IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 25, 2021, accepted February 14, 2022, date of publication February 22, 2022, date of current version March 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3153482

Multidimensional Population Health Modeling: A Data-Driven Multivariate Statistical Learning Approach

ZHIYUAN WEI, ADIL BARAN NARIN, AND SAYANTI MUKHERJEE
Department of Industrial and Systems Engineering, University at Buffalo—The State University of New York, Buffalo, NY 14260, USA

Corresponding author: Sayanti Mukherjee (sayantim@buffalo.edu)

ABSTRACT Population health is multidimensional in nature, having complex relationships with the various health determinants. However, most previous studies investigate a single dimension of population health using linear models, failing to capture the nonlinearity in the data and interdependence of multiple dimensions in health outcomes. In this paper, we propose a data-driven multivariate statistical learning approach to simultaneously model various aspects of population health—characterizing the length and quality of life—as a function of health behaviors, clinical care, socioeconomic factors, physical environment, and demographics. We also propose a novel percentile-based variable selection for multivariate regression, without compromising the model's generalization performance. We demonstrate the applicability of our proposed data-driven methodological framework using the New York State as a case study. Leveraging cross-validation techniques and statistical hypothesis tests, the results indicate that multivariate tree boosting method outperforms the traditionally-used univariate linear regression model and random forest in modeling multidimensional population health. The variable importance heat-map illustrates the relative influence of the key health determinants on the various dimensions of population health. Partial dependence plots are used to quantify the marginal effects and the nonlinear relationships between the health outcomes and health inputs. Our results show that teen birth rate is strongly associated with both length of life (e.g., child mortality) and quality of life (e.g., physically unhealthy days). Socioeconomic status is the key indicator to predict child and infant mortality. Our proposed framework can be used as a decision support tool for accurately assessing and predicting multivariate population health.

INDEX TERMS Data-driven framework, multivariate tree boosting, multidimensional population health, variable selection.

I. INTRODUCTION

Human health and wellbeing is the key to a thriving and equitable society [1]. Population health is often conceptualized as the health status and health outcomes within a group of people, instead of considering the individual health at a time [2]. In the last decade, efforts have been made to enhance the overall population health by not only improving the overall or mean population health of a community, but also eliminating health disparities within that population [3]–[5]. Recently, the "Healthy People 2030", developed by the U.S. Department of Health and Human Services Advisory Committee on National Health Promotion and Disease Prevention for 2030, sets data-driven national objectives to improve health and well-being over the next decade [1]. Even though increasing attention has been paid to enhancing population health, several challenges still exist towards quantitatively assessing population health.

Most importantly, population health has a multidimensional construct, characterized by length of life (a.k.a., mortality) and quality of life (a.k.a., morbidity) [5], [6]. Although a number of studies focus on assessing population health, most of them consider only one of the dimensions such as health-related quality of life [7], [8], life expectancy [9] or the mortality rate [10] in their analysis. Such a siloed approach, however, fails to provide the comprehensive and holistic picture of the health outcomes within a population.

In addition, the recent studies show that the relationships between health outcomes and health variables such as social

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

and economic factors are highly complex and nonlinear [11]–[13]. However, there is a lack of systematic quantitative approach in modeling the complex nonlinear characteristics of population health [14], [15]. The traditionally-used linear models often fall short in accurately estimating the health outcomes [5], [6], [16], thus underestimating the risks to population health.

Moreover, population health is affected by a wide range of factors including health behaviors, clinical care, social and economic conditions, physical environment and demographics [16], [17]. Presently, a weighted average technique is used to account for the relative contribution of such factors on population health. However, the weighting mechanism is simplistic where the weights of the factors in contribution to the health outcomes are often assumed to be equally distributed, or are assigned based on expert opinions [4], [18], [19]. Recognizing the complex interactions of the health determinants to population health, such traditional weighing mechanisms could lead to a biased estimate of the population health, leading to sub-optimal decision making.

To address the above-mentioned challenges and gaps, in this paper, we propose to develop a **novel data-driven multivariate framework** to model the population health at county-level leveraging the advanced statistical learning theory, while considering its multidimensional construct. Specifically, in this proposed framework, nine different dimensions of population health outcomes characterizing length of life and quality of life are considered. Population health of a county is then modeled as function of a wide range of factors including health behaviors, clinical care, social and economic characteristics, physical environment and demographics. A suite of statistical learning models including linear regression and non-linear ensemble tree-based models along with their multivariate and univariate versions are implemented to evaluate the population health. Additionally, we propose a percentile-based variable selection method for multivariate analysis without compromising the generalization performance of the model. Finally, we present the visualization tools including a variable importance heat-map and partial dependence plots of the key predictors to explain the underlying relationships of the important variables with the population health outcomes.

We implemented our proposed framework for the state of New York (NY) as a case study because [20]:

- NY state's population is the fourth-largest in the United States;
- NY state has one of the most diverse demographic characteristics;
- the minority populations in NY state increased as a percentage of the total population (39.8% in 2006 to 44.5% in 2016);
- NY population is aging—e.g., the median age of NY population increased from 37.4 years in 2006 to 23.4 in 2016; in fact, the percentage of population aged 65 and over increased to 15.3% from 13.1%, while that aged 19 and younger decreased to 23.8% from 26.3%.

Therefore, NY presents a unique case to study the populations health and its key determinants. Although we present the applicability of our proposed framework for the state of NY, this methodological framework is generalized enough that can be implemented to any other states in the US or states of other nations, provided data are available.

The major contributions of our proposed data-driven framework are two-fold. First, it provides a systematic approach to develop and select a model that can best capture the complex relationships of the various determinants of health with the multidimensional population health. Second, it helps to identify and evaluate the key health factors affecting the overall health outcomes, which can help policymakers, community leaders, and public health officials in informed decision making towards improving overall population health.

The rest of this paper is organized as follows. Section II presents the literature review regarding population health assessment, highlighting the gaps in the body of knowledge. Section III presents the collection and preprocessing of the data. Section IV introduces our proposed research methodology. Section V provides the results including model comparison and selection, variable selection, and statistical inference. Finally, the discussion and conclusion are exhibited in Section VI and VII respectively. The Appendix provides additional information for variable descriptions and results on hypothesis testing.

II. LITERATURE REVIEW

The concept of multidimensional population health provides a holistic picture of integrating all major aspects of health outcomes and various health determinants. This concept recently has gained more and more attention to better understand and improve overall population health [21], [22]. Generally speaking, length of life and quality of life describe different aspects of population health, and a combination of these two characteristics can provide more comprehensive and holistic analysis of health outcomes of the population [5], [6].

Length of life is one of the critical dimensions of population health [5], [16]. Various studies indicated that length of life can be represented by the overall life expectancy that measures how long people can live on average [9], [23]. In recent years, the average life expectancy in the United States has been declining [24], [25], even falling behind other wealthy countries [10]. Besides life expectancy, mortality rate is another metrics for length of life. Preston *et al.* found that different age groups exhibit different mortality rates. For example, in 2017 the US witnessed a sharp increase in mortality rate among adults aged 20 to 34, in contrast to a decline in mortality rates among people aged 85 and up, compared to other similar European countries [10].

Quality of life, on the other hand, indicates the extent to which people feel physically, mentally, socially, and emotionally healthy [5], [16]. Quality of life, therefore, is measured by various characteristics of a population such as number of unhealthy days, prevalence of diabetes within the population, low birth weight, psychological distress, etc. However, most of the existing studies have focused on evaluating a single measure depicting the quality of life. For instance, Slabaugh *et al.* used number of healthy days to measure quality of life in order to understand both mental and physical well-being of a population [26]. Prevalence of diabetes within a population, as another measure of quality of life, has been studied in various existing literature [27]–[29]. The epidemic of diabetes is one of the most prevalent and costly chronic diseases in the US, which adversely affects the quality of life in majority of the population [30], [31]. Longer duration of any type of diabetes was found to be correlated with poor quality of life [27], [32]. Some other studies included low birthweight [33] and psychological distress [34] in the evaluation of quality of life.

In addition to being multidimensional in nature, recent studies demonstrate that the associations between population health and the various health determinants are complex and exhibit nonlinear characteristics, and that the linear models are inadequate to capture such nonlinear behaviors [11]–[13]. However, most of the existing studies in the population health assessment domain have used linear models to analyze the relationships between the health determinants and the population health outcomes. For example, to determine the relationships between health-related behaviors and the health outcomes of people with chronic conditions, Hinnell *et al.* implemented Logistic regression model. The authors found that physical inactivity was significantly correlated with epilepsy disorder [35]. Another study by Khedmat *et al.*, which aimed to predict health-related quality of life, leveraged ordinal logistic regression to characterize the socioeconomic and demographic conditions associated with the perceived physical and mental health conditions of the participants [7]. Veen *et al.*, in a different study, used linear regression analysis to investigate the degree of variance in quality of life that contributed to the risk of facial palsy [8]. In another study, Michel *et al.* examined the disparities of age and gender in the quality of life among teenagers using multilevel regression analyses. The authors found that girls reported a significant deterioration in quality of life than boys with increasing age [36]. Thus, although linear models are predominantly used to model population health outcomes because of their easier interpretability and low computational cost, the theory of these models are founded on a set of rigid assumptions regarding the underlying distribution of the data such as linearity and normality [37]. However, in reality such assumptions often do not hold, leading to poor generalization performance of the model [38]. Recently, data-driven techniques are developed and demonstrated to have potential values in discovering the complex nonlinear relationships in the field of public health [39], [40].

Population health is also affected by a number of factors. For example, socioeconomic determinants of health including unemployment and income are found to be strongly associated with the population health [41]. Moriarty *et al.* showed that people from poor socioeconomic condition are likely to suffer from more unhealthy days, impacting their quality of life—an important dimension of population health [42]. Lin *et al.* separately investigated the relationships between demographic backgrounds, and the physical and mental wellbeing of US adults aged over 65 years. The authors found that people living in the rural areas are likely to suffer more from poor physical health rather than mental illness [43]. Wu *et al.* examined a range of demographic and socioeconomic factors in relation to the mental wellbeing of adolescents. The authors pointed out that African-American youths, girls, and children from low income family are the most vulnerable to mental illness [44]. Physical environment also plays a critical role in shaping population health. Samet *et al.* investigated the relationships between air pollutants and mortality rates in the major metropolitan areas in the US. The authors found that the level of PM_{10} was positively linked to a higher risk of death from all causes [45]. Zanobetti *et al.* examined the association between air temperature and mortality rate, and concluded that a 10°F increase in air temperature was associated with an 1.8% increase in mortality rate [46]. Thus, although it is evident that population health is dependent on a multitude of factors, most of the studies focused on understanding the association of a particular type of a determinant/factor on population health, neglecting their complex interactions and the holistic effects of such factors on the population health.

It is noteworthy that some of the recent studies have developed the population health assessment framework, integrating all the different types of health determinants to the population health outcomes. For example, Kindig *et al.* proposed a framework to assess the population health outcomes by considering five determinants including medical care, individual behaviors, social environment, physical environment and genetics. However, the authors assumed that all the determinants are equally weighted, i.e., they make equal contributions to the health outcomes [4]. Similarly, the County Health Ranking (CHR) framework developed by the University of Wisconsin Population Health Institute and the Robert Wood Johnson Foundation, assumes a fixed weight for all the health determinants. The CHR framework delineates four main categories of health determinants including health behaviors, clinical care, socioeconomic factors, and physical environment [16], [17], where the weight of each health determinant is determined by expert knowledge, and is fixed for all counties [16]. However, the interactions of health determinants to population health are highly complex [47], and the simple weighing mechanisms can lead to sub-optimal decision making where the decision makers cannot adequately prioritize the key risk factors affecting the population health that needs attention.

To summarize, the gaps in the current body of literature aiming to evaluate the population health are as follows:

- the state-of-the-art modeling approach undermines the importance of the multidimensional aspect of the population health;

- the traditionally-used linear models cannot capture the nonlinearities and the complex interactions between the various factors and the population health, thereby underestimating the risk factors of population health;
- most of the existing studies use silo-ed approaches, i.e., focus on the effects of a specific type of a determinant, instead of leveraging a holistic approach that can consider a wide range of predictors for the analysis; and,
- finally, few of the studies that have considered a range of various factors in assessing population health, used a fixed or equal weighting system assuming that all the predictors have equal contribution towards estimating population health, which is often times not true.

To the best of our knowledge, no previous studies have simultaneously modeled all major dimensions of population health as a predictive function of a wide range of health determinants. Therefore, to address this research gap, we propose a systematic data-driven multivariate approach to model the multiple dimensions of population health (such as quality of life and length of life) as a function of various determinants of health, leveraging advanced statistical learning algorithms. Our framework can not only provide a robust way to select the optimal model for understanding and predicting the population health outcomes, but also help decision makers identify and quantify the focal health factors influencing population health.

III. DATA COLLECTION AND PRE-PROCESSING

In this section, we describe the detailed data preprocessing procedures for response and input variables used in this study. The population health data are collected at the county-level for each of 62 counties in New York State during 2010 to 2020. The mean of population per county is 0.3 million, and the total population across all counties is 19.5 million (2020 U.S. Census).

