the nature of these variables and their association with disease may also be better understood through models that account for both traditional and social risk factors. Because machine learning is advantageous when considering higher numbers of predictors and the complex mediation and interactions involved in SDH variables and their action on CVD, with appropriate efforts for interpreting the effect of variables, it can also be used for prioritizing SDH in electronic health records. For instance, no studies included in this review accounted for the social processes associated with the continuity of socioeconomic conditions across the life course (e.g., conditions during childhood versus those later in life). These are relevant because SDH can have a cumulative impact over time on the development of major CVD risk factors (dyslipidemia, hypertension, and smoking), which are likely to be important in the development of CVD.[51] Finally, models that incorporated SDH and machine learning for CVD prediction also reflect limitations of many machine learning algorithms that have been highlighted recently, namely being developed on homogenous populations. For example, in the Biobank study,[16] the cohort was 94% White. The lack of diversity in studies in this review is evident when examining the locations of the corresponding authors of papers that fulfilled all inclusion criteria and were included in the review. Although the investigators did not restrict the paper selection by geographic focus, authors are located most frequently in just a few locations, with low-income countries and locations not represented at all. This is particularly striking given the high and increasing CVD burden and the changing socioenvironmental circumstances in low-income countries and regions; furthermore, the variance of SDH will be decreased with less diversity.[52]

Limitations

This review was limited in several aspects. First, included studies evaluated different types of cardiovascular outcomes; therefore, the heterogeneity of the outcome metrics makes it difficult to compare the machine learning performance across different studies. The populations in the included studies also represent samples from different data sources, hospitals, and countries, which illustrates the wide range of studies and further directions for specific studies. For example, the types of SDH relevant to populations in one geography may not be as prominent in other locations. A purposefully broad approach behind the selection of papers to identify innovative ideas and provide an overview of the state-of-art machine learning application in population health was adopted. Third, most studies did not evaluate external validity, making it inconclusive about applying the algorithms in other populations or healthcare settings. Fourth, the review was also limited to studies published in English, which might have created some bias in the articles that were ultimately retained for the analysis.

In general, this review shows that there is room to more systematically and comprehensively incorporate SDH into CVD risk prediction. This could be due to a lack of SDH data: even when collected, the measures are not comprehensive or standardized. For example, if race is conceptualized as a proxy for variables, such as SES or cultural factors, better ways to measure these social/cultural factors should be investigated. Clear identification of potential mediating and moderating factors in the relevant pathways (for example, sense of personal control or social support) will further inform model and public health intervention design. Improved constructs will also help in the incorporation of variables to represent the built and social environments that were not well represented in current studies. Inclusion of such information in the electronic health record would make it easily accessible for machine learning studies and increase sample size to reduce overfitting of models. This will also improve the performance of risk prediction because current risk scores overestimate risk, particularly for low SES and vulnerable groups.[19,53]

CONCLUSIONS

Alongside the recent growth of work on algorithmic fairness, which is broadly concerned with the statistical parity of algorithms for different groups,[54] including individual- and community-level SDH can help to better understand and disentangle where disparities are rooted, for example, if there are differences in outcomes between men and women on the basis of prediction and allocation of treatments/resources or on the basis of unequal SDH. The use of a prediction model that includes SDH should be closely linked to clinical practices. Ideally, accounting for SDH can activate risk mitigation beyond primary care by improving healthcare teams' ability to understand upstream factors impacting patients' health and the ability to act on care recommendations, by identifying patients in need of referral to community resources, and by informing the provision and funding of community resources.[19,55–57] Including SDH also

provides an opportunity to consider how structural data may also be considered, beyond individual-level attributes, in algorithmic fairness.[58] Finally, results emphasize the need for studies that include more diverse populations to improve CVD prediction in diverse settings,[52] in particular those where disease risk is increasing.

ACKNOWLEDGMENTS

The authors thank Dorice Vieira and Dr. Rajesh Vedanthan for valuable help with the search process. The authors acknowledge funding from the National Science Foundation (IIS-1845487).

The funders had no role in the decision to publish the study.

YZ contributed to study methodology, data curation, and visualization and writing (original draft preparation and reviewing and editing) of this paper. NM and EPW contributed to the study methodology and data curation and writing (reviewing and editing) of this paper. SHC contributed to the study conceptualization and methodology and writing (reviewing and editing) of this paper. RC contributed to the study conceptualization, methodology, and supervision; writing (reviewing and editing) of this paper; and the study funding acquisition.

