Preventive Medicine Reports 20 (2020) 101238

Contents lists available at ScienceDirect

Preventive Medicine Reports

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

Machine learning to identify and understand key factors for provider-patient discussions about smoking

Liangyuan Hu [*], Lihua Li, Jiayi Ji

Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA
The Institute for Healthcare Delivery, Mount Sinai Health System, New York, NY, USA

ARTICLE INFO

Keywords:
Smoking cessation
Bayesian machine learning
Variable selection
Survey data

ABSTRACT

We sought to identify key determinants of the likelihood of provider-patient discussions about smoking and to understand the effects of these determinants. We used data on 3666 self-reported current smokers who talked to a health professional within a year of the time the survey was conducted using the 2017 National Health Interview Survey. We included wide-ranging information on 43 potential covariates across four domains, demographic and socio-economic status, behavior, health status and healthcare utilization. We exploited a principled nonparametric permutation based approach using Bayesian machine learning to identify and rank important determinants of discussions about smoking between health providers and patients. In the order of importance, frequency of doctor office visits, intensity of cigarette use, length of smoking history, chronic obstructive pulmonary disease, emphysema, marital status were major determinants of disparities in provider-patient discussions about smoking. There was a distinct interaction between intensity of cigarette use and length of smoking history. Our analysis may provide some insights into strategies for promoting discussions on smoking and facilitating smoking cessation. Health care resource usage, smoking intensity and duration and smoking-related conditions were key drivers. The “usual suspects”, age, gender, race and ethnicity were less important, and gender, in particular, had little effect.

1. Introduction

Cigarette smoking is the leading preventable cause of death in the United States (The Health Consequences of Smoking - 50 Years of Progress: A Report of the Surgeon General, 2014). Smoking causes more than 480,000 deaths – that is nearly one in five deaths – each year in the US. Smoking harms nearly every organ of the body, causes many diseases including cardiovascular disease (CVD), respiratory disease, cancer and other health risks (The Health Consequences of Smoking - 50 Years of Progress: A Report of the Surgeon General, 2014; How Tobacco Smoke Causes Disease: What It Means to You, 2010; Women and Smoking: A Report of the Surgeon General, 2001). There are both immediate and long-term health benefits of quitting for all smokers, including lowered heart rate and blood pressure, reduced risk of various heart and lung diseases and cancer diseases. Quitting smoking also reduces the excess risk of many diseases related to second-hand smoke in children (The Health Consequences of Smoking: A Report of the Surgeon General. Office of the Surgeon General (US), Office on Smoking and Health (US), Atlanta (GA): Centers for Disease Control and Prevention (US), 2004; Mahmud and Feely, 2003).

Despite the fact that considerable progress has been made in reducing cigarette smoking and the prevalence of cigarette smoking has reached the lowest point of 14% in the U.S. adult population since the Surgeon General’s report on smoking in 1964, smoking remains to be the leading preventable cause for many diseases (Samet, 2013; Smoking Cessation: A Report of the Surgeon General, 2020; Critchley and Capewell, 2003). To help smokers quit, the U.S. Public Health Service’s Clinical Practice Guideline, Treating Tobacco Use and Dependence, recommends provider-patient discussions about smoking as one of the five ‘A’s approaches (Toll et al., 2014; Treating Tobacco Use and Dependence, 2008; Helping Smokers Quit, 2008). The guideline suggests providers:1) ask about tobacco use at every visit; 2) advise all tobacco users to quit; 3) assess readiness to quit; 4) assist tobacco users with a quit plan; and 5) arrange follow-up visits.

Although studies have shown provider-patient discussions about smoking has a positive effect on smoking cessation, this

[*] Corresponding author at: Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, One Gustave L. Levy Place, Box 1077, New York, NY 10029, USA.
E-mail address: liangyuan.hu@mssm.edu (L. Hu).

https://doi.org/10.1016/j.pmedr.2020.101238
Received 15 August 2020; Received in revised form 7 October 2020; Accepted 20 October 2020
Available online 5 November 2020

recommendation has not been widely adopted in clinical practice (Kruger et al., 2012; Solberg et al., 2001; Bao et al., 2006). Based on nationally representative samples from the National Health Interview Survey (NHIS), the prevalence of provider-patient discussions about smoking ranged from 51.3% in 2011 to 55.4% in 2015 (Huo et al., 2020). Identifying key factors for and understanding how they are associated with the likelihood of provider-patient discussions about smoking are important to improving the discussion rate and to reducing cigarette smoking.

Existing studies have linked the likelihood of a discussion on smoking between physician and patient to several factors, including sex, age, race/ethnicity, education, health insurance, and health condition (Bao et al., 2006; Huo et al., 2020). For example, Hispanics were found to be less likely to receive consulting than non-Hispanic white (Cokkinides et al., 2008; Lopez-Quintero et al., 2006; Centers for Disease Control and Prevention. Quitting smoking among adults - United States, 2011). The presence of respiratory conditions was shown to be a strong predictor of provider-patient discussions about smoking (Bao et al., 2006; Huo et al., 2020). Gender and insurance coverage were also identified as contributing factors, for example, uninsured male smokers tend to be less likely asked for smoking (Huo et al., 2020; King et al., 2013). While these factors were based on domain knowledge on the survey questions, there could also be important variables overlooked because still much remains unknown about smoking cession. With the advancement in statistical machine learning, innovative and principled variable selection procedures have been made accessible to applied researchers working with healthcare data. Leveraging these new developments to identity key determinants of the provider-patient discussions about smoking in a more holistic way may complement current research about smoking cessation (Bao et al., 2006; Huo et al., 2020).

This study exploited a state-of-the-art machine learning technique, Bayesian Additive Regression Trees (BART), to identify key determinants of the likelihood of provider-patient discussions about smoking using a nationally representative survey. We also evaluated the associations between the identified major determinants and interactions thereof and the likelihood of provider-patient discussions about smoking. Comparisons with two commonly used logistic regression based variable selection approaches were performed. Results from our study suggest that health care resource usage, smoking intensity and duration and smoking-related conditions were the most important factors.

2. Methods

2.1. Data source

We used publicly available data from the 2017 NHIS. NHIS is an ongoing, yearly, cross-sectional, in-person household survey using a multistage sampling design to survey approximately 87,500 persons in 35,000 households representative of the civilian non-institutionalized population in the US. The response rate was 67.9%, 98.9%, and 80.9% for household, family, and sample adult components, respectively (Centers for Disease Control and Prevention, 2016).

Among 33,028 persons interviewed for the sample adult questions, 4163 adults were self-reported current smokers and talked to a health professional within a year of the time the survey was conducted. We then excluded 497 individuals who had missing values in key baseline variables. Our final analysis data set included 3666 adult smokers. Fig. 1 shows a flowchart of the sample selection for our analysis.

