DIABETES TECHNOLOGY & THERAPEUTICS
Volume 21, Number 5, 2019

DOI: 10.1089/dia.2018.0390

ORIGINAL ARTICLE

Application of Machine Learning to Identify Clustering of Cardiometabolic Risk Factors in U.S. Adults

Xiyue Liao, PhD,[1] David Kerr, DM,[2] Jessikah Morales, BA,[2] and Ian Duncan, FSA[1]

Abstract

Aims: The aim of this study is to compare some machine learning methods with traditional statistical parametric analyses using logistic regression to investigate the relationship of risk factors for diabetes and cardiovascular (cardiometabolic risk) for U.S. adults using a cross-sectional data from participants in a wellness improvement program.

Methods: Logistic regression was used to find the relationship between individual risk factors, predictor and cardiometabolic risk. Supervised machine learning methods were used to predict risk and produce a ranking of variables' importance. A clustering method was used to identify subpopulations of interest. Predictors were divided into those that are nonmodifiable and those that are modifiable.

Results: The population comprised 217,254 adults of whom 8.1% had diabetes. Using logistic regression, six variables were identified to be negatively related and eleven were positively related to cardiometabolic risk. Three supervised machine learning classifiers (random forest, gradient boosting, and bagging) were applied with average AUC to be 0.806. Each classifier also produced a ranking of variables' importance. Four subgroups were identified with a *k*-medoid clustering algorithm, which were mainly distinguished by gender and diabetes status.

Conclusions: The study illustrates that machine learning is an important addition to traditional logistic regression in terms of identifying important cardiometabolic risk factors and ranking their importance and the potential for interventions based on lifestyle and medications at an individual level.

Keywords: Diabetes, Machine learning, Prediction, Variable importance, Supervised learning, Clustering.

Introduction

The increase in the number of individuals developing type 2 diabetes (T2D) continues to add a significant financial burden to health care systems in the United States and elsewhere.[1,2] Established risk factors for T2D include a family history of diabetes, excess weight gain, inactivity, and a history of gestational diabetes. T2D is also more common in certain racial and ethnic groups and is associated with serious long-term complications, including premature death, heart attacks, heart failure, stroke, and renal failure.[3–5] Most notably the risk of cardiovascular complications is driven, in part, by clustering of factors, including insulin resistance, obesity, abnormal plasma lipid profiles, and high blood pressure in addition to hyperglycemia. There is also increasing awareness of the contribution of nontraditional factors related to social factors, including income, education, and culture, as well as environmental factors.[6]

Recently, the use of big data analyses has suggested that there are subtypes of individuals with T2D who may have different trajectories related to long-term risk of diabetes-related complications.[7,8] At an individual level, risk factors for T2D and the associated cardiovascular complications can be stratified into modifiable through lifestyle changes (e.g., avoiding excess weight gain) and nonmodifiable (e.g., family history). Furthermore, and within the hierarchy of modifiable risk factors, it is likely that these can be differentially influential between and within individuals. Therefore, the aim of this study was to assess the contribution of potentially modifiable and nonmodifiable risk factors by applying machine learning to a cross-sectional dataset of U.S. adults participating in a wellness improvement program.

[1]Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, California.
[2]Sansum Diabetes Research Institute, Santa Barbara, California.

Methods

Dataset

The dataset source was a U.S. based provider of a workplace health promotion and wellness programs, The Vitality Group (TVG; www.thevitalitygroup.com). Program participants are employees and dependents of employers that contract with The Vitality Group for incentives for participating in physical activity and other healthy behaviors, which are then exchangeable for rewards. Physical activity levels are self-reported in an annual health risk assessment but are also verified throughout the year either by device or gym utilization. Gym visits are verified through a GPS mobile application: a person has to be at the gym location for at least 30 min (the user interface is through a countdown timer on the application). Visits recorded in this manner give rise to ''standard workouts,'' but if the participant is using a device at the gym an advanced workout may be recorded using the device. Sedentary hours are referred to the time spent awake but inactive. The type of diabetes was not specified in the dataset but among U.S. adults with a diagnosis of diabetes, T2D accounts for more than 90% of cases.[4]

Participants also record a number of self-reported health-related factors (presence of chronic diseases, including heart disease; alcohol consumption and smoking behavior; and so on) in addition to clinical (laboratory) measures that are recorded either at employer-sponsored health fairs or reported by attending physicians. Stress was reported by participants using a validated psychological distress scale (K10).[9]

For this study we focused on a cohort of 217, 254 people who were in this study for 1 year between 2012 and 2015. Figure 1 is a flowchart for the procedure of data preprocessing, model building, and model evaluation.