A. DATA PREPROCESSING

The health-related population data used in this study are provided by County Health Rankings & Roadmaps (CHR) program [48], and weather information are collected through National Center for Environmental Information (NOAA) [49]. Then, we conducted two following steps to process the collected raw data.

The first step is to address missing values, which is a common issue in data-driven analytics [50]. We found that few health features (e.g., income inequality, insufficient sleep) in NY had more than half of data missing in our study period 2010–2020. This was mostly due to small reported sample size or other uncertainties [5]. To avoid introducing bias in the analysis, we only kept the variables that had at least 50% of non-missing observations. For the variable that had less than or equal to 50% of the observations missing, we imputed the missing values leveraging the multivariate imputation by chained equations (MICE) technique using the R package *'mice'* [51].

The second step is to detect the correlation between variables. The Pearson correlation coefficient (denoted as ρ) between two variables in each category of predictors and responses is calculated. If two variables are demonstrated to exceed a correlation of 0.90, the one with higher percentage of missing values is excluded from our analysis. For example, we found that the premature age-adjusted mortality is highly correlated with years of potential life lost (YPLL) with $\rho = 0.93$, and the proportion of missing data for the premature age-adjusted mortality is 27% compared to 0% missing values in YPLL. Then, the YPLL is kept in the analysis. To identify the highly correlated variables are essential for statistical inference to avoid "masking effect", which has been successfully used in previous research studies [52], [53].

B. RESPONSE VARIABLES

To capture the multifaceted length of life in a population, we included three measures of mortality, namely, years of potential life lost (YPLL), child mortality, and infant mortality. Here, YPLL can capture overall mortality of the population. Child and infant mortality can inform policymakers to design prevention strategies for meeting United Nations' Sustainable Development Goals by 2030 in the reduction of preventable deaths of newborns and children under age five [54].

To characterize the quality of life, six measures are used including poor or fair health, poor physical health days, poor mental health days, low birth weight, diabetes, and HIV prevalence. These variables have been used and validated by the CHR framework to assess the quality of life in population [5], [16]. Note that, CHR program also provides additional measures of quality of life including physical and mental distress among adults. Since these two variables have missing data for six years (i.e., over 50% of the data is missing), we did not include them in our study.

The summary statistics of the total of nine multivariate responses are displayed in Table 1. The description of all response variables is exhibited in Table 5 from Appendix A, and their dependencies and distributions are depicted in Fig. 13 from Appendix B.

C. INPUT VARIABLES

A wide range of input variables in relation to population health are analyzed in our study. These health factors are grouped into five categories: health behaviors, clinical care, social and economic factors, physical environment and demographics, based on the CHR framework [48]. Each variable has been validated by researchers and experts to satisfy several criteria such as: a) their importance to population health, b) their applicability to future population health, and c) availability and reliability of the indicators [5]. Additionally, we collected weather-related variables (precipitation, snowfall, etc.) from NOAA, and added to the category

TABLE 1. Summary statistics of the response variables.

Response Variables (Unit)	Mean (SD)	Max	Min
YPLL (years per 100k)	6141 (926)	8969	3925
Child mortality (per 100k)	45 (10)	75	22
Infant mortality (per 100k)	617 (131)	970	301
Poor or fair health (%)	14.4 (3.0)	28.7	6.3
Poor physical health days (per month)	3.6 (0.57)	6.5	2.2
Poor mental health days (per month)	3.6 (0.58)	5.7	1.6
Low birthweight (%)	7.3 (0.9)	10.0	4.2
Diabetes (%)	9.6 (1.3)	15.5	5.3
HIV prevalence (per 100k)	354 (386)	2391	31

of physical environment for better characterizing the impacts of physical surroundings on population health. All the input variables used in our analysis are displayed in Table 2, and the detailed description of variables are presented in Table 5 from Appendix A.

To summarize, we have 67 variables (9 responses variables, 58 input variables including 57 health measures and one proxy variable 'year') for each county in New York State from 2010 to 2020, that are used for model development.

IV. METHODOLOGY

A. RESEARCH FRAMEWORK

In this study, we propose a data-driven multivariate framework to assess and predict the multidimensional population health outcomes, leveraging the advanced predictive algorithms. The framework is depicted in Fig. 1, where each component is explained as follows.

The first component in the framework is data collection and pre-processing (described in Section III). Specifically, nine different variables are used as health outcomes to represent length of life (mortality) and quality of life (morbidity) in population. Input variables including health behaviors, clinical care, social and economic factors, physical environment and demographics are used in the analysis. Then, a sequence of data pre-processing procedures are implemented including screening for missing data and correlated variables. Additionally, we applied the min-max normalization to scale the response values to the range [0, 1] so that the performance of various models are comparable across response variables [55]. Finally, the data are aggregated for each county at a given year, which is used for model development.

The second component in the framework is model development, evaluation, and selection as described in Sections IV-B. The final model selection consists of two steps—**Step-1**: a library of multivariate models including multivariate linear regression, multivariate random forest, and multivariate tree boosting model have been developed. Following that, the best multivariate model is chosen based on the generalization performance; **Step-2**: based on the multivariate model selection, the corresponding univariate model is implemented; following this, the generalization performances were again compared to select the final model for statistical inferencing. Note that, generalization performance is obtained through 5-fold cross validation approach. Further, hyper-parameter tuning process is executed based on the selected model. Specifically, our proposed two-step model selection framework aims to test two hypotheses as follows:

- **Hypothesis (1)**: multivariate tree-based models can better capture the complex nonlinear effects and interactions between population health and the various predictor variables than the traditionally used linear model (outcome of **Step-1**);
- **Hypothesis (2)**: the multivariate model can better predict the multi-dimensional population health outcomes simultaneously, compared to corresponding univariate models for capturing each health outcome separately (outcome of **Step-2**).

The third component in the framework is model interpretation and inference as presented in Section IV-C. To further reduce model complexity without sacrificing the generalization performance, the percentile-based variable selection for multivariate regression (PVS-MR) approach is proposed to select the optimal subset of features. The statistical insights of key factors related to health outcomes are provided, with the help of the variable importance heatmap and partial dependence plots.

B. MODEL IMPLEMENTATION

In this section, different types of models are introduced including parametric and nonparametric models, multivariate and univariate models. Specifically, we implement tree-based models, including random forest and gradient boosted model in the context of univariate and multivariate constructs, and compare their performances to that of the traditionally-used linear regression model. The rationale for selecting tree-based models is that they are able to capture the non-linearity and the complex interactions between the health outcomes and the corresponding health factors [13], [40]. Then, we provide the details of the multivariate tree boosting model. The generalization performance is also presented to help select the final model.

1) PARAMETRIC VS. NON-PARAMETRIC MODELS

The objective of supervised learning is to estimate the function f that maps the input vector X to the response Y. Mathematically it can be written as $Y = f(X) + \epsilon$, where ϵ is the irreducible error term that captures unobserved heterogeneity from the data [56], [57].

Supervised learning algorithms vary differently in their degree of complexity and flexibility depending on the construction of function f [58], [59]. Broadly speaking, parametric and non-parametric are the common types of learning models. The widely-used generalized linear regression belongs to the family of parametric models, where the model parameters are predetermined and estimated from data. One of the major advantages of parametric model is easier interpretability from explicit formulae of the function [60]; however, it comes at the cost of predictive accuracy. On the contrary, non-parametric model does not

TABLE 2. List of input variables in the model.

Categories	Input Variables (Unit)
Health Behaviors (13 measures)	Adult smoking (%), Adult obesity (%), Food environment index (Index), Physical inactivity (%), Access to exercise opportunities (%), Excessive drinking (%), Alcohol-impaired driving deaths (%), Chlamydia rate (per 100k), Teen births (per 100k), Food insecurity (%), Limited access to healthy foods (%), Drug overdose deaths (per 100k), Motor vehicle crash deaths (per 100k)
Clinical Care (11 measures)	Uninsured (%), Primary care physicians (population ratio), Dentists (population ratio), Mental health providers (population ratio), Preventable hospital stays (per 100k Medicare enrollees), Diabetes monitoring (%), Mammography screening (%), Uninsured adults (%), Uninsured children (%), Health care costs (USD), Other primary care providers (population ratio)
Socioeconomic Factors (11 measures)	High school graduation (%), Some college (%), Unemployment (%), Children in poverty (%), Children in single-parent households (%), Social associations (per 100k), Violent crime (per 100k), Injury deaths (per 100k), Median household income (USD), Free lunch (%), Homicides (per 100k)
Physical Environment (10 measures)	Air pollution - PM2.5 ($\mu g\, m^{-3}$), Driving alone to work (%), County type (index), Average wind speed (dm/s), Cooling degree days (°F), Heating degree days (°F), DX90 (days per year), DT32 (days per year), Precipitation (mm), Snowfall (mm)
Demographics (12 measures)	Population, Below 18 years of age (%), 65 and older (%), Non-Hispanic African American (%), American Indian and Alaskan Native (%), Asian (%), Native Hawaiian/Other Pacific Islander (%), Hispanic (%), Non-Hispanic white (%), not proficient in English (%), Females (%), Rural (%)

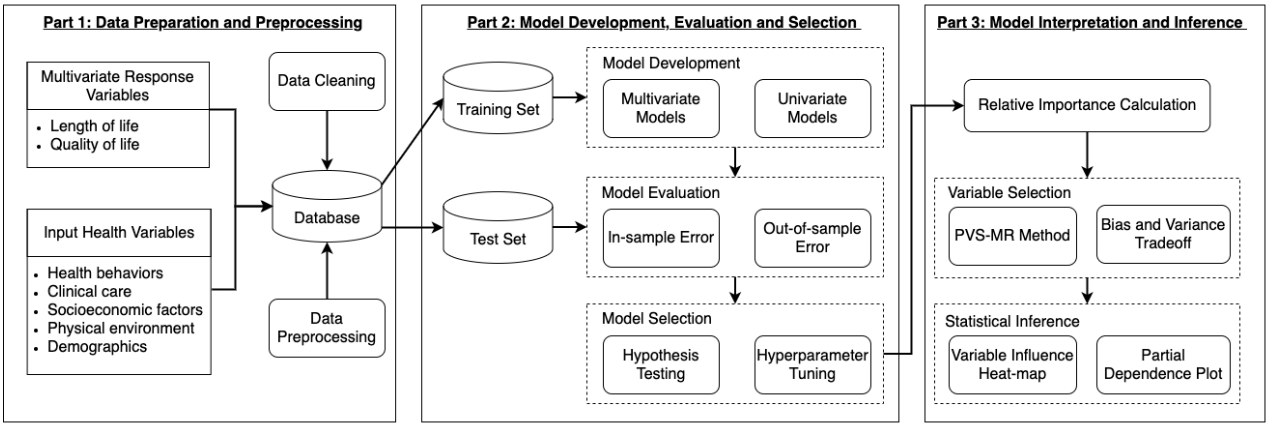


FIGURE 1. Schematic depicting the proposed framework for modeling multivariate population health.

require prior knowledge about the form of mapping function, and has the flexibility to fit in any structure of the data. Thus, it can better capture the complex nonlinear relationships but comes at the cost of interpretability [37], [58]. Grounded in ensemble theory, non-parametric tree ensembles are robust to outliers and noise in the data, and exhibit superior predictive accuracy [37], [61]. Therefore, in this paper we implemented ensemble tree-based models including random forest and gradient tree boosting in our analysis.

2) MULTIVARIATE VS. UNIVARIATE MODELS

Depending on the number of response variables in the analysis, the statistical model can be either univariate or multivariate. Specifically, multivariate model involves multivariate response $Y \in \mathbb{R}^{N\times Q}$ where $N \geq 1$ is the total observations and $Q > 1$ denotes the cardinality of response variables, and it can be reduced to univariate regression when $Q = 1$. Multivariate analysis is often implemented in the cases where the covariances between multiple response variables are dependent on a set of input variables [62]. By utilizing the covariance structure of the response variables, it can allow us to predict multivariate response simultaneously, and to capture the variability of responses in order to improve model's predictive accuracy [62], [63]. Therefore, in this study, we implemented both multivariate and univariate models to investigate if the covariance between the nine different response variables representing population health contributes to the overall accuracy of the population health assessment and prediction model.

3) MULTIVARIATE TREE BOOSTING ALGORITHM

The gradient boosted trees [64] (viz., gradient boosting machines [65]) leverage the gradient boosting technique to iteratively fit ensembles of decision trees to minimize the loss function, and it can be applied for both classification and regression problems. Specifically, gradient boosting algorithm constructs several decision trees sequentially, where each tree is grown by utilizing information (i.e., residuals) from the previous trees in each iteration [66]. Essentially, the model performance can be "boosted" by adding more penalty on bad predictions [67]. The gradient boosted tree model has several advantages: (1) it does not require any pre-determined form of the function, (2) the nonlinear effects and interactions among variables can be captured in the process of growing trees, and (3) the tree-building process uses previous information (i.e., each tree is grown sequentially on residuals), which helps the model to be efficient and robust [64], [67].