No financial disclosures were reported by the authors of this paper.

SUPPLEMENTAL MATERIAL

Supplemental materials associated with this article can be found in the online version at https://doi.org/10.1016/j.amepre.2021.04.016.

REFERENCES

1. WHO. Cardiovascular Diseases (CVDs). Geneva, Switzerland: WHO; Published May 17, 2017. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds. Accessed March 8, 2021.
2. Deaton C, Froelicher ES, Wu LH, Ho C, Shishani K, Jaarsma T. The global burden of cardiovascular disease. *Eur J Cardiovasc Nurs.* 2011;10 (suppl 2):S5–S13. https://doi.org/10.1016/S1474-5151(11)00111-3.
3. Heidenreich PA, Albert NM, Allen LA, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. *Circ Heart Fail.* 2013;6(3):606–619. https://doi.org/10.1161/HHF.0b013e318291329a.
4. Critchley J, Liu J, Zhao D, Wei W, Capewell S. Explaining the increase in coronary heart disease mortality in Beijing between 1984 and 1999. *Circulation.* 2004;110(10):1236–1244. https://doi.org/10.1161/01.CIR.0000140668.91896.AE.
5. Worth RM, Kato H, Rhoads GG, Kagan K, Syme SL. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: mortality. *Am J Epidemiol.* 1975;102 (6):481–490. https://doi.org/10.1093/oxfordjournals.aje.a112186.
6. Patel JV, Vyas A, Cruickshank JK, et al. Impact of migration on coronary heart disease risk factors: comparison of Gujaratis in Britain and their contemporaries in villages of origin in India. *Atherosclerosis.* 2006;185 (2):297–306. https://doi.org/10.1016/j.atherosclerosis.2005.06.005.
7. Hedlund E, Kaprio J, Lange A, et al. Migration and coronary heart disease: a study of Finnish twins living in Sweden and their co-twins residing in Finland. *Scand J Public Health.* 2007;35(5):468–474. https://doi.org/10.1080/14034940701256875.
8. Levenson JW, Skerrett PJ, Gaziano JM. Reducing the global burden of cardiovascular disease: the role of risk factors. *Prev Cardiol.* 2002;5 (4):188–199. https://doi.org/10.1111/j.1520-037x.2002.00564.x.
9. Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study [published correction appears in *Lancet.* 2020;395(10226):784]. *Lancet.* 2020;395(10226):795–808. https://doi.org/10.1016/S0140-6736(19)32008-2.
10. WHO. Closing the gap in a generation: health equity through action on the social determinants of health: commission on social determinants of health. Geneva, Switzerland: WHO; Published August 27, 2008. https://www.who.int/publications/i/item/WHO-IER-CSDH-08.1. Accessed March 8, 2021.
11. Havranek EP, Mujahid MS, Barr DA, et al. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the American Heart Association. *Circulation.* 2015;132(9):873–898. https://doi.org/10.1161/CIR.0000000000000228.
12. Joseph P, Leong D, McKee M, et al. Reducing the global burden of cardiovascular disease, part 1: the epidemiology and risk factors. *Circ Res.* 2017;121(6):677–694. https://doi.org/10.1161/CIRCRESAHA.117.308903.
13. Tillmann T, Pikhart H, Peasey A, et al. Psychosocial and socioeconomic determinants of cardiovascular mortality in Eastern Europe: a multicentre prospective cohort study. *PLoS Med.* 2017;14(12): e1002459. https://doi.org/10.1371/journal.pmed.1002459.
14. He X, Matam BR, Bellary S, Ghosh G, Chattopadhyay AK. CHD risk minimization through lifestyle control: machine learning gateway. *Sci Rep.* 2020;10(1):4090. https://doi.org/10.1038/s41598-020-60786-w.
15. Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ.* 2019;364: I886. https://doi.org/10.1136/bmj.l886.
16. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One.* 2019;14(5):e0213653. https://doi.org/10.1371/journal.pone.0213653.
17. Dimopoulos AC, Nikolaidou M, Caballero FF, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Med Res Methodol.* 2018;18(1):179. https://doi.org/10.1186/s12874-018-0644-1.
18. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *J Am Heart Assoc.* 2018;7(22):e009476. https://doi.org/10.1161/JAHA.118.009476.
19. Cook NR, Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA Intern Med.* 2014;174 (12):1964–1971. https://doi.org/10.1001/jamainternmed.2014.5336.
20. Rose S. Intersections of machine learning and epidemiological methods for health services research. *Int J Epidemiol.* 2021;49(6):1763–1770. https://doi.org/10.1093/ije/dyaa035.
21. Caballero FF, Soulis G, Engchuan W, et al. Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the ATHLOS project. *Sci Rep.* 2017;7(1):43955. https://doi.org/10.1038/srep43955.
22. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM Popul Health.* 2018;4:95–99. https://doi.org/10.1016/j.ssmph.2017.11.008.
23. Kreatsoulas C, Anand SS. The impact of social determinants on cardiovascular disease. *Can J Cardiol.* 2010;26(suppl C):8C–13C. https://doi.org/10.1016/s0828-282x(10)71075-8.
24. Bhatnagar A. Environmental determinants of cardiovascular disease. *Circ Res.* 2017;121(2):162–180. https://doi.org/10.1161/CIRCRESAHA.117.306458.
25. Cheng I, Ho WE, Woo BK, Tsiang JT. Correlations between health insurance status and risk factors for cardiovascular disease in the elderly Asian American population. *Cureus.* 2018;10(3):e2303. https://doi.org/10.7759/cureus.2303.
26. Fang J, Yuan K, Gindi RM, Ward BW, Ayala C, Loustalot F. Association of birthplace and coronary heart disease and stroke among U.S.