Preprocessing and imputation of data

For missing values, we used the R package MICE (Multivariate Imputation by Chained Equations) to impute missing

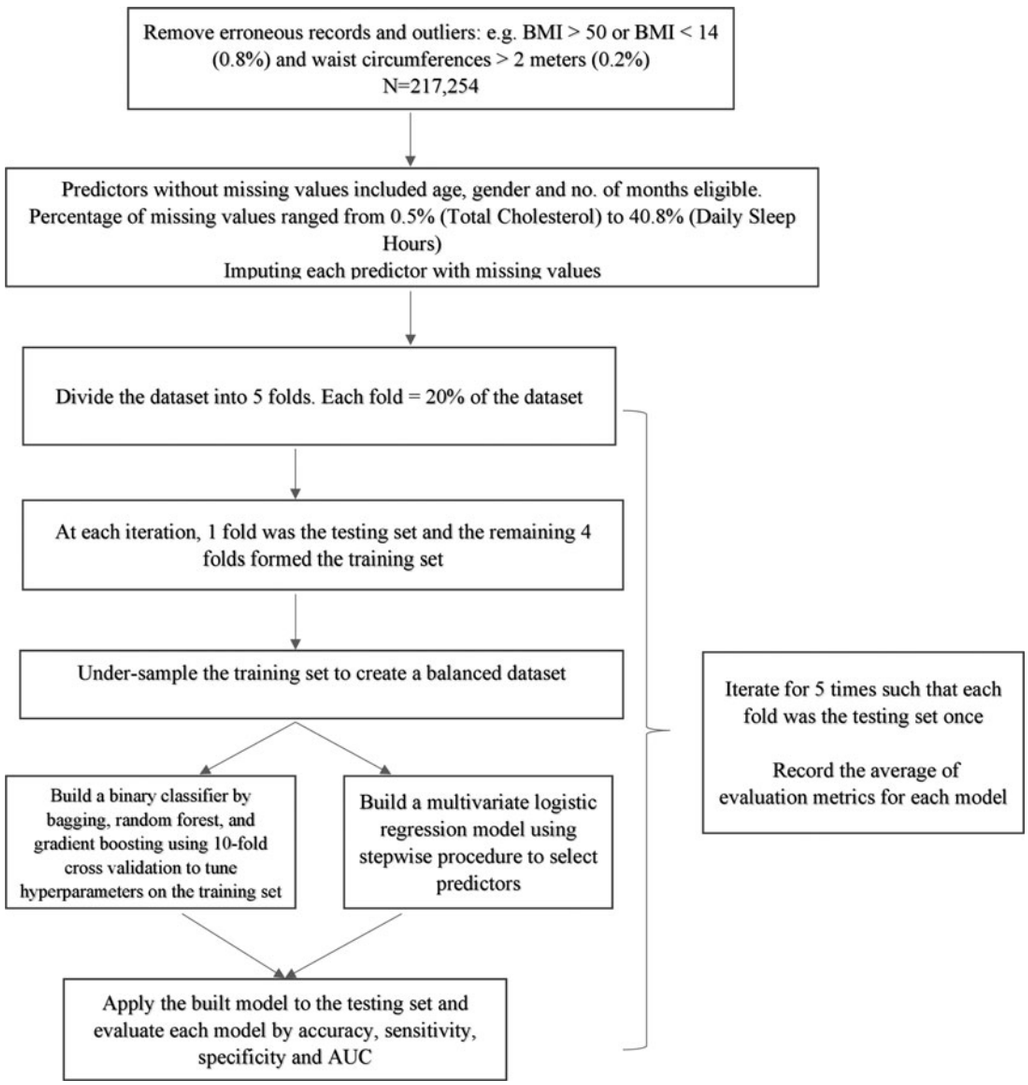

FIG. 1. Flowchart showing data preprocessing, model building, and model evaluation.

values. MICE generates multiple imputations for incomplete multivariate data by Gibbs sampling.[10]

Statistical analysis

Multivariate logistic regression was used to assess the association between explanatory factors and diabetes risk. We used the stepwise algorithm to choose important predictors,[11] which would help avoid overfitting and increase the interpretability of coefficients.

Machine learning methods

To understand the factors influencing cardiometabolic risk, we applied two different machine learning methods: supervised and unsupervised learning.

Data imbalance

The dataset was highly skewed showing 8% of participants with and 92% of participants without a reported diagnosis of diabetes. We used R package ROSE[12] to address the data imbalance problem when building the binary classifier.

Supervised learning methods

Supervised learning builds a function mapping of a set of input attributes such as age, gender, body mass index (BMI), and so on to a labelled output, such as the response variable ''diabetes'' labeled as 0 (without diabetes) and 1 (with diabetes). Since the diabetes variable is binary, the model is a classifier. The analysis in this subsection was performed with the R package caret.[14]

Three methods of machine learning were applied (Bagging, Random Forest, and Gradient Boosting) to derive a predictive model.[15]

To avoid the prediction performance being affected by single random split of the dataset, we used nested cross-validation, in which an outer 5-fold cross validation loop to split the data into training and test folds and an inner 10-fold cross validation loop were done for each classifier to choose the tuning parameters that maximized the area under the ROC curve (AUC) on the training fold. A receiver operating characteristic (ROC) curve is created by plotting sensitivity (true positive rate) against 1-specificity (false positive rate). AUC is the area under the ROC curve, and it represents degree of separability. When a model makes random guesses, AUC will be 0.5. When AUC gets closer to 1.0, the model is better at predicting diabetes cases as diabetes cases and nondiabetes cases as nondiabetes cases. R package pROC[16] was used for AUC computation.