The multivariate tree boosting algorithm is an extension of the gradient boosted tree algorithm, and it helps to predict the multivariate response simultaneously, given a set of input variables. Particularly, the multivariate tree boosting algorithm sequentially fits the regression trees to simultaneously minimize the loss function $\mathcal{L}$ for each response variable and maximize the covariance discrepancy $\mathcal{D}$ in the multivariate response. Here, the mean squared error (a.k.a., squared L_2 norm) is represented as loss function, which is given by

$$\mathcal{L}(Y, \hat{Y}) = \frac{1}{NQ} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2, \tag{1}$$

where $\hat{Y}_i$ is the predicted value of the multivariate response using the i-th observation, and Y_i is the actual value. In addition, covariance discrepancy $\mathcal{D}$ can be mathematically written as [68]

$$\mathcal{D}_{m,q} = ||\widehat{\sum}_{(m-1)} - \widehat{\sum}_{(m,q)}||, \tag{2}$$

where $\widehat{\sum}_{(m-1)}$ is the covariance matrix of the response variables at the previous step $m - 1$, and $\widehat{\sum}_{(m,q)}$ is the covariance matrix at the current step m after fitting a decision tree to the response variable $y^{(q)}$. The discrepancy $\mathcal{D}_{(m,q)}$ is used to measure the amount of covariance explained in the multivariate responses by the predictors selected by the tree to fit in $y^{(q)}$ in step m. Basically, $\mathcal{D}$ denotes the improvement in model fitting in the sample covariance matrix during each iteration [68]. The steps of multivariate ensemble tree boosting algorithm are displayed in Algorithm (1).

Algorithm 1 Multivariate Ensemble Tree Boosting Algorithm [68]

```
for $m$ in $1, \cdots, M$ steps (regression trees) do
    for $q$ in $1, \cdots, Q$ response variables do
        train tree $m^{(q)}$ to residuals, and assess the covariance discrepancy $\mathcal{D}_{m,q}$ as in (2).
    end for
    Select the response $y^{(q)}$ corresponding to the regression tree with the maximum covariance discrepancy $D_{m,q}$.
    Update residuals by subtracting the predictions of the tree fitted to $y^{(q)}$, multiplied by step-size.
end for
```

4) MODEL EVALUATION

The generalization performance of a learned model is evaluated leveraging a robust cross-validation technique [57]. In this study, we leveraged a 5-fold cross validation technique, where the data are randomly divided into five different folds of roughly equal size. During each of the five cross-validation loops, the model is trained using the four folds (i.e., 80% of data) and tested using the remaining fold (i.e., 20% of data). This procedure is repeated five times until every fold is utilized for model testing. During each loop, the in-sample goodness-of-fit performance is estimated using the training set, while the out-of-sample predictive accuracy is evaluated using the test set. This procedure can guarantee every single data point is being utilized for both training and test, to produce the generalized and robust model results. Note that, to control other factors (i.e., sampling bias) that could affect model performance, all the models are fitted into the same training set, and tested using identical test set during each cross validation loop.

We apply root mean square error ($RMSE$) and R^2 to be representative of in-sample and out-of-sample model performance metrics. The overall model performance is evaluated through the averaged $RMSE$ and R^2 across all validation loops for each population health dimension. The model with the highest R^2, and the lowest $RMSE$ for both in-sample and out-of-sample is then selected as the final model, which is used to conduct the statistical inference. Mathematically, the $RMSE$ and R^2 are given by

$$\begin{aligned} RMSE(q) &= \frac{1}{K} \sum_{k=1}^{K} \sqrt{\frac{1}{u} \sum_{i=1}^{u} \left(y_{i,k}^{(q)} - \hat{y}_{i,k}^{(q)} \right)^2} \\ R^2(q) &= \frac{1}{K} \sum_{k=1}^{K} \left(1 - \frac{\sum_{i=1}^{u} \left(y_{i,k}^{(q)} - \hat{y}_{i,k}^{(q)} \right)^2}{\sum_{i=1}^{u} \left(y_{i,k}^{(q)} - \bar{y}_{k}^{(q)} \right)^2} \right), \end{aligned} \tag{3}$$

where q represents the q-th response variable to be calculated, k is the number of times cross validation performed, u indicates the number of observations from either in-sample or out-of-sample. For response q under the k-th iteration, $\hat{y}_{i,k}^{(q)}$ describes the predicted value at the observation i, $y_{i,k}^{(q)}$ is the actual value, and $\bar{y}_{k}^{(q)}$ is the mean value of the response q.

C. STATISTICAL INFERENCE

Inferential statistical techniques are provided here based on the final multivariate tree boosting model. First, variable importance is introduced to assess how useful health predictor is at predicting population health outcome variables. Second, we proposed a feature selection approach for multivariate analysis. Third, the partial dependence analysis is presented to quantify the marginal effect of health determinants on the health outcomes.

1) VARIABLE IMPORTANCE

To identify and evaluate the impacts of key predictors on population health outcome variables, the relative importance of each input variable is measured. A predictor with larger relative importance is considered to be of significance in contributing to the overall model performance. In ensemble tree-based methods, the relative importance can be computed in two steps [65]. First, the importance of a predictor j

in a single tree T is measured by the number of times this predictor is used for splitting in growing of the tree, weighted by the squared improvement of the model as a result of each split, as denoted $\widehat{I_j^2}(T)$ in (4). Secondly, the relative importance can be obtained by averaging all importance over ensembles of trees $\{T_m\}_1^M$, as denoted $\widehat{I_j^2}$ in (5). Mathematically, the importance in a single tree can be written as:

$$\widehat{I_j^2}(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 \mathbb{1}(v_t = j). \tag{4}$$

Here, the summation is over the non-terminal nodes t of the J-terminal node tree T, v_t is the splitting variable associated with node t, and $\hat{i}_t^2$ is the improvement in squared error as a result of the split in the tree. For an ensemble of trees, the relative importance of the predictor j can be given by

$$\widehat{I_j^2} = \frac{1}{M} \sum_{m=1}^{M} \widehat{I_j^2}(T_m), \tag{5}$$

where M is the number of trees in the model. More details can be referred to [65].

2) VARIABLE SELECTION

The variable selection (a.k.a., feature selection) is used to select a subset of original features to reduce the dimensions of data, and it is computationally expensive [69], [70]. The naive brute-force method for feature selection aims to exhaustively search over all possible combinations of variables, which require massive computation power [71]. Some greedy search strategies, which are computationally advantageous, can be generally classified as either forward selection or backward elimination, which can iteratively add (or delete) one variable at a time to (or from) the existing subset [72]. Variable importance scores can also help in key variable selection based on the ranking of variable importance obtained from the model [73].

Most of the feature selection methods are developed for univariate models, which fall short in selecting key features in the context of multivariate analysis. Therefore, in this paper, we propose a novel approach for feature selection, named **Percentile-based Variable Selection for Multivariate Regression (PVS-MR)**, to strategically select the optimal subset of input variables for multivariate analysis without compromising the generalization performance of the model. The proposed PVS-MR algorithm leverages backward elimination to iteratively eliminate variables, based on the quantile of the variables' relative importance obtained from multivariate tree boosting model. Specifically, during each model run, one or more variables can be removed from the model if their relative importance is universally below a pre-determined quantile threshold across all responses. This process is repeatedly executed with a small increment in threshold to its upper bound. Finally, the optimal subset of features can be obtained based on the generalization performance of the model. The PVS-MR algorithm is displayed in Algorithm (2).

Note that the proposed PVS-MR is particularly developed for multivariate analysis, and has the following advantages. First, it is computationally efficient as one or more variables are deleted at each iteration. Second, the model's generalization performance can be guaranteed, in the sense that model error is evaluated in each step of variable selection. Third, variables are assessed simultaneously across all responses in the context of multivariate analysis. Finally, decision makers can adjust their threshold and increment to control the process of feature selection.

Algorithm 2 Percentile-Base d Variable Selection for Multivariate Regression (PVS-MR)

```
Initialization: $\mathcal{J} = \{$All input features$\}$, threshold $p$,
upper bound $U$, and increment $\delta$.
while $p \leq U$ do
    Run multivariate model with $\mathcal{J}$ and evaluate model
    performance. Let $J = |\mathcal{J}|$.
    for $q$ in $1, \cdots, Q$ response variable do
        for $j$ in $1, \cdots, J$ input variable do
            Calculate $\widehat{I_{qj}^2}$ based on (5).
        end for
        Create a flag variable
        $\theta_{qj} = \begin{cases} 0, & \text{if quantile of } \widehat{I_{qj}^2}, \forall j, \text{ is below } p \\ 1, & \text{Otherwise.} \end{cases}$
    end for
    for $j$ in $1, \cdots, J$ input variable do
        if $\sum_q \theta_{qj} = 0$ then
            Delete variable $j$, and update $\mathcal{J} \leftarrow \mathcal{J} - \{j\}$.
        end if
    end for
    $p \leftarrow p + \delta$.
end while
```

3) PARTIAL DEPENDENCE ANALYSIS

The partial dependence plot (PDP) is a widely-used method to assess the marginal effects of an input variable to the response variable in non-parametric statistical learning models. Presence of non-linearity can be easily detected leveraging PDPs, where the estimated values of the function are produced by changing the value of the predictor, while keeping rest of predictors constant (i.e., *ceteris paribus* condition) [65]. Mathematically, the partial dependency can be written as:

$$\hat{f}_j(x_j) = E_{x_{-j}}[\hat{f}(x_j, x_{-j})] = \int \hat{f}(x_j, x_{-j}) dP(x_{-j}). \tag{6}$$

Here, x_{-j} represents all the predictor variables except j. The estimated partial dependency of predictor x_j is calculated by integrating the function $\hat{f}$ when x_j is fixed and x_{-j} varies over its marginal distribution. The PDPs could inform the changing direction and functional form of the marginal effect

TABLE 3. Performance comparisons among multivariate models.

Response Variables	Goodness of Fit						Predictive Accuracy					
	Linear Model[a]		Random Forest[b]		Boosting[c]		Linear Model[a]		Random Forest[b]		Boosting[c]	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
YPLL	0.80	0.08	0.86	0.07	0.99	0.02	-6.46	0.50	0.71	0.10	0.79	0.08
Child mortality	0.50	0.14	0.75	0.10	0.98	0.03	-4.88	0.47	0.48	0.14	0.58	0.13
Infant mortality	0.56	0.13	0.77	0.09	0.98	0.03	-5.30	0.49	0.53	0.13	0.57	0.13
Poor or fair health	0.67	0.08	0.81	0.06	0.99	0.01	-5.85	0.35	0.60	0.08	0.73	0.07
Poor physical health days	0.50	0.09	0.77	0.06	0.98	0.02	-5.16	0.32	0.48	0.09	0.64	0.08
Poor mental health days	0.44	0.10	0.77	0.07	0.98	0.02	-4.98	0.34	0.48	0.10	0.60	0.09
Low birthweight	0.70	0.08	0.87	0.06	0.99	0.01	-6.23	0.41	0.73	0.08	0.86	0.06
Diabetes	0.68	0.07	0.80	0.06	0.98	0.02	-5.89	0.34	0.59	0.08	0.70	0.07
HIV prevalence	0.87	0.06	0.93	0.04	1.00	0.01	-7.11	0.45	0.85	0.06	0.96	0.03

[a] The multivariate linear model is implemented using R package *'lm'*.
[b] The multivariate random forest is implemented using R package *'MultivariateRandomForest'*.
[c] The multivariate tree boosting is implemented using R package *'mvtboost'*.

of the predictor on the response. Thus, it can be used to facilitate the model inference [68].

V. RESULTS

In this section, we present the results from our case study to illustrate the applicability of our proposed data-driven multivariate framework for modeling the multidimensional population health outcomes.

A. MODEL SELECTION

In this section, first, we discuss and compare the performance of the various multivariate models, following which we select the multivariate model that outperforms all the other models in terms of their generalization performance. Second, we discuss and compare the performance of the selected multivariate model and the corresponding univariate model, following which we select the final model for statistical inferencing.

1) COMPARING PERFORMANCE AMONG MULTIVARIATE MODELS

Table 3 displays the generalization performances of the three multivariate models—linear, random forest and tree-boosting. The *RMSE* and R^2 are calculated for each response variable in terms of both in-sample goodness of fit and out-of-sample predictive accuracy. The goodness of fit indicates how well the model fits the data, and predictive accuracy describes the prediction power of the model in an unseen dataset. Higher values of R^2, conversely lower values of *RMSE*, indicate superior performance of the model. In Table 3, we observe that the tree-based models (multivariate random forest and multivariate tree boosting) demonstrate superior performance on average for each of the nine response variables, over the multivariate linear regression model. This pattern is observed across both the goodness of fit and predictive accuracy performances. Additionally, we implemented t-test to verify if this pattern is statistically significant. Our results show that $p < 0.01$ for each of the response variables (shown in Table 6 from Appendix C), indicating that there is a significant statistical difference between the performances of the linear versus the non-linear ensemble tree-based models. Therefore, the **Hypothesis-1** holds good, concluding that non-parametric tree-based models better capture the non-linear effects and interactions between population health outcomes and the predictors than the parametric linear model.

We can also observe from Table 3 that multivariate tree boosting model outperforms the multivariate random forest model. Specifically, multivariate tree boosting model has improved the goodness-of-fit by 24% over linear regression, and the predictive accuracy by 11% over random forest. This illustrates that the multivariate tree boosting method best models the county-level population health outcomes. Additionally, we compared the model results to the null (a.k.a. mean-only) model, where the average values of each response variable is used to fit the data.

