SSM - Population Health 15 (2021) 100836

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects

Shiho Kino [a,b,*], Yu-Tien Hsu [a], Koichiro Shiba [c], Yung-Shin Chien [a], Carol Mita [d], Ichiro Kawachi [a], Adel Daoud [e,f,g,h]

[a] Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b] Department of Social Epidemiology, Kyoto University, Kyoto, Japan
[c] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[d] Countway Library of Medicine, Harvard University, Boston, MA, USA
[e] Center for Population and Development Studies, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
[f] Department of Sociology and Work Science, University of Gothenburg, Sweden
[g] The Division of Data Science and Artificial Intelligence of the Department of Computer Science and Engineering, Chalmers University of Technology, Sweden
[h] Institute for Analytical Sociology, Linköping University, Sweden

ARTICLE INFO

Keywords:
Review
Machine learning
Social determinants of health

ABSTRACT

Background: Machine learning (ML) has spread rapidly from computer science to several disciplines. Given the predictive capacity of ML, it offers new opportunities for health, behavioral, and social scientists. However, it remains unclear how and to what extent ML is being used in studies of social determinants of health (SDH).
Methods: Using four search engines, we conducted a scoping review of studies that used ML to study SDH (published before May 1, 2020). Two independent reviewers analyzed the relevant studies. For each study, we identified the research questions, Results, data, and algorithms. We synthesized our findings in a narrative report.
Results: Of the initial 8097 hits, we identified 82 relevant studies. The number of publications has risen during the past decade. More than half of the studies (n = 46) used US data. About 80% (n = 66) utilized surveys, and 70% (n = 57) employed ML for common prediction tasks. Although the number of studies in ML and SDH is growing rapidly, only a few studies used ML to improve causal inference, curate data, or identify social bias in predictions (i.e., algorithmic fairness).
Conclusions: While ML equips researchers with new ways to measure health outcomes and their determinants from non-conventional sources such as text, audio, and image data, most studies still rely on traditional surveys. Although there are no guarantees that ML will lead to better social epidemiological research, the potential for innovation in SDH research is evident as a result of harnessing the predictive power of ML for causality, data curation, or algorithmic fairness.

Introduction

Machine learning (ML) algorithms are increasingly used to model sociological, psychological, and biological processes (Bi et al., 2019a; Luo et al., 2016; Molina & Garip, 2019; Mullainathan and Spiess, 2017; Wiemken and Kelley, 2020; Wilkerson and Casas, 2017). ML is a subdiscipline of computer science, which studies how algorithms learn automatically through experience (Hastie, Tibshirani, & Friedman, 2009; Jordan & Mitchell, 2015). This subdiscipline lies at the intersection of computer science and statistics and is known in popular discourse as "artificial intelligence" (Hastie et al., 2009; Jordan & Mitchell, 2015).

An algorithm, f, is a set of instructions that defines how a computer learns from experience. *Learning* implies that algorithms have identified how to reliably predict an outcome (Y) from a set of predictors (X). The algorithm operates on the predictors $f(X)$ to reproduce (predict) a new outcome $\widehat{Y}$ as close as possible to the original outcome, Y. That is, $Y = f(X) + e$, where e is an error term and $f(X) = \widehat{Y}$. Because the algorithm targets outcome Y, this type of learning is called supervised learning. In contrast, unsupervised learning is a process in which an algorithm lacks such an outcome and is tailored instead to summarize a (large) set of predictors (Z) to one (or a few) (Hastie et al., 2009). The more data an

* Corresponding author. Floor 2, Science Frontier Laboratory, Yoshida-konoe-cho, Sakyo-ku, Kyoto-shi, Kyoto, 606-8501, Japan.
E-mail address: kino.shiho.63y@st.kyoto-u.ac.jp (S. Kino).

https://doi.org/10.1016/j.ssmph.2021.100836
Received 2 March 2021; Received in revised form 15 May 2021; Accepted 1 June 2021
Available online 5 June 2021

algorithm experiences, the better trained it is to summarize information and predict outcomes from data it has not encountered. Thus, through experience, ML algorithms learn to identify how Y is statistically related to X or to extract meaningful information from Z.

With the growth of the number of data sources and their size, ML has been gaining traction in several disciplines. For example, an increasing number of studies in biomedicine (Luo et al., 2016), economics (Mullainathan and Spiess, 2017), political science (Wilkerson and Casas, 2017), and sociology (Molina & Garip, 2019) use ML. Although these studies indicate the usefulness of ML, in epidemiology, ML has been mainly applied to analyze clinical questions, such as predicting intensive care unit patient survival, suggesting personalized treatments, and predicting the onset of disease (Miotto et al., 2018; Wiemken and Kelley, 2020). However, the diffusion of ML to other major subdisciplines of epidemiology, such as social epidemiology, seems to occur at a slower rate (Bi et al., 2019a).

Social epidemiology, which is a part of science in epidemiology and public health, is defined as the study of social determinants of health (SDH). The focus of social epidemiology is to analyze how, for instance, income, education, working conditions, race, ethnicity, and neighborhoods affect health outcomes and health inequality (Berkman, Kawachi, & Glymour, 2014). As SDH contain many complex relationships among biological, psychological, and sociological mechanisms, this subdiscipline would benefit from statistical tools that can inductively identify these interactions by pre-specified variables. Interaction is *effect modification*, where the effect of the exposure differs with the levels of other covariates (independent variables). These covariates that influence the effect are called *effect modifiers*. Although conventional parametric models—models for which an investigator specifies the statistical relationships between the dependent and independent variables by defining a functional form (e.g., linearity)—can test such complex interactions, it is generally difficult to identify and model these interactions correctly based solely on subject matter knowledge. ML can be an alternative approach for reasoning complex interactions as it only requires investigators to specify a set of covariates that may serve as effect modifiers. Because the scholar has to specify what interactions to test deductively, a parametric model will not suggest any other interactions by the rest of the covariates in the model (Athey, Tibshirani, & Wager, 2019). Additionally, parametric models lack the full statistical flexibility of nonparametric ML algorithms (Hastie et al., 2009). To give an example, whereas ML algorithms can analyze data with more covariates than data rows, linear models fail in such situations (Hastie et al., 2009).

Despite the capability of ML to learn from large data more optimally than commonly used methods, ML suffers from weaknesses that may render it unsuitable for SDH analyses. One general weakness of ML, or rather a critique, is that it might be an instance of a statistical technique that adds little scientific value to social epidemiology (Bi et al., 2019a; Wiemken and Kelley, 2020). While the validity of this critique is best evaluated in each specific empirical case of SDH research, another weakness of ML is that many algorithms are black boxes with low degrees of interpretability compared to commonly used linear models (Hastie et al., 2009; Wiemken and Kelley, 2020). By "black box," we mean an algorithm that consists of complex procedures and thus may be difficult to decipher for an untrained scholar. In other words, the algorithm lacks transparency for the user and therefore leads to lower interpretability (Hastie et al., 2009; Lipton, 2018). For example, in a linear model, the parameters β_0 (i.e., conditional mean) and β_1 (i.e., the slope or marginal effect of x_i on y_i) have clear interpretations in $y_i = \beta_0 + \beta_1 x_i + e_i$ because the model is transparent. For many ML algorithms, there are no such parameters readily available for interpretation. Lastly, because ML is mainly tailored for prediction, it may not be immediately apparent how its algorithms can be used for causal inference, which is a fundamental goal of most epidemiological research (Bi et al., 2019b). Given the strength (predictive power) and weaknesses of ML, a valuable research question is to identify to what extent and how ML is used in SDH research.

Based on how other disciplines have employed ML algorithms, we identify four useful ways researchers can harness the predictive-power of these algorithms (Luo et al., 2016; Molina & Garip, 2019; Mullainathan and Spiess, 2017; Wilkerson and Casas, 2017). First, Hastie et al. (Hastie et al., 2009) showed how scholars can use unsupervised ML to "reduce high-dimensional datasets" (i.e., data with millions of columns and rows) to a manageable size and supervised ML to synthesize data Y, e.g., the level of brain activity from brain images. Second, Mullainathan and Spiess (Mullainathan and Spiess, 2017) suggest that ML can be applied for the "prediction of policy problems." Policy problems can be condensed into anticipating future events, such as predicting the onset of morbidity or mortality. Third, based on the usual assumptions stipulated for causal inference (Hernán & Robins, 2020), Van der Laan and Rose (Van der Laan and Rose, 2011) and others (Athey et al., 2019; Goin et al., 2020; Künzel et al., 2019; Platt et al., 2018) show how the predictive power of ML benefits causal inference. Fourth, another relevant way ML is being used is to reveal social and ethnic biases hidden in data and sometimes committed by algorithms (Kusner et al., 2017; Obermeyer et al., 2019). In summary, these four domains of ML application—data curation, pure prediction, causal inference, and algorithmic fairness —all rely on the predictive power of ML, but they mobilize it for different goals.

ML for pure prediction is when an algorithm, f, is used to predict an outcome Y based on a set of covariates X without ascribing any causal value to these covariates (Kleinberg et al., 2018a; Mullainathan and Spiess, 2017). As the goal is to predict an outcome as accurately as possible, pure prediction is a supervised learning task. Pure prediction is the traditional method of using ML in computer science (Hastie et al., 2009). Examples relevant in social epidemiology are analyses that predict firearm violence in communities (Goin, Rudolph, & Ahern, 2018), women's height from their socio-economic characteristics (Daoud, Kim, & Subramanian, 2019), or the onset of cancer from social factors (Hanson et al., 2019). Compared to commonly used methods in SDH research, a key advantage of ML for pure prediction is that the algorithm finds the optimal functional form between the outcome and the covariates, $f(X) = \widehat{Y}$, without guidance from the scholar. In other words, although conventional methods are routinely used for prediction, they are generally outperformed by the predictive power of ML in new samples because ML is tailored for out-of-sample prediction (Hastie et al., 2009).

ML for algorithmic fairness is the use and development of algorithms to identify bias in data and decision-making (Kusner et al., 2017). Scholars debate how ML may unintentionally contribute to social injustice in computerized-decision-making systems (Kleinberg et al., 2018b). For instance, in predictive policing, algorithms are used to classify suspect individuals with the help of facial recognition algorithms and predicting crime-hot spots from geographically tagged information (Chan & Bennett Moses, 2016). Although these algorithms are helpful for sorting through vast amounts of information, they have also been shown to reproduce human biases (Kleinberg et al., 2018b). Consequently, when sensitive attributes such as race, ethnicity, and gender are involved in a study, similar issues arise in SDH research. These issues are particularly relevant when scholars analyze health systems that rely on algorithmically powered decision-making (Obermeyer et al., 2019). For example, Daoud et al. show how using the logic of algorithmic fairness—combined with causal inference—can provide a formal approach to making trade-offs between macroeconomic performance and population health when they are at odds with each other (Daoud, Herlitz, & Subramanian, 2007).

Notwithstanding, if ML serves to amplify societal biases, then this would run directly counter to the normative goals of SDH research—to not only explain social inequalities but also reduce them (Berkman et al., 2014). To give an example, algorithmic fairness studies have recently shown that a trade-off exists between developing predictive and fair

algorithms (Kleinberg et al., 2018b; Loftus et al., 2018). Thus, SDH scholars would benefit from engaging with the growing field of algorithmic fairness, which identifies how algorithms become biased and how to mitigate this bias (Kusner et al., 2017; Loftus et al., 2018). While the number of studies on algorithmic fairness in computer science is increasing rapidly (Kusner et al., 2017), it is unclear to what extent SDH scholars have researched this aspect of ML.

ML for causal inference is when algorithms are used to enhance the evaluation of how treatment affects an outcome. Under the potential-outcome framework, causal effect is quantified by comparing two potential outcomes for an individual had they taken the treatment (Y_1) or not taken it (Y_0) (Hernán & Robins, 2020; Rubin, 1974). ML consists mainly of algorithms tailored for statistical (learning) inference (Hastie et al., 2009); thus, any leap to causal inference has to rely on the plausibility of a set of identifiability assumptions, which require subject-matter knowledge and cannot be assessed via data-driven approaches (Hernán et al., 2002; Hernán & Robins, 2020). In observational data, one such essential assumption is that of a conditional exchangeability: a treatment (T) is as good as randomly assigned to treatment (Y_1) and control (Y_0) groups conditional on a set of observed confounders (Z), mathematically, that is, $Y_1, Y_0 \perp T|Z$. Determining such a set of Z is a key challenge in causal inference from observational data; however, using ML does not overcome the challenge in identification because it is a tool for statistical learning.