The outcome variable is an indicator of whether current smokers and their providers discussed about smoking during the year prior to the time of survey. It was defined based on participants' responses to the question of 'During the past 12 months, has a doctor or other health professionals talked to you about your smoking?' Those whose answers were 'Refused', 'Not ascertained' or 'Don't know' were considered as missing.

We included 43 potential covariates across four domains. The *demographic and socio-economic status* domain included age, gender, education, current employment status, family annual income, number of

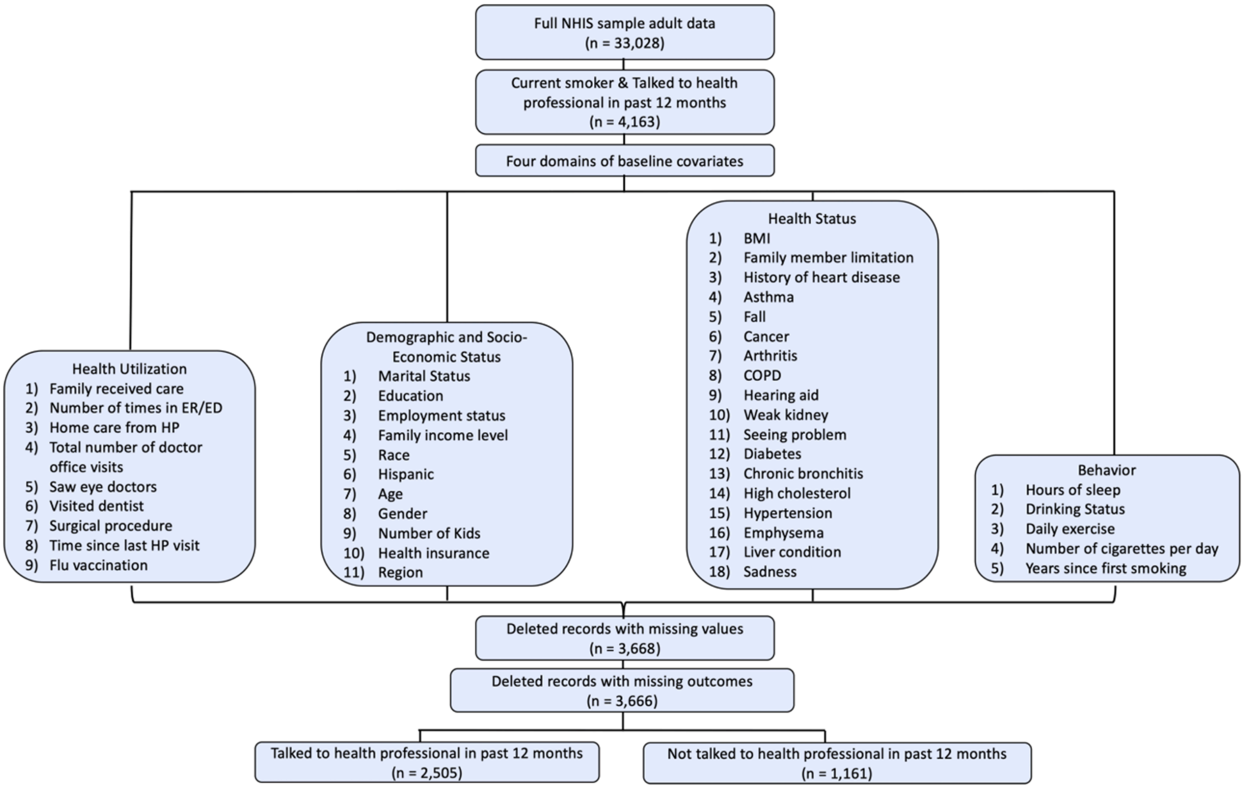

Fig. 1. Data selection procedures.

children, race, marital status, health insurance, Hispanic and region.

Derived from the NHIS adult sample, *behavior* variables included drinking, daily exercise, hours of sleep, number of cigarettes per day and years since first smoking. *Health status* was drawn from the adult file and family file. We dichotomized the co-morbidity variables as yes/no for heart disease, asthma, falling, arthritis, chronic obstructive pulmonary disease (COPD), hearing aid, kidney problem, eye vision problem, diabetes, chronic bronchitis, high cholesterol, hypertension, emphysema, liver condition and mental problem. To define the cancer variable,

we categorized it into tobacco-related cancers, other cancers and no cancer (Predefined SEER Stat Variables for Calculating the Number of Associated Cancers for Selected Risk Factors). We also included BMI and family member's functional status (limited, not limited). The *health utilization* category consisted of variables indicating whether family member received care more than 10 times from doctor, whether visited dentist, whether saw eye doctor, number of times in ER/ED, whether received home care from health professionals, total number of office visits, whether had surgical procedure, time since last seen a health professional (6–12 months or 0–6 months ago), and whether had a flu shot. All variables were assessed during the year prior to time of survey. Table S1 in Supplemental Materials describes the summary statistics of these variables. Among the 43 potential covariates, five were continuous, age, hours of sleep, number of cigarettes per day, years since first smoking, body mass index.

2.2. Bayesian additive regression trees

We used a generative probabilistic model, Bayesian Additive Regression Trees (BART), that has grown to be influential over the past years in the field of statistical machine learning (Chipman et al., 2010). BART is a nonparametric Bayesian modeling technique using decision trees. Tree-based regression models can address complex dependence structures among and distributional shapes of covariates while keeping the underlying assumptions as weak as possible. BART is a "sum-of-trees" model with tree parameters treated in a formal statistical model rather than just algorithmically. A prior is placed on the parameters and the posterior distributions of the parameters are computed using Markov chain Monte Carlo (MCMC). Regularization priors are used to hold back the fit of each tree allowing for only a small contribution to the overall fit, consequently preventing overfitting (Chipman et al., 2010; Hill, 2011). BART has been shown to have better predictive performance compared to other supervised learning methods, including random forests, boosted models and neural nets, in a variety of study settings (Chipman et al., 2010; Hill, 2011; Lu et al., 2018; Hu et al., 2020a, 2020b, 2020c). For a binary outcome, BART uses probit regression and models the functional form $f(\cdot)$ using an auxiliary latent variable z_i, and the outcome can be viewed as an indicator for whether the latent variable is positive, i.e., $y_i = I(z_i > 0)$. The latent variable has a truncated normal distribution centered at a user-supplied value μ_0, and takes a positive value when $y_i = 1$ and a negative value when $y_i = 0$ from $N(\mu_0 + f(x_i), 1)$. Details of BART have been described elsewhere (Chipman et al., 2010).