Comparing the results, we found that multivariate tree boosting model significantly reduces the training and test errors, i.e., improves the goodness of fit and predictive accuracy by 91% and 54% in YPLL, 87% and 36% in child mortality, 86% and 35% in infant mortality, 90% and 48% in poor or fair health, 87% and 40% in poor physical health days, 88% and 37% in poor mental health days, 93% and 62% in low birth weight, 88% and 45% in diabetes, 96% and 80% in HIV prevalence, over the null model. Therefore, we selected multivariate tree boosting model as the best multivariate model to compare with its corresponding univariate model.

2) COMPARING PERFORMANCE BETWEEN MULTIVARIATE AND UNIVARIATE MODELS

Table 4 shows the in-sample and out-of-sample model performance of the multivariate tree boosting and the corresponding univariate tree boosting models. The corresponding t-test results are shown in Table 6 in Appendix C. The results show multivariate tree boosting model performs significantly better than the corresponding univariate model across all response variables in terms of goodness of fit, but this significance is relatively weak with respect to predictive accuracy. Specifically, in case of predicting the responses YPLL, infant mortality, diabetes and HIV prevalence, there is no significance in predictive accuracy between multivariate and univariate models. But for the remaining five responses, the multivariate model is statistically better than the corresponding univariate model in predictive performance. Overall we concluded that multivariate tree boosting model

TABLE 4. Performance comparisons between multivariate and univariate models.

Response Variables	Goodness of Fit				Predictive Accuracy			
	Univariate Model[d]		Multivariate Model[e]		Univariate Model[d]		Multivariate Model[e]	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
YPLL	0.89	0.06	0.99	0.02	0.76	0.09	0.79	0.08
Child mortality	0.76	0.09	0.98	0.03	0.52	0.13	0.58	0.13
Infant mortality	0.78	0.09	0.98	0.03	0.53	0.13	0.57	0.13
Poor or fair health	0.84	0.05	0.99	0.01	0.67	0.08	0.73	0.07
Poor physical health days	0.80	0.06	0.98	0.02	0.54	0.09	0.64	0.08
Poor mental health days	0.77	0.07	0.98	0.02	0.50	0.10	0.60	0.09
Low birthweight	0.90	0.05	0.99	0.01	0.81	0.07	0.86	0.06
Diabetes	0.83	0.05	0.98	0.02	0.65	0.08	0.70	0.07
HIV prevalence	0.97	0.02	1.00	0.01	0.93	0.04	0.96	0.03

[d] The univariate gradient boosted tree model is implemented using R package *'gbm'*.

[e] The multivariate tree boosting is implemented using R package *'mvtboost'*.

outperforms the univariate tree boosting models on average, confirming that our **Hypothesis-2** holds good.

B. VARIABLE SELECTION

After selecting the final model, we then applied a grid search technique to fine tune two parameters in the multivariate tree boosting model, namely the number of trees T and depth of each of the trees D. The rationale for choosing these two parameters are: a) they are critical for controlling the generalization performance of the model, and b) they can be easily modified by decision-makers to modulate computational complexity [68], [74]. Instead of searching the whole two-dimensional space $\mathbb{Z}_+^2$ to find the optimal values, we limited the bound for parameter T is the range of [500, 5000] with a 500 unit grid-step, and the bound for parameter D is the range of [2, 20] with a grid-step of 2 units. In total, there are $100 = 10 \times 10$ combinations for those two parameters. Among those, we aim to find the optimal values of T and D that can achieve the minimal generalization error of the model. Leveraging this method, we finally determined $T = 3500$ and $D = 16$ in the model, which are used for further variable selection and statistical inferencing.

We applied the proposed PVS-MR method on the final model to select a subset of important variables, and to further reduce model complexity by mitigating the curse of dimensionality. The training error (denoted by black curve) and test error (denoted by red curve) shown in Fig. 2, are obtained by averaging the *RMSE* across all the nine response variables using the final multivariate tree boosting model. From the Fig. 2, the model with exact 29 variables (denoted by blue vertical line) yields the least generalization error. Additionally, we observed that the model performance improves while downsizing the number of input variables; but after a certain threshold as further variables are removed, the model performance deteriorates. This indicates that optimal selection of variables plays a pivotal role in accurately and sufficiently modeling the multidimensional population health. Thus, without compromising model generalization performance, we included only 29 variables as the key influencing factors of the population health outcomes in the model.

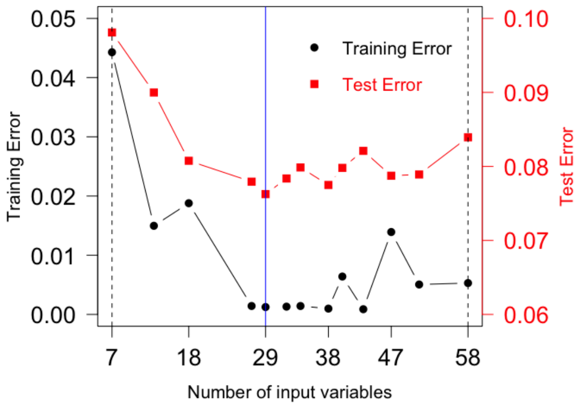
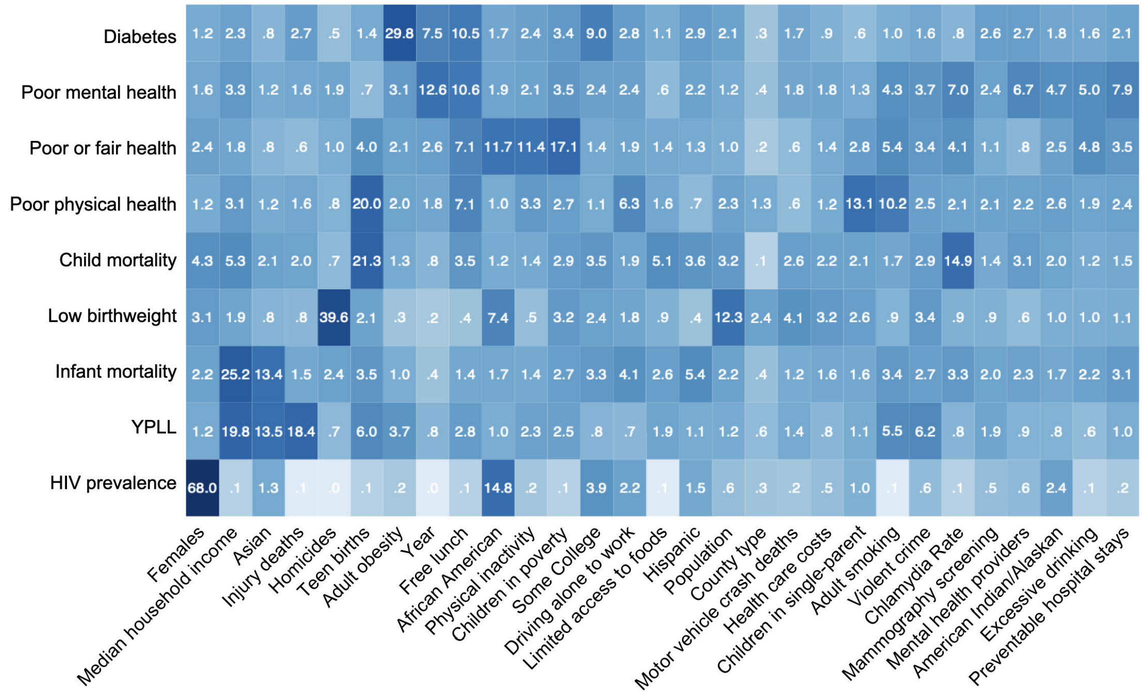


FIGURE 2. Variable selection using the PVS-MR method. The training error and test error are the averaged RMSE across all response variables. The blue vertical line indicates the optimal number of variables selected by PVS-MR model with the minimal generalization error.

C. STATISTICAL INFERENCING

1) VARIABLE IMPORTANCE

Fig. 3 presents the heat-map of relative importance of the selected 29 key predictor variables (indicated on the horizontal axis) to health outcomes (indicated on the vertical axis). Each number in the heat-map indicates the relative importance obtained from (5), as a fraction out of 100 for each health outcome variable. The larger numbers associated with the darker blue in grid cells, indicate higher degree of importance, while smaller numbers displayed in the lighter blue grid cells, represent lower degree of importance. For example, two most influencing factors for HIV prevalence are female population (with a relative importance of 68) and African American (with a relative importance of 14.8), which are consistent with the previous research findings [75], [76]. Even though some of the health factors have relatively small importance scores to the response variables such as excessive drinking and preventable hospital stays (as shown in the last two columns in the heat-map), these predictors are deemed to be important for the overall population health of a region, as they contribute to the overall performance of the multivariate model. This is established by the fact that



	Females	Median household income	Asian	Injury deaths	Homicides	Teen births	Adult obesity	Year	Free lunch	African American	Physical inactivity	Children in poverty	Some College	Driving alone to work	Limited access to foods	Hispanic	Population	County type	Motor vehicle crash deaths	Health care costs	Children in single-parent	Adult smoking	Violent crime	Chlamydia Rate	Mammography screening	Mental health providers	American Indian/Alaskan	Excessive drinking	Preventable hospital stays
Diabetes	1.2	2.3	.8	2.7	.5	1.4	29.8	7.5	10.5	1.7	2.4	3.4	9.0	2.8	1.1	2.9	2.1	.3	1.7	.9	.6	1.0	1.6	.8	2.6	2.7	1.8	1.6	2.1
Poor mental health	1.6	3.3	1.2	1.6	1.9	.7	3.1	12.6	10.6	1.9	2.1	3.5	2.4	2.4	.6	2.2	1.2	.4	1.8	1.8	1.3	4.3	3.7	7.0	2.4	6.7	4.7	5.0	7.9
Poor or fair health	2.4	1.8	.8	.6	1.0	4.0	2.1	2.6	7.1	11.7	11.4	17.1	1.4	1.9	1.4	1.3	1.0	.2	.6	1.4	2.8	5.4	3.4	4.1	1.1	.8	2.5	4.8	3.5
Poor physical health	1.2	3.1	1.2	1.6	.8	20.0	2.0	1.8	7.1	1.0	3.3	2.7	1.1	6.3	1.6	.7	2.3	1.3	.6	1.2	13.1	10.2	2.5	2.1	2.1	2.2	2.6	1.9	2.4
Child mortality	4.3	5.3	2.1	2.0	.7	21.3	1.3	.8	3.5	1.2	1.4	2.9	3.5	1.9	5.1	3.6	3.2	.1	2.6	2.2	2.1	1.7	2.9	14.9	1.4	3.1	2.0	1.2	1.5
Low birthweight	3.1	1.9	.8	.8	39.6	2.1	.3	.2	.4	7.4	.5	3.2	2.4	1.8	.9	.4	12.3	2.4	4.1	3.2	2.6	.9	3.4	.9	.9	.6	1.0	1.0	1.1
Infant mortality	2.2	25.2	13.4	1.5	2.4	3.5	1.0	.4	1.4	1.7	1.4	2.7	3.3	4.1	2.6	5.4	2.2	.4	1.2	1.6	1.6	3.4	2.7	3.3	2.0	2.3	1.7	2.2	3.1
YPLL	1.2	19.8	13.5	18.4	.7	6.0	3.7	.8	2.8	1.0	2.3	2.5	.8	.7	1.9	1.1	1.2	.6	1.4	.8	1.1	5.5	6.2	.8	1.9	.9	.8	.6	1.0
HIV prevalence	68.0	.1	1.3	.1	.0	.1	.2	.0	.1	14.8	.2	.1	3.9	2.2	.1	1.5	.6	.3	.2	.5	1.0	.1	.6	.1	.5	.6	2.4	.1	.2

FIGURE 3. **Relative importance of the key predictors on the multivariate health outcomes.**

deletion of these variables can deteriorate the overall model generalization performance as depicted in Fig. 2.

From Fig. 3, the key factors that are significantly associated with more than one health outcome can be also identified. For example, the median household income is the most significant factor in contributing to both infant mortality and YPLL, indicating that economic condition is strongly linked to the length of life of the population, on average. Another health variable, teen birth rate, plays the most critical role in predicting both poor physical health days and child mortality. This highlights that adolescent pregnancy significantly affects both quality of life and length of life.

