MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190258

Using Machine Learning to Integrate Socio-Behavioral Factors in Predicting Cardiovascular-Related Mortality Risk

Hanyin Wang[a], Yikuan Li[b], Hongyan Ning[a], John Wilkins[a], Donald Lloyd-Jones[a], Yuan Luo[a]

[a] Feinberg School of Medicine, Northwestern University, Chicago, IL, U.S.A
[b] Department of EECS, Northwestern University, Chicago, IL, U.S.A

Abstract

Cardiovascular disease is prevalent and associated with significant mortality rate. Robust lifetime risk stratification for cardiovascular disease is important for effective prevention, early diagnoses, targeted intervention, and improved prognosis. Health disparities, manifested as socio-behavioral factors, are believed to have multiple effects throughout life with great complexity. Multiple studies investigated lifetime cardiovascular-related mortality risk prediction focusing on subjects' pathophysiology and intervention profiles. In this study, we applied machine learning algorithms and focused on integrating socio-behavioral factors to pathophysiology and intervention profiles to predict cardiovascular-related mortality risk. Our results showed that multiple machine learning algorithms can predict risk with reasonable accuracy, using mixed types of features. Particularly, socio-behavioral factors contributed significantly to the improved accuracy of mortality risk prediction. Feature analysis identified the odds ratio of socio-behavioral factors for cardiovascular-related mortality and offered potential insights on how they impact subjects' long-term outcomes. Our results call for further investigation of this important topic.

Keywords:
Risk Factors, Cardiovascular Diseases, Machine Learning

Introduction

Cardiovascular Disease (CVD) is a major cause of disability and premature death throughout the world and substantially contributes to the escalating costs of health care, mortality, and morbidity [1]. According to the report from World Health Organization, an estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths [2]. In the US, the mortality of cardiovascular disease is as high as 0.25 (95% CI 0.10-0.63) [3]. In order to reduce the physical, emotional, and economical burden caused by cardiovascular-related mortality, it is important to develop robust lifetime risk stratification algorithm that is accurate and that generalizes at a population level.

Previous studies revealed the association between cardiovascular disease and demographics, physiology, comorbidity, and medication features, such as body mass index [4], chronic kidney diseases [5], hypertension and obesity [6], anemia [7], etc. Social determinants and behavioral factors, such as education level, income level, smoking status, and diet habit, act upon populations at all times, and they have also been found to have wide-ranging health related effects across the life course of a population [8, 9]. For example, higher income is linked to better health outcomes, including lower prevalence of cardiovascular disease and lower mortality risk among populations with cardiovascular disease. With growing awareness of their importance, previous studies have explored relationships between social determinants and risks for developing cardiovascular diseases [10-13]. In addition to social determinants, behavioral factors, such as diet [14-16] and physical activity [17, 18], are investigated to be related to various mortality and cardiovascular-related outcomes. For example, poor diet, lack of physical activity, and cigarette smoking are all linked to premature deaths [19]. Unfortunately, a majority of existing studies focused on linking social and behavioral factors to risks for developing specific diseases, leaving systematic analysis on these factors' relations with mortality risks a largely uncharted territory. On the other hand, pooling the social determinants and behavioral factors together and utilizing more powerful approaches, such as machine learning [20-22], may shed more light on the landscape and yield a more robust mortality-risk prediction model.

In this study, we apply different machine learning models to predict cardiovascular-related mortality risks by integrating social behavioral factors with subjects' pathophysiology and intervention profiles as features. We aim to provide a comprehensive understanding of the socio-behavioral landscape and use the additional knowledge to improve cardiovascular-related mortality prediction. Our results demonstrate the impact of social and behavioral factors on cardiovascular health at population level and the feasibility towards more robust lifetime risk stratification by integrating them.

Methods

Study population

In this study, we used data from The Cardiovascular Disease Life Risk Pooling Project (LRPP) [23]. LRPP was designed as an individual-level pooled dataset from 20 US community-based cardiovascular disease cohorts that include demographics, physiological test results, medication status, socio-behavioral factors, and mortality indicators (for mortality of cardiovascular disease, coronary heart disease, and mortality resulting from all causes). From the overall sample size of 1,097,178 observations of 277,296 subjects, 40,711 completed cases with all predicting features available were selected for each mortality in this study. Shown in Table 1 is the data distribution after splitting training and test set into approximately 7:3 ratio.