Although ML alone cannot identify causal effects (Glymour, Zhang, & Spirtes, 2019; Peters, Janzing, & Schölkopf, 2017), it can support causal inference in two main ways when the identifiability assumptions based on subject-matter knowledge are reasonable. First, ML supports causal inference for population average treatment effects by finding the optimal statistical functional form among the outcomes, exposures, and covariates, without imposing restrictive parametric modeling assumptions (Hernán et al., 2002; Schuler and Rose, 2017; VanderWeele, 2019). When the identifiability assumptions (e.g., conditional exchangeability) hold, causal effect estimation reduces to the prediction of conditional expectations or probabilities. ML can enhance such an estimation process by allowing more flexible modeling. For example, the SuperLearner algorithm developed by Van der Laan and his colleagues estimates the performance of a set of algorithms—this set is known as an ensemble—where each algorithm produces a causal estimate (Van der Laan and Rose, 2011). These estimates are then weighted according to their out-of-sample performance. Through such an ensemble, an SDH scholar can use virtually any combination of parametric and nonparametric models to produce robust causal estimates. For instance, Ahern et al. (Ahern et al., 2016) used SuperLearner to estimate the relationship between childhood adversity and the onset of mental disorders by race and ethnicity.

Second, recent developments in the application of ML to estimate heterogeneous treatment effects offer opportunities for SDH researchers to identify subpopulations that are more or less likely to benefit from a certain treatment targeting SDH (Athey et al., 2019; Braveman, Egerter, & Williams, 2011). Although the identifiability assumptions based on subject-matter knowledge are still required, ML can contribute to estimating effect heterogeneity (i.e., how exposure effects differ across subgroups with effect modifiers, such as different population characteristics) by allowing flexible estimation of conditional expectations. This is so because subgroup-specific exposure effects, under the identifiability assumptions, can be quantified by comparing two alternative potential outcomes *conditional on* pre-specified covariates. For example, ML allows scholars to identify socioeconomic subpopulations that are likely to benefit the most from quitting smoking. Even if an intervention improves population health *on average*, assessment of the heterogeneous treatment effects may demonstrate that people with lower socioeconomic backgrounds or racial minorities benefit less from the intervention (Athey et al., 2019). In this case, the population-wide intervention may result in *widening* the existing health disparity. Although several statisticians advocate the use of ML to identify social stratification and health inequalities (Athey et al., 2019; Imai & Ratkovic, 2013; Künzel et al., 2019), a key task is to determine to what extent this approach has become popular in SDH research.

ML for data curation is the practice of using algorithms to synthesize new data (Hastie et al., 2009). ML can be used to construct new covariates using unstructured information sources (such as text, audio, or images) or structured information from traditional surveys (Molina & Garip, 2019). By "structured information," we mean information that is systematically organized using well-defined variables. In contrast, "unstructured information" is when data lacks such systematic organization. For example, when scholars combine satellite images with survey data to train an algorithm to measure the distribution of resources within and across neighborhoods, is one example of how one can use ML to produce new data (Jean et al., 2016).

Another example is when ML determines psychological profiles and mood from language processing in social media (Inkster et al., 2016; Kosinski et al., 2015). A third example is the following: neighborhood food and built environments are often assessed with the help of ML to identify features such as proximity to grocery stores, fast-food restaurants, greenness, and walkability (Duncan & Kawachi, 2018; Leal et al., 2012). Because these features tend to be geographically clustered and highly correlated with each other, it is difficult to disentangle the effects of specific characteristics. In such case, ML can be used to assess "patterns" of multiple neighborhood environment characteristics more holistically rather than to focus on any single dimension (Jones & Huh, 2014). Even in traditional surveys, ML can find complex interactions among many variables and combine them into a single measure (Hastie et al., 2009). As these examples show, ML allows researchers to creatively combine many different types of data to produce new variables that can be used for further analyses. By evaluating how SDH scholars have been using ML for data curation, our goal is to assemble a repertoire of examples for future SDH research.

Although ML may provide statistical innovations for social epidemiology, the application of ML to SDH studies must be based on substantive social-epidemiological theories, a requirement no different from other scientific studies (Krieger, 2011).

Using a scoping review, our study identifies how and to what extent scholars have used ML to study SDH (social epidemiology), which is one of the major segments of epidemiology/public health. A scoping review systematically maps previous research and thus identifies gaps in existing knowledge. Such a review of the application of ML to study SDH provides a vantage point for future research. While ML can be utilized in the many fields of epidemiology and public health, our focus of this review is not the entirety of ML applications to epidemiology/public health but the use of ML in the field of social epidemiology to study SDH. Besides identifying the frequency of studies in each of the four approaches to ML, our review records a number of other characteristics relevant to SDH research when combined with data science and artificial intelligence. Most notably, we report the algorithms, identify data types, and discuss key Results.

Methods

Protocol and registration

We conducted a scoping review on the application of machine learning algorithms to the research of social determinants of health. Our protocol was drafted utilizing the *Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols for a Scoping Review* (PRISMA-ScR) (Tricco et al., 2018).

Eligibility criteria

No language limits or year restrictions were applied to include as many relevant papers as possible. Papers were excluded if they did not fit into the conceptual framework of this study. Notably, we excluded

clinical studies if the studies had no direct SDH relevance. Clinical studies can be defined by the study setting (e.g., within the hospital) and the situation of an interaction with a doctor but not by clinical outcomes. To give an example, we considered the study of Prayaga et al. as eligible even though it used clinical data, because it focused on SDH (Prayaga et al., 2019). We included peer-reviewed publications to include only reliable evidence, and we did not include conference abstracts, letters, or notes.

Information sources

Studies discussing the application of ML to SDH studies were identified by searching Medline/PubMed (National Library of Medicine), EMBASE (Elsevier), and Web of Science (Thomson Reuters). Controlled vocabulary terms (i.e., MeSH and EMTREE) were included when available and appropriate. The search strategies were designed and executed by a reference and education librarian (CM).

Search

Titles and abstracts were searched using keywords and controlled vocabulary terms, including combinations of terms for machine learning, social determinants of health, and algorithms. The exact search terms used for each of the databases and Results are provided in Appendix. We also included relevant studies found by hand searching, which is recommended by the Cochrane Handbook for Systematic Reviews (Higgins et al., 2020, p. 2020; Lefebvre et al., 2019).

Selection of sources of evidence

To increase the consistency among reviewers, two independent reviewers (YTH and YSC) screened the same publications from the title and abstract and then performed the full-text screening. We resolved disagreements on study selection and data extraction by consensus and discussion with other reviewers (IK, AD, and SK).

Data charting process

The reviewers jointly developed the data charting to determine which variables to extract. Three reviewers (YTH, YSC, and SK) charted the data, discussed the Results, and updated the data charting form.

Data items

We evaluated each study based on its theme, as defined in the Introduction (algorithmic fairness, causal inference, pure prediction, and data curation), data characteristics (type, source, country, size, year, outcomes, and covariates), methodology (study design, algorithm's functional class, and algorithmic approach), and main findings.

Synthesis of Results

Table 1 summarizes our analysis using 10 components. *Authors* and *Year* specify the author and date of the study. *Aim* is the verbatim goal of the study. *Theme* captures the four ways ML may be mobilized for SDH (pure prediction, algorithmic fairness, causal inference, and data curation). *Data Type* is the structure of the data sources. The relevant types are survey, text (e.g., health records), accelerometer information, maps, video, images, and audio (Bi et al., 2019a; Hastie et al., 2009; Jordan & Mitchell, 2015). *Country* defines the empirical population of the study. *Number of Rows* signifies the first dimension of the data; because many studies did not reveal the *Number of Columns*, we could not define the second dimension of the data. Together, these two dimensions define the size of the dataset each study used. *Outcomes* define the *Y* variable of interest. If the study uses dimension reduction or unsupervised learning—that is, no predefined outcome—we categorized that Outcome as "N/A." *Algorithm* captures the ML models used in the study. *Main Findings* is a verbatim quote of a key passage summarizing the Results of the study.

Results

Selection of sources of evidence

In the initial search, 8097 references were retrieved from the database searches (3437 from PubMed/Medline (NLM), 2234 from EMBASE (Elsevier), and 2407 from Web of Science (Thomson Reuters)), resulting in 5826 unique records for screening. We also included 22 studies for screening found by manual searching. From the title and abstract screening, we excluded 5672 studies, retaining 154 studies for full-text review. During the full-text assessment for eligibility using the inclusion and exclusion criteria, 72 studies were excluded for the following reasons: ineligible application of ML methods (n = 32), ineligible study aim (n = 26), ineligible research methods (n = 6), commentary and conference abstract (n = 5), and ineligible study setting (n = 3). We identified 82 studies for data extraction. The flowchart is presented in Fig. 1.

Characteristics of sources of evidence (studies).

The included studies are listed in Table 1, with the columns defined in Methods in subsection Synthesis of Results. Appendix Table 1 provides additional information, such as journal and study design. As Fig. 2 shows, we found that the number of publications increased during the past decade, with accelerating growth in the last three years. As expected, few papers (n = 2) were published in the 1990s before the boom of the Internet Age (Barnes, Welte, & Dintcheff, 1991; Boerstler & de Figueiredo, 1991).

Results of individual sources of evidence

Regarding the origin of the data sources, more than half of the studies (n = 46, 56%) used US data, followed by global data (n = 9, 11%). Several studies were based on data from Australia (n = 5) (Bentley et al., 2018; Cramb, Mengersen, & Baade, 2011; Handley et al., 2014; Hu et al., 2009, 2010), the UK (n = 3) (Engchuan et al., 2019; Penny and Smith, 2012; Suel et al., 2019), Iran (n = 2) (Bastaminia, Rezaei, & Saraei, 2017; Darvishi et al., 2017), and Brazil (n = 2) (Ambriola Oku et al., 2020; Chiavegatto Filho et al., 2018).

Of the 82 studies, 80% used survey data (n = 66), followed by text (n = 11) and map and image (n = 4) data. Conway et al. used text in electronic health records to extract social risk factors (Conway et al., 2019), Robson and Boray combined data mining and prediction algorithms to predict the length of stay in the hospital (Robson and Boray, 2019), Prayaga et al. evaluated the efficacy of sending text message reminders to Medicare patients (Prayaga et al., 2019), and Crossley et al. used natural language processing to identify health literacy in patients with diabetes (Crossley et al., 2020). Larkin (Larkin & Hystad, 2019) used a map (e.g., Google Street View) with deep convolutional neural networks to estimate environmental exposures using images. One study applied a deep learning approach to street images in order to measure spatial distributions of neighborhood characteristics (Suel et al., 2019). This study concluded that the application of deep learning could better predict inequalities in income and living environment than inequalities in crime and self-reported health.

About 96% of the studies used supervised methods, two employed unsupervised methods (Abarca-Alvarez, Reinoso-Bellido, & Campos-Sánchez, 2019; Crossley et al., 2020), and one study used a semi-supervised method (Nguyen et al., 2017a). Among supervised methods, classification and regression trees (CART) were the most common (n = 20), followed by random forests (n = 17).

There was substantial variation in data size, ranging from 30 cases (Leach et al., 2016) up to 80 million cases (Nguyen et al., 2016a, 2017b). Because few studies provide information about the number of variables

Table 1
Summary of the included studies.