2) PARTIAL DEPENDENCE ANALYSIS

To quantify the association of the health determinants with the population health outcomes, partial dependence plot (PDP) is implemented. PDP is used to unravel the marginal effect of a particular variable by keeping other variables constant (*ceteris paribus* condition). Note that, there is a total of 261 potential combinations (i.e., 29 predictors variables for each of the nine outcome variables) of PDPs, that can be constructed. For the sake of brevity, we only explained the PDPs of the top two key predictors for each of the nine population health outcome variables, based on the relative importance scores in Fig. 3. Figs.4–12 depict the PDPs of the top two important predictors for each of the nine health outcome variables.

a: DIABETES

Fig. 4(a) exhibits that the number of people diagnosed with diabetes is positively correlated with adult obesity, i.e., as the percentage of adults with obesity problem increases, the prevalence of diabetes in the community also increases. This finding is consistent with the previous studies where the prevalence of obesity and overweight is closely linked to the diabetes [77]–[79]. In addition to the correlation, our analysis reveals that this relationship is nonlinear. We observe that when the adult obesity rate $\leq 25\%$, the prevalence of diabetes is insensitive to the obesity rate within the community. However, the diabetes prevalence rate sharply increases at adult obesity rate $\geq 25\%$. Thus, threshold for adult obesity rate can be considered to be 25%, i.e., efforts to be given such that the overall percentage of the population suffering from obesity is $<25\%$. Another second-most important predictor of diabetes is free lunch. This predictor describes the percentage of children enrolled in public school who are eligible for free or reduced priced lunch; thus, it is often used as an indicator for family poverty levels [80]. From Fig. 4(b), we observe that the prevalence of diabetes in a community increases as the percentage of school children eligible to

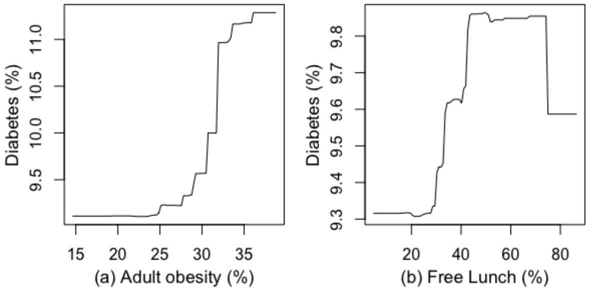


FIGURE 4. **Partial dependence plots of top two factors (adult obesity and free lunch) on diabetes in population.**

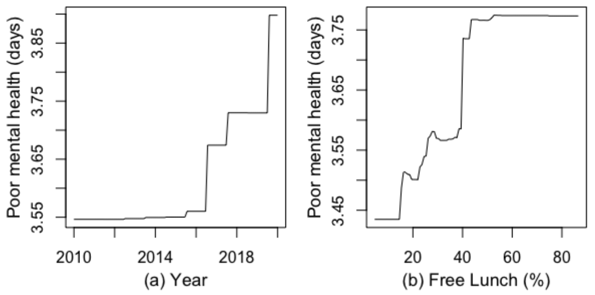


FIGURE 5. Partial dependence plots of top two factors (year and free lunch) on poor mental health.

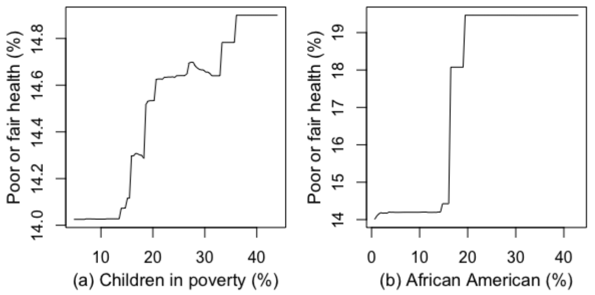


FIGURE 6. Partial dependence plots of top two factors (children in poverty and African American) on poor or fair health.

get free meals crosses the 30% threshold. This finding is in line with a previous study showing that children attending schools located in deprived areas had the higher prevalence of diabetes compared to those in non-deprived areas [81]. In this study, schools were viewed as being situated in a deprived region, if over 21% of children received free lunch [81].

b: POOR MENTAL HEALTH

Our results shows that the two most significant predictors of poor mental health days are the variables "year" and "free lunch". Fig. 5(a) shows that the population mental health is worsening over time. Especially after the year of 2017, prevalence of poor mental health has significantly increased within the communities. This re-establishes the fact that mental health burden is crippling the established market economies such as the US, where an estimated 26% of Americans aged 18 and older—about 1 in 4 adults—suffers from a clinical mental disorder in a given year [82]. Free lunch as another important predictor of poor mental health. Fig. 5(b) displays the positive association between poor mental health days and the percentage of children receiving free lunch in their schools. As the percentage of children receiving free lunch ≥20% in a community, the mental health days, on average, increases from 3.45 days to 3.75 days. Since higher values of free lunch indicates higher poverty rates in a community, our results indicate that population in deprived areas suffer from poor mental health. In fact, a previous study found that children receiving free or reduced-priced lunch were more likely to suffer mental illness such as depression, anxiety and pessimism [83].

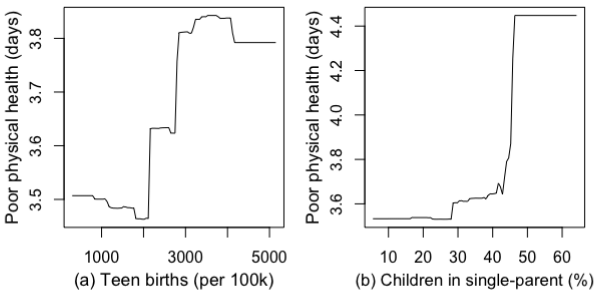


FIGURE 7. Partial dependence plots of top two factors (teen birth rate and children in single-parent households) on poor physical health days.

c: POOR OR FAIR HEALTH

Our analysis indicate the two most important predictors of the poor or fair population health to be "the percentage of children in poverty" and the "African American population". It can be observed from Fig. 6(a) that as the percentage of children in poverty increases to ≥15%, the percentage of overall population in a community suffering from poor/fair health condition increases to 14.8%, from 14.0%. This finding is in line with the existing study that shows poverty to be the key risk factor of child health and development [84], which in turn is a key determinant of the overall poor health of a population in a community. Racial disparities in population health is a well-studied area. Our results show that higher percentages of people reporting poor or fair health are from communities having higher percentages of African American population (≥15%) on average. Fig. 6(b) re-emphasizes that racial health disparities exist in population health.

d: POOR PHYSICAL HEALTH

Our results indicate that the number of poor physical health days reported by the population in a county is positively associated with "teen births", (i.e., adolescent pregnancy) (see From Fig. 7(a)) and "percentages of children from a single-parent family" (see Fig. 7(b)). More specifically, communities with ≥1500 teen births per 100K of the population or number of children from single-parent families ≥30% of the population, witness a sharp increase in the number of poor health days experienced by the population on average. It is established in literature that adolescent pregnancy could have adverse social and economic impacts on mothers, children, their families [85], [86], which can be linked to the overall poor physical wellbeing of a population in that community, on average. In addition, as the percentage of children living with their single parent increases in a community, the number of poor physical health days reported by the population in a community on average also increases. This finding is also consistent with the previous studies [87], [88].

e: CHILD MORTALITY RATE

The top two predictors of child mortality rate are found to be "teen births" and "Chlamydia rate". Fig. 8(a) shows that the child mortality rate is positively associated with the teen birth

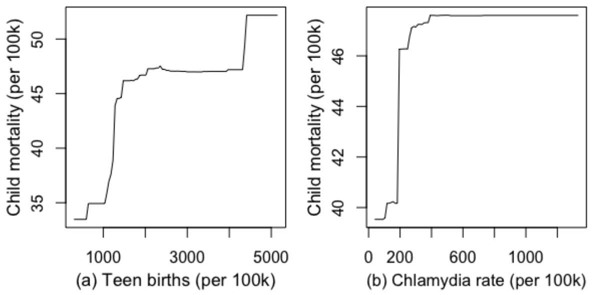


FIGURE 8. **Partial dependence plots of top two factors (teen birth rate and chlamydia rate) on child mortality.**

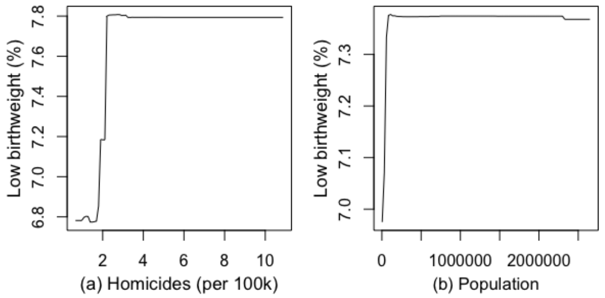


FIGURE 9. **Partial dependence plots of top two factors (homicides and population) on low birthweight.**

rate, which is consistent with the previous studies [89], [90]. We found that as the teen birth rate increases to ≥500 per 100K population in a community, the child mortality rate increases from 35—50 per 100K of the population. In fact, the burden of child mortality still remains unevenly distributed globally [91]. On the other hand, chlamydia rate is also found to be positively correlated with child mortality. As observed from Fig. 8(b), with increasing Chlamydia rate, the child mortality rate increases monotonically from 40—48 per 100K of the population in a community, on average. Our finding is in line with the previous research outcomes that established chlamydia infection to be the most widely reported sexually transmitted disease in the US, especially among females aged from 15 to 24 [92]. Chlamydia infection increases the risk of still births and infant mortality rates significantly [93], [94].

f: LOW BIRTH WEIGHT

The two most important factors of low birth weight—one of the dimensions of population health—are found to be "rates of homicides" and "population of a county". The positive association between low birth weight and homicide is exhibited in Fig. 9(a). We found that as the prevalence of homicide increases beyond 2 per 100K population in a county, the percentage of low birth weight grows rapidly from 6.8% to 7.8%. This positive correlation between the homicide rate and low birth weight can be attributed to the presence of a confounding variable such as socioeconomic condition of a region that influences both low birthweight and homicides. For example, communities with lower socioeconomic status have been witnessing higher homicide rates in general [95]. On the other hand, low birth weight babies are mostly outcomes of adolescent pregnancy or poor socioeconomic status of a region [96]. Prenatal poverty, in fact, is a key determinant of low birthweight [97]–[99]. Population size is also strongly correlated with low birth weight rate, which is displayed in Fig. 9(b). This strong correlation between low birthweight and population density may arise due to a confounding factor such as air pollution. Higher population density is a surrogate for urban and suburban areas which experience higher levels of air pollution. Air pollution, in turn, has a confounding effect on public health and has a strong link with low birth weight [100], [101].

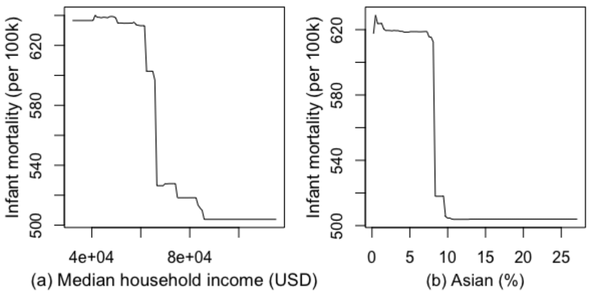


FIGURE 10. **Partial dependence plots of top two factors (median household income and Asian population) on infant mortality.**

g: INFANT MORTALITY RATE

Factors such as "median household income" and "percentage of Asian population" are found to be the top two predictors of infant mortality rate. Fig. 10(a) shows that the median household income has negative correlation with the infant mortality rate, indicating that income inequality plays a critical role in affecting the survival and health of newborns. Specifically, communities with median household income ≤80, 000 USD, witness a sharp increase in infant mortality rates from 500—640 per 100K population, on average. This finding is in line with a meta-analysis study which found that a significant inverse relationship exists between household income and mortality rate among infants and children [102]. Similarly, proportion of Asian population in a community is another key predictor showing negative association with infant mortality rate, as shown in Fig. 10(b). This observation is consistent with findings from an existing study that shows infant mortality rate is lower among Asian population, compared to the White population [103].

h: YEARS OF POTENTIAL LIFE LOST (YPLL)

This dimension of population health is a measure of premature mortality, that provides an estimate of the average years a person would have lived if they had not died prematurely [104]. Our analysis indicates that "median household income" and "injury deaths (per 100K)" are the top two significant predictors of YPLL. Fig. 11(a) shows negative relationship between median household income and

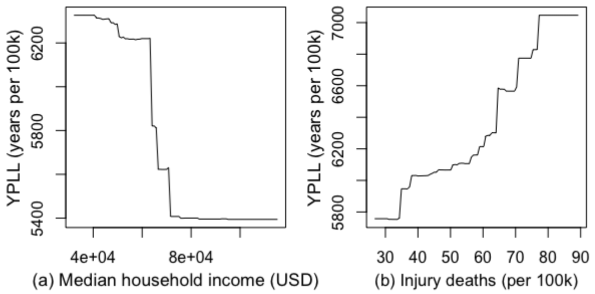


FIGURE 11. Partial dependence plots of top two factors (median household income and injury deaths) on YPLL.

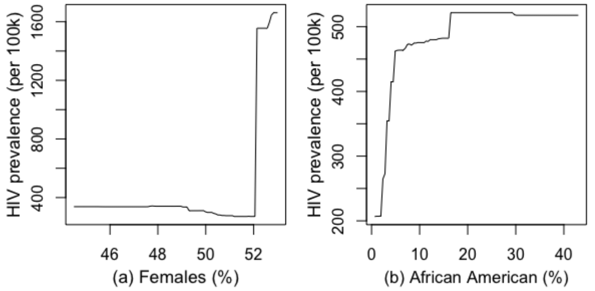


FIGURE 12. Partial dependence plots of top two factors (female population and African American) on HIV prevalence.