We randomly split our data into a training set and a test set, with a ratio of 7:3. We tuned operational parameters of BART on the training data using cross-validation and selected the optimal cross-validated tuning parameters for the BART model (Chipman et al., 2010). We applied the tuned model on both the training and test sets to report model performance, and on the full data for variable selection. We used a traditional and widely used metric for binary outcomes, area under the curve (AUC), to assess the overall classification accuracy of a model (Steyerberg et al., 2010).

2.3. Variable selection using BART-Machine

We implemented a variable selection procedure, BART-Machine, developed by Bleich et al., to select a parsimonious set of most influential determinants for the provider-patient discussions about smoking (Bleich et al., 2014). This method performs favorably compared to variable selection using random forests' "importance scores" (Bleich et al., 2014). BART-Machine uses the *variable inclusion proportion (VIP)*, i.e., the proportion of times each variable is selected as a splitting rule divided by the total number of splitting rules in building the model, as the measure of variable importance. We briefly describe the variable selection procedure: a) Compute the VIP for each covariate from the BART model fitted to the observed data, b) Permute the response variable and rebuild the model and compute the VIPs for all covariates, which we refer to as the "null" VIPs. Repeat this process 100 times to create a null permutation distribution of the VIPs, c) Include a covariate if its VIP from the observed data exceeds the $1-\alpha$ quantile of the distribution of the null VIPs. Following convention in the BART-Machine literature, we chose $\alpha = 0.05$ (Hu et al., 2020c; Bleich et al., 2014).

For a factor variable, variable selection is performed individually on dummy variables for the factor, and the dummy variables' inclusion proportions are aggregated to measure the VIP of the factor variable. We provided a flowchart of the BART-Machine algorithm in Fig. 2. Detailed descriptions have been provided elsewhere (Hu et al., 2020c; Bleich et al., 2014). The variable selection procedure was implemented using the R package bartMachine (Kapelner et al., 2016).

Following the selection of major determinants, we then examined interaction effects with a BART model. Variables were considered to interact in a given tree only if they appeared together in a contiguous downward path from the top to the bottom of the tree. We computed the total number of interactions for each pair of variables by summing across trees and MCMC iterations, from which relative importance of each interaction was evaluated.

2.4. Variable selection using penalized logistic regression

To compare with BART-Machine, we also implemented penalized logistic regression with least absolute shrinkage and selection operator (LASSO) to select most important factors (Santosa and Symes, 1986; Tibshirani, 1996). We considered two variants of this approach, one without accounting for survey weights (LR-LASSO) and one taking survey weights into the penalized likelihood (McConville et al., 2017) (Survey-LR-LASSO). LASSO is a regularization technique, which shrinks the coefficients of certain variables towards zero via a penalty term added to the logistic regression model. LASSO-related regularization methods have been descried in many studies (Candes and Tao, 2007; Efron et al., 2004; Friedman et al., 2008; Hastie et al., 2004; Park and Hastie, 2007; Yuan and Lin, 2006; Zou and Hastie, 2005). We implemented the method using the R software package *glmnet* (Friedman et al., 2010). We chose the shrinkage parameter, which controls the strength of shrinkage and variable selection, based on the model deviance. It is a goodness-of-fit statistic that measures differences in deviances between the full model and a nested model (Candes and Tao, 2007; Efron et al., 2004; Friedman et al., 2008; Hastie et al., 2004; Park and Hastie, 2007; Yuan and Lin, 2006; Zou and Hastie, 2005). A smaller value of deviance indicates better goodness-of-fit. A subset of covariates were selected from the final model with the optimal value of shrinkage parameter determined.

We apportioned the data into training and test sets, with an 70–30 split. We tuned the shrinkage parameter on the training data using cross-validation and selected the optimal shrinkage for LR-LASSO and Survey-LR-LASSO. We reported both the training and test set performances and applied the tuned model on the full data for variable selection.

2.5. Assessing performance of BART-Machine for variable selection

To assess the ability of BART-Machine to identify most important determinants, we compared two BART models – one with the full set of 43 covariates, and the other with seven covariates selected by BART-Machine, versus logistic regression based approaches, LR-LASSO and

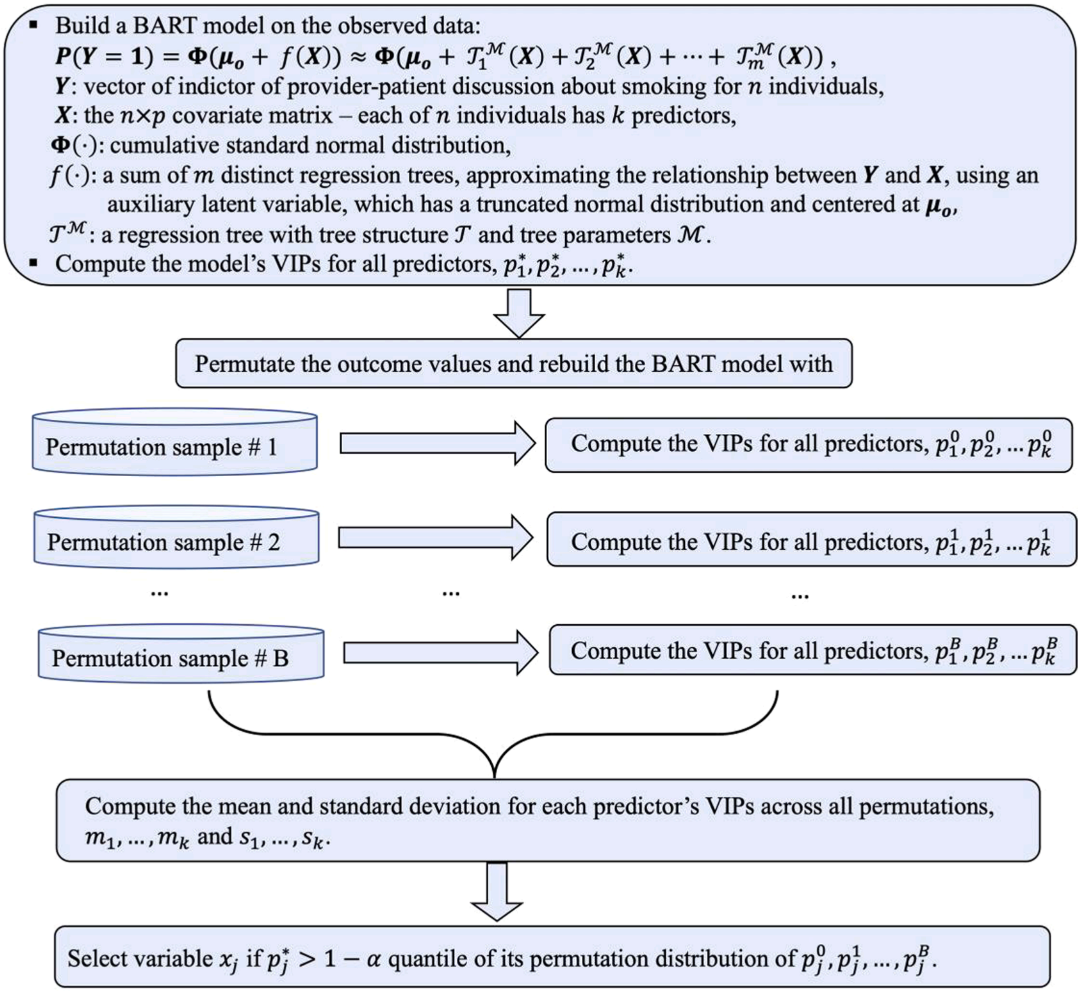
▪ Build a BART model on the observed data:
$P(Y = 1) = \Phi(\mu_o + f(X)) \approx \Phi(\mu_o + \mathcal{T}_1^{\mathcal{M}}(X) + \mathcal{T}_2^{\mathcal{M}}(X) + \cdots + \mathcal{T}_m^{\mathcal{M}}(X))$,
Y: vector of indictor of provider-patient discussion about smoking for n individuals,
X: the $n \times p$ covariate matrix – each of n individuals has k predictors,
$\Phi(\cdot)$: cumulative standard normal distribution,
$f(\cdot)$: a sum of m distinct regression trees, approximating the relationship between Y and X, using an auxiliary latent variable, which has a truncated normal distribution and centered at μ_o,
$\mathcal{T}^{\mathcal{M}}$: a regression tree with tree structure $\mathcal{T}$ and tree parameters $\mathcal{M}$.
▪ Compute the model's VIPs for all predictors, $p_1^*, p_2^*, \ldots, p_k^*$.
Permutate the outcome values and rebuild the BART model with
Permutation sample # 1
Compute the VIPs for all predictors, $p_1^0, p_2^0, \ldots p_k^0$
Permutation sample # 2
Compute the VIPs for all predictors, $p_1^1, p_2^1, \ldots p_k^1$
...
...
...
Permutation sample # B
Compute the VIPs for all predictors, $p_1^B, p_2^B, \ldots p_k^B$
Compute the mean and standard deviation for each predictor's VIPs across all permutations, $m_1, \ldots, m_k$ and $s_1, \ldots, s_k$.
Select variable x_j if $p_j^* > 1 - \alpha$ quantile of its permutation distribution of $p_j^0, p_j^1, \ldots, p_j^B$.