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
(Abarca-Alvarez et al., 2019)	2019	To assess the connection between social vulnerability and its urban and dwelling context by a decision model.	Prediction	Survey	Spain	5381	Social vulnerability	ANNs[a], decision tree	There is a connection and relationship between demographic and social vulnerability phenomena and the residential configuration of Andalusia.
(Abirami et al., 2020)	2020	To analyze how socio-economic and socio-cultural factors play a role in the initiation and cultivation of addictive behaviors and use a machine learning approach to predict the early onset of such behaviors.	Prediction	Survey	Global	176	Smoking and alcohol habit	Gaussian naïve Bayes, SVM[a], logistic regression algorithms	Logistic Regression to be the best performing classifier to predict both drinking and smoking habits.
(Adeyinka et al., 2019)	2019	To examine spatial patterns of country-level stillbirth rates and determine the influence of social determinants of health on spatial patterns of stillbirth rates.	Data curation	Survey	Global	194	Stillbirth rate	Bayesian networks	The Bayesian network model suggests strong dependencies between stillbirth rate and gender inequality index, geographic regions, and skilled birth attendants during delivery.
(Allali et al., 2010)	2010	To evaluate the impact of educational attainment on the prevalence of osteoporosis and peripheral fractures and to develop a simple algorithm using a tree-based approach with education level and clinical data.	Prediction	Survey	Morocco	356	Bone mineral density	CART[a], CTA[a]	A lower level of education was associated with significantly lower bone mineral densities at the lumbar spine and the hip sites and with a higher prevalence of osteoporosis at these sites in a dose-response manner.
(Ambriola Oku et al., 2020)	2019	To explore potential confounders in an adolescent public health dataset of a developing country by using a combination of machine learning methods and graph analysis.	Prediction	Survey	Brazil	102301	Health status	Gradient boosting machines	The proposed approach might be a useful tool to obtain novel insights on the association between variables and to identify general factors related to health conditions.
(Ahern, Balzer, & Galea, 2015)	2015	To examine the relation of neighborhood alcohol outlet density and norms around drunkenness with alcohol	Prediction	Survey	United States	N/A	Alcohol use disorder	SuperLearner algorithm	The neighborhood environment shapes alcohol use disorder. Despite the lack of additive interaction, each exposure had a substantial relationship with alcohol use disorder. Their findings suggest that alteration of outlet density and norms together would likely be more effective than either one alone.
(Ahern et al., 2016)	2016	To examine the relations between childhood adversities and mental disorders by race/ethnicity in the National Comorbidity SurveyAdolescent Supplement	Causal inference, Prediction	Survey	United States	N/A	Mental disorders	SuperLearner algorithm	Among adversities, physical abuse, emotional abuse, and sexual abuse had the strongest associations with mental disorders. Of all outcomes, behavior disorders were most strongly associated with adversities.
(Bai et al., 2020)	2020	To explore the relationship between individual social capital and functional ability.	Prediction	Survey	China	N/A	Activity function	CART	Subjects with lower social participation and lower social connection had an increased risk of functional disability.
(Barnes et al., 1991)	1991	To apply the classification and regression trees method to classify abstainers and drinkers according to interactions among ten sociodemographic factors.	Prediction	Survey	United States	5952	Alcohol use	CART	Low rates of drinking were shown for low-income women with less than high school education.
(Bastaminia et al., 2017)	2017	To quantify and compare different dimensions of social and economic resilience in Bam and	Prediction	Survey	Iran	660	Social and economic resilience	Feedforward multilayer perceptron ANN	The social component, namely, social capital, was the most important determinant of resilience.

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
		Rudbar with a descriptive-analytical method.							
(Basu & King, 2013)	2013	To provide critical insights into the vast heterogeneity of disability within India.	Prediction	Survey	Global	7150	Disability score	Regression tree analysis	Having two or more symptomatic NCDs[a] was a key factor correlated with disability. The prevalence of symptomatic, undiagnosed NCDs was highest among the lowest two wealth quintiles of Indian adults, contrary to prior hypotheses of increased NCDs with wealth. Women and persons from rural populations were also disproportionately affected by non-diagnosed NCDs, with high out-of-pocket health care expenditures increasing the probability of remaining symptomatic from NCDs.
(Basu & Narayanaswamy, 2019)	2019	To develop a model for predicting whether a person with T2DM has uncontrolled diabetes (hemoglobin A1c $\geq$ 9%), incorporating individual and area-level (census tract) covariates	Prediction	Survey	United States	1015808	Type 2 diabetes mellitus	LASSO regression, Ridge regression, RF[a]	Machine learning models improved upon risk prediction.
(Bellavia et al., 2020)	2020	To apply Logic regression in a study evaluating the association between occupational history and the risk of amyotrophic lateral sclerosis (ALS), and discuss advantages of the method as well as drawbacks and practical issues relevant for epidemiological research.	Prediction	Survey	Denmark	37972	Incidence of ALS	Logic regression	Logic regression may represent a useful methodology in several epidemiological studies dealing with a high number of covariates and is one of the few available approaches to investigate patterns of multiple binary covariates as they relate to a given outcome, which can offer several advantages in terms of both computation and interpretation.
(Bentley et al., 2018)	2018	To examine the cumulative effect of additional years and tenure security of social housing on mental health in a large cohort of lower-income Australians.	Prediction	Survey	Australia	4777	Mental health	Marginal structural models with machine learning-generated weights	The more transitions people made in/out of social housing, the greater the impact on mental health and psychological distress.
(Berkowitz et al., 2019)	2019	To determine if area-level resources, defined as organizations that assist individuals with meeting health-related social needs, are associated with lower levels of cardiometabolic risk factors.	Prediction	Survey	United States	123355	Body mass index	RF	Resources associated with lower BMI included more food resources, employment resources, and nutrition resources.
(Bhavsar et al., 2018)	2018	To assess whether knowledge of neighborhood socioeconomic status improves the prediction of health outcomes.	Prediction	Survey	United States	90097	Use of health care services and hospitalizations due to accidents, asthma, influenza, myocardial infarction, and stroke.	Random survival forest	Information on neighborhood socioeconomic status may not contribute much more to risk prediction above and beyond what is already provided by electronic health record data.
(Bodnar et al., 2020)	2020	To estimate associations between fruit and vegetable intake relative to total energy intake and adverse pregnancy outcomes.	Causal inference	Survey	United States	7572	Adverse pregnancy outcomes	Targeted maximum likelihood estimation,	The differences in Results between Super Learner with TMLE and logistic regression suggest that dietary synergy, which is accounted for in machine learning, may

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
								SuperLearner algorithm	play a role in pregnancy outcomes.
(Boerstler & de Figueiredo, 1991)	1991	To identify potential high users of services among low-income psychiatric outpatients using CART.	Prediction	Survey	United States	382	Potential high users of services	CART	Discharge from inpatient psychiatric treatment right before admission to outpatient psychiatric treatment is the most powerful predictor.
(Brondeel et al., 2016)	2016	To present and apply a method that makes predictions for trips reported in a household travel survey based on the data from a GPS and accelerometer data collection conducted in the same geographical context.	Data curation, Prediction	Survey, Accelerometer data	Global	82084	Transport-related physical activity	RF	The education level had a positive association with transport-related physical activity (T-MVPA). Household income had a negative association with T-MVPA, especially for those people without a motorized vehicle.
(Cairney et al., 2014)	2014	To explore complex interactions between different social determinants and their impact on mental healthcare use.	Prediction	Survey	Canada	10600	Mental health visits with a primary care provider or geriatrician	CART	Income adequacy plays an important role among women, while marital status is of greater importance among men for mental health services utilization.
(Chiavegatto Filho et al., 2018)	2018	To use machine learning algorithms to predict life expectancy at birth and then compare health-related characteristics of the under- and overachievers.	Prediction	Survey	Brazil	3052	Life expectancy at birth	ANNs, RFs, gradient boosted trees, least squares, Ridge and LASSO regressions, SVMs	Overachievers presented better Results regarding primary health care. Underachievers performed more cesarean deliveries and mammographies and had more life-support health equipment.
(Choi, Fram, & Frongillo, 2017)	2017	To identify patterns of characteristics that distinguish very low food security households in the United States from other households.	Prediction	Survey	United States	13351	Food security	CART	Household experiences of VLFS were associated with different predictors for different types of households and often occurred at the intersection of multiple characteristics spanning unmet medical needs, poor health, disability, limitation, depressive symptoms, low income, and food assistance program participation.
(Choi et al., 2018)	2018	To investigate the probability of suicide death using baseline characteristics and simple medical facility visit history.	Prediction	Survey	South Korea		Suicidal rate	SVMs, ANNs	Male gender, older age, lower-income, medical aid, and disability were linked to increased risk for suicide death at 10-year follow-up.
(Choi et al., 2019)	2019	To use network modeling to characterize co-occurring psychosocial risks to maternal and child health among at-risk pregnant women.	Prediction	Survey	South Africa	200	Distress about pregnancy	Network analysis	Unintended pregnancy was strongly tied to distress about pregnancy. Distress about pregnancy was most central in the network and was connected to antenatal depression and HIV stigma
(Conway et al., 2019)	2019	To present a new and highly configurable rule-based clinical natural language processing system designed to automatically extract information that requires inferencing from clinical notes.	Algorithmic fairness, Prediction	Text (Electronic health records)	United States	N/A	Housing situation, living alone, and social support	NLP[a]	The algorithm is highly accurate in extracting and classifying the three variables of interest (housing situation, living alone, and social support).
(Cramb et al., 2011)	2011	To identify the complex interplay of area-level factors associated with the high area-specific incidence of Australian priority cancers using a classification and regression tree approach.	Prediction	Survey	Australia	186075	Cancer incidence	CART	HL suffered more negative health outcomes and had higher healthcare service utilization.
(Crossley et al., 2020)	2020	To better understand HL from a linguistic perspective and to open	Data curation		United States	1050577	Hospitalization	Discriminant function analysis	The developed model predicts human ratings of HL with ~80% accuracy.

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
		new research areas to enhance population management and individualized care.		Text (Electronic health records), Survey					Validation indicates that lower HL patients are more likely to be nonwhite and have lower educational attainment. In addition, patients with lower
(Daoud et al., 2019)	2019	To predict women's height from their socioeconomic status	Prediction	Survey	Global	N/A	Height	LASSO, Ridge regression, generalized additive model, Bayesian neural network, bagged CART, RF	There were no relevant non-linear relationships between SES and women's height.
(Daoud & Johansson, 2019)	2019	To examine the pathways of economic austerity propagate through families' living conditions and societies' structural and political characteristics.	Causal inference	Survey	Global	1940734	Child poverty	Generalized RF	The International Monetary Fund (IMF) program affects children residing in the middle of the social stratification more than compared to their peers residing in both the top and bottom of this stratification; for those children residing in societies that have selected into IMF programs and have historically spent most on education, are at a higher risk of falling into poverty.
(Darvishi et al., 2017)	2017	To provide an empirical model of predicting low back pain by considering the occupational, personal, and psychological risk factor interactions in workers population employed in industrial units using an ANNs approach.	Prediction	Survey	Iran	92	Low back pain severity	Neural network model	The mean classification accuracy of the developed neural networks for the testing and training phase data was about 88% and 96%, respectively. In addition, the mean classification accuracy of both training and testing data was 92%, indicating much better Results compared with other methods.
(DiGuiseppi et al., 2020)	2020	To identify predictors of youths' first episode of homelessness during the 12 months after substance use treatment entry.	Prediction	Survey	United States	20069	The first episode of homelessness	Lasso machine learning regression	The adolescents who were older, male, reported more victimization experiences, mental health problems, family problems, deviant peer relationships, and substance use problems were more likely to report experiencing homelessness.
(Engchuan et al., 2019)	2019	To evaluate the determinants of health in aging using machine learning methods and to compare the accuracy with traditional methods.	Prediction	Survey	United Kingdom	6209	Health metrics	RF, deep learning, linear model	Health-trend, physical activity, and personal-fitted variables were the main predictors of health. The performance of the RF method was similar to the traditional linear model, but RF significantly outperformed deep learning.
(Fan et al., 2019)	2019	To study the social and economic data and the relationship between opioid drug abuse situations.	Prediction	Survey	United States	8344	Opioid	Grey relation analysis	The deviation of prediction is reduced from (−10.99%, 22.33%) to (−8.29%, 2.81%), making the modified value closer to the measured value.
(Filikov et al., 2020)	2020	To build and evaluate a novel framework termed Stratified Cascade Learning and used it for predicting the risk of hospitalization.	Prediction	Survey	United States	14300	Hospitalization risk	Stratified cascade learning mode	The stratified cascade learning model does not improve either the area under the curve or the negative predictive value of the basic classifier but materially improves accuracy and specificity measures at the expense of lowering sensitivity for the "predictable" subset.
(Friel, Newell, & Kelleher, 2005)	2005	To identify the socioeconomic, sociodemographic, and health-related lifestyle behavior profile	Prediction	Survey	Ireland	6539	Fruit and vegetable consumption	CART	Irish people do not comply with the dietary recommendations, but this varies greatly by social circumstance.