In terms of computation time, for example, on a laptop with a 2.16GHz dual-core Intel(R) Celeron(R) CPU, it took 6 min to build a random forest classifier in one inner loop of cross validation. The speed for gradient boosting and bagging is close.

Unsupervised learning method

Unsupervised learning is used to identify underlying groups in a dataset. One most commonly used unsupervised learning method is clustering. To find subpopulations in this dataset, a *k*-medoid clustering algorithm was applied: the partitioning around medoids (PAM) algorithm, where medoid is the data point chosen as the ''center'' of a cluster.[16–18] Note that for the clustering analysis, the balanced training set obtained by undersampling was still used. The analysis in this subsection was performed with the R package cluster.[16]

To apply the PAM algorithm, a metric must be chosen to define the distance between data points. Since the dataset is a mixture of both continuous and categorical variables, the

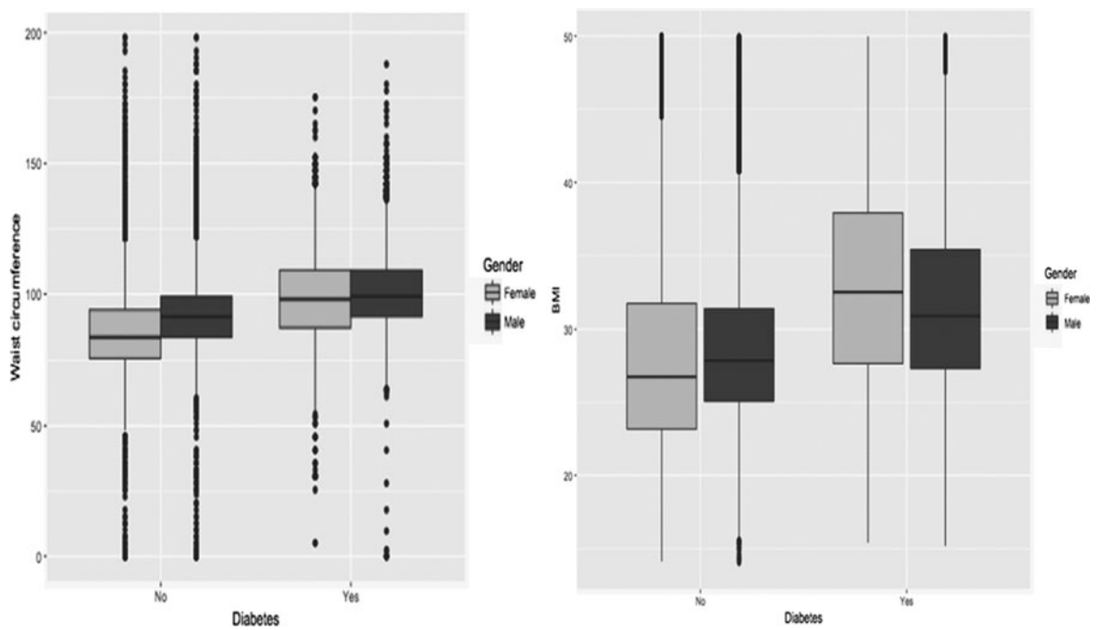

FIG. 2. Boxplot of waist circumference (left panel) and BMI (right panel) by gender and self-reported diagnosis of diabetes. BMI, body mass index.

Table 1. Odds Ratios Using a Logistic Regression Model for Potentially Modifiable Predictors

Predictors	*Units*	*OR (95% CI)*	P
BMI	kg/m^2	1.036 (1.029–1.044)	<0.0001
Waist circumference	cm	1.021 (1.018–1.024)	<0.0001
Triglycerides	mmol/L	1.553 (1.499–1.610)	<0.0001
Systolic BP	mmHg	1.016 (1.014–1.019)	<0.0001
Diastolic BP	mmHg	0.991 (0.987–0.995)	<0.0001
Other health-related activities[a]	count	0.935 (0.926–0.947)	<0.0001
Kessler stress score		1.015 (1.008–1.022)	<0.0001
Alcohol (no. of drinks/week)		0.965 (0.957–0.972)	<0.0001
Daily sedentary time	h	1.016 (1.009–1.023)	<0.0001
Total cholesterol	mmol/L	0.729 (0.706–0.752)	<0.0001
Weekly verified standard workouts	count	0.974 (0.957–0.991)	0.003

[a]Online courses, participation in employer sponsored health events, flu shots, screening, etc. (0–12/week).
BMI, body mass index; BP, blood pressure; CI, confidence interval; OR, odds ratio; SD, standard deviation.