Fig. 2. Variable selection algorithm using BART-Machine.

Survey-LR-LASSO. We computed AUC increase per variable to answer the question of how much gain do we get for adding each variable suggested by a variable selection approach. The AUC increase per variable is defined as (AUC$_{\text{model}}$ - AUC$_{\text{null}}$)/Number of Variables$_{\text{model}}$, where AUC$_{\text{null}}$ is the AUC from the null model, i.e., intercept only model, which is equivalent to 0.5. Methods that give larger AUC increase per variable without sacrificing the overall predictive accuracy are preferred (Bleich et al., 2014).

2.6. *Evaluating the covariate-outcome relationships via survey logistic regression*

Finally, as all machine learning methods, BART-Machine has limited interpretability due to its "black-box" nature, to strengthen study findings, we further fitted a survey logistic regression model to quantify the effects of each key determinant and the top ranked interaction on the likelihood of provider-patient discussions about smoking. Survey logistic regression is an appropriate approach for binary outcomes from survey data with survey elements incorporated, including survey strata, primary sampling unit and weights. Since the sample studied is a subpopulation of the entire survey, we properly calculated the standard errors by assigning sample weight zero to those individuals outside the subpopulation and selected individuals in the subpopulation retained their original weights (Graubard and Korn, 1996).

3. Results

Table S1 in Supplemental Materials summarizes the baseline characteristics, categorized in four domains, for 3666 individuals considered in our analysis, 68.3% (2505) of whom reported to have had provider-patient discussions about smoking during the year prior to time of survey. Compared to those with no discussions, the respondents who had such discussions were less likely to be Hispanic and more likely to be older and unemployed, have insurance and use healthcare resource more. They tended to smoke more, have longer smoking history and have chronic conditions, including diseases including CVD, current asthma, arthritis, diabetes, chronic bronchitis, high cholesterol, hypertension and emphysema.

As shown in Fig. 3, the BART-Machine algorithm identified, for the likelihood of provider-patient discussions about smoking, six most important determinants: total number of doctor's visits in the past 12 months, number of smoked cigarettes per day, years since first smoking, ever had COPD, ever had emphysema and marital status. The relative importance of these six variables based on the observed data exceeded their respective threshold values (the tips of the vertical lines), determined from the "null" distributions for VIPs estimated from the permutated data. Ever had emphysema and marital status were selected corresponding to a relaxed selection criterion, $\alpha = 0.1$, while the other four variables were identified using a more stringent threshold, $\alpha = 0.05$. Each variable's inclusion proportion estimated from the BART model built on the observed data compares the observed relative importance of the covariates. Healthcare utilization usage, intensity of