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
		of adults who comply with the recommended four or more servings per day of fruit and vegetables.							
(Fu et al., 2007)	2007	To investigate the differences in culture, attitudes, and social networks between Australian and Taiwanese men and women and to identify the factors that predict midlife men and women's quality of life in both countries.	Prediction	Survey	Global	715	Quality of life	CART	People who had higher levels of horizontal individualism and collectivism, positive attitudes, and better social support had better psychological, social, physical, and environmental health, while it emerged that vertical individualists with competitive characteristics would experience a lower quality of life.
(Goin et al., 2018)	2018	To use machine learning to identify an optimal set of predictors for urban interpersonal firearm violence rates using a broad set of community characteristics.	Prediction	Survey	United States	N/A	Firearm violence	Random forest analysis	The top 5 covariates with the highest variable importance – the black isolation index, the black segregation index, the percent of households receiving food stamps, the percent of men age 65+ with high school education, and the percent never married – achieved 0.708 average R-squared and average MSE alone.
(Goin et al., 2020)	2020	To examine the association between community firearm violence and risk of preterm birth.	Causal inference, Prediction	Survey	United States	2084417	Preterm birth	SuperLearner algorithm	Firearm violence was associated with the risk of preterm delivery, and this association was partially mediated by infection and substance use.
(Gray, Schvey, & Tanofsky-Kraff, 2020)	2019	To assess the relative associations of demographic, psychological, behavioral, and cognitive variables with body mass index in a nationally representative sample of youth.	Prediction	Survey	United States	4524	Body mass index	LASSO regression	Stimulant medications and demographic factors were most strongly associated with body mass index.
(Hamad et al., 2019)	2019	To examine differences in sociodemographic characteristics, health, nutritional status, and food purchasing behavior between new and existing recipients of SNAP after the recession.	Prediction	Survey	United States	21806	Household-level nutritional characteristics	LASSO regression	Given that new recipients are generally better off than existing recipients, it may be more impactful from a public health perspective to instead intervene among those existing recipients who may have more long-standing challenging socioeconomic circumstances.
(Handley et al., 2014)	2014	To determine long-term risk profiles for suicidal ideation among a community sample of older adults using a decision tree approach, with a focus on the role of physical, social, and psychological risk factors, and their interactions	Prediction	Survey	Australia	2160	Suicidal ideation	CART	Psychological factors are important for predicting suicidal ideation. Both physical and social factors significantly improved the predictive ability of the model.
(Hanson et al., 2019)	2019	To measure the relative importance of race compared to health care and social factors on prostate cancer-specific mortality.	Prediction	Survey	United States	514878	Prostate cancer mortality	RFs	Tumor characteristics at diagnosis were the most important factors for prostate cancer mortality. Across all groups, race was less than 5% as important as tumor characteristics and only more important than health care and social factors in 2 of the 18 groups. Although race had a significant impact, health care and social factors known to be associated with racial

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
									disparities had greater or similarly important effects across all ages and stages.
(Herrera-Ibatá et al., 2015)	2015	To develop a computational algorithm for network epidemiology to map structure-activity data of HAART-drugs cocktails over complex networks of AIDS epidemiology and socioeconomic factors.	Prediction	Survey	United States	131252	The probability of AIDS could be halted in a county with a HAART cocktail	Linear neural network	The machine-learning algorithms could be useful as an initial form of screening for the prediction of effective drugs in preclinical assays for the treatment of HIV in different populations of U.S. counties with a given AIDS epidemiological prevalence. However, the models did not appear to be effective when using socioeconomic factors to predict the efficacy of the treatment of HIV.
(Higgins et al., 2016)	2016	To characterize cumulative risk associated with co-occurring risk factors for cigarette smoking.	Prediction	Survey	United States	114426	Smoking status	CART	The effects associated with common risk factors for cigarette smoking are independent, cumulative, and generally summative.
(Higgins et al., 2017)	2017	To examine risk factors for using full-flavor versus other cigarette types, including socioeconomic disadvantage and other risk factors for tobacco use or tobacco-related adverse health impacts.	Prediction	Survey	United States	114426	Tobacco use or tobacco-related adverse health impacts	CART	The use of full-flavor cigarettes is overrepresented in socioeconomically disadvantaged and other vulnerable populations and associated with an increased risk of nicotine dependence.
(Hu et al., 2009)	2009	To explore the spatial distribution of notified cryptosporidiosis cases and identified major socioeconomic factors associated with the transmission of cryptosporidiosis in Brisbane, Australia.	Data curation	Survey, Digital base map from the Australian Bureau of Statistics	Australia	N/A	Incidence of Cryptosporidiosis infection	Spatial CART	A spatial CART model shows that the relative risk for cryptosporidiosis transmission was 2.4 when the value of the social economic index for areas was over 1028 and the proportion of residents with low educational attainment in a statistical local area exceeded 8.8%.
(Hu et al., 2010)	2010	To examine the impact of social economic and weather factors on cryptosporidiosis and explored the possibility of developing such a model using social economic and weather data in Queensland, Australia.	Prediction	Survey	Australia	N/A	Monthly incidence of cryptosporidiosis	Spatiotemporal CART	Spatiotemporal CART models based on social economic and weather variables can be used for predicting the outbreak of cryptosporidiosis in Queensland, Australia.
(Jamei et al., 2017)	2017	To build a model to predict all-cause 30-day readmission risk, and added block-level census data as proxies for social determinants of health	Prediction	Survey	United States	323813	30-day hospital readmission	ANN	Neural networks are great candidates to capture the complexity and interdependency of various data fields in electronic health records.
(Kanerva et al., 2018)	2018	To explore the mutual importance or hierarchy of sociodemographic and lifestyle-related risk factors of being overweight using RF.	Prediction	Survey	Finland	4757	Overweight	RF	RF did not demonstrate higher power in variable selection compared to linear regression in our study. The features of RF are more likely to appear beneficial in settings with a larger number of predictors.
(Kraamwinkel et al., 2019)	2019	To analyze treatment heterogeneity of maternal agency on severe child undernutrition and how this effect plays out in the context of armed conflict.	Causal inference, Prediction	Survey	Nigeria	48613	Severe child malnutrition	Bayesian additive regression tree	Maternal education decreases severe child undernutrition, but only when mothers acquire ten years of education or higher. This protective effect remains even during the exposure of armed conflict.
(Larkin and Hystad, 2019)	2019	To evaluate the potential of Google Street View (GSV) as a novel green space measure for epidemiological studies.	Data curation, Prediction	Image	United States	254	Green View Index, normalized difference vegetation index	Green screen algorithm	GSV-based measures captured unique information about green space exposures. We further developed a GVI:NDVI ratio, which was associated with the amount of

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
									vertical green space in an image. The GVI and GVI:NDVI ratio were weakly related to neighborhood socioeconomic status and are therefore less susceptible to confounding in health studies compared to other green space measures.
(Leach et al., 2016)	2016	To determine which built environment characteristics contributed to the classification of African American women as having four or more CVD risk factors at optimal levels.	Prediction	Survey	United States	30	CVD[a]	CART	The classification and regression trees identified participants with few, low-quality neighborhood physical activity resources and who were older than 55 as least likely to have four or more CVD risk factors at optimal levels
(Lewis and McCormick, 2012)	2012	To present an epidemiologic systems approach for identifying potential determinants of diarrhea in children under five years.	Causal inference, Prediction	Survey	Pakistan	18202	Diarrhea	Bayesian network modeling	The only access to a dry pit latrine (protective), access to an atypical water source (protective), and no formal garbage collection (unprotective) were directly dependent on the presence of diarrhea.
(Li et al., 2019)	2019	To identify and ranks predictors of cardiovascular health at the neighborhood level in the United States.	Prediction	Survey	United States	27066	Neighborhood CVD and stroke prevalence	RF	Demographics, health behaviors, and prevention measures explained the vast majority of the variance: 93.2% for CVD and 96.0% for stroke.
(Luo et al., 2016)	2016	To examine the feasibility of inferring regional health outcomes from socio-demographic data through national censuses and community surveys.	Prediction	Survey	United States	N/A	Angina or CHD, heart attack, stroke, diabetes, hypertension, obesity	Regression with stepwise feature selection, group LASSO, RF, Gaussian process regression	The model had sufficient external validity to make good predictions, based on demographics alone, for areas not included in the model development.
(Maksimov and Artamonova, 2013)	2013	To evaluate the influence of work environment on the risk of arterial hypertension development in employees of various social groups.	Prediction	Survey	Russia	3664	Arterial hypertension	CTA	Quetelet index, age, and occupation are highly significant for arterial hypertension prediction in the working-age population. Occupation is very significant for arterial hypertension prediction in middle-aged patients.
(Matthay et al., 2019)	2019	To evaluate whether the Operation Peacemaker Fellowship, an innovative firearm violence-prevention program implemented in Richmond, California, was associated with reductions in firearm and non-firearm violence.	Prediction	Survey	United States	N/A	Experiencing non-firearm and firearm violence	Generalization of the synthetic control method	The program was associated with reductions in firearm violence (annually, 55% fewer deaths and hospital visits, 43% fewer crimes) but also unexpected increases in non-firearm violence (annually, 16% more deaths and hospital visits, 3% more crimes). These associations were unlikely to be attributable to chance for all outcomes except non-firearm homicides and assaults in crime data
(Matusik, Aska-Mierzejewska, & Chrzanowska, 2011)	2011	To assess the usefulness of the decision trees method as a research method of multidimensional associations between menarche and socioeconomic variables.	Prediction	Survey	Poland	2354	Menstruation appearance at the age of 12–14	CART	The strongest discriminatory power was attributed to the number of children in a family and the mother's and then father's educational level.
Meng et al. (Nguyen et al., 2017a)	2017	To examine openly shared substance-related tweets to estimate prevalent sentiment around substance use and identify popular substance use activities.	Data curation	Text (tweets), Survey	United States	79848992	Substance use behaviors	Maximum entropy text classifier	More convenience stores in a zip code were associated with higher percentages of tweets about alcohol. Larger zip code population size and higher percentages of African Americans and Hispanics were associated with fewer tweets about