adults: National Health Interview Survey, 2006 to 2014. *J Am Heart Assoc.* 2018;7(7):e008153. https://doi.org/10.1161/JAHA.117.008153.
27. Lapane KL, Lasater TM, Allan C, Carleton RA. Religion and cardiovascular disease risk. *J Relig Health.* 1997;36(2):155–164. https://doi.org/10.1023/A:1027444621177.
28. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. *Emerg Artif Intell Appl Comput Eng.* 2007;160(1):3–24. https://http://www.informatica.si/index.php/informatica/article/viewFile/148/140. Accessed March 8, 2021.
29. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–444. https://doi.org/10.1038/nature14539.
30. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017 [published correction appears in *Lancet.* 2019;393(10190):e44]. *Lancet.* 2018;392(10159):1736–1788. https://doi.org/10.1016/S0140-6736(18)32203-7.
31. Chen PC, Liu Y, Peng L. How to develop machine learning models for healthcare. *Nat Mater.* 2019;18(5):410–414. https://doi.org/10.1038/s41563-019-0345-0.
32. Cheon S, Kim J, Lim J. The use of deep learning to predict stroke patient mortality. *Int J Environ Res Public Health.* 2019;16(11):1876. https://doi.org/10.3390/ijerph16111876.
33. Jabbar M, Deekshatulu B, Chndra P. Alternating decision trees for early diagnosis of heart disease. In: Paper presented at: International Conference on Circuits, Communication, Control and Computing; November 21–22, 2014. https://doi.org/10.1109/CIMCA.2014.7057816.
34. McGeachie M, Ramoni RLB, Mychaleckyj JC, et al. Integrative predictive model of coronary artery calcification in arteriosclerosis. *Circulation.* 2009;120(24):2448–2454. https://doi.org/10.1161/CIRCULATIONAHA.109.865501.
35. Rasmy L, Wu Y, Wang N, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform.* 2018;84:11–16. https://doi.org/10.1016/j.jbi.2018.06.011.
36. Chen J, Li H, Luo L, et al. Machine learning-based forecast of hemorrhagic stroke healthcare service demand considering air pollution. *J Healthc Eng.* 2019;2019:7463242. https://doi.org/10.1155/2019/7463242.
37. Illing B, Gerstner W, Brea J. Biologically plausible deep learning—but how far can we go with shallow networks? *Neural Netw.* 2019;118:90–101. https://doi.org/10.1016/j.neunet.2019.06.001.
38. Harrison RF, Kennedy RL. Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. *Ann Emerg Med.* 2005;46(5):431–439. https://doi.org/10.1016/j.annemergmed.2004.09.012.
39. Hae H, Kang SJ, Kim WJ, et al. Machine learning assessment of myocardial ischemia using angiography: development and retrospective validation. *PLoS Med.* 2018;15(11):e1002693. https://doi.org/10.1371/journal.pmed.1002693.
40. Chu C, Hsu AL, Chou KH, Bandettini P, Lin C, Alzheimer's Disease Neuroimaging Initiative. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage.* 2012;60(1):59–70. https://doi.org/10.1016/j.neuroimage.2011.11.066.
41. Bishop CM. *Bayesian methods for Neural Networks*. Birmingham, United Kingdom: Neural Computing Research Group, Department of Computer Engineering and Applied Mathematics, Aston University. https://www.microsoft.com/en-us/research/wp-content/uploads/1995/01/NCRG_95_009.pdf. Published 1995. Accessed May 6, 2021.
42. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform.* 2002;35(5–6):352–359. https://doi.org/10.1016/s1532-0464(03)00034-0.
43. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ Res.* 2017;121(9):1092–1101. https://doi.org/10.1161/CIRCRESAHA.117.311312.
