SSM - Population Health 4 (2018) 95–99

Contents lists available at ScienceDirect

SSM - Population Health

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

Short report

Machine learning approaches to the social determinants of health in the health and retirement study

Benjamin Seligman[a,*], Shripad Tuljapurkar[b], David Rehkopf[c]

[a] Department of Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA
[b] Department of Biology, Stanford University, Stanford, CA 94305, USA
[c] Department of Medicine, School of Medicine, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Keywords:
Social determinants of health
Machine learning
Neural network
Biomarkers
Health and retirement study

ABSTRACT

Background: Social and economic factors are important predictors of health and of recognized importance for health systems. However, machine learning, used elsewhere in the biomedical literature, has not been extensively applied to study relationships between society and health. We investigate how machine learning may add to our understanding of social determinants of health using data from the Health and Retirement Study.

Methods: A linear regression of age and gender, and a parsimonious theory-based regression additionally incorporating income, wealth, and education, were used to predict systolic blood pressure, body mass index, waist circumference, and telomere length. Prediction, fit, and interpretability were compared across four machine learning methods: linear regression, penalized regressions, random forests, and neural networks.

Results: All models had poor out-of-sample prediction. Most machine learning models performed similarly to the simpler models. However, neural networks greatly outperformed the three other methods. Neural networks also had good fit to the data (R^2 between 0.4–0.6, versus < 0.3 for all others). Across machine learning models, nine variables were frequently selected or highly weighted as predictors: dental visits, current smoking, self-rated health, serial-seven subtractions, probability of receiving an inheritance, probability of leaving an inheritance of at least $10,000, number of children ever born, African-American race, and gender.

Discussion: Some of the machine learning methods do not improve prediction or fit beyond simpler models, however, neural networks performed well. The predictors identified across models suggest underlying social factors that are important predictors of biological indicators of chronic disease, and that the non-linear and interactive relationships between variables fundamental to the neural network approach may be important to consider.

1. Introduction

Biomedical practice and research often generate large quantities of data, from administrative records to molecular information. While how to “learn from data” is not a new challenge, the scale of data has prompted interest in algorithm driven approaches to analysis and interpretation. Due to the large number of loci studied and relative lack of *a priori* knowledge relevant to a particular disease, genomic research has been both a major user and source of innovation in these methods (Risch and Merikangas, 1996). These approaches have also been used in environmental health and nutrition, identifying environmental contaminants that have strong associations with diabetes (Patel, Bhattacharya, & Butte, 2010), adverse lipid profiles (Patel et al., 2012) as well as micronutrient associations with hypertension (Tzoulaki et al., 2012). They have been used to study pediatric obesity and mortality as well (Rehkopf and Laraia, 2011; Patel et al., 2013). Similarly, there is a proliferation of “-omics” approaches to studying disease, such as metabolomics (Trygg, Holmes, & Lundstedt, 2007; Wang et al., 2011; Wishart, 2016; Fearnley and Inouye, 2016) and epigenomics (Emes and Farrell, 2012; Lee et al., 2012; Horvath, 2013), which seek to understand biochemical pathway and genetic regulatory bases of disease respectively.

By contrast, research on the social determinants of health has usually focused on hypothesis-driven models to understand how factors such as poverty and education contribute to health. This has aided in understanding causal mechanisms underlying social determinants' effects on health. This focus on causation has perhaps in some ways been one reason why there has been a limited use of machine learning, although efforts to bring causal inference to machine learning are making great strides (van der Laan and Rose, 2011; Varian, 2014; Athey and

[*] Correspondence to: 1245. 16th Street, Ste. 309, Santa Monica, CA 90404, USA.
E-mail address: bseligman@mednet.ucla.edu (B. Seligman).

https://doi.org/10.1016/j.ssmph.2017.11.008
Received 26 July 2016; Received in revised form 15 November 2017; Accepted 16 November 2017

Imbens, 2015) with compelling results (Ahern, Balzer, & Galea, 2015).

Like genomic studies, many social science studies also generate large quantities of data. There is a role for machine learning to explore these data as hypothesis generation and validation of theory (Raftery, 1995; Sala-I-Martin, 1997; Hendry and Krolzig, 2004; Glymour and Osypuk, 2013). In addition to traditional survey data, information such as credit scores and social networks have predictive power for health and add to our understanding of how social determinants may operate; (Christakis and Fowler, 2007; Israel et al., 2014) integrating multiple sources of data will increase the scale of potentially useful datasets. It is important to understand how methods commonly used to analyze and interpret "big data" may be applied to the social determinants of health. Two questions about machine learning methods are particularly relevant: first, do they lead to substantially better predictions than models based on established theory about the social determinants of health, and second, do they enhance our understanding of how social determinants may result in differences in health outcomes?

We compare four major regression based methods in machine learning with both a minimal and a theory-driven model. We assess the performance of each in predicting four health-related biomarkers using data from a large social science survey. Secondarily, we also consider the interpretability of the models. The answers to our study question are relevant both to professionals managing social, educational, or health service data systems as well as scientists exploring high-dimensional social data.