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
									substance use and underage engagement. Zipcodeeconomic disadvantage was associated with fewer alcohol tweets but more drug tweets.
(Mooney et al., 2017)	2017	To evaluate neighborhood influences on physical activity among older adults, analogous, in a genetic context, to a genome-wide association study.	Prediction	Survey	United States	3497	Physical activity	LASSO regression, RF	Only neighborhood socioeconomic status and disorder measures were associated with total activity and gardening, whereas a broader range of measures was associated with walking.
(Nayak et al., 2016)	2016	To evaluate health problems in income-based groups.	Prediction	Survey	United States	3604	Self-rate health	CART	More risk factors for self-rate health problems and chronic burden indicators were associated with SRH in lower-income groups
(Nollen et al., 2016)	2016	To explore interactions between demographic, tobacco, and psychosocial factors to identify cigarette smokers at highest risk for alternative tobacco product use from a racially/ethnically and socioeconomically diverse sample of adult smokers across the full smoking spectrum.	Prediction	Survey	Global	2376	Concurrent alternative tobacco product use	CTA	Alcohol for men and age, race/ethnicity, and discrimination for women increased the probability of alternative tobacco product use.
(Nguyen et al., 2016b)	2016	To use publicly available, geotagged Twitter data to create neighborhood indicators for happiness, food, and physical activity for three large counties: Salt Lake, San Francisco, and New York.	Data curation	Text (tweets), Survey	United States	2.8 million	Neighborhood happiness, diet, and physical activity	NLP	Happy tweets, healthy food references, and physical activity references were less frequent in census tracts with greater economic disadvantage and higher proportions of racial/ethnic minorities and youths
(Nguyen et al., 2016a)	2016	To build, from geotagged Twitter data, a national neighborhood database with area-level indicators of well-being and health behaviors.	Data curation	Text (tweets), Survey	United States	80 million	Happy, food, and physical activity tweets, Mortality, chronic condition (obesity, diabetes, high cholesterol), self-rated health	NLP	Census tract factors like percentage African American and economic disadvantage were associated with lower census tract happiness. Urbanicity was related to a higher frequency of fast-food tweets. Greater numbers of fast-food restaurants predicted a higher frequency of fast-food mentions. Surprisingly, fitness centers and nature parks were only modestly associated with a higher frequency of physical activity tweets. Greater state-level happiness, positivity toward physical activity, and positivity toward healthy foods, assessed via tweets, were associated with lower all-cause mortality and prevalence of chronic conditions such as obesity and diabetes and lower physical inactivity and smoking, controlling for state median income, median age, and percentage white non-Hispanic.
(Nguyen et al., 2017b)	2017	To leverage geotagged Twitter data to create national indicators of the social environment, with small-area indicators of prevalent sentiment and social modeling of	Data curation, Prediction	Text (tweets), Survey	United States	80 million	National indicators of the social environment	NLP	Twitter indicators of happiness, food, and physical activity were associated with lower premature mortality, obesity, and physical inactivity. Alcohol-use tweets

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
		health behaviors, and to test associations with county-level health outcomes, while controlling for demographic characteristics.							predicted higher alcohol-use–related mortality.
Nguyen et al. (Meng et al., 2017)	2017	To create zip code level indicators of happiness, food, and physical activity culture from geolocated Twitter data to examine the relationship between these neighborhood characteristics and obesity and diabetes diagnoses.	Data curation, Prediction	Text (tweets), Survey	United States	1.86 million	Obesity and diabetes prevalences	NLP	Individuals living in zip codes with the greatest percentage of happy and physically active tweets had lower obesity prevalence after accounting for individual age, sex, nonwhite race, Hispanic ethnicity, education, and marital status, as well as zip code population characteristics. More happy tweets and lower caloric density of food tweets in a zip code were associated with the lower individual prevalence of diabetes.
(Nguyen et al., 2017c)	2017	First, to build a national food environment database from geotagged Twitter and Yelp data. Second, to test associations between state food environment indicators and health outcomes.	Data curation, Prediction	Text (tweets, Yelp listing), Survey	United States	79,848,992	Mortality, chronic condition (obesity, diabetes, high cholesterol), self-rated health	NLP	A one standard deviation increase in caloric density of food tweets was related to higher all-cause mortality (+46.50 per 100,000), diabetes (+0.75%), obesity (+1.78%), high cholesterol (+1.40%), and fair/poor self-rated health (2.01%). More burger Yelp listings were related to a higher prevalence of diabetes (+0.55%), obesity (1.35%), and fair/poor self-rated health (1.12%). More alcohol tweets and Yelp bars and pub listings were related to higher state-level binge drinking and heavy drinking, but lower mortality and lower percent reporting fair/poor self-rated health.
(Nguyen et al., 2017a)	2017	To construct built environment indicators using computer vision techniques and publicly available Google Street View images and to examine relationships between neighborhood built environments, demographic characteristics of residents, and health outcomes.	Data curation, Prediction	Map (Google Street View), survey	United States	N/Aa	Obesity and diabetes	Convolutional neural networks	Computer vision models had an accuracy of 86%–93% compared with manual annotations. Individuals living in zip codes with the greenest streets, crosswalks, and commercial buildings/apartments had relative obesity prevalences that were 25%–28%lower and relative diabetes prevalences that were 12%–18%lower than individuals living in zip codes with the least abundance of these neighborhood features.
(Obermeyer et al., 2019)	2019	To show that a widely used algorithm exhibits significant racial bias under the healthcare context.	Algorithmic fairness	Text (Electronic health records)	United States	N/A	Patients who will derive the greatest benefit from the high-risk care management	Prediction algorithms (unspecified)	Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients
(Özge et al., 2006)	2006	To evaluate the most important sociodemographic factors on the	Prediction	Survey	Turkey	3304	Smoking status	CART	The significantly important factors that affect current smoking in these age groups

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
		smoking status of high school students using a broad randomized epidemiological survey.							were increased by household size, late birth rank, certain school types, low academic performance, increased second-hand smoking, and stress.
(Penny and Smith, 2012)	2012	To examine the health-related quality of life in a cohort of individuals with irritable bowel syndrome and to explore the use of several data-mining methods to identify which socio-demographic and irritable bowel syndrome symptoms are most highly associated with impaired health-related quality of life.	Prediction	Survey	United Kingdom	494	Quality of life	CTA, ANN	Psychological morbidity and socio-demographic factors such as marital status and employment status also have a major influence on health-related quality of life in irritable bowel syndrome.
(Platt et al., 2018)	2018	To examine the associations between 11 childhood adversities and intelligence, using targeted maximum likelihood estimation	Causal inference, Prediction	Survey	United States	10073	Nonverbal score from the Kaufman Brief Intelligence Test	SuperLearner algorithm	The largest associations were observed for deprivation-type experiences, including poverty and low parental education, which were related to reduced intelligence. Although lower in magnitude, threat events related to intelligence included physical abuse and witnessing domestic violence. Violence prevention and poverty-reduction measures would likely improve childhood cognitive outcomes.
(Prayaga et al., 2019)	2019	To evaluate the efficacy of an SMS-based refill reminder solution using conversational artificial intelligence	Data curation, Prediction	Text	United States	273356	Medication refill request	Neural network multilayer perceptron	There are sharp differences in the likelihood to reply to a refill reminder and complete a refill request via SMS based on demographic and socioeconomic factors. We found a strong association between refill request rates and patient language, age, race/ethnicity, and SDOH levels, and these differences may contribute to health disparities and impact health outcomes in Medicare patients.
(Robson and Boray, 2019)	2019	To demonstrate how collective use of data mining and prediction algorithms to analyze socioeconomic population health data can stand beside classical correlation analysis in routine data analysis.	Data curation	Text (Website)	United States	N/A	Hospital length of stay	Hyperbolic Dirac net	The combined use of tools and modes of use described in this paper appears capable of adding significant value to the analysis of socioeconomic health data.
(Seligman, Tuljapurkar, & Rehkopf, 2018)	2018	To investigate how machine learning may add to our understanding of social determinants of health using data from the Health and Retirement Study.	Prediction	Survey	United States	N/A	Blood pressure, body mass index, waist circumference, and telomere length	Repeated linear regressions, penalized linear regressions, RFs, neural networks.	Dental visits, current smoking, self-rated health, serial-seven subtractions, probability of receiving an inheritance, probability of leaving an inheritance of at least $10,000, number of children ever born, African-American race, and gender are highly weighted predictors.
(Shimony-Kanat and Benbenishty, 2018)	2018	To characterize trauma-related falls in infants and toddlers aged 0–3 years over a 4-year period and develop a risk stratification model of causes of fall injuries.	Prediction	Survey	Israel	2277	Trauma-related falls in infants and toddlers	CART	The leading determinants of fall injuries in children below the age of 3 years are age, ethnicity, and low socioeconomic status.
(Shin et al., 2018)	2018	To empirically test the contribution of social components	Prediction	Survey	United States	3678	Asthma patients at risk of a hospital	RF, SVM	Socio-markers in the Memphis study area aggregated on the ZIP code level can be

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Aim	Theme	Data Type	Country	Number of rows	Outcomes	Algorithm	Main Findings
		versus more traditional symptom-related features in the prediction of health outcomes.					revisit after an initial visit		reliable predictors of pediatric asthma patients at risk of hospital revisit within a year.
(Sow et al., 2019)	2019	To solve health-perspective problems by understanding socioeconomic factors which affect children's health and how they influence malaria and anemia.	Causal inference, Prediction	Survey	Global	6935	Malaria and anemia among children	ANNs, SVM, k nearest neighbors, RFs, naive Bayes	ANNs gave the best Results of 94.74% and 84.17% accuracy for malaria and anemia prediction, respectively.
(Suel et al., 2019)	2019	To apply a deep learning approach to street images for measuring spatial distributions of income, education, unemployment, housing, living environment, health, and crime.	Data curation, Prediction	Image	United Kingdom	525860	Income, education, unemployment, housing, living environment, health, and crime	Deep learning (not specific)	The application of deep learning to street imagery predicted inequalities better in some outcomes (i.e., income, living environment) than others (i.e., crime, self-reported health).
(Torres et al., 2018)	2018	To estimate the associations between having an adult child migrant and depressive symptoms among middle-aged and older adults in Mexico followed over an 11-year period.	Prediction	Survey	United States	11,806	Depressive symptoms	Targeted maximum likelihood estimation	Associations between having an adult child migrant and depressive symptoms varied by respondent gender, family size, and the location of the child migrant.
(Torres et al., 2020)	2020	To evaluate the association between adult child US migration status and change in cognitive performance scores.	Prediction	Survey	United States	5972	Cognitive performance	Targeted maximum likelihood estimation	For women, having an adult child in the United States was associated with a steeper decline in verbal memory scores and overall cognitive performance. There were mostly null associations for men.
(Yu et al., 2017)	2017	To explore the racial disparity in obesity considering not only the individual behavior but also geospatially derived environmental risk factors.	Prediction	Survey	United States	5240	Obesity	Multiple additive regression trees	Multiple additive regression trees (MART) performed better than generalized linear models. MART explained a larger proportion of the racial disparity in obesity. However, there remained disparities that cannot be explained by factors collected in this study.

[a] ANN, artificial neural network; CART, classification and regression trees; CTA, classification tree analysis; CVD, cardiovascular disease; NCD, non-communicable disease; NLP, natural language processing; RF, random forest; SVM, support vector machine.

used, our review cannot identify the total size of the data. Additionally, 53 studies (65%) employed a cross-sectional study design, 24 (29%) used panel data, and 5 (6%) used other study designs.

The studies focused on a range of different health outcomes. Thirty-nine studies used clinical health outcomes (e.g., body mass index, cardiovascular diseases, and cancer incidence), followed by health-related behaviors (n = 19), use of health services (n = 7), social determinants of health (n = 5), mental health (n = 5), mortality (n = 3), self-rated health/quality of life (n = 3), and living environment (n = 1).

Synthesis of Results

The majority of the studies employed ML for pure prediction (n = 57), and the remaining studies used ML for data curation (n = 15), causal inference (n = 8), and algorithmic fairness (n = 2).

Several studies using ML for pure prediction discussed how ML methods could enhance predicting outcomes from SDH factors. First, some studies showed how ML could be applied to various data sources that traditional statistical methods cannot handle well. These studies used high-dimensional longitudinal health data (Engchuan et al., 2019), hierarchical data (Fu et al., 2007), and zero-inflated data (Hu et al., 2010).

Second, other prediction studies discussed how ML algorithms can flexibly model non-linear relationships as well as possible interactions among variables (Bai et al., 2020; Barnes et al., 1991; Bellavia et al., 2020; Cairney et al., 2014; Hanson et al., 2019; Kanerva et al., 2018; Li et al., 2019; Nollen et al., 2016). Third, they discussed how ML trades off bias and variance in estimating the association between exposure and predicting outcomes in new data without overfitting (Fu et al., 2007; Platt et al., 2018). ML generalizes better to new data (out-of-sample) than traditional models, meaning that its predictions are more accurate than non-ML methods (Boerstler & de Figueiredo, 1991; Darvishi et al., 2017; Engchuan et al., 2019; Fu et al., 2007; Li et al., 2019). Finally, one study discussed how ML could easily handle missing values (Fu et al., 2007).

Although a few studies evaluated the strengths and weaknesses of using ML in the context of SDH research, most studies did not. Most studies used ML for prediction: they added SDH variables to a list of other predictors without substantive reasons for how the inclusion of SDH will improve the predictive performance. A minority of the studies explicitly discussed different reasons the addition of SDH variables improved the prediction (Boerstler & de Figueiredo, 1991; Darvishi et al., 2017; Engchuan et al., 2019; Fu et al., 2007; Li et al., 2019).