44. Sitar-tăut A, Zdrenghea D, Pop D, Sitar-tăut D. Using machine learning algorithms in cardiovascular disease risk evaluation. *J Appl Comput Sci Math.* 2009;1(3):29–32. https://www.researchgate.net/profile/Dan_Andrei_Sitar-Taut/publication/26635430_Using_Machine_Learning_Algorithms_in_Cardiovascular_Disease_Risk_Evaluation/links/5ea01c62299bf13079b20c80/Using-Machine-Learning-Algorithms-in-Cardiovascular-Disease-Risk-Evaluation.pdf. Accessed March 8, 2021.
45. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One.* 2017;12(4):e0174944. https://doi.org/10.1371/journal.pone.0174944.
46. Ahmad MA, Eckert C, Teredesai AM. Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; 2018 Aug 15. Washington, DC. New York: Association for Computing Machinery, 2018. https://doi.org/10.1145/3233547.3233667.
47. Alpert MA, Omran J, Bostick BP. Effects of obesity on cardiovascular hemodynamics, cardiac morphology, and ventricular function. *Curr Obes Rep.* 2016;5(4):424–434. https://doi.org/10.1007/s13679-016-0235-6.
48. Meade TW, Imeson J, Stirling Y. Effects of changes in smoking and other characteristics on clotting factors and the risk of ischaemic heart disease. *Lancet.* 1987;2(8566):986–988. https://doi.org/10.1016/s0140-6736(87)92556-6.
49. Institute of Medicine. Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. Washington, DC: The National Academies Press, 2014. https://doi.org/10.17226/18709.
50. Alley DE, Asomugha CN, Conway PH, Sanghavi DM. Accountable health communities—addressing social needs through Medicare and Medicaid. *N Engl J Med.* 2016;374(1):8–11. https://doi.org/10.1056/NEJMp1512532.
51. Peasey A, Bobak M, Kubinova R, et al. Determinants of cardiovascular disease and other non-communicable diseases in Central and Eastern Europe: rationale and design of the HAPIEE study. *BMC Public Health.* 2006;6(1):255. https://doi.org/10.1186/1471-2458-6-255.
52. Harper S, Lynch J, Smith GD. Social determinants and the decline of cardiovascular diseases: understanding the links. *Annu Rev Public Health.* 2011;32:39–69. https://doi.org/10.1146/annurev-publhealth-031210-101234.
53. Yadlowsky S, Hayward RA, Sussman JB, McClelland RL, Min YI, Basu S. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann Intern Med.* 2018;169(1):20–29. https://doi.org/10.7326/M17-3011.
54. Kleinberg J, Ludwig J, Mullainathan S, Rambachan A. Algorithmic fairness. *AEA Pap Proc.* 2018;108:22–27. https://doi.org/10.1257/pandp.20181018.
55. Garg A, Toy S, Tripodis Y, Silverstein M, Freeman E. Addressing social determinants of health at well child care visits: a cluster RCT. *Pediatrics.* 2015;135(2):e296–e304. https://doi.org/10.1542/peds.2014-2888.
56. Gottlieb LM, Tirozzi KJ, Manchanda R, Burns AR, Sandel MT. Moving electronic medical records upstream: incorporating social determinants of health. *Am J Prev Med.* 2015;48(2):215–218. https://doi.org/10.1016/j.amepre.2014.07.009.
57. DeVoe JE, Bazemore AW, Cottrell EK, et al. Perspectives in primary care: a conceptual framework and path for integrating social determinants of health into primary care practice. *Ann Fam Med.* 2016;14(2):104–108. https://doi.org/10.1370/afm.1903.
58. Mhasawade V, Chunara R. Causal multi-level fairness. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21); 2021 May 19–21. https://doi.org/10.1145/3461702.3462587.
59. Lund Crick, Brooke-Sumner C, Baingana F, et al. Social determinants of mental disorders and the Sustainable Development Goals: a systematic review of reviews. *The Lancet Psychiatry.* 2018;5(4):357–369.