YPLL, indicating that prevalence of premature death is less in families from high income group. Specifically, communities with median household income $\leq 80,000$ USD witness higher risk of premature deaths; however, risk of premature death is insensitive to median household income $\geq 80,000$ USD. In fact, previous studies indicated that higher rates of income inequality is strongly linked to the preventable or immediate death rates across the US cities [105]. Number of injury deaths is another key predictor of premature deaths. As can be observed from Fig. 11(b), as the number of injury deaths increases, the prevalence of YPLL in the community also increases monotonically, which is expected. In fact, the unintentional injury deaths contribute to be the number one leading cause of premature death among people aged below 44 years, in the US [106].

i: HIV PREVALENCE

Prevalence of HIV is one of the key dimensions of population health. Our analysis indicates that female population is strongly associated with the HIV prevalence. From Fig. 12(a), it can be noticed that communities with higher proportions of females ($\geq 52\%$), witness a sharp increase in HIV prevalence. This sharp increase can be attributed to the right-skewed distribution of HIV prevalence in the state of NY (see Fig. 13 in Appendix B). In fact, NY is ranked top in the US with the highest female HIV infection rate in 2017, where the Bronx County in NY alone contributes to the 27% of the total HIV cases in the state of NY [107]. In addition, racial disparity is again observed where the African American population is significantly affected by the HIV. Fig. 12(b) shows that increased prevalence of HIV is associated with higher percentages of African Americans in a community, which is in line with the previous findings [108].

VI. DISCUSSIONS

The data-driven framework developed in this paper could potentially help enhance the overall population health of a community (e.g., county), by transforming information from routinely collected data into informed decisions. To facilitate decision making, this paper provides variable importance heat-map and partial dependence plots to identify and assess the associations of the various factors with the multidimensional population health that would help communicate data-driven results to the policy makers. The heat-map of variable importance reveals the key health inputs that jointly influence various dimensions of population health. It is beneficial for state and federal health agencies to identify focal variable(s) and develop informed public health intervention strategies to enhance the overall population health considering its various dimensions. For example, teen birth rate needs to pay more attention, because it is strongly associated with both the quality of life (e.g., poor physical health) and length of life (e.g., child mortality). Similarly, the free lunch variable (i.e., the percentage of children who are eligible for free lunch) significantly correlates with both the incidence of diabetes and poor mental health in a community. The partial dependence plots are used to assess the marginal effects and quantify the complex nonlinear relationships of the essential health determinants and the various dimensions of population health. It could provide a holistic picture of the overall trend in population health with respect to changes in specific health determinant, given all the other factors remain constant (ceteris paribus condition). For instance, child mortality is more sensitive to an increase in teen births compared to other factors such as chlamydia rate in the community. Additionally, years of potential life loss (YPLL) is highly correlated with injury deaths in a nonlinear form, where the growth rate of the YPLL becomes faster as the number of injury deaths increases. The racial disparity gap still persists in New York State. African American groups are more likely to suffer from poor of fair health and high HIV infections, which could further deteriorate their quality of life. This finding is consistent with the previous research [108].

Although we applied our framework to the New York State, our framework is generalized enough that can be easily applied to other regions of interest, provided adequate data is available. The model selection and variable selection techniques will also remain unchanged. However, the associations between the health determinants and the health outcomes might be different in reflection of the regional health disparities. To validate different statistical learning models, we implemented a robust cross-validation technique to evaluate the generalization performance of the models which is used as an input of the model selection process. Our hypotheses of the superior performance of multivariate tree boosting model over linear method and

TABLE 5. Description of variables used in the model.

Variables	Description
Response Variables	
Premature Death (YPLL)	Years of potential life lost before age 75 per 100,000 population (age-adjusted).
Child Mortality	Number of deaths among children under age 18 per 100,000 population.
Infant Mortality	Number of all infant deaths (within 1 year), per 100,000 live births.
Poor or Fair Health	Percentage of adults reporting fair or poor health (age-adjusted).
Poor Physical Health Days	Average number of physically unhealthy days reported in past 30 days (age-adjusted).
Poor Mental Health Days	Average number of mentally unhealthy days reported in past 30 days (age-adjusted).
Low Birthweight	Percentage of live births with low birthweight (<2,500 grams).
Diabetes Prevalence	Percentage of adults aged 20 and above with diagnosed diabetes.
HIV Prevalence	Number of people aged 13 years and older living with a diagnosis of human immunodeficiency virus (HIV) infection per 100,000 population.
Health Behaviors	
Adult Smoking	Percentage of adults who are current smokers.
Chlamydia Rate	Number of newly diagnosed chlamydia cases per 100,000 population.
Teen Births	Number of births per 100,000 female population ages 15-19.
Adult Obesity	Percentage of the adult population (age 20 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m^2.
Food Environment Index	Index of factors that contribute to a healthy food environment, from 0 (worst) to 10 (best).
Physical Inactivity	Percentage of adults age 20 and over reporting no leisure-time physical activity.
Access to Exercise Opportunities	Percentage of population with adequate access to locations for physical activity.
Food Insecurity	Percentage of population who lack adequate access to food.
Limited Access to Healthy Foods	Percentage of population who are low-income and do not live close to a grocery store.
Excessive Drinking	Percentage of adults reporting binge or heavy drinking.
Alcohol-Impaired Driving Deaths	Percentage of driving deaths with alcohol involvement.
Drug Overdose Deaths	Number of drug poisoning deaths per 100,000 population.
Motor Vehicle Crash Deaths	Number of motor vehicle crash deaths per 100,000 population.
Clinical Care Measures	
Uninsured	Percentage of population under age 65 without health insurance.
Primary Care Physician	Ratio of population to primary care physicians.
Dentist	Ratio of population to dentists.
Mental Health Providers	Ratio of population to mental health providers.
Uninsured Adults	Percentage of adults under age 65 without health insurance.
Uninsured Children	Percentage of children under age 19 without health insurance.
Other Primary Care Providers	Ratio of population to primary care providers other than physicians.
Preventable Hospital Stays	Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees.
Health care costs	Amount of price-adjusted Medicare reimbursements per enrollee.
Mammography Screening	Percentage of female Medicare enrollees ages 65-74 that received an annual mammography screening.
Diabetes Monitoring	Percentage of diabetic Medicare enrollees ages 65-75 that receive HbA1c monitoring
Socioeconomic Measures	
High School Graduation	Percentage of ninth-grade cohort that graduates in four years.
Some College	Percentage of adults ages 25-44 with some post-secondary education.
Unemployment	Percentage of population ages 16 and older unemployed but seeking work.
Children In Poverty	Percentage of people under age 18 in poverty.
Median Household Income	The income where half of households in a county earn more and half of households earn less.
Free Lunch	Percentage of children enrolled in public schools that are eligible for free lunch.
Children in Single Parent Households	Percentage of children that live in a household headed by single parent.
Social Association	Number of membership associations per 100,000 population.
Violent Crime	Number of reported violent crime offenses per 100,000 population.
Injury Deaths	Number of deaths due to injury per 100,000 population.
Homicides	Number of deaths due to homicide per 100,000 population.
Physical Environment Measures	
Air Pollution Particulate Matter	Average daily density of fine particulate matter in micrograms per cubic meter (PM2.5).
Driving Alone to Work	Percentage of the workforce that drives alone to work.
Cooling Degree Days (CLDD)	A yearly total of heating degree days – computed when daily average temperature is above 18.3°C/65°F.
Heating Degree Days (HTDD)	A yearly total of heating degree days – computed when daily average temperature is less than 18.3°C/65°F..
DX90	Number of days with maximum temperature >= 32.2°C/90°F.
DT32	Number of days with minimum temperature <= 0°C/32°F.
Precipitation (PRCP)	Total annual precipitation in millimeters.
Snowfall (SNOW)	Total annual snowfall in millimeters.
Average Wind Speed (AWND)	Annual average wind speed in tenths of meters per second.
County Type	Urban-Rural Classification of the 2013 (1-6)
Demographics	
% Hispanic	Percentage of population that is Hispanic.
% Non-Hispanic Black	Percentage of population that is non-Hispanic Black or African American.
% Non-Hispanic White	Percentage of population that is non-Hispanic White.
% Asian	Percentage of population that is Asian.
% American Indian & Alaska Native	Percentage of population that is American Indian or Alaska Native.
% Native Hawaiian/Other Pacific Islander	Percentage of population that is Native Hawaiian or Other Pacific Islander.
% of Below 18 Years Old	Percentage of population below 18 years of age.
% of 65 and Older	Percentage of population ages 65 and older.
Population	Resident population.
% Females	Percentage of population that is female.
% Not Proficient in English	Percentage of population that is not proficient in English.
% Rural	Percentage of population living in a rural area.

univariate model are being validated by conducting statistics significance tests. This model selection process ensures that the model results have low bias as well as low variance, and are generalized enough to be equally valid for any future data provided that the unobserved heterogeneities remain the same [57], [58].

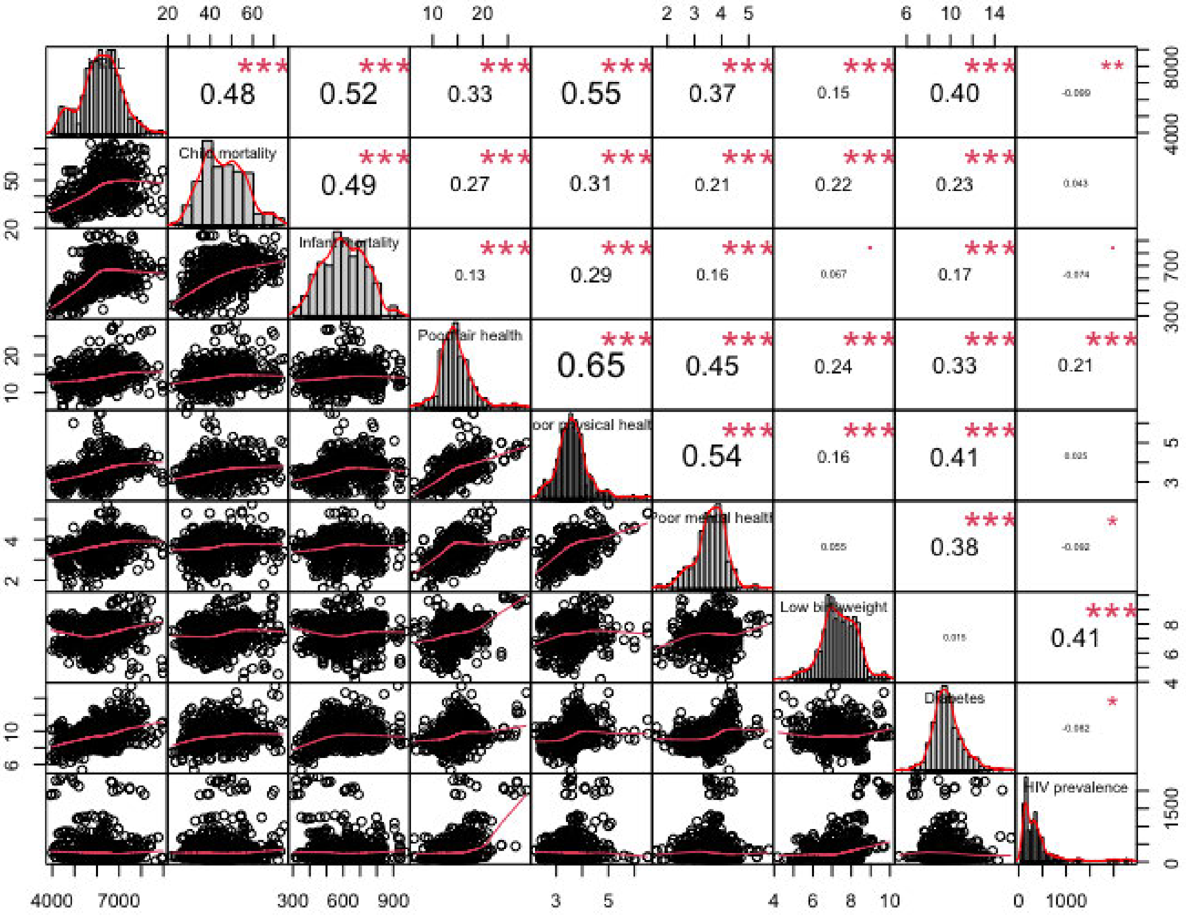


FIGURE 13. Dependencies among multivariate response variables. The numbers are the Spearman correlation coefficients, indicating the degree of two variables are correlated.

VII. CONCLUSION

Accurate prediction and evaluation of population health plays a vital role in the development of a thriving and equitable society. In this paper, we propose a data-driven multivariate framework to simultaneously model nine dimensions (outcomes) of population health—characterizing the length of life and quality of life—as a nonlinear function of health behaviors, clinical care, socioeconomic factors, physical environment and demographics. We also developed a novel percentile-based variable selection for multivariate regression (PVS-MR) method for dimension reduction, without compromising model's generalization performance. To validate different statistical learning models, we implemented an iterative cross validation approach to generalize model's performance, and a statistical significance test for model's results. Furthermore, variable importance heat-map and partial dependence plots are provided to inform decision makers for understanding underlying health determinants and pathways in population.