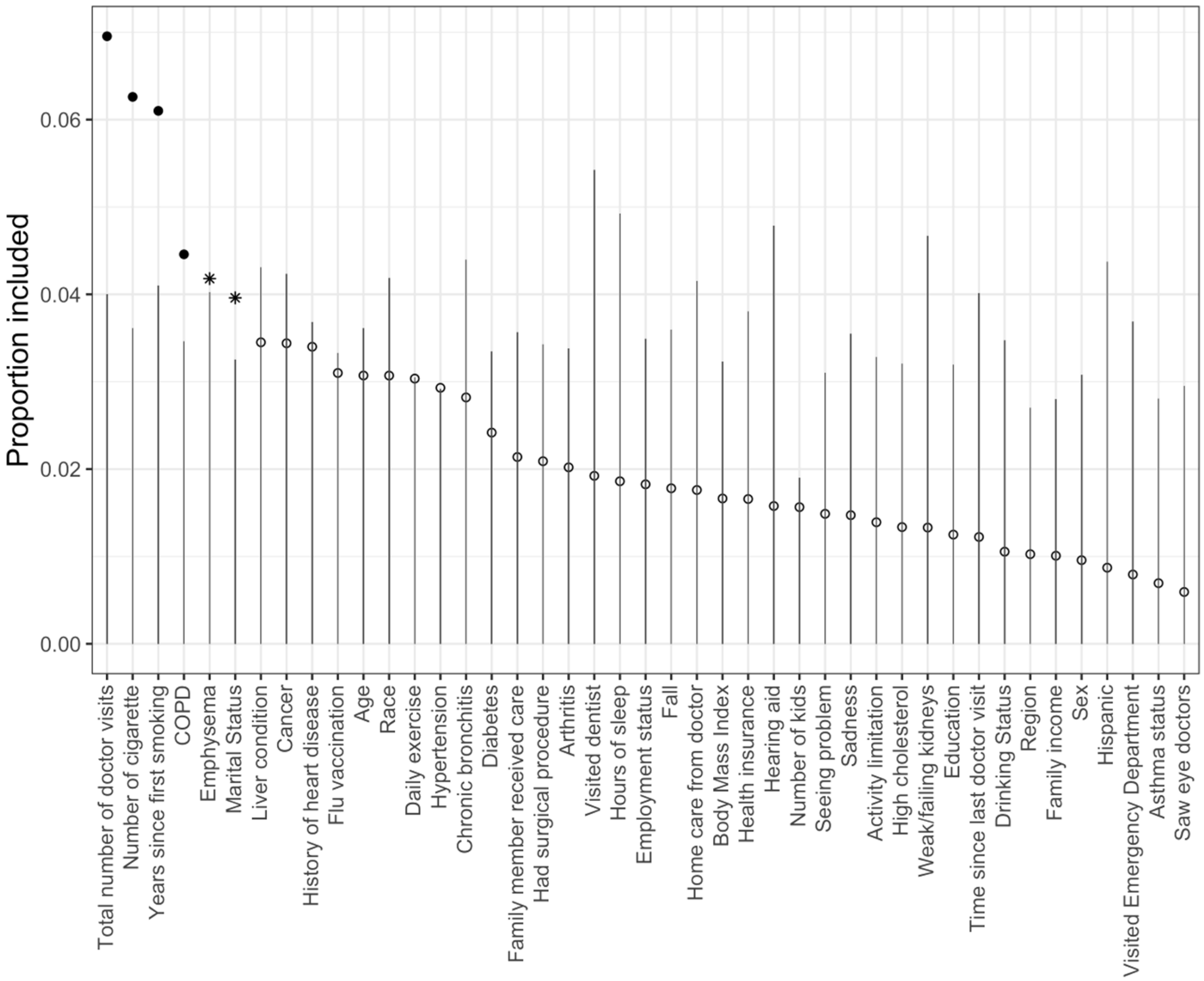

Fig. 3. Visualization of BART-Machine variable selection. The lines are threshold levels for the variable selection algorithm described in Fig. 1 corresponding to $\alpha = 0.1$. Variables passing this threshold are displayed as solid dots and asterisks. Solid dots represent variables selected with more stringent rule $\alpha = 0.05$ and asterisks correspond to those with less stringent rule $\alpha = 0.1$. Open dots correspond to variables that are not selected.

smoking and smoking history appeared to be the most important variables as they had the largest VIPs; whereas gender and ethnicity were among the least important variables. Fig. S1 in Supplemental Materials shows top five interaction terms computed from BART-Machine averaged across 25 model constructions. The relative importance is most distinct for the interaction between years since first smoking and number of cigarettes smoked per day.

Table 1 displays operational parameters for BART, LR-LASSO and Survey-LR-LASSO, and the AUC on both training and test sets. The results show that both the BART full model (with 43 variables) and reduced BART model with selected 7 variables had substantially higher AUC than their logistic regression counterparts, on both training and test sets. Survey-LR-LASSO had similar performance compared to LR-LASSO. Table S2 in Supplemental Materials summarizes the selected variables and AUC gain per variable for each method. The advantage of BART-Machine was supported by the higher AUC with fewer selected variables as well as larger AUC gain per variable. In addition, BART-Machine was able to detect the important interaction between years since first smoking and number of cigarettes smoked per day, which the two logistic regression based approaches ignored. The other six variables selected by BART-Machine were also selected by the other two approaches.

The comparison of the BART model and two logistic regression based models, each with two versions – one with full set of covariates and one with reduced set of covariates selected by their respective variable

Table 1
Operational parameters for BART and logistic regression with LASSO, and out-of-sample AUC on the training set and AUC on the test set from the optimal model for each method. Confidence intervals for the AUC were computed from 100 replications.

	Tuning parameters	Values considered	Training set out-of-sample AUC	Test set AUC
BART (Full model)	# trees$m$ μprior:k	50, 200 1,2,3,5	0.753 (0.745, 0.761)	0.761 (0.743, 0.779)
BART (7 variables)	# trees$m$ μprior:k	50, 200 1,2,3,5	0.730 (0.722, 0.738)	0.741 (0.724, 0.758)
LR-LASSO	Shrinkage λ in range 0–1	e^{-6}, $e^{-5.9}$, …,0	0.695 (0.682, 0.708)	0.709 (0.682, 0.736)
LR	Shrinkage λin range 0–1	0	0.716 (0.701, 0.731)	0.722 (0.701, 0.743)
Survey-LR-LASSO	Shrinkage λ in range 0–1)	e^{-6}, $e^{-5.9}$, …,0	0.685 (0.673, 0.698)	0.698 (0.672, 0.725)
Survey-LR	Shrinkage λ in range 0–1	0	0.707 (0.692, 0.721)	0.713 (0.691, 0.734)

selection algorithms – appears in Fig. S2 in Supplemental Materials. The results show that the two BART models gave highest similar AUC and BART-Machine selected the most parsimonious set of variables, which led to a significantly larger AUC increase per variable, substantiating that it selected the most influential determinants.

Fig. 4 displays the point estimates and 95% confidence intervals for each main and interaction effect. Longer smoking history and higher intensity of cigarette use were associated with higher likelihood of having provider-patient discussions about smoking, with the increase of odds by 24% (95% CI,13%–36%) for every 10 more years since first smoking, and by 40% (95% CI: 21%–60%) for every 5 more cigarettes daily. More use of healthcare resources was linked to higher chance of discussions, with an odds ratio (OR) of 3.08 (95% CI: 2.40–3.94) and 2.19 (95% CI, 1.77–2.70) for more than 5 office visits during the past 12 months and 2–5 visits, respectively, compared to no office visit. The odds of being asked by doctors about smoking for individuals with COPD were 2.36 (95% CI: 1.64–3.40) times that of those without. Married respondents and those with emphysema were also found to be associated with higher likelihood of the discussion, although the 95% CIs contained one, which is consistent with the relaxed threshold ($\alpha = 0.1$) used in our BART-Machine algorithm. Positive sign of the interaction term between years since first smoking and the number of cigarettes per day indicate that these two variables together accelerated the likelihood of a provider-patient discussions about smoking.

In addition, the regression model using the restricted cubic splines for continuous variables (number of cigarettes smoked per day and number of years since first smoking) yielded similar results (see Fig. S3 in Supplemental Materials).