2. Methods

2.1. Data

Data were from the Health and Retirement Study (HRS), a rolling cohort of men and women 50 years old and above and their spouses begun in 1992, with biennial follow-up and periodic recruitment of eligible new participants; this analysis incorporates only primary participants, not spouses (Health and Retirement Study, RAND public use dataset, 2014). 15,784 participants had medical examinations with anthropometry and blood biomarker measurement in either 2006 or 2008 (Crimmins, Guyer, & Langa, 2008). We investigate four outcomes that are biological markers of chronic disease risk: systolic blood pressure (SBP, N = 13,784), body mass index (BMI, N = 13,568), waist circumference (N = 13,995), and telomere length (N = 5808). Telomere length was measured from buccal cells collected from a smaller subsample of HRS respondents than the other measurements. Biologically implausible values were removed as described in Supplementary Table 1; the logarithm of values for telomere length was taken following removal of biologically implausible values to eliminate skew. Distributions of each biomarker are given in Supplementary Figure 1. The first three are associated with a variety of health risks, including cardiovascular disease, stroke, and diabetes. Further, BMI and waist circumference are related measures, both intended to assess adiposity. We consider telomere length as a novel biomarker that may have associations with health, in particular with cardiovascular disease (Haycock et al., 2014), but for which connections to health are less established.

Social and economic data on participants were taken from the RAND HRS data file version N for the wave prior to the measurement of the biomarkers (RAND, 2014). The RAND HRS Data file is an easy to use longitudinal data set based on the HRS data. It was developed at RAND with funding from the National Institute on Aging and the Social Security Administration. Among the variables included are information on individual, spousal, and household income and wealth, education, family structure, receipt of Social Security and other benefits, and health behaviors.

Variables from all sections of the survey were initially included. These are predominantly social and economic data, including health and health insurance, family structure, income, pensions, Social Security, and employment. Based on *a priori* criteria there were categories of variables that we did not include in our analysis: 1) variables with more than 10% missing values; 2) subject, household, and wave identifiers; 3) death variables; and 4) biometric or certain health variables from the RAND dataset that were duplicates of data from the biomarker file or were closely associated with them (i.e. BMI, cholesterol, height, weight, hypertension, diabetes, stroke, heart disease, lung disease, and number of conditions). Binary and categorical variables were then mode-imputed for missing data and categorical variables converted into dummy variables. Variables with a variance less than 0.0475 (equivalent to a binary variable with at least 95% of values in one category) were then removed. The resulting 458 variables were then standardized and missing values of continuous variables were mean-imputed.

2.2. Analytic methods

To assess different machine learning methods' ability to predict the biomarkers of interest, we first considered two OLS regression models. The first was minimal and included gender, age, and age squared. The second was based on current understanding of social determinants of health, particularly that education and economic position have demonstrated associations with health. This theory-based model was parsimonious and included, as linear variables, household income, household wealth, and two binary variables indicating a high school-level education and less than a high school-level education, in addition to the parameters in the minimal model.

We next consider four machine learning algorithms: repeated linear regressions - akin to genome-wide association studies (GWAS), penalized linear regressions (Hastie, 2009), random forests (Breiman, 2001), and neural networks (Kriesel, 2007). These cover parametric and non-parametric approaches, with varying abilities to account for non-linearity. While it is not possible to consider all machine learning algorithms, in addition to the broad coverage offered by these algorithms, all have been used in the medical literature (Patel et al., 2010; Rehkopf and Laraia, 2011; Horvath, 2013; Kapetanovic, Rosenfeld, & Izmirlian, 2004; Sato et al., 2005; Goldstein et al., 2010) and penalized regressions and random forests are particularly commonly-taught methods (Hastie, 2009; Bishop, 2006). These four also offer some prospect for interpretation rather than being completely "black box" approaches.

Approaches similar to GWAS have been previously used in studies surveying many potential disease predictors (Patel et al., 2013, 2015) however this is the first attempt to systematically analyze the associations between a broad range of social measures and biomarkers of health. For brevity we refer to this as SWAS, for society-wide association study. Similar to GWAS, for each biomarker Y we screen for adjusted associations with the candidate predictor X using the following model:

$$Y_i = \alpha + \beta_{Gender} Gender_i + \beta_{Age} Age_i + \beta_{Age\hat{}2} Age_{i}^{2} + \beta_k X_i + \varepsilon_i$$

where subscript i denotes one of the subjects in the dataset, βs are regression coefficients, α is the y-intercept, and ε is an independent, normally-distributed error term with a mean of 0. P-values for β_k were deemed significant if they were below a Bonferroni-corrected α = 0.05. Those variables with statistically significant β_k were then included in a final linear regression model of the biomarker Y.

LASSO is a penalized regression, adding the sum of the absolute values of the coefficients in the model to the residual sum of squares, as in this formula for a linear regression (Hastie, 2009):

$$RSS_{LASSO} = \frac{1}{2}\sum_{i=1}^{N}(Y_i - \alpha - \sum_{k=1}^{P}\beta_k X_{i,k})^2 + \lambda \sum_{k=1}^{P}|\beta_k|$$

Where i, Y, X, α, and β are as defined above, k denotes the different variables included in the model, P is the total number of variables in the model, N is the total number of subjects in the model, and λ is a weight