Using NY state as a case study, we established the applicability of the framework, and quantified the associations linking mortality and mobility to some of key influencing factors in NY. Our numerical analyses suggest that the multivariate tree boosting algorithm best captures non-linearity relationships and interdependence of population health across multiple dimensions. Our findings also indicate that socioeconomic factors and health behaviors are the most predictors of population health in NY.

This study is focused towards modeling various aspects of population health and assessing the key determinants of population health at county-level. Clearly there are some future work directions for multidimensional health studies. First, as individual-level relevant data become available, similar data-driven methodological frameworks can be applied to further study the key determinants of individual health. Such studies can help the state and federal health agencies to design individual-level health intervention strategies. Second, the correlations between health factors and health outcomes revealed in this paper only imply association, and not causation. Future research could utilize the key influencing factors identified in this work along with longitudinal studies, to better examine the casual mechanism of how the key factors affect population health.

.

APPENDIX A VARIABLES DESCRIPTION

See Table 5 for description of all variables in the model.

TABLE 6. Hypothesis testing results of model selection for all response variables.

	p-value (Hypothesis-1)		p-value (Hypothesis-2)	
Response Variables	Goodness of fit	Predictive Accuracy	Goodness of fit	Predictive Accuracy
YPLL	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.50
Child mortality	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.02^{*}
Infant mortality	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.60
Poor or fair health	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$
Poor physical health days	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.03^{*}
Poor mental health days	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.04^{*}
Low birthweight	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.06^{*}
Diabetes	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.14
HIV prevalence	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	0.20

Significance levels: $^{**}p < 0.01$, $^{*}p < 0.1$.

APPENDIX B CORRELATIONS BETWEEN RESPONSE VARIABLES

See Fig. 13 for the correlation and distribution for nine multivariate response variables.

APPENDIX C HYPOTHESIS TESTING

See Table 6 for the statistical results for our two hypotheses.

REFERENCES

[1] *Healthy People 2030*, U.S. Dept. Health Hum. Services Office Disease Prevention Health Promotion, Rockville, MD, USA, 2021.

[2] D. M. Berwick, T. W. Nolan, and J. Whittington, "The triple aim: Care, health, and cost," *Health Affairs*, vol. 27, no. 3, pp. 759–769, May 2008.

[3] NYS Department of Health Strategic Report. *Promotion, Access, Response, Quality, Research*. Accessed: Jul. 1, 2021. [Online]. Available: https://www.health.ny.gov/

[4] D. A. Kindig, Y. Asada, and B. Booske, "A population health framework for setting national and state health goals," *Jama*, vol. 299, no. 17, pp. 2081–2083, 2008.

[5] P. L. Remington, B. B. Catlin, and K. P. Gennuso, "The county health rankings: Rationale and methods," *Population Health Metrics*, vol. 13, no. 1, p. 11, Dec. 2015.

[6] M. Hendryx, M. M. Ahern, and K. J. Zullig, "Improving the environmental quality component of the county health rankings model," *Amer. J. Public Health*, vol. 103, no. 4, pp. 727–732, Apr. 2013.

[7] H. Khedmat, G.-R. Karami, V. Pourfarziani, S. Assari, M. Rezailashkajani, and M. Naghizadeh, "A logistic regression model for predicting health-related quality of life in kidney transplant recipients," in *Transplantation Proc.*, vol. 39, no. 4, 2007, pp. 917–922.

[8] M. M. van Veen, J. Tavares-Brito, B. M. van Veen, J. R. Dusseldorp, P. M. N. Werker, P. U. Dijkstra, and T. A. Hadlock, "Association of regional facial dysfunction with facial Palsy–Related quality of life," *JAMA Facial Plastic Surgery*, vol. 21, no. 1, pp. 32–37, Jan. 2019.

[9] J. M. McGinnis, "Income, life expectancy, and community health: Underscoring the opportunity," *Jama*, vol. 315, no. 16, pp. 1709–1710, 2016.

[10] S. H. Preston and Y. C. Vierboom, "Excess mortality in the United States in the 21st century," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 16, Apr. 2021, Art. no. e2024850118.

[11] Z. Wei and S. Mukherjee, "Health-behaviors associated with the growing risk of adolescent suicide attempts: A data-driven cross-sectional study," *Amer. J. Health Promotion*, vol. 35, no. 5, pp. 688–693, Jun. 2021.

[12] S. Mukherjee, N. Botchwey, and E. F. Boamah, "Towards mental wellbeing in cities: A data-driven learning from mental health—Environment Nexus," in *Proc. 30th Eur. Saf. Rel. Conf. 15th Probabilistic Saf. Assessment Manage. Conf.*, 2020, pp. 1–8.

[13] S. Mukherjee, E. F. Boamah, P. Ganguly, and N. Botchwey, "A multilevel scenario based predictive analytics framework to model the community mental health and built environment Nexus," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, Dec. 2021.

[14] J. Bircher and S. Kuruvilla, "Defining health by addressing individual, social, and environmental determinants: New opportunities for health care and public health," *J. Public Health Policy*, vol. 35, no. 3, pp. 363–386, Aug. 2014.

[15] H. Rutter, N. Savona, K. Glonti, J. Bibby, S. Cummins, D. T. Finegood, F. Greaves, L. Harper, P. Hawe, L. Moore, M. Petticrew, E. Rehfuess, A. Shiell, J. Thomas, and M. White, "The need for a complex systems model of evidence for public health," *Lancet*, vol. 390, no. 10112, pp. 2602–2604, Dec. 2017.

[16] C. M. Hood, K. P. Gennuso, G. R. Swain, and B. B. Catlin, "County health rankings: Relationships between determinant factors and health outcomes," *Amer. J. preventive Med.*, vol. 50, no. 2, pp. 129–135, 2016.

[17] T. J. Anderson, D. M. Saman, M. S. Lipsky, and M. N. Lutfiyya, "A cross-sectional study on health differences between rural and non-rural U.S. counties using the county health rankings," *BMC Health Services Res.*, vol. 15, no. 1, p. 441, Jun. 2015.

[18] J. M. McGinnis, P. Williams-Russo, and J. R. Knickman, "The case for more active policy attention to health promotion," *Health Affairs*, vol. 21, no. 2, pp. 78–93, Mar. 2002.

[19] D. M. Cutler, A. B. Rosen, and S. Vijan, "The value of medical spending in the United States, 1960–2000," *New England J. Med.*, vol. 355, no. 9, pp. 920–927, 2006.

[20] *U.S. Department of Health and Human Services*. New York State Health Assessment 2018. Accessed: Sep. 13, 2021. [Online]. Available: https://www.health.ny.gov/prevention/

[21] L. Jensen, S. M. Monnat, J. J. Green, L. M. Hunter, and M. J. Sliwinski, "Rural population health and aging: Toward a multilevel and multidimensional research agenda for the 2020s," *Amer. J. Public Health*, vol. 110, no. 9, pp. 1328–1331, Sep. 2020.

[22] L. Flores-Payan, D. M. Hernández-Corona, and T. González-Heredia, "Multidimensional analysis of health in mexico: Implementation of fuzzy sets," *BMC Public Health*, vol. 21, no. 1, pp. 1–13, Dec. 2021.

[23] E. A. Dobis, H. M. Stephens, M. Skidmore, and S. J. Goetz, "Explaining the spatial variation in American life expectancy," *Social Sci. Med.*, vol. 246, Feb. 2020, Art. no. 112759.

[24] A. S. Venkataramani, R. O'Brien, and A. C. Tsai, "Declining life expectancy in the United States: The need for social policy as health policy," *JAMA*, vol. 325, no. 7, pp. 621–622, 2021.

[25] S. Harper, C. A. Riddell, and N. B. King, "Declining life expectancy in the United States: Missing the trees for the forest," *Annu. Rev. Public Health*, vol. 42, no. 1, pp. 381–403, Apr. 2021.

[26] S. L. Slabaugh, M. Shah, M. Zack, L. Happe, T. Cordier, E. Havens, E. Davidson, M. Miao, T. Prewitt, and H. Jia, "Leveraging health-related quality of life in population health management: The case for healthy days," *Population Health Manage.*, vol. 20, no. 1, pp. 13–22, Feb. 2017.

[27] R. R. Rubin and M. Peyrot, "Quality of life and diabetes," *Diabetes/Metabolism Res. Rev.*, vol. 15, no. 3, pp. 205–218, 1999.

[28] J. T. Coffey, M. Brandle, H. Zhou, D. Marriott, R. Burke, B. P. Tabaei, M. M. Engelgau, R. M. Kaplan, and W. H. Herman, "Valuing health-related quality of life in diabetes," *Diabetes Care*, vol. 25, no. 12, pp. 2238–2243, Dec. 2002.

[29] O. Solli, K. Stavem, and I. S. Kristiansen, "Health-related quality of life in diabetes: The associations of complications with EQ-5D scores," *Health Quality life outcomes*, vol. 8, no. 1, pp. 1–8, 2010.

[30] M. Y. M. Leung, N. P. Carlsson, G. A. Colditz, and S.-H. Chang, "The burden of obesity on diabetes in the United States: Medical expenditure panel survey, 2008 to 2012," *Value Health*, vol. 20, no. 1, pp. 77–84, Jan. 2017.

[31] M. K. Palmer and P. P. Toth, "Trends in lipids, obesity, metabolic syndrome, and diabetes mellitus in the United States: An NHANES analysis (2003–2004 to 2013–2014)," *Obesity*, vol. 27, no. 2, pp. 309–314, Feb. 2019.

[32] R. E. Glasgow, L. Ruggiero, E. G. Eakin, J. Dryfoos, and L. Chobanian, "Quality of life and associated characteristics in a large national sample of adults with diabetes," *Diabetes Care*, vol. 20, no. 4, pp. 562–567, Apr. 1997.

[33] S. Saigal, "Self-perceived health status and health-related quality of life of extremely low-birth-weight infants at adolescence," *JAMA: J. Amer. Med. Assoc.*, vol. 276, no. 6, pp. 453–459, Aug. 1996.

[34] J. Fellinger, D. Holzinger, U. Dobner, J. Gerich, R. Lehner, G. Lenz, and D. Goldberg, "Mental distress and quality of life in a deaf population," *Social Psychiatry Psychiatric Epidemiology*, vol. 40, no. 9, pp. 737–742, Sep. 2005.

[35] C. Hinnell, J. Williams, A. Metcalfe, S. B. Patten, R. Parker, S. Wiebe, and N. Jetté, "Health status and health-related behaviors in epilepsy compared to other chronic conditions—A national population-based study," *Epilepsia*, vol. 51, no. 5, pp. 853–861, May 2010.

[36] G. Michel, C. Bisegger, D. C. Fuhr, and T. Abel, "Age and gender differences in health-related quality of life of children and adolescents in Europe: A multilevel analysis," *Qual. Life Res.*, vol. 18, no. 9, pp. 1147–1157, Nov. 2009.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[38] E. J. Schmitt and H. Jula, "On the limitations of linear models in predicting travel times," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2007, pp. 830–835.

[39] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–15, Dec. 2019.

[40] S. Mukherjee and Z. Wei, "Suicide disparities across metropolitan areas in the U.S.: A comparative assessment of socio-environmental factors using a data-driven predictive approach," *PLoS ONE*, vol. 16, no. 11, Nov. 2021, Art. no. e0258824.

[41] V. Etches, J. Frank, E. D. Ruggiero, and D. Manuel, "Measuring population health: A review of indicators," *Annu. Rev. Public Health*, vol. 27, no. 1, pp. 29–55, Apr. 2006.

[42] D. G. Moriarty, R. Kobau, M. M. Zack, and H. S. Zahran, "Tracking healthy days—A window on the health of older adults," *Preventing Chronic Disease*, vol. 2, no. 3, pp. 1–8, 2005.

[43] Y.-H. Lin, A. C. McLain, J. C. Probst, K. J. Bennett, Z. P. Qureshi, and J. M. Eberth, "Health-related quality of life among adults 65 years and older in the United States, 2011–2012: A multilevel small area estimation approach," *Ann. Epidemiol.*, vol. 27, no. 1, pp. 52–58, Jan. 2017.

[44] P. Wu, C. W. Hoven, P. Cohen, X. Liu, R. E. Moore, Q. Tiet, N. Okezie, J. Wicks, and H. R. Bird, "Factors associated with use of mental health services for depression by children and adolescents," *Psychiatric Services*, vol. 52, no. 2, pp. 189–195, Feb. 2001.

[45] J. M. Samet, F. Dominici, F. C. Curriero, I. Coursac, and S. L. Zeger, "Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994," *New England J. Med.*, vol. 343, no. 24, pp. 1742–1749, Dec. 2000.

[46] A. Zanobetti and J. Schwartz, "Temperature and mortality in nine US cities," *Epidemiology*, vol. 19, no. 4, p. 563, 2008.

[47] J. F. Figueroa, A. B. Frakt, and A. K. Jha, "Addressing social determinants of health: Time for a polysocial risk score," *Jama*, vol. 323, no. 16, pp. 1553–1554, 2020.