Several studies used ML for data curation of text (Crossley et al., 2020; Meng et al., 2017; Nguyen et al., 2016a, 2016b, 2017a, 2017b, 2017c; Prayaga et al., 2019; Robson and Boray, 2019), surveys (Adeyinka, Olakunde, & Muhajarine, 2019), maps or street images (Hu et al., 2009; Larkin & Hystad, 2019; Nguyen et al., 2017a; Suel et al., 2019), and accelerometer data (Brondeel, Pannier, & Chaix, 2016). These studies highlighted practical ML applications, e.g., estimating health literacy from the text at the individual patient level (Crossley et al., 2020) or reducing socioeconomic inequalities in medication refills (Prayaga et al., 2019).

For instance, Hu et al. used a digital base map for data curation and discussed that spatial CART could handle a wide variety of data structures (i.e., hierarchical and non-linear relationships) with fewer assumptions than other traditional methods (Hu et al., 2009). Furthermore, Suel et al. studied street images and concluded that street imagery could complement traditional survey-based and administrative data sources for high-resolution urban surveillance to measure inequalities and monitor policy impacts (Suel et al., 2019).

The few studies that used ML for causal inference used it to identify determinants of health in low-income families (Lewis and McCormick, 2012), to classify malaria and anemia cases among children (Sow et al., 2019), to capture the effect heterogeneity of armed conflict on children (Kraamwinkel et al., 2019), to estimate heterogeneous treatment effects of economic crises on children (Daoud & Johansson, 2019), to examine the associations between 11 childhood adversities and intelligence (Platt et al., 2018), to examine the association between community firearm violence and risk of preterm birth (Goin et al., 2020), to examine the relationship between childhood adversities and mental disorders by race and ethnicity (Ahern et al., 2016), and to estimate the associations between dietary intake and adverse pregnancy outcomes (Bodnar et al., 2020). For example, Kraamwinkel et al. used Bayesian additive regression trees to identify which subgroups of children were most at risk during conflicts as a function of mother's educational attainment (Kraamwinkel et al., 2019). Daoud and Johansson used generalized random forest to quantify the average treatment effect of International Monetary Fund policies on child poverty as well as effect heterogeneity (Daoud & Johansson, 2019). Lewis and McCormick showed that Bayesian networks could separate variables directly and indirectly associated with the outcome variables and help in causal reasoning (Lewis and McCormick, 2012).

Two studies used ML for algorithmic fairness. One study used ML to extract information from a clinical natural language processing system to identify sensitive attributes such as race, sex, and age (Conway et al., 2019). The second study quantified how commercial prediction algorithms perpetrate racial biases in the US health system (Obermeyer et al., 2019). Obermeyer et al. discussed in their study that researchers need to produce new labels with a deep understanding of the domain, identify and extract relevant data elements, and improve the capacity to iterate and experiment (Obermeyer et al., 2019).

Discussion

In this scoping review, we identified 82 primary SDH studies employing ML algorithms published before May 1, 2020. We found that the number of publications sharply increased in the past three years. Currently, most studies use US data and rely on cross-sectional surveys.

Of the four ML themes, the pure prediction was by far the most common goal of ML use among the studies we identified (70%, n = 57). Notably, ML application to other themes, despite its potential advantages for SDH researchers, was surprisingly scarce: 18% used ML for data curation (n = 15), 10% for causal inference (n = 8), and 2% for algorithmic fairness (n = 2). Consequently, as most ML applications in social epidemiology mobilize ML for pure prediction, researchers entering the SDH-ML field are likely to have a higher chance of advancing SDH by applying ML as done in the three remaining themes than producing yet another ML-for-prediction study.

Nonetheless, ML-for-prediction studies have their advantages and will continue playing a central role in SDH. For such studies, we recommend analyzing and explaining the added predictive power of using SDH variables or abstaining from including them at all. Despite the fact that social-determinant covariates were used in many studies, the authors of these studies rarely discussed whether the inclusion of SDH improved prediction. The relative importance of SDH over clinical characteristics in predicting health outcomes has been long debated in epidemiology. To give an example, the Framingham risk score—the most widely used algorithm for cardiovascular disease prediction—does not incorporate socioeconomic status (e.g., educational attainment and household income) (Fiscella, Tancredi, & Franks, 2009; Fiscella & Tancredi, 2008; Franks et al., 2010; Hammond et al., 2020; Jolobe, 2009), despite the demonstrated importance of socioeconomic status for health outcomes—as shown by a large number of SDH studies (Berkman et al., 2014; Havranek et al., 2015; Kreatsoulas & Anand, 2010).

ML for algorithmic fairness will likely play an increasingly important role for SDH. Given that SDH often plays an explanatory and normative role, researchers are advised to analyze how ML can bias SDH Results and public-health policy recommendations. The study of algorithmic fairness serves both ambitions. One way of using ML is to train an algorithm to detect bias in health records (Conway et al., 2019; Obermeyer et al., 2019). Another way is to identify the causes and

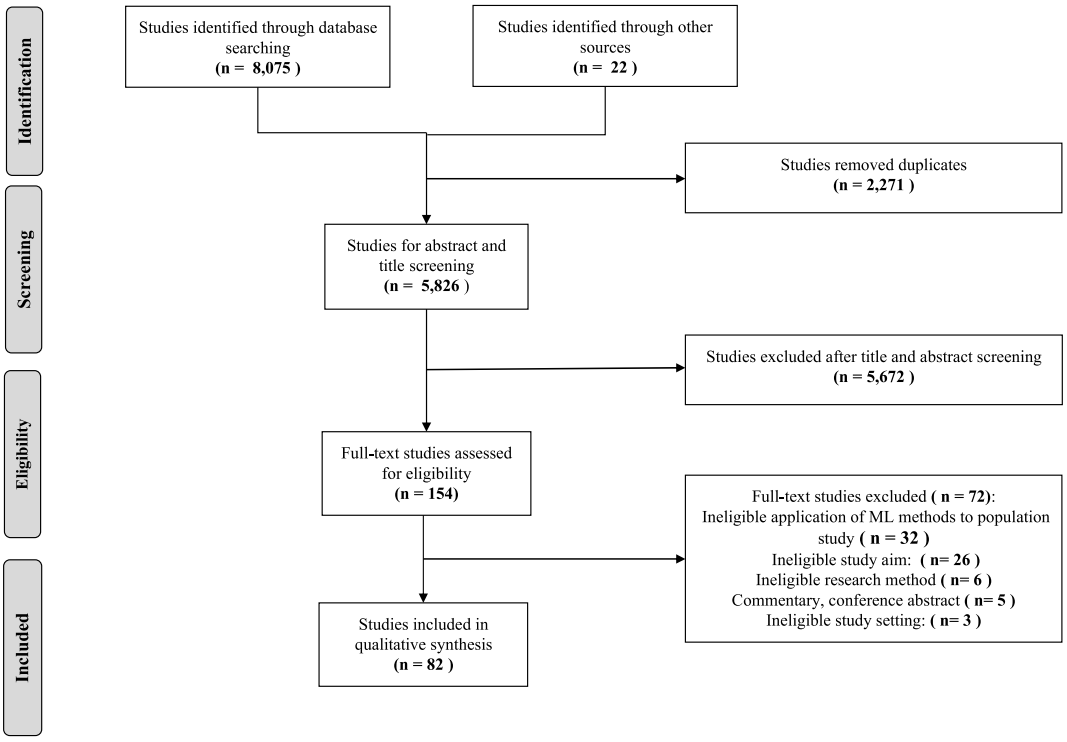

Fig. 1. Flowchart of literature search for scoping review on the traction of machine learning in the social determinants of health.

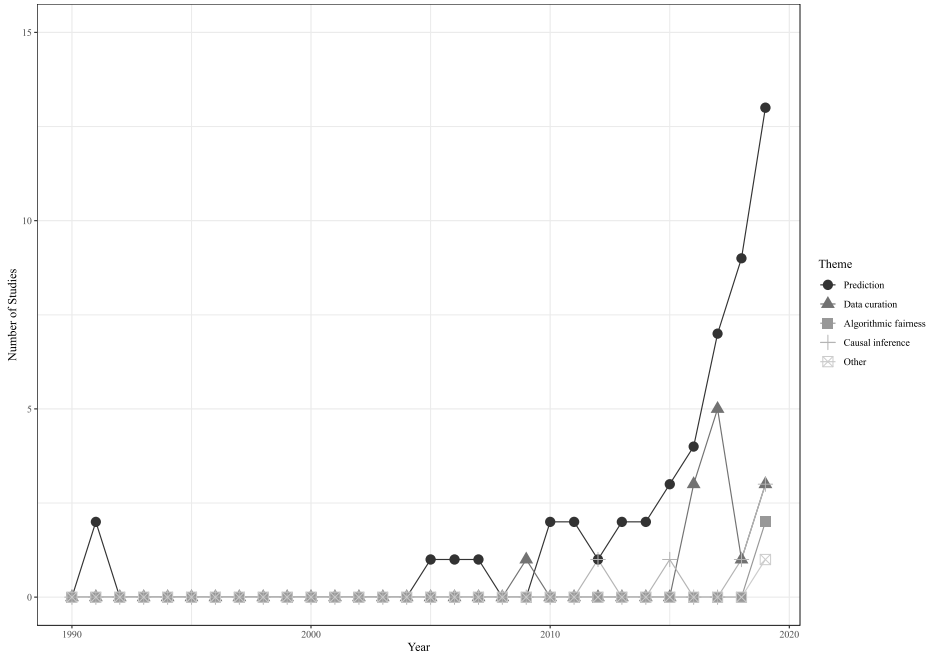

Fig. 2. Trend in the number of included studies by themes in machine learning.

consequences of health policy based on a combination of algorithmic fairness and ML. For example, Daoud et al. combined causal inference and algorithmic fairness to achieve both goals. Using distributive ethics—the principles of achieving fair allocation of scarce resources—their study shows how algorithmic thinking helps make trade-offs between the impact on economic and population health (Daoud et al., 2007). Similar approaches provide a new way to combine ML with substantive ethical theories for SDH.

The application of ML for causal inference is another future path that SDH researchers should engage in. As discussed in the Introduction, one type of combination of ML and causal inference is the estimation of heterogeneous treatment effects (Kusner et al., 2017). Researching how the effects of different SDH exposures on population health differ across subpopulations would enable researchers to move beyond estimating the average effects of SDH and gain a more granular view of health disparities (Athey et al., 2019). Lacking a method to find these granular combinations, scholars and policymakers may widen inequality in health outcomes if they deploy a one-size-fits-all type of intervention

(Berkman et al., 2014).

Scholars traditionally use parametric models with product terms to test the existence of impact (effect) heterogeneity (Kusner et al., 2017). Although these parametric models are useful for deductively testing the existence of a specific interaction formulated by the researcher, these deductive models do not test the wide range of possible effect heterogeneity among all the covariates and the exposure (Athey et al., 2019). However, ML for causal inference does this type of inductive test-and-search operation (Athey et al., 2019). Lewis and McCormick used Bayesian networks to identify determinants of health in low-income families with a number of effect modifiers (Lewis and McCormick, 2012). Consequently, when SDH researchers aim to test a specific effect heterogeneity, parametric models remain the best choice (Kusner et al., 2017). When SDH researchers aim to search for possible heterogeneity in the effects of exposure, ML offers the most flexible framework for finding and testing it.

Heterogeneous effect estimation is also helpful when resources are scarce for implementing health policy interventions (Kusner et al., 2017). ML can save resources by first identifying the subpopulations for which a particular intervention is most effective and then predicting the most useful timing of the intervention. For instance, generalized random forest was employed to quantify the effect heterogeneity and the average treatment effect of policies on child poverty (Daoud & Johansson, 2019). In addition, the Bayesian additive regression tree helped to capture the effect heterogeneity of armed conflict on children (Kraamwinkel et al., 2019). Consequently, ML for causal inference is a valuable way of employing these algorithms for SDH studies.