Gower's coefficient was used as the distance metric, which handles mixed data types well.[19]

Results

Participants were aged 18–80 years (mean 43 ± 12 years, 54.7% female). Of the total population of 217,254, 17,554 (8.1%) had diabetes, and the prevalence of diabetes was higher for males (9.4% vs. 7.0%). Participant demographics and health outcomes are shown in the Appendix Tables A1 and A2.

We compared the distribution of waist circumference and BMI for both genders, separately for the group with diabetes and the group without diabetes in Figure 2. In both groups, average waist circumference of female participants was smaller compared with male participants, but the difference was larger in the group without diabetes. The average BMI of female participants was lower compared with male participants in the group without diabetes, but it was higher compared with male participants in the group with diabetes.

Figure 3 shows a comparison of fasting plasma glucose (FPG) levels by age and BMI suggesting that for both genders, FPG level increases as age or BMI increases, and females tend to have a lower FPG than males given the same age and BMI.

Logistic regression

Tables 1 and 2 show estimated odds ratios of important predictors for the logistic regression model. Predictors were divided into those that are nonmodifiable (such as age and sex) and those that are modifiable either by behavior change (e.g., waist circumference and BMI) or medication + behavior change (e.g., cholesterol and blood pressure).

Supervised learning

Table 3 shows the average of four evaluation metrics for all three machine learning classifiers described above together with the traditional logistic regression model. Overall, the three classifiers had similar performances although random forest was slightly better than another two methods. Each metric value by random forest is slightly better than that by

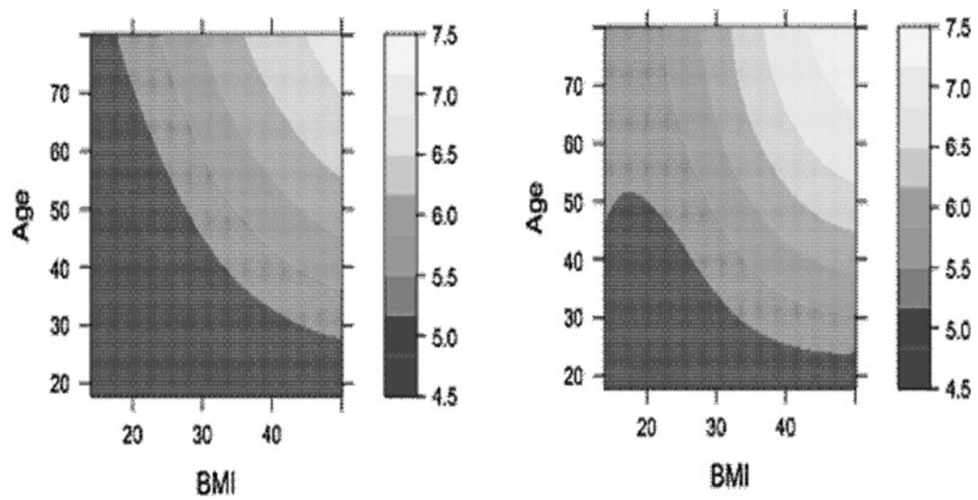

FIG. 3. Contour surface plot of FPG against age and BMI for female ($n = 118,855$, left) and male ($n = 98,399$, right). This figure shows a comparison of FPG levels by age and BMI suggesting that for both genders, FPG level increases as age or BMI increases, and females tend to have a lower FPG than males given the same age and BMI. FPG, fasting plasma glucose.

Table 2. Odds Ratios Using a Logistic Regression Model for Nonmodifiable Predictors

Predictors	*Units*	*OR (95% CI)*	P
Age	18–80 Years	1.052 (1.047–1.053)	<0.0001
Heart disease	Yes/no	1.485 (1.218–1.819)	<0.0001
No. of months eligible[a]	Months	1.012 (1.004–1.020)	0.003
Education[b]	1		
	2	0.930 (0.754–1.145)	0.495
	3	0.885 (0.727–1.076)	0.224
	4	0.750 (0.610–0.922)	0.007
Chronic lung disease	Yes/no	1.722 (1.073–2.847)	0.023
Gender	Male/female	1.073 (1.007–1.144)	0.033

[a]No. of months in the year that the individual participated in the Vitality program (max. 12/year).
[b]1: Did not complete high school, 2: High school completed, 3: College degree, 4: Postgraduate degree

logistic regression. Variable importance ranking by each machine learning method is shown in Table 4.

Interpretation of supervised learning models

Machine learning methods measure a variable's importance by computing the increase of the model's prediction error after permuting the variable. A variable is important if permuting its values increases the prediction error; otherwise, it is considered unimportant.[20,21]

There was no significant disagreement between the three machine learning models in terms of which variables were important. There were, however, differences between variable importance in the logistic regression and machine learning models. Lipids (a modifiable risk factor) were not significant in the logistic regression model, although they were in the machine learning models. Eating fruits and vegetables, sleep, and exercise are important in the machine learning models but not in the logistic regression model. Figure 4 shows the relative ''importance'' of each risk factor by three machine learning models.