[48] *County Health Rankings & Roadmaps. Measures & Data Sources*. Accessed: Jul. 1, 2021. [Online]. Available: https://www.countyhealthrankings.org/

[49] *National Oceanic and Atmospheric Administration. Climate Data Online Data Tools*. Accessed: Jul. 1, 2021. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datatools

[50] M. M. Rahman and D. N. Davis, "Machine learning-based missing value imputation method for clinical datasets," in *IAENG Transactions on Engineering Technologies*. Dordrecht, The Netherlands: Springer, 2013, pp. 245–257.

[51] S. Van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *J. Of Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[52] S. Mukherjee and R. Nateghi, "Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States," *Energy*, vol. 128, pp. 688–700, Jun. 2017.

[53] S. Mukherjee and R. Nateghi, "A data-driven approach to assessing supply inadequacy risks due to climate-induced shifts in electricity demand," *Risk Anal.*, vol. 39, no. 3, pp. 673–694, Mar. 2019.

[54] *United Nations. Ensure Healthy Lives and Promote Well-Being for All at All Ages*. Accessed: Jul. 1, 2021. [Online]. Available: https://www.un.org/sustainabledevelopment/

[55] R. Obringer, S. Mukherjee, and R. Nateghi, "Evaluating the climate sensitivity of coupled electricity-natural gas demand using a multivariate framework," *Appl. Energy*, vol. 262, Mar. 2020, Art. no. 114419.

[56] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer, 2001.

[57] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.

[58] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer Series in Statistics), vol. 1, no. 10. New York, NY, USA: Springer, 2001.

[59] R. Nateghi, "Multi-dimensional infrastructure resilience modeling: An application to hurricane-prone electric power distribution systems," *IEEE Access*, vol. 6, pp. 13478–13489, 2018.

[60] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL, USA: CRC Press, 2020.

[61] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 161–168.

[62] J. Klett, *Applied Multivariate Analysis*. New York, NY, USA: McGraw, 1972.

[63] C. J. Huberty and M. D. Petoskey, "Multivariate analysis of variance and covariance," in *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Amsterdam, The Netherlands: Elsevier, 2000, pp. 183–208.

[64] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *J. Animal Ecol.*, vol. 77, no. 4, pp. 802–813, Jul. 2008.

[65] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[66] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

[67] Z. Quan and E. A. Valdez, "Predictive analytics of insurance claims using multivariate decision trees," *Dependence Model.*, vol. 6, no. 1, pp. 377–407, Dec. 2018.

[68] P. J. Miller, G. H. Lubke, D. B. McArtor, and C. Bergeman, "Finding structure in data using multivariate tree boosting," *Psychol. methods*, vol. 21, no. 4, p. 583, 2016.

[69] J. Fan and R. Li, "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," Tech. Rep. math/0602133, 2006.

[70] D. Foster, H. Karloff, and J. Thaler, "Variable selection is hard," in *Proc. Conf. Learn. Theory*, 2015, pp. 696–709.

[71] R. Chen, J. G. Andrews, and R. W. Heath, Jr., "Efficient transmit antenna selection for multiuser mimo systems with block diagonalization," in *Proc. IEEE IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2007, pp. 3499–3503.

[72] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[73] S. Mukherjee, R. Nateghi, and M. Hastak, "A multi-hazard approach to assess severe weather-induced major power outage risks in the U.S.," *Rel. Eng. Syst. Saf.*, vol. 175, pp. 283–305, Jul. 2018.

[74] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6673–6690, Jan. 2017.

[75] H. I. Hall, R. Song, P. Rhodes, J. Prejean, Q. An, L. M. Lee, J. Karon, R. Brookmeyer, E. H. Kaplan, M. T. McKenna, and R. S. Janssen, "Estimation of HIV incidence in the United States," *Jama*, vol. 300, no. 5, pp. 520–529, 2008.

[76] N. El-Bassel, N. A. Caldeira, L. M. Ruglass, and L. Gilbert, "Addressing the unique needs of African American women in HIV prevention," *Amer. J. Public Health*, vol. 99, no. 6, pp. 996–1001, Jun. 2009.

[77] M. A. Lazar, "How obesity causes diabetes: Not a tall tale," *Science*, vol. 307, no. 5708, pp. 373–375, Jan. 2005.

[78] P. Hossain, B. Kawar, and M. El Nahas, "Obesity and diabetes in the developing world—A growing challenge," *New England J. Med.*, vol. 356, no. 3, pp. 213–215, 2007.

[79] P. Dandona, "Inflammation: The link between insulin resistance, obesity and diabetes," *Trends Immunol.*, vol. 25, no. 1, pp. 4–7, Jan. 2004.

[80] C. Gundersen, B. Kreider, and J. Pepper, "The impact of the national school lunch program on child health: A nonparametric bounds analysis," *J. Econometrics*, vol. 166, no. 1, pp. 79–91, Jan. 2012.

[81] S. Brophy, A. Rees, G. Knox, J. Baker, and N. E. Thomas, "Child fitness and Father's BMI are important factors in childhood obesity: A school based cross-sectional study," *PLoS ONE*, vol. 7, no. 5, May 2012, Art. no. e36597.

[82] *Mental Health Disorder Statistics*, Johns Hopkins Univ. Med., Baltimore, MD, USA, 2021.

[83] C. B. R. Evans, P. R. Smokowski, and K. L. Cotter, "Cumulative bullying victimization: An investigation of the dose–response relationship between victimization and the associated mental health outcomes, social supports, and school experiences of rural adolescents," *Children Youth Services Rev.*, vol. 44, pp. 256–264, Sep. 2014.

[84] J. L. Aber, N. G. Bennett, D. C. Conley, and J. Li, "The effects of poverty on child health and development," *Annu. Rev. Public Health*, vol. 18, no. 1, pp. 463–483, May 1997.

[85] F. Sinić, "Adolescent pregnancy is a serious social problem," *J. Gynecol. Res. Obstetrics*, vol. 4, pp. 006–008, Apr. 2018.

[86] G. M. Kassa, A. O. Arowojolu, A. A. Odukogbe, and A. W. Yalew, "Prevalence and determinants of adolescent pregnancy in Africa: A systematic review and meta-analysis," *Reproductive Health*, vol. 15, no. 1, pp. 1–17, Dec. 2018.

[87] I. Lut, J. Woodman, A. Armitage, E. Ingram, K. Harron, and P. Hardelid, "Health outcomes, healthcare use and development in children born into or growing up in single-parent households: A systematic review study protocol," *BMJ Open*, vol. 11, no. 2, Feb. 2021, Art. no. e043361.

[88] D. Nishioka, J. Saito, K. Ueno, and N. Kondo, "Single-parenthood and health conditions among children receiving public assistance in Japan: A cohort study," *BMC Pediatrics*, vol. 21, no. 1, pp. 1–9, Dec. 2021.

[89] M. G. Phipps, M. Sowers, and S. M. DeMonner, "The risk for infant mortality among adolescent childbearing groups," *J. Women's Health*, vol. 11, no. 10, pp. 889–897, Dec. 2002.

[90] M. Adeolu, O. Akpa, A. Adeolu, and I. Aladeniyi, "Environmental and socioeconomic determinants of child mortality: Evidence from the 2013 Nigerian demographic health survey," *Amer. J. Public Health Res.*, vol. 4, no. 4, pp. 41–134, 2016.

[91] S. Cha and Y. Jin, "Have inequalities in all-cause and cause-specific child mortality between countries declined across the world?" *Int. J. Equity Health*, vol. 19, no. 1, pp. 1–13, Dec. 2020.

[92] N. B. Johnson, L. D. Hayes, K. Brown, E. C. Hoo, and K. A. Ethier, "CDC national health report: Leading causes of morbidity and mortality and associated behavioral risk and protective factors—United States, 2005–2013," Centers Disease Control, Tech. Rep., 2014, vol. 63, no. 4.

[93] T. Nyári, M. Woodward, G. Mészáros, J. Karsai, and L. Kovács, "Chlamydia trachomatis infection and the risk of perinatal mortality in Hungary," De Gruyter, Tech. Rep., 2001, pp. 55–59, vol. 29.

[94] A. J. Warr, J. Pintye, J. Kinuthia, A. L. Drake, J. A. Unger, R. S. McClelland, D. Matemo, L. Osborn, and G. John-Stewart, "Sexually transmitted infections during pregnancy and subsequent risk of stillbirth and infant mortality in kenya: A prospective study," *Sexually Transmitted Infections*, vol. 95, no. 1, pp. 60–66, Feb. 2019.

[95] C. Cheon, Y. Lin, D. J. Harding, W. Wang, and D. S. Small, "Neighborhood racial composition and gun homicides," *JAMA Netw. Open*, vol. 3, no. 11, Nov. 2020, Art. no. e2027591.

[96] P. Blumenshine, S. Egerter, C. J. Barclay, C. Cubbin, and P. A. Braveman, "Socioeconomic disparities in adverse birth outcomes: A systematic review," *Amer. J. preventive Med.*, vol. 39, no. 3, pp. 263–272, 2010.

[97] J. W. Collins, J. Wambach, R. J. David, and K. M. Rankin, "Women's lifelong exposure to neighborhood poverty and low birth weight: A population-based study," *Maternal Child Health J.*, vol. 13, no. 3, pp. 326–333, May 2009.

[98] K. W. Strully, D. H. Rehkopf, and Z. Xuan, "Effects of prenatal poverty on infant health: State earned income tax credits and birth weight," *Amer. Sociol. Rev.*, vol. 75, no. 4, pp. 534–562, Aug. 2010.

[99] J. E. Watson, R. S. Kirby, K. J. Kelleher, and R. H. Bradley, "Effects of poverty on home environment: An analysis of three-year outcome data for low birth weight premature infants," *J. Pediatric Psychol.*, vol. 21, no. 3, pp. 419–431, 1996.

[100] O. Laurent, J. Hu, L. Li, M. J. Kleeman, S. M. Bartell, M. Cockburn, L. Escobedo, and J. Wu, "Low birth weight and air pollution in california: Which sources and components drive the risk?" *Environ. Int.*, vols. 92–93, pp. 471–477, Jul. 2016.

[101] C. Li, M. Yang, Z. Zhu, S. Sun, Q. Zhang, J. Cao, and R. Ding, "Maternal exposure to air pollution and the risk of low birth weight: A meta-analysis of cohort studies," *Environ. Res.*, vol. 190, Nov. 2020, Art. no. 109970.

[102] B. O'Hare, I. Makuta, L. Chiwaula, and N. Bar-Zeev, "Income and child mortality in developing countries: A systematic review and meta-analysis," *J. Roy. Soc. Med.*, vol. 106, no. 10, pp. 408–414, Oct. 2013.

[103] H. W. Morrow, G. F. Chávez, P. P. Giannoni, and R. S. Shah, "Infant mortality and related risk factors among Asian Americans," *Amer. J. Public Health*, vol. 84, no. 9, pp. 1497–1500, Sep. 1994.

[104] J. W. Gardner and J. S. Sanborn, "Years of potential life lost (YPLL)—What does it measure?" *Epidemiology*, vol. 1, no. 4, pp. 322–329, Jul. 1990.

[105] C. R. Ronzio, "The politics of preventable deaths: Local spending, income inequality, and premature mortality in U.S. cities," *J. Epidemiol. Community Health*, vol. 58, no. 3, pp. 175–179, Mar. 2004.

[106] (2019). CDC. *Ten Leading Causes of Death*. United States. [Online]. Available: https://wisqars-viz.cdc.gov:8006/lcd/home/

[107] Kaiser Family Foundation. *Women and HIV in the United States*. [Online]. Available: https://www.kff.org/hivaids/fact-sheet/

[108] J. Prejean, R. Song, A. Hernandez, R. Ziebell, T. Green, F. Walker, L. S. Lin, Q. An, J. Mermin, A. Lansky, and H. I. Hall, "Estimated hiv incidence in the United States, 2006–2009," *PLoS ONE*, vol. 6, no. 8, p. e17502, 2011.

ZHIYUAN WEI received the M.S. degree in industrial economics from the University of International Business and Economics (UIBE), Beijing, China, and the M.S. degree in industrial engineering from University at Buffalo—The State University of New York, Buffalo, NY, USA, where he is currently pursuing the Ph.D. degree in industrial engineering. His research interests include data analytics and operations research with a focus on public health.

ADIL BARAN NARIN received the B.S. degree in industrial engineering from Istanbul Bilgi University, in June 2018, and the M.S. degree in industrial and systems engineering with an operations research concentration from University at Buffalo—The State University of New York, in May 2021. His research interests include the domains of machine learning, operations research, data science, and data-driven decision making.

SAYANTI MUKHERJEE received the Ph.D. degree in civil engineering and the M.S. degree in economics from Purdue University, USA, and the M.S. degree in civil engineering from Iowa State University, USA. She is currently an Assistant Professor in industrial and systems engineering with University at Buffalo—The State University of New York, USA. She conducts cutting-edge interdisciplinary research to address the resiliency and sustainability challenges related to the vulnerable communities and socio-technical systems leveraging advanced machine learning algorithms and robust optimization techniques. More specifically, she develops quantitative models to examine systemic/stochastic impacts of various chronic/acute shocks on the interdependent socio-technical systems, develop risk-informed decision models, and investigate cost-effective adaptation measures to advance resilience and sustainability of our communities and critical infrastructure systems.