In addition to estimating heterogeneous effects, the ML-based estimation approach (e.g., SuperLearner) aids social epidemiologists in relaxing restrictive assumptions of parametric models (Van der Laan and Rose, 2011). Ahern et al. employed the SuperLearner to examine the relationship between childhood adversities and mental disorders by race and ethnicity (Ahern et al., 2016). By using an ensemble—a bundle of many different ML algorithms—a scholar avoids committing to one functional form, often a linear model, and thus achieves a more robust estimation. Such ensembles also help avoid researcher discretion (e.g., p-hacking and subjective model specifications) and set SDH research on a more solid statistical foundation. Thus, ML for more robust causal inference is another application of ML that has large potential; yet, as our review shows, it is underutilized (Ahern et al., 2016).

Lastly, ML for data curation provides an exciting path for SDH researchers to produce new data (Hastie et al., 2009). From our review, we found that most SDH articles use surveys as the main source for their analysis. While surveys remain an essential data source, text archives, satellite images, or even audio channels offer a new way to extract information from previously inaccessible sources. For example, Conway et al. used electronic health records to measure social risk factors (Conway et al., 2019), and Crossley et al. used natural language processing to determine the health literacy of diabetes patients (Crossley et al., 2020). Given the increasingly rapid rate of digitization of all aspects of society and health care, we expect that many new data sources will be available for SDH researchers. Researchers that know how to capitalize on these new sources for SDH will likely have a research edge over researchers using only traditional methods and data sources.

In summary, the 82 studies identified by this scoping review exemplify the various ways ML can contribute to the development of social-epidemiological research. Although commonly used methods will continue to play a pivotal role in the social-epidemiological toolbox, the role of ML will likely increase.

Limitations

This scoping review has four limitations. First, "social determinants of health" include a wide range of definitions (Berkman et al., 2014; Catalyst, 2017; Marmot et al., 2008). Although we went through several iterations of relevant terms to capture and define the social determinants of health (see Appendix for detailed information on our search strategy), some relevant studies may still have been missed. Second, a similar definitional limitation applies to "machine learning." As new algorithms are continuously being developed under new brands, our ML search list might not have captured all relevant articles. Third, because our search required that an article uses at least one term from both the ML and SDH lists in the title, abstract, or keywords, it did not include articles that use these terms only in the full text. For this reason, we complemented our bibliometric search with a manual (less strict) search to identify potential articles omitted by the bibliometric search. Fourth, we included only peer-reviewed papers, and including conference papers would have expanded the number of reviewed studies.

Conclusions

We conducted a scoping review to summarize how and to what extent ML has been used in the studies of SDH. Our review produced five conclusions, with their corresponding research opportunities.

First, of the vast number of yearly SDH publications, only 82 studies produced research using ML. While the number of ML-SDH studies is increasing, especially in the past three years, there is a clear opportunity to fuel this research further.

Second, as most articles used US data, one future direction is to explore public-health issues from other world regions. Comparing how the same problems (e.g., predicting child mortality) produced varying predictions in different populations is an intriguing problem to analyze—in ML, called transfer learning.

Third, most studies use tabular (structured) data. Another future possibility is to leverage other data formats, such as text, audio, and image data, since ML equips researchers with new techniques to measure health outcomes and their determinants from these non-conventional sources.

Fourth, most studies used cross-sectional surveys and employed supervised ML algorithms. Using longitudinal data is therefore an extension for SDH. Unpacking the temporal scope of an SDH issue is useful for not only using ML to predict (correlation) what next years health outcomes will be (e.g., mortality levels) but also using ML to evaluate (causation) the effect of public-health interventions on these outcomes.

Fifth, the majority of the 82 studies use ML for prediction, and thus, future studies using these approaches have an opportunity to innovate SDH further. Our four-category framework offers a mental model for researchers on how to use ML for SDH issues. Although there are no guarantees that ML will lead to better social-epidemiological research, the potential for innovation in SDH is evident from the several advances in artificial intelligence. Our framework provides a methodological basis for harnessing the predictive power of ML for SDH.

Ethical statement

There were no human subjects in this article and ethical approval is not applicable.

Acknowledgment

This study was supported by Grant-in-Aid for JSPS Fellows (JP20J01910).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssmph.2021.100836.

References

Abarca-Alvarez, F. J., Reinoso-Bellido, R., & Campos-Sánchez, F. S. (2019). Decision model for predicting social vulnerability using artificial intelligence. *ISPRS International Journal of Geo-Information, 8*(12), 575.

Abirami, M., et al. (2020). A classification model to predict onset of smoking and drinking habits based on socio-economic and sociocultural factors. *J Amb Intel Hum Comp*, 1–9.

Adeyinka, D. A., Olakunde, B. O., & Muhajarine, N. (2019). Evidence of health inequity in child survival: Spatial and bayesian network analyses of stillbirth rates in 194 countries. *Scientific Reports, 9*(1), 1–11.

Ahern, J., Balzer, L., & Galea, S. (2015). The roles of outlet density and norms in alcohol use disorder. *Drug and Alcohol Dependence, 151*, 144–150.

Ahern, J., et al. (2016). Racial/ethnic differences in the role of childhood adversities for mental disorders among a nationally representative sample of adolescents. *Epidemiology, 27*(5), 697–704.

Allali, F., et al. (2010). Educational level and osteoporosis risk in postmenopausal Moroccan women: A classification tree analysis. *Clinical Rheumatology, 29*(11), 1269–1275.

Ambriola Oku, A. Y., et al. (2020). Potential confounders in the analysis of Brazilian adolescent's health: A combination of machine learning and graph theory. *International Journal of Environmental Research and Public Health, 17*(1), 90.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics, 47*(2), 1148–1178.

Bai, Z., et al. (2020). Relationship between individual social capital and functional ability among older people in anhui province, China. *International Journal of Environmental Research and Public Health, 17*(8), 2775.

Barnes, G. M., Welte, J. W., & Dintcheff, B. (1991). Drinking among subgroups in the adult population of New York state: A classification analysis using CART. *Journal of Studies on Alcohol, 52*(4), 338–344.

Bastaminia, A., Rezaei, M. R., & Saraei, M. H. (2017). *Evaluating the components of social and economic resilience: After two large earthquake disasters rudbar 1990 and bam 2003.* Jàmbá. *Journal of Disaster Risk Studies, 9*(1), 1–12.

Basu, S., & King, A. C. (2013). Disability and chronic disease among older adults in India: Detecting vulnerable populations through the WHO SAGE study. *American Journal of Epidemiology, 178*(11), 1620–1628.

Basu, S., & Narayanaswamy, R. (2019). A prediction model for uncontrolled type 2 diabetes mellitus incorporating area-level social determinants of health. *Medical Care, 57*(8), 592–600.

Bellavia, A., et al. (2020). The use of logic regression in epidemiologic studies to investigate multiple binary exposures: An example of occupation history and amyotrophic lateral sclerosis. *Epidemiologic Methods*, 1((open-issue)).

Bentley, R., et al. (2018). The impact of social housing on mental health: Longitudinal analyses using marginal structural models and machine learning-generated weights. *International Journal of Epidemiology, 47*(5), 1414–1422.

Berkman, L. F., Kawachi, I., & Glymour, M. M. (2014). *Social epidemiology*. Oxford University Press.

Berkowitz, S. A., et al. (2019). Association between access to social service resources and cardiometabolic risk factors: A machine learning and multilevel modeling analysis. *BMJ Open, 9*(3). e025281.

Bhavsar, N. A., et al. (2018). Value of neighborhood socioeconomic status in predicting risk of outcomes in studies that use electronic health record data. *JAMA Netw. Open, 1*(5). e182716-e182716.

Bi, Q., et al. (2019a). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology, 188*(12), 2222–2239.

Bi, Q., et al. (2019b). What is machine learning? *A primer for the epidemiologist, 188*(12), 2222–2239.

Bodnar, L. M., et al. (2020). Machine learning as a strategy to account for dietary synergy: An illustration based on dietary intake and adverse pregnancy outcomes. *American Journal of Clinical Nutrition, 111*(6), 1235–1243.

Boerstler, H., & de Figueiredo, J. M. (1991). Prediction of use of psychiatric services: Application of the CART algorithm. *Journal of Mental Health Administration, 18*(1), 27–34.

Braveman, P., Egerter, S., & Williams, D. R. (2011). The social determinants of health: Coming of age. *Annual Review of Public Health, 32*, 381–398.

Brondeel, R., Pannier, B., & Chaix, B. (2016). Associations of socioeconomic status with transport-related physical activity: Combining a household travel survey and accelerometer data using random forests. *J Transp Health, 3*(3), 287–296.

Cairney, J., et al. (2014). Exploring the social determinants of mental health service use using intersectionality theory and CART analysis. *Journal of Epidemiology & Community Health, 68*(2), 145–150.

Catalyst, N. (2017). Social determinants of health (SDOH). *NEJM Catalyst, 3*(6).

Chan, J., & Bennett Moses, L. (2016). Is big data challenging criminology? *Theoretical Criminology, 20*(1), 21–39.

Chiavegatto Filho, A. D. P., et al. (2018). Overachieving municipalities in public health: A machine-learning approach. *Epidemiology, 29*(6), 836–840.

Choi, S. B., et al. (2018). Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. *Journal of Affective Disorders, 231*, 8–14.

Choi, K. W., et al. (2019). Mapping a syndemic of psychosocial risks during pregnancy using network analysis. *International Journal of Behavioral Medicine, 26*(2), 207–216.

Choi, S. K., Fram, M. S., & Frongillo, E. A. (2017). Very low food security in US households is predicted by complex patterns of health, economics, and service participation. *Journal of Nutrition, 147*(10), 1992–2000.

Conway, M., et al. (2019). Moonstone: A novel natural language processing system for inferring social risk from clinical narratives. *Journal of Biomedical Semantics, 10*(1), 1–10.

Cramb, S. M., Mengersen, K. L., & Baade, P. D. (2011). Identification of area-level influences on regions of high cancer incidence in queensland, Australia: A classification tree approach. *BMC Cancer, 11*(1), 311.

Crossley, S. A., et al. (2020). Developing and testing automatic models of patient communicative health literacy using linguistic features: Findings from the ECLIPPSE study. *Health Communication*, 1–11.

Daoud, A., Herlitz, A., & Subramanian, S. (2007). arXiv. *Combining distributed ethics and causal Inference to make trade-offs between austerity and population health* (Vol. 15550), 2020.

Daoud, A., & Johansson, F. (2019). *Estimating treatment heterogeneity of international monetary fund programs on child poverty with generalized random forest.* SocArXiv.

Daoud, A., Kim, R., & Subramanian, S. (2019). Predicting women's height from their socioeconomic status: A machine learning approach. *Social Science & Medicine, 238*, 112486.

Darvishi, E., et al. (2017). Prediction effects of personal, psychosocial, and occupational risk factors on low back pain severity using artificial neural networks approach in industrial workers. *J Manipulative Physiol Ther, 40*(7), 486–493.

DiGuiseppi, G. T., et al. (2020). Predictors of adolescents' first episode of homelessness following substance use treatment. *Journal of Adolescent Health, 66*(4), 408–415. https://doi.org/10.1016/j.jadohealth.2019.11.312

Duncan, D. T., & Kawachi, I. (2018). *Neighborhoods and health*. Oxford, UK: Oxford University Press.

Engchuan, W., et al. (2019). Sociodemographic indicators of health status using a machine learning approach and data from the English Longitudinal Study of Aging (ELSA). *Medical Science Monitor, 25*, 1994.

Fan, W., et al. (2019). Research and prediction of opioid crisis based on BP neural network and Markov chain. *AIMS MATH, 4*(5), 1357.

Filikov, A., et al. (2020). Use of Stratified Cascade Learning to predict hospitalization risk with only socioeconomic factors. *Journal of Biomedical Informatics, 104*, 103393.

Fiscella, K., & Tancredi, D. (2008). Socioeconomic status and coronary heart disease risk prediction. *Journal of the American Medical Association, 300*(22), 2666–2668.