Table 3. Model Evaluation Metrics for Four Machine Learning Methods

Model	*Accuracy*	*Sensitivity*	*Specificity*	*AUC*
Random forest	0.727	0.724	0.769	0.818
Bagging	0.724	0.725	0.712	0.801
Gradient boosting	0.726	0.725	0.729	0.799
Logistic regression	0.722	0.722	0.728	0.791

AUC, area under the curve.

Unsupervised learning

The *k*-medoid algorithm is partitioned, breaking the dataset up into groups or clusters allowing for minimization of the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In other words, objects in the cluster have more in common with each other than they do with objects assigned to another cluster. The information for each cluster is summarized by the medoid (center) of the cluster.

Tables 5 and 6 include the variables in each medoid. The number of people assigned to each is approximately equal: the numbers from cluster 1 to cluster 4 are 7680, 6247, 6640, and 7335, respectively. The percentage of self-reported diabetes cases from cluster 1 to cluster 4 is 1.3%, 4.5%, 98.1%, and 100%, respectively. Cluster 1 (female) and 2 (male) are nondiabetes clusters; cluster 3 (female) and 4 (male) are diabetes clusters. The ''worst'' combination of all characteristics is shown by the medoid of cluster 4. Once again these are separated into potentially modifiable and nonmodifiable variables.

Table 4. Comparison of Logistic Regression and Machine Learning to Determine Importance of Individual Potentially-Modifiable Risk Factors for Cardiometabolic Disease

Predictors	*Logistic regression*	*Random forest*	*Gradient boosting*	*Bagging*
Waist circumference	+	+	+	+
Triglycerides	+	+	+	+
BMI	+	+	+	+
HDL cholesterol	—	+	+	+
Systolic BP	+	+	+	+
LDL cholesterol	—	+	+	+
Total cholesterol	+	+	+	+
Diastolic BP	+	+	+	+
Daily sedentary time	+	+	+	+
Other activities	+	+	+	+
Kessler Stress Score	+	+	+	+
Alcohol (no. of drinks/week)	+	+	+	+
Daily servings of fruits and vegetables	—	+	+	+
Daily hours of sleep	—	+	+	+
Weekly verified standard workouts	—	+	+	+
Weekly verified light workouts	—	+	—	+
Tobacco use	—	+	+	+
Weekly self-reported workouts	—	+	+	+

Important predictors: important (+), unimportant (—).
HDL, high-density lipoprotein; LDL, low-density lipoprotein.

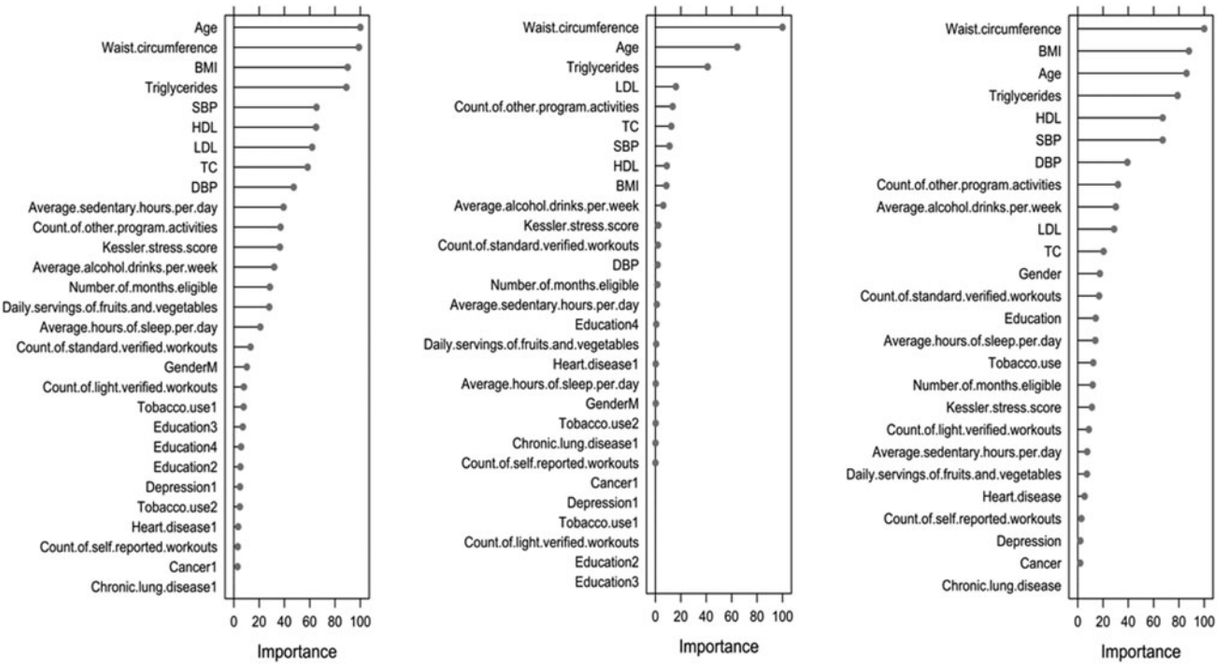

FIG. 4. Risk factors' importance ranking by three machine learning methods: random forest (left), gradient boosting (middle), and bagging (right).