4. Discussions

This study used a state-of-the-art Bayesian machine learning approach, BART-Machine to identify and investigate key determinants of the propensity of a physician discussing about smoking with a patient, leveraging a large-scale NHIS dataset with information on health utilization, demographic and socioeconomic status, health status and behaviors collected from more than 3600 individuals in the US. This approach has three advantages. First, the variable selection is formulated within the Bayesian paradigm, thereby avoiding the multiple testing issue (Gelman et al., 2012). Second, it is a nonparametric permutation-based approach and is not hypothesis driven, consequently not susceptible to issues related to type I errors. Third, in addition to the rankings of the variable importance, we also evaluated the AUC of the full BART model versus the reduced model with only selected covariates included. The two AUCs were similar (0.75 and 0.73), indicating that the selected variables were strongly associated with the outcome. A comparison of the BART based model and penalized logistic regression based methods demonstrates that BART-Machine selected the most parsimonious set of important covariates while maintaining a higher model predictive accuracy. A disadvantage of this approach is the computational cost associated with running BART models on multiple (e.g., 100) permutation sets. However, parallel computing on multiple cores can be used to speed up computation.

The prevalence rate of the provider-patient discussions about smoking is 68%. To investigate disparities in the discussion rate, we identified major determinants of the likelihood of the discussions. Smokers who had higher intensity of cigarette use or longer smoking history, who were married, who had COPD or emphysema and who visited doctors more tended to have a higher chance of receive patient-provider discussions about smoking. Longer smoking history and higher

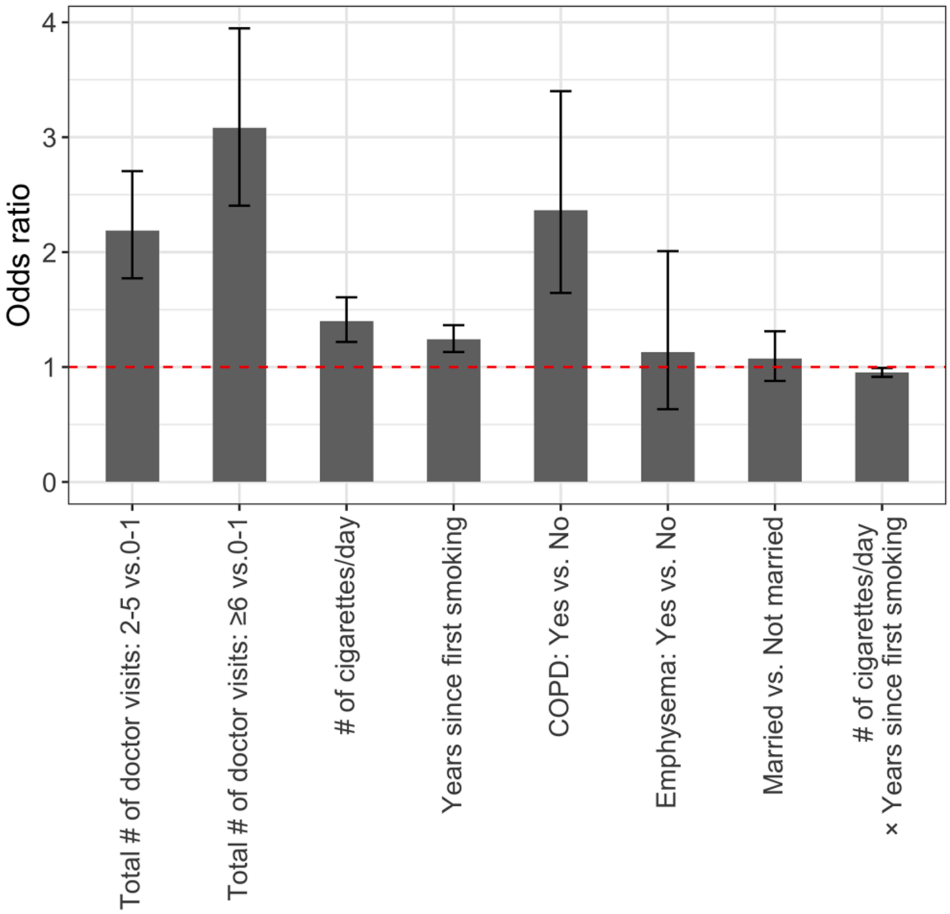

Fig. 4. Effect estimates and 95% confidence intervals (CI) for six key determinants and one most important interaction. For continuous variables, effect estimates represent changes in odds ratio per 10 years increase in years since first smoking and per five more cigarettes smoked per day. For factor variables, effect estimates compare odds for different levels to the reference level. The dashed red line corresponds to an odds ratio of one. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

smoking frequency jointly accelerated the likelihood of a provider-patient discussions about smoking. The BART-Machine algorithm was able to quantify the relative importance of each variable by keeping track of variable inclusion frequencies. The total number of doctor visits, number of cigarettes per day and years since first smoking were ranked as the top three factors, followed by COPD, emphysema and marital status.

These results agree with findings in the literature. Huo et al. found that the number of office visits to doctors and other healthcare professionals and whether a respiratory condition was diagnosed are two strong determinants (Huo et al., 2020). However, the authors did not identify emphysema as the key co-morbidity factors. Bao et al. found that healthcare providers are more likely to discuss smoking with heavier smokers and with those who had smoking-related conditions (Bao et al., 2006). Our study suggests that smoking-related complications and behaviors play a much larger role in a consultation about smoking between doctors and patients, and that age, race and sex, which were frequently deemed as important factors, had lower level of importance in explaining the variability in provider-patient discussions about smoking. Particularly, sex was among the lowest-ranking variables while age and race were much higher in the rankings. A number of prior studies also demonstrated that sex is not significantly associated with the likelihood of provider-patient discussions about smoking (Henley et al., 2019; Huo et al., 2020; Zhang et al., 2019) Multiple factors may contribute to the discrepancy between findings from our study and prior literature. First, the data source may be different. Second, the set of candidate covariates considered may be different. Third, the analysis approaches used are different. We used a nonparametric permutation-based approach, which kept the underlying assumptions as weak as possible (Wasserman, 2006; Mazumdar et al., 2020). Prior work frequently used main effects logistic regression, which makes stronger assumptions about dependence structures among and distributional shapes of covariates.