Table 1– Data distribution

	Training set				Test set				All data			
	Total	Death	Non-event	% death	Total	Death	Non-event	% death	Total	Death	Non-event	%death
CVD	28,497	1,772	26,725	6.22%	12,214	846	11,368	6.93%	40,711	2,618	38,093	6.43%
CHD	28,497	1,772	26,725	6.22%	12,214	846	11,368	6.93%	40,711	2,618	38,093	6.43%
TOT	28,497	6,048	22,449	21.22%	12,214	2,703	9,511	22.14%	40,711	8,751	31,960	21.50%

CVD: mortality due to cardiovascular diseases; CHD: mortality due to coronery heart disease; TOT: all causes mortality

Statistical analysis

In total, 25 features were chosen from the entire LRPP dataset. They belong to the following domains: demographics, physiological test results, medication status, and socio-behavioral factors. Features were grouped into 3 different combinations, physiological test result (Phy), physiological test results combined with medication status (Phy+Med), and further including socio-behavioral factors (Phy+Med+Soc). While demographics were included in each group (Table 2).

Four supervised machine learning algorithms were applied for prediction of the three outcomes, including Naïve Bayes (NB), Logistic Regression (LR), Linear Support Vector Machine (SVM), and Random Forest (RF). Training and test set were split to approximately 7:3. Due to the imbalance of the data (incidence rate at approximately 6%), parameter *'class_weight'* was set to *'balanced'* for LR, SVM and RF. To reduce over-fitting and variance, five-fold cross validation was performed on training set. Best parameters for each model, tuned through cross validation, were then applied on test set to evaluate the predictive performance of each classifier. The area under the receiver operating characteristic curve (AUC) as well as predicted probability were acquired from each algorithm.

The AUCs of the three models are shown in Table 3, ranging from the lowest to the highest. Considering that the distribution of the predicted probability was unknown, permutation test [24] was applied to explore the difference between the probability of correct prediction of every two models (ie. tested between Phy and Phy + Med, Phy+Med and Phy+Med+Soc).

In RF, coefficient *'feature_importance'* was extracted to identify the importances of features. Generalized linear regression (logit link function) was employed to find the odds ratio (OR) to evaluate risk for mortality of each socio-behavioral factor. Data cleaning, permutation test, and generalized linear regression were performed by R 3.3.1. *'scikit learn'* package was used for machine learning algorithms.

Results

AUCs of different algorithms and models

Table 3 shows the AUCs yielded from each of the four machine learning algorithms as applied to predict the three outcomes, with the three combination of the features. P-value 1 was derived from the permutation test between the feature combination of Phy and Phy+Med, while P-value 2 shows the significance of difference between the feature combination of Phy+Med and Phy+Med+Soc. Result demonstrates that AUCs are in a range from 0.7263 to 0.8941 across all the models.

Table 2– Predictors features

Categories	Variable names	Description
Demographics	SEX	Gender
	AGE	Calculated age at exam, round to 0 .1
	BMI	Body mass index, kg/m^2
Physiological tests	SBP	Systolic blood pressure, mmHg
	DBP	Diastolic blood pressure, mmHg
	LDLCHL	LDL cholesterol, mg/dL
	HDLCHL	HDL cholesterol, mg/dL
	TOTCHL	Total cholesterol, mg/dL
	TRIGLY	Triglycerides, mg/dL
	GLU	Glucose level
	CVD_BASELINE	Prevalent of cardiovascular disease at enrollment
Medication statues	HXDIAB	Diabetic Medication Status
	RXCHL	Medication of lowering lipid
	RXHYP	Medication of anti-hypertension
Socio-behavioral factors	RACE	Race/ethnicity
	EDU	School education
	SMOKER	Current smoker
	FORSMOKER	Former smoker
	DRINKER	Current drinker
	ALCO_ML_DAY	Alcohol intake, mL /day
	DASH	Dietary Approaches to Stop Hypertension score
	aHEI	Alternate Healthy Eating Index score 2010
	aMed	Alternate Med diet score (all cohorts median)
	primary PA	Primary physical activity
	MVPA	Moderate to vigorous physical activity
Outcomes	CVD_DTH	Death caused by cardiovascular diseases
	CHD_DTH	Death caused by coronary heart disease
	TOT_DTH	All causes death