Discussion

The increase in the number of individuals developing diabetes has added a significant financial burden to health care systems.[22] In the United States, diabetes care already accounts for 1 of every 4 dollars spent on health care.[23] Therefore, the most cost-effective approach to reducing the burden of T2D and the associated serious cardiovascular complications is to increase the focus on prevention.

A number of risk factors associated with the development of T2D and the cardiovascular complications are well established, including obesity, hypertension, and hyperlipidemia. However, given that rates of diabetes and complications vary between and within different populations, there is growing interest in other contributing factors including genetics, the environment, and sociocultural factors.[24] For an individual, risk factors can be stratified into nonmodifiable (e.g., family history and age) and modifiable (e.g., blood pressure).

In this study we used both traditional parametric methods and newer machine learning methods to analyze cardiometabolic data from a cohort of members of a large U.S.-based provider of a workplace health promotion and wellness programs. We compared three supervised machine learning methods: random forest, bagging, and gradient boosting, and a traditional statistical method: parametric multivariate logistic regression. Parametric models have strong model specification, which may not be guaranteed; and machine learning methods such as random forest can be considered as nonparametric modeling methods, which avoid misspecification and hence reduce bias. Logistic regression may not be the best statistical technique if there are significant correlations between some of the regressor variables; and machine learning methods such as random forest remedy this

Table 5. Characteristics of the Medoids of Four Clusters for Nonmodifiable Risk Factors

	Medoid			
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
Diabetes	*No*	*No*	*Yes*	*Yes*
Gender	Female	Male	Female	Male
Age (years)	44	39	55	48
Chronic lung disease	No	No	No	No
Cancer	No	No	No	No
Heart disease	No	No	No	No
Depression	No	No	No	No
No. of months eligible	8	8	12	8
Education	College degree	College degree	College degree	College degree

Table 6. Characteristics of the Medoids of Four Clusters for Potentially Modifiable Risk Factors

	Medoid			
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
Diabetes	*No*	*No*	*Yes*	*Yes*
Daily sedentary time (h)	7	7	9	7
Average daily servings of fruits and vegetables	3	2	3	2
Average hours of sleep/day	7	7	7	7
Kessler Stress Score	13	12	12	13
Tobacco use	Never smoked	Never smoked	Never smoked	Never smoked
Alcohol (no. of drinks/week)	2	2	1	2
BMI	24.4	29.2	33.8	32.2
Waist circumference (cm)	80.8	96.5	104.1	99.1
Systolic BP (mmHg)	118	125	118	131
Diastolic BP (mmHg)	62	70	70	82
Total cholesterol (mmol/L)	5.17	4.65	4.53	4.40
HDL cholesterol (mmol/L)	1.76	1.11	1.16	1.14
LDL cholesterol (mmol/L)	2.92	2.95	2.61	2.20
Triglycerides (mmol/L)	1.06	1.31	1.60	2.30
Weekly verified standard workouts	0	0	0	0
Weekly verified light workouts	0	0	0	0
Other activities	5	5	5	5

by decreasing correlation, using only a subset of predictors to create splits within the data.

The number of potentially modifiable risk factors and the ability to track changes in them over time will inevitably increase as new digital ecosystems become available.[25] In the future, expectations will shift to technology including machine learning to produce meaningful personal benefits from using these technologies, mainly as a consequence of advances in sensor technology (especially miniaturization, increased power, and improvements in esthetics), smartphone computing capability, and artificial intelligence. Applying machine learning to determine weighted risk for individual factors and their changes with time using digital health technologies has the potential to be more impactful for disease prevention.

There are some limitations to our analyses. First, we were not able to stratify the diagnosis for the type of diabetes, although it is likely that this population reflects the current status of diabetes in the background U.S. population, where the overwhelming majority has a diagnosis of T2D. Second, we did not have access to the use of individual medications, namely, therapies for hyperlipidemia and hypertension as well as glucose-lowering medicines for those with diabetes. Given the cross-sectional nature of our analyses, we were therefore unable to determine the impact of these changes over the individual risk factors over time. Third, there are limitations to self-reported data although, as mentioned earlier, with the increased availability of new health technologies for automatic data capture, the impact of this may wane with time.[26] Fourth, we also did not have access to information on race and ethnicity. Currently in the United States, over 29 million people are uninsured, with substantial inequalities in access to care along economic, gender, and racial lines persisting.[27] Previous studies have documented that racial/ethnic minority groups also received low quality of care, including preventative health services, compared to their White counterparts, and that racial/ethnic minority groups have higher rates of diabetes-related complications.[24]

As the present study shows, big datasets are often plagued by missing data. Some corrective measures may be taken; although these cannot compensate for complete and accurate data collection, such corrective measures can provide a safeguard when dealing with missing data.