Key determinants of the likelihood of provider-patient discussions about smoking, identified by the BART-Machine algorithm, may provide some insights into strategies for promoting discussions on smoking (Institute of Medicine, Committee on the Learning Health Care System in America). For example, while the number of doctor visits may correlate with one's health condition, it also reflects access to healthcare. On the one hand, health care providers may offer brief discussions or counselling about smoking in a healthcare encounter to those who have access to healthcare (Bao et al., 2006; Shelley et al., 2005; Tong et al., 2006). On the other hand, increasing healthcare access and reducing barriers to healthcare for disadvantaged population may help facilitate provider-patient discussions about smoking. Our results suggest that doctors tended to consult more with heavy smokers about smoking cessation. Encouraging such discussions between providers and light smokers who may be at risk for heavy smoking (e.g., asking about the smoking habits of parents and other caregivers) can help reduce cigarette smoking (Role of the physician in smoking prevention, 2001).

Discovering the subset of covariates that are most influential on the outcome is challenging, especially when the number of relevant variables is sparse relative to the total number of available variables, which requires rigorous techniques to tease out noises and identify real effects. Existing studies that assess the determinants of provider-patient discussions on smoking often rely on a set of variables selected *a priori*, or less robust variable selection procedures such as univariate analysis and stepwise selection, which depend on statistical tests that may lead to biased and erroneous results (Hurvich and Tsai, 1990).

We considered a wide range of potential factors for the likelihood of provider-patient discussions on smoking, and exploited a principled permutation-based variable selection approach using Bayesian machine learning to identify most important factors. The feature of "unblack-boxing" interactions supplied by BART-Machine provided us with an opportunity to gain insights into the synergistic effects of the major determinants, which are often ignored in public health research. Judging on the basis of AUC increase from 0.5 of a null model, the seven variables selected by BART-Machine accounted for 91% of the increase there can be with all 43 variables, corroborating that the major determinants were identified. Finally, a survey logistic regression was conducted to quantify the associations between the major determinants and interactions thereof and the probability of provide-patient discussions about smoking.

This study has several limitations, which provide potential avenues for future research. First, our study findings are correlational and not causal in nature due to the cross-sectional data and survey design. However, our results identified key determinants of a provider-patient discussions about smoking and can potentially stimulate future research about causality using longitudinal data with causal inference techniques (Hu and Hogan, 2019; Hu et al., 2018) Second, survey data is self-reported, susceptible to self-reporting bias. Future studies need to overcome the biases through use of adjustment methods such as conducting internal or external validation study and using Martin-Larsen Approval Motivation score (Althubaiti, 2016). Developing sensitivity analysis strategies for evaluating the impact of different self-reporting mechanisms could also be a worthwhile future contribution (Hogan et al., 2014) (ref). In addition, findings from our study are generalizable insofar as we used a nationally reprehensive sample drawn from the NHIS data. However, the BART-Machine variable selection algorithm was not able to account for survey elements. This limitation should motivate the development in survey methodology. Third, the number of doctor visits may mediate the effect of healthcare access on discussions about smoking. Examining the association of healthcare access, number of doctor visits and provider-patient discussions about smoking would be an important contribution to smoking cessation. Finally, there could be other important variables that were included in our analysis. It is possible that provider variables have as great or greater an effect on the likelihood of a discussion about cessation. However, the NHIS only included measured characteristics on patients. Aspects of the patient-provider relationship, such as same/different race, may also have an effect. Despite the potential unmeasured variables, by considering wide-ranging information across multiple domains and using an innovative and principled machine learning approach, we believe the scope and depth of our analysis can complement current research and encourage more innovative investigations in the area of smoking cessation.

5. Conclusions

Principled nonparametric permutation-based variable selection approaches, like BART-Machine, can provide insights into the key factors of discussions about smoking between health providers and patients in a data-driven and reproducible way. Health care resource usage, smoking intensity and duration and smoking-related conditions were key determinants, whereas the "usual suspects", age, gender, race and ethnicity were less important, and gender, in particular, had little effect.

Funding source

Liangyuan Hu was supported by a methodology award ME 2017C3 9041 from the Patient-Centered Outcomes Research Institute and a grant R21CA245855 from the National Cancer Institute.

CRediT authorship contribution statement

Liangyuan Hu: Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Lihua Li:** Data curation, Writing - review & editing. **Jiayi Ji:** Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2020.101238.

References

The Health Consequences of Smoking - 50 Years of Progress: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion,Office on Smoking and Health, 2014. Accessed April 20, 2017.

How Tobacco Smoke Causes Disease: What It Means to You. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2010. Accessed April 20, 2017.

Women and Smoking: A Report of the Surgeon General. Rockville (MD): U.S. Department of Health and Human Services, Public Health Service, Office of the Surgeon General, 2001 Accessed April 20, 2017.

The Health Consequences of Smoking: A Report of the Surgeon General. Office of the Surgeon General (US), Office on Smoking and Health (US), Atlanta (GA): Centers for Disease Control and Prevention (US), 2004.

Mahmud, A., Feely, J., 2003. Effect of smoking on arterial stiffness and pulse pressure amplification. Hypertension 41 (1), 183–187.

Samet, J., 2013. Tobacco smoking. The leading cause of preventable disease worldwide. Thoracic Surgery Clin. 23, 103–112.

Smoking Cessation: A Report of the Surgeon General. Centers for Disease Control and Prevention, 2020. https://www.cdc.gov/tobacco/data_statistics/sgr/2020-smoking-cessation/index.html#full-report.

Critchley, J.A., Capewell, S., 2003. Smoking cessation for the secondary prevention of coronary heart disease. Cochrane Database Syst. Rev. 4.

Toll, B.A., Rojewski, A.M., Duncan, L.R., et al., 2014. "Quitting smoking will benefit your health": the evolution of clinician messaging to encourage tobacco cessation. Clin Cancer Res. 20 (2), 301–309.

Treating Tobacco Use and Dependence: 2008 Update. Agency for Healthcare Research and Quality. https://www.ahrq.gov/prevention/guidelines/tobacco/index.html.

Helping Smokers Quit: A Guide for Clinicians. U.S. Department of Health and Human Services, 2008. https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/clinicians-providers/guidelines-recommendations/tobacco/clinicians/references/clinhlpsmkqt/clinhlpsmksqt.pdf.

Kruger, J., Shaw, L., Kahende, J., Frank, E., 2012. Health care providers' advice to quit smoking, National Health Interview Survey, 2000, 2005, and 2010. Prev Chronic Dis. 9, E130.

Solberg, L.I., Boyle, R.G., Davidson, G., Magnan, S.J., Carlson, C.L., 2001. Patient satisfaction and discussion of smoking cessation during clinical visits. Mayo Clin. Proc. 76 (2), 138–143.

Bao, Y., Duan, N., Fox, S.A., 2006. Is some provider advice on smoking cessation better than no advice? An instrumental variable analysis of the 2001 National Health Interview Survey. Health Serv. Res. 41 (6), 2114–2135.