Fiscella, K., Tancredi, D., & Franks, P. (2009). Adding socioeconomic status to Framingham scoring to reduce disparities in coronary risk assessment. *American Heart Journal, 157*(6), 988–994.

Franks, P., et al. (2010). Including socioeconomic status in coronary heart disease risk estimation. *The Annals of Family Medicine, 8*(5), 447–453.

Friel, S., Newell, J., & Kelleher, C. (2005). Who eats four or more servings of fruit and vegetables per day? Multivariate classification tree analysis of data from the 1998 survey of lifestyle, attitudes and nutrition in the republic of Ireland. *Public Health Nutrition, 8*(2), 159–169.

Fu, S.-Y. K., et al. (2007). The relationship between culture, attitude, social networks and quality of life in midlife Australian and Taiwanese citizens. *Maturitas, 58*(3), 285–295.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics, 10*, 524.

Goin, D. E., et al. (2020). Mediation of firearm violence and preterm birth by pregnancy complications and health behaviors: Addressing structural and post-exposure confounding. *American Journal of Epidemiology, 189*(8), 820–831. https://doi.org/10.1093/aje/kwaa046

Goin, D. E., Rudolph, K. E., & Ahern, J. (2018). Predictors of firearm violence in urban communities: A machine-learning approach. *Health & Place, 51*, 61–67.

Gray, J. C., Schvey, N. A., & Tanofsky-Kraff, M. (2020). Demographic, psychological, behavioral, and cognitive correlates of BMI in youth: Findings from the Adolescent Brain Cognitive Development (ABCD) study. *Psychol Med, 50*(9), 1539–1547.

Hamad, R., et al. (2019). Comparing demographic and health characteristics of new and existing SNAP recipients: Application of a machine learning algorithm. *American Journal of Clinical Nutrition, 109*(4), 1164–1172.

Hammond, G., et al. (2020). Social determinants of health improve predictive accuracy of clinical risk models for cardiovascular hospitalization. In *Annual cost, and death*. Circ Cardiovasc Qual Outcomes.

Handley, T. E., et al. (2014). Predictors of suicidal ideation in older people: A decision tree analysis. *American Journal of Geriatric Psychiatry, 22*(11), 1325–1335.

Hanson, H. A., et al. (2019). The relative importance of race compared to health care and social factors in predicting prostate cancer mortality: A random forest approach. *The Journal of Urology, 202*(6), 1209–1216.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Havranek, E. P., et al. (2015). Social determinants of risk and outcomes for cardiovascular disease: A scientific statement from the American heart association. *Circulation, 132*(9), 873–898.

Hernán, M. A., et al. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology, 155*(2), 176–184.

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if* (Vol. 2020). Boca Raton: Chapman & Hill/CRC.

Herrera-Ibatá, D. M., et al. (2015). Mapping chemical structure-activity information of HAART-drug cocktails over complex networks of AIDS epidemiology and socioeconomic data of US counties. *Biosystems, 132*, 20–34.

Higgins, S. T., et al. (2016). Co-occurring risk factors for current cigarette smoking in a US nationally representative sample. *Preventive Medicine, 92*, 110–117.

Higgins, S. T., et al. (2017). Socioeconomic disadvantage and other risk factors for using higher-nicotine/tar-yield (regular full-flavor) cigarettes. *Nicotine & Tobacco Research, 19*(12), 1425–1433.

Higgins, J., et al. (2020). *Cochrane Handbook for systematic reviews of interventions*. version 6.1 (updated September 2020). Cochrane.
Hu, W., Mengersen, K., & Tong, S. (2009). Spatial analysis of notified cryptosporidiosis infections in Brisbane, Australia. *Annals of Epidemiology, 19*(12), 900–907.
Hu, W., Mengersen, K., & Tong, S. (2010). Risk factor analysis and spatiotemporal CART model of cryptosporidiosis in Queensland, Australia. *BMC Infectious Diseases, 10*(1), 311.
Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics, 7*(1), 443–470.
Inkster, B., et al. (2016). A decade into facebook: Where is psychiatry in the digital age? *Lancet Psychiatry, 3*(11), 1087–1090.
Jamei, M., et al. (2017). Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PloS One, 12*(7). e0181173.
Jean, N., et al. (2016). Combining satellite imagery and machine learning to predict poverty. *Science, 353*(6301), 790–794.
Jolobe, O. (2009). Adding socioeconomic status to Framingham scoring might also improve stroke risk evaluation in young adults with hypertension. *American Heart Journal, 158*(3), e35. author reply e37.
Jones, M., & Huh, J. (2014). Toward a multidimensional understanding of residential neighborhood: A latent profile analysis of los angeles neighborhoods and longitudinal adult excess weight. *Health & Place, 27*, 134–141.
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.
Kanerva, N., et al. (2018). Suitability of random forest analysis for epidemiological research: Exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design. *Scandinavian Journal of Public Health, 46*(5), 557–564.
Kleinberg, J., et al. (2018a). Human decisions and machine predictions. *Quarterly Journal of Economics, 133*(1), 237–293.
Kleinberg, J., et al. (2018b). Algorithmic fairness. In *Aea papers and proceedings*.
Kosinski, M., et al. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70*(6), 543.
Kraamwinkel, N., et al. (2019). The influence of maternal agency on severe child undernutrition in conflict-ridden Nigeria: Modeling heterogeneous treatment effects with machine learning. *PloS One, 14*(1). e0208937.
Kreatsoulas, C., & Anand, S. S. (2010). The impact of social determinants on cardiovascular disease. *Canadian Journal of Cardiology, 26*, 8C–13C.
Krieger, N. (2011). *Epidemiology and the people's health: Theory and context*. Oxford University Press.
Künzel, S. R., et al. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences, 116*(10), 4156–4165.
Kusner, M. J., et al. (2017). Counterfactual fairness. In *Advances in neural information processing systems*.
Larkin, A., & Hystad, P. (2019). Evaluating street view exposure measures of visible green space for health research. *Journal of Exposure Science and Environmental Epidemiology, 29*(4), 447–456.
Leach, H. J., et al. (2016). An exploratory decision tree analysis to predict cardiovascular disease risk in African American women. *Health Psychology, 35*(4), 397.
Leal, C., et al. (2012). Multicollinearity in associations between multiple environmental features and body weight and abdominal fat: Using matching techniques to assess whether the associations are separable. *American Journal of Epidemiology, 175*(11), 1152–1162.
Lefebvre, C., et al. (2019). Searching for and selecting studies. *Cochrane Handbook for systematic reviews of interventions*, 67–107.
Lewis, F. I., & McCormick, B. J. (2012). Revealing the complexity of health determinants in resource-poor settings. *American Journal of Epidemiology, 176*(11), 1051–1059.
Li, Y., et al. (2019). Unhealthy behaviors, prevention measures, and neighborhood cardiovascular health: A machine learning approach. *Journal of Public Health Management and Practice, 25*(1), E25–E28.
Lipton, Z. C. (2018). The mythos of model interpretability. *arXiv preprint arXiv: 1606.03490, 16*(3), 31–57.
Loftus, J. R., et al. (2018). *Causal reasoning for algorithmic fairness*. arXiv preprint arXiv: .05859.
Luo, W., et al. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research, 18*(12), e323.
Maksimov, S. A., & Artamonova, G. V. (2013). Modeling of arterial hypertension's risk in occupational groups. *Russian Open Medical Journal, 2*(1), 0104-0104.
Marmot, M., et al. (2008). Closing the gap in a generation: Health equity through action on the social determinants of health. *Lancet, 372*(9650), 1661–1669.
Matthay, E. C., et al. (2019). Firearm and nonfirearm violence after operation peacemaker fellowship in richmond, California, 1996–2016. *American Journal of Public Health, 109*(11), 1605–1611.
Matusik, S., Aska-Mierzejewska, T., & Chrzanowska, M. (2011). Socioeconomic determinants of menarche in rural polish girls using the decision trees method. *Journal of Biosocial Science, 43*(3), 257.
Meng, H.-W., et al. (2017). National substance use patterns on Twitter. *PloS One, 12*(11). e0187691.
Miotto, R., et al. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics, 19*(6), 1236–1246.
Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology, 45*, 27–45. https://doi.org/10.1146/annurev-soc-073117-041106
Mooney, S. J., et al. (2017). *Contextual correlates of physical activity among older adults: A neighborhood environment-wide association study (NE-WAS)*. AACR.
Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *The Journal of Economic Perspectives, 31*(2), 87–106.
Nayak, S., et al. (2016). Characteristics associated with self-rated health in the CARDIA study: Contextualising health determinants by income group. *Prev Med Rep, 4*, 199–208.
Nguyen, Q. C., et al. (2016a). Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. *JMIR Public Health Surveill, 2*(2), e158.
Nguyen, Q. C., et al. (2016b). Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography, 73*, 77–88.
Nguyen, Q., et al. (2017a). Neighborhood looking glass: 360 degree automated characterization of the built environment for neighborhood effects research. In *APHA 2017 annual meeting & expo (nov. 4-Nov. 8)*. American Public Health Association.
Nguyen, Q. C., et al. (2017b). Geotagged US tweets as predictors of county-level health outcomes, 2015–2016. *American Journal of Public Health, 107*(11), 1776–1782.
Nguyen, Q., et al. (2017c). Social media indicators of the food environment and state health outcomes. *Public Health, 148*, 120–128.
Nollen, N. L., et al. (2016). Adult cigarette smokers at highest risk for concurrent alternative tobacco product use among a racially/ethnically and socioeconomically diverse sample. *Nicotine & Tobacco Research, 18*(4), 386–394.
Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.
Özge, C., et al. (2006). Which sociodemographic factors are important on smoking behaviour of high school students? The contribution of classification and regression tree methodology in a broad epidemiological survey. *Postgraduate Medical Journal, 82*(970), 532–541.
Penny, K. I., & Smith, G. D. (2012). The use of data-mining to identify indicators of health-related quality of life in patients with irritable bowel syndrome. *Journal of Clinical Nursing, 21*(19pt20), 2761–2771.
Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.
Platt, J. M., et al. (2018). Targeted estimation of the relationship between childhood adversity and fluid intelligence in a US population sample of adolescents. *American Journal of Epidemiology, 187*(7), 1456–1466.
Prayaga, R. B., et al. (2019). Impact of social determinants of health and demographics on refill requests by Medicare patients using a conversational artificial intelligence text messaging solution: Cross-sectional study. *JMIR mHealth uHealth, 7*(11). e15771.
Robson, B., & Boray, S. (2019). Studies in the use of data mining, prediction algorithms, and a universal exchange and inference language in the analysis of socioeconomic health data. *Computers in Biology and Medicine, 112*, 103369.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688.
Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology, 185*(1), 65–73.
Seligman, B., Tuljapurkar, S., & Rehkopf, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. *SSM Popul Health, 4*, 95–99.
Shimony-Kanat, S., & Benbenishty, J. (2018). Age, ethnicity, and socioeconomic factors impacting infant and toddler fall-related trauma. *Pediatric Emergency Care, 34*(10), 696–701.
Shin, E. K., et al. (2018). Sociomarkers and biomarkers: Predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *NPJ Digit Med, 1*(1), 1–5.
Sow, B., et al. (2019). Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. *Informatics for Health and Social Care*, 1–13.
Suel, E., et al. (2019). Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports, 9*(1), 1–10.
Torres, J. M., et al. (2018). Longitudinal associations between having an adult child migrant and depressive symptoms among older adults in the Mexican Health and Aging Study. *International Journal of Epidemiology, 47*(5), 1432–1442.
Torres, J. M., et al. (2020). Adult child US migration status and cognitive decline among older parents who remain in Mexico. *American Journal of Epidemiology, 189*(8), 761–769. https://doi.org/10.1093/aje/kwz277
Tricco, A. C., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473.
VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology, 34*(3), 211–219.
Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media.
Wiemken, T. L., & Kelley, R. R. (2020). Machine learning in epidemiology and health outcomes research. *Annual Review of Public Health, 41*, 21–36.
Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science, 20*, 529–544.
Yu, Q., et al. (2017). Exploring racial disparity in obesity: A mediation analysis considering geo-coded environmental factors. *Spat Spatio-temporal Epidemiol, 21*, 13–23.