R packages exist to implement machine learning algorithms that enable the process of analyzing large datasets and help to reveal important explanatory factors. The vision for the future of such models may likely be an integrated one. In such a closed integrated system, data will be collected in real-time from patients, processed through models such as those discussed in this study, and then immediately fed back to the patient in the form of information that focuses on the patient's modifiable risks. We are still some way from having such as closed system, but models such as those discussed in this study will be an important component.

Acknowledgment

The study was funded by a Society of Actuaries Centers of Actuarial Excellence Research Grant.

Author Disclosure Statement

No competing financial interests exist.

References

1. Cho NH, Shaw JE, Karuranga S, et al.: IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diab Res Clin Pract 2018;138:271–281.
2. Bommer C, Heesemann E, Sagalova V, et al.: The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study. Lancet Diab Endocrinol 2017; 5:423–430.
3. NIH: National Institute of Diabetes and Digestive and Kidney Diseases: Risk factors for type 2 diabetes 2016. [cited November 11, 2018]. https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes

4. Xu G, Liu B, Sun Y, et al.: Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study. BMJ 2018;362:k1497.
5. Meagher P, Adam M, Civitarese R, et al.: Heart failure with preserved ejection fraction in diabetes: mechanisms and management. Can J Cardiol 2018;34:632–643.
6. Bowe B, Xie Y, Li T, et al.: Analysis of the Global Burden of Disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. Kidney Int 2018;567–581.
7. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol 2018;6:361–369.
8. Li L, Cheng W-Y, Glicksberg BS, et al.: Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Trans Med 2015;7:311ra174.
9. Kessler RC, Andrews G, Colpe LJ, et al.: Short screening scales to monitor population prevalences and trends in nonspecific psychological distress. Psych Med 2002;32:959–976.
10. van Buuren S, Groothuis-Oudshoorn K: mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011; 45:63.
11. Hastie TJ, Pregibon D: Generalized linear models. In: Chambers S, Hastie TJ, eds. Statistical Models. London; Chapman & Hall/CRC, 1992.
12. Lunardon N, Menardi G, Torelli N: ROSE: a package for binary imbalanced learning. The R Journal 2014;6:79–89.
13. Menardi G, Torelli N: Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 2017; 28:92–122.
14. Kuhn M: caret: Classification and Regression Training. R package version 6.0-77. 2017.
15. Hastie T, Tibshirani R, Friedman J: Elements of Statistical Learning: Data Mining Inference and Prediction. Springer Series in Statistics. New York, NY: Springer Verlag, 2001.
16. Mächler M, Rousseeuw P, Struyf A, et al.: cluster: ''Finding Groups in Data'': Cluster Analysis Extended. CRAN, 2017.
17. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ: Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J Math Model Algor 2006; 5:475–504.
18. Robin X: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:1–8.
19. Gower JC: A general coefficient of similarity and some of its properties. Biometrics 1971;27:857–871.
20. Breiman L: Random forests. Mach Learn 2001;45:5–32.
21. Friedman JH: Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–1232.
22. Ogurtsova K, da Rocha Fernandes JD, Huang Y, et al.: IDF Diabetes Atlas: global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract 2017;128: 40–50.
23. American Diabetes Association: Economic costs of diabetes in the U.S. in 2017. Diabetes Care 2017;41:917–928.
24. Golden SH, Brown A, Cauley JA, et al.: Health disparities in endocrine disorders: biological, clinical, and nonclinical factors—An Endocrine Society Scientific Statement. J Clin Endocrinol Metab 2012;97:E1579–E1639.
25. Kerr D, Axelrod C, Hoppe C, Klonoff DC: Diabetes and technology in 2030: a utopian or dystopian future? Diabet Med 2018;35:498–503.
26. Greenfield TK, Bond J, Kerr WC: Biomonitoring for improving alcohol consumption surveys: the new gold standard? Alcohol Res 2014;36:39–45.
27. Gaffney A, McCormick D: The Affordable Care Act: implications for health-care equity. Lancet 2017;389:1442–1452.