Huo, J., Chung, T.H., Kim, B., Deshmukh, A.A., Salloum, R.G., Bian, J., 2020. Provider-Patient discussions about smoking and the impact of lung cancer screening guidelines: NHIS 2011–2015. J Gen Intern Med. 35 (1), 43–50.

Cokkinides, V.E., Halpern, M.T., Barbeau, E.M., Ward, E., Thun, M.J., 2008. Racial and ethnic disparities in smoking-cessation interventions: analysis of the 2005 National Health Interview Survey. Am. J. Prev. Med. 34 (5), 404–412.

Lopez-Quintero, C., Crum, R.M., Neumark, Y.D., 2006. Racial/ethnic disparities in report of physician-provided smoking cessation advice: analysis of the 2000 National Health Interview Survey. Am. J. Public Health 96 (12), 2235–2239.

Centers for Disease Control and Prevention. Quitting smoking among adults - United States, 2001-2010, 2011. Morbidity Mortality Weekly Report (MMWR) Rep. 60(44), 1513–1519.

King, B.A., Dube, S.R., Babb, S.D., McAfee, T.A., 2013. Patient-reported recall of smoking cessation interventions from a health professional. Prev. Med. 57 (5), 715–717.

Centers for Disease Control and Prevention. 2016 National Health Interview Survey, Survey Description. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2016/srvydesc.pdf. Accessed April 27, 2020.

Predefined SEER Stat Variables for Calculating the Number of Associated Cancers for Selected Risk Factors. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/uscs/public-use/predefined-seer-stat-variables.htm.

Chipman, H.A., George, E.I., McCulloch, R.E., 2010. BART: Bayesian additive regression trees. Ann. Appl. Stat. 4 (1), 266–298.

Hill, J.L., 2011. Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat. 20 (1), 217–240.

Lu, M., Sadiq, S., Feaster, D.J., Ishwaran, H., 2018. Estimating individual treatment effect in observational data using random forest methods. J. Comput. Graph. Stat. 27 (1), 209–219.

Hu, L., Gu, C., Lopez, M., Ji, J., Wisnivesky, J., 2020. Estimation of causal effects of multiple treatments in observational studies with a binary outcome. Stat. Methods Med. Res. 29 (11), 3218–3234.

Hu, L., Liu, B., Li, Y., 2020. Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: A Bayesian machine learning approach. Prev. Med. 141, 106240.

Hu, L., Ji, J., Liu, B., Li, Y., 2020. Tree-based machine learning to identify and understand major determinants for stroke at the neighborhood level. J. Am. Heart Assoc., e016745 https://doi.org/10.1161/JAHA.120.016745.

Steyerberg, E.W., Vickers, A.J., Cook, N.R., et al., 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass). 21 (1), 128–138.

Bleich, J., Kapelner, A., George, E.I., Jensen, S.T., 2014. Variable selection for BART: An application to gene regulation. Ann. Appl. Stat. 8 (3), 1750–1781.

Kapelner, A., Bleich, J., bartMachine,, 2016. Machine learning with Bayesian additive regression trees. J. Stat. Softw. 1 (4), 2016.

Santosa, F., Symes, W.W., 1986. Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput. 7 (4), 1307–1330.

Tibshirani, R., 1996. Regression Shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288.

McConville, K.S., Breidt, F.J., Lee, T.C.M., Moisen, G.G., 2017. Model-assisted survey regression estimation with the lasso. J. Survey Stat. Methodol. 5 (2), 131–158.

Candes, E., Tao, T., 2007. The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. 35 (6), 2313–2351.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32 (2), 407–451.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

Hastie, T., Rosset, S., Tibshirani, R., Zhu, J., 2004. The entire regularization path for the support vector machine. J. Mach. Learn. Res. 5, 1391–1415.

Park, M.Y., Hastie, T., 2007. L1-regularization path algorithm for generalized linear models. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 69 (4), 659–677.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. B 68, 49–67.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol). 67 (2), 301–320.

Friedman, J.H., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 1 (1), 2010.

Graubard, B.I., Korn, E.L., 1996. Survey inference for subpopulations. Am. J. Epidemiol. 144 (1), 102–106.

Gelman, A., Hill, J., Yajima, M., 2012. Why we (usually) don't have to worry about multiple comparisons. J. Res. Educ. Effectiveness 5 (2), 189–211.

Henley, S.J., Asman, K., Momin, B., et al., 2019. Smoking cessation behaviors among older U.S. adults. Prev. Med. Rep. 16, 100978.

Zhang, L., Babb, S., Schauer, G., Asman, K., Xu, X., Malarcher, A., 2019. Cessation Behaviors and Treatment Use Among U.S. Smokers by Insurance Status, 2000&#x2013;2015. Am. J. Prev. Med. 57(4), 478–486.

Wasserman, L., 2006. All of Nonparametric Statistics. Springer, New York.

Mazumdar, M., Lin, J.-Y.J., Zhang, W., et al., 2020. Comparison of statistical and machine learning models for healthcare cost data: a simulation study motivated by Oncology Care Model (OCM) data. BMC Health Serv. Res. 20(1), 350.

Institute of Medicine, Committee on the Learning Health Care System in America, 2013. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington, DC: National Academies Press.

Shelley, D., Cantrell, J., Faulkner, D., Haviland, L., Healton, C., Messeri, P., 2005. Physician and dentist tobacco use counseling and adolescent smoking behavior: results from the 2000 national youth tobacco survey. Pediatrics 115 (3), 719.

Tong, E.K., Ong, M.K., Vittinghoff, E., Pérez-Stable, E.J., 2006. Nondaily smokers should be asked and advised to quit. Am. J. Prev. Med. 30 (1), 23–30.

Role of the physician in smoking prevention, 2001. Paediatr. Child Health 6 (2), 89–109.

Hurvich, C.M., Tsai, C.-L., 1990. The impact of model selection on inference in linear regression. Am. Stat. 44 (3), 214–217.

Hu, L., Hogan, J.W., 2019. Causal comparative effectiveness analysis of dynamic continuous-time treatment initiation rules with sparsely measured outcomes and death. Biometrics 75 (2), 695–707.

Hu, L., Hogan, J.W., Mwangi, A.W., Siika, A., 2018. Modeling the causal effect of treatment initiation time on survival: Application to HIV/TB co-infection. Biometrics 74 (2), 703–713.

Althubaiti, A., 2016. Information bias in health research: definition, pitfalls, and adjustment methods. J. Multidiscip. Healthc. 9, 211–217.

Hogan, J.W., Daniels, M.J., Hu, L., 2014. A Bayesian perspective on assessing sensitivity to assumptions about unobserved data. In: Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A., Verbeke, G. (Eds.), Handbook of Missing Data Methodology. CRC Press, Boca Raton, FL, pp. 405–434.