Table 3– AUCs of the three models and four algorithms

	Phy	PM	*P-val 1*	PMS	*P-val 2*
Mortality due to cardiovascular disease					
NB	0.7436	0.7579	*0.001*	0.8048	*0.010*
LR	0.8510	0.8554	*0.001*	0.8752	*0.001*
SVM	0.8510	0.8558	*0.298*	0.8754	*0.001*
RF	0.8461	0.8535	*0.001*	0.8941	*0.001*
Mortality due to coronary heart disease					
NB	0.7436	0.7579	*0.001*	0.8048	*0.010*
LR	0.8510	0.8554	*0.001*	0.8752	*0.001*
SVM	0.8510	0.8558	*0.298*	0.8754	*0.001*
RF	0.8461	0.8535	*0.001*	0.8941	*0.001*
Mortality result from all causes					
NB	0.7263	0.7349	*0.001*	0.7864	*0.001*
LR	0.8214	0.8247	*0.001*	0.8541	*0.001*
SVM	0.8208	0.8241	*0.001*	0.8538	*0.001*
RF	0.8325	0.8393	*0.001*	0.8921	*0.001*

PM: combination of Phy+Med; PMS: combination of Phy+Med+Soc

Considering NB algorithm as the baseline, overall result illustrates significant improvement in AUCs when applying LR, SVM, and RF algorithms. The Phy combination consisted of the 11 features from Demographics and Physiological tests categories. NB algorithm provided AUCs between 0.7263 and 0.7436 based on these features. RF showed a better performance within this combination that yielded AUCs of 0.8461 for prediction of mortality risk of CVD and CHD. The most promising algorithms for this model were LR and SVM which achieved AUCs of 0.8510 for prediction regarding CVD and CHD mortality. When adding the Medication status to the feature collection, we observed an elevation in all AUCs compared to the Phy model. Meanwhile, p-values ($\leq$0.001) generated from permutation test also confirmed the statistically significant improvement on the probability of correct prediction, except for SVM algorithm for CVD and CHD mortality risk prediction. Last but not least, much better performance was observed when integrating the Socio-behavioral factors into the feature combination. All AUCs yielded from RF algorithm became above 0.80, the highest AUC reached 0.8941. Comparing the AUCs to the previous model, this model with socio-behavioral factors demonstrated the best performance. Results from permutation test for difference between the correct prediction probability of Phy+Med and Phy+Med+Soc also helped verify the significant improvement.

Feature analysis

Shown in Figure 1 are the feature importances derived from RF for different feature combinations. In the combinations without socio-behavioral factors (Phy and Phy+Med), age served as the most important feature for classification. Blood pressure, BMI, triglycerides level, and the three kinds of cholesterol levels also played relatively important roles in the first two Phy and Phy+Med [25]. The pattern was different in the third feature combination (Phy+Med+Soc), race performed as the most important feature for all three outcomes, followed by age and SBP. Meanwhile, aHEI score also appeared to be an essential feature for classification. Other important factors shown in the previous two feature combinations still remained on top.