Address correspondence to:
Xiyue Liao, PhD
Department of Statistics and Applied Probability
University of California
Santa Barbara, CA 93106

E-mail: liao@pstat.ucsb.edu

(*Appendix follows* →)

Appendix Table A1. Female Participants (N=18,855), Self-Reported Diagnoses of Diabetes and Heart Disease, and Measurements of Body Mass Index, Waist Size, Stress, Weekly Alcohol Consumption, Fasting Plasma Glucose Levels, and Time Spent Inactive

			Disease prevalence (%)		*Average measures*					
Year	*Age*	N	*Diabetes*	*Heart disease*	*BMI*	*Waist size (cm)*	*Stress score*	*Alcohol (no./week)*	*Sedentary (h/day)*	*FPG (mmol/L)*
	<40	4948	3.0	0.3	27.3±6.5	83.3±14.9	14.4±5.3	1.9±4.1	7.8±3.6	4.9±0.8
2012	40–60	7013	8.3	1.1	28.8±6.6	87.64±14.7	13.7±4.8	1.9±4.2	7.6±3.6	5.3±1.1
	>60	1477	15.1	3.3	29.4±6.2	90.44±14.6	13.3±5.2	1.9±4.0	7.2±3.4	5.5±1.3
	<40	8653	2.9	0.3	27.6±6.7	83.5±14.9	14.6±5.5	1.5±2.6	8.9±4.4	4.8±0.8
2013	40–60	10,902	8.8	1.2	29.2±6.8	88.3±15.0	13.9±5.1	1.6±2.9	8.9±4.2	5.2±1.1
	>60	2195	15.1	3.7	29.6±6.3	90.3±14.3	13.3±4.3	1.4±2.6	8.7±4.1	5.5±1.3
	<40	20,012	3.0	0.3	27.6±6.7	84.4±15.3	14.7±4.3	1.4±2.4	9.5±3.8	4.8±0.7
2014	40–60	26,049	9.1	0.9	28.9±6.6	88.3±15.0	13.9±3.8	1.3±2.6	9.3±3.8	5.2±1.1
	>60	4756	16.1	2.1	28.9±6.3	89.9±14.6	13.2±3.2	1.2±2.4	8.8±3.7	5.5±1.2
	<40	14,211	3.2	0.3	27.4±6.6	84.1±15.2	14.8±4.8	1.7±2.5	9.6±4.3	4.8±0.7
2015	40–60	15,934	8.1	0.8	29.0±6.6	88.9±15.2	14.0±4.2	1.7±2.9	9.5±4.3	5.2±1.1
	>60	2705	13.5	2.1	29.0±6.2	90.4±14.2	13.4±3.7	1.7±3.5	9.3±4.2	5.5±1.2

Data are expressed as percentage and mean±SD.
SD, standard deviation.

Appendix Table A2. Male Participants (N=8,399) Self-Reported Diagnoses of Diabetes and Heart Disease and Measurements of Body Mass Index, Waist Size, Stress, Weekly Alcohol Consumption, Fasting Plasma Glucose Levels, and Time Spent Inactive

			Disease prevalence (%)		*Average measure*					
Year	*Age*	N	*Diabetes*	*Heart disease*	*BMI*	*Waist size (cm)*	*Stress score*	*Alcohol (no./week)*	*Sedentary (h/day)*	*FPG (mmol/L)*
	<40	4924	9.9	2.5	29.0±5.4	89.7±11.3	13.1±4.4	3.6±6.0	7.2±3.6	5.1±0.9
2012	40–60	7066	9.3	2.5	29.3±5.2	93.9±11.7	13.1±3.9	3.9±6.8	7.2±3.4	5.6±1.2
	>60	1725	9.5	1.9	29.3±4.9	95.6±11.4	13.1±3.4	4.0±6.0	7.1±3.2	5.9±1.4
	<40	6232	3.7	0.3	28.3±5.5	89.3±12.2	14.1±4.8	3.3±5.0	8.4±4.2	5.1±0.9
2013	40–60	7315	10.3	2.0	29.4±5.3	93.4±12.9	13.3±4.2	3.2±4.5	8.4±3.8	5.5±1.3
	>60	1427	19.2	7.3	29.4±5.0	95.3±13.0	12.8±3.7	3.4±6.1	7.6±3.6	5.9±1.5
	<40	16,128	3.7	0.2	28.3±5.4	90.8±12.8	14.1±3.9	3.0±4.3	8.9±3.6	5.1±0.8
2014	40–60	20,065	12.4	1.6	29.3±5.2	94.8±12.9	13.2±3.3	2.7±4.4	8.4±3.4	5.5±1.3
	>60	4,180	21.9	5.5	28.9±4.9	96.5±13.1	12.6±2.8	2.6±4.1	7.7±3.2	5.8±1.4
	<40	12,844	4.3	0.3	28.3±5.5	91.1±13.1	14.2±4.3	3.3±4.7	8.8±4.0	5.0±0.8
2015	40–60	13,849	12.0	1.7	29.6±5.4	96.0±13.4	13.3±3.6	3.3±4.9	8.4±3.7	5.5±1.3
	>60	2644	17.4	6.3	29.3±5.0	97.4±12.8	12.7±2.9	3.3±5.1	7.9±3.6	5.7±1.4

Data are expressed as percentage and mean±SD.
BMI, body mass index.