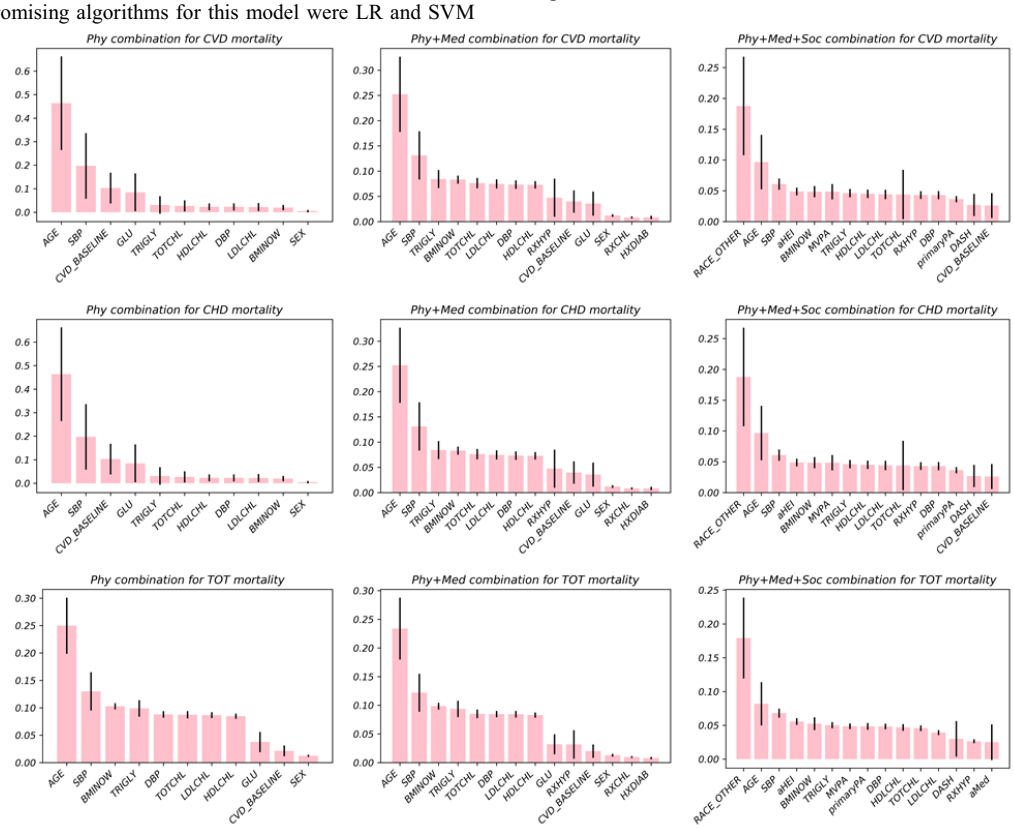

Figure 1– Feature importances for different feature combinations derived from Random Forest. For combination Phy+Med+Soc,only the first 15 highest ranked features were shown in the figure.

Discussions, Limitations and Future Work

With machine learning approaches, we were able to detect the hidden patterns within larg-scale data. Among the four algorithms applied in this study, LR, SVM and RF demonstrated reasonable accuracy. Among them, RF appeared to be the best for solving the current problem.

Significant improvements were observed from those machine learning configurations when adding socio-behavioral factors as well as medication status into predicting models. It should not be surprising to observe medication status (BP medication, lipids medication, and DM medication) as an important predictor, since only the subjects with such health disorders would take the medicine. Meanwhile, blood pressure disorders [26, 27], dyslipidemia [28, 29], and diabetes mellitus [30] were also found to be physiologically related with mortality caused by cardiovascular diseases. Eleven socio-behavioral factors were integrated to the previous models, covered the aspects of race, education, smoking, diet, and physical activity. To understand which sociobehavioral factors in particular have a higher risk for cardiovascular-related mortality, a further general linear regression (logit link function) was performed (Table 4). Comparing to Asian, subjects with race in African American and White were found to have odds ratio greater than 1 with p-values satisfying significant level, which indicates a higher mortality risk for African American and White population. Moreover, current smokers also experienced a higher risk than those who were not currently smoking. Compared with the educational level below, high school, higher educational level (high school or above) reduces the risk. Subjects with higher dietary scores (DASH and aHEI) were also found to be undertaking higher risk.

Previous studies with consideration of social factors were generally focused on social determinants [31] while ignored the importance of behavior. This study showed that socio-behavioral factors, such as smoking and diet, can also be essential predictors for mortality risk of cardiovascular diseases and is worthwhile to be included into predictive models. Similarly, socio-behavioral factors might also be worth considering for risk prediction of other health outcomes.

The LRPP study is a pooled project from a large variety of cohorts. Therefore, missing rate is relatively high for part of the features especially the socio-behavioral factors. This limited the size of training and test set and likely introduced biases. In future work, we will use modern imputation methods [32-34] to impute the missing data and scale our analysis on larger cohorts. Continuing from what we found in this study, we are planning to perform further analysis considering time sequence. With the dimension of time, we will generate more sophisticated models for prediction. At the same time, the approaches we demonstrated in this study could be applied to other CVD outcomes as well as other diseases.

Conclusions

In the present study, we applied multiple machine learning algorithms and integrated socio-behavioral factors with pathophysiology and medication profiles of subjects to predict cardiovascular-related mortality. All the machine learning algorithms, including Naïve Bayes, Logistic Regression, Linear Support Vector Machine, and Random Forest yielded reasonable accuracy. In addition to pathophysiological and interventional profiles, socio-behavioral factors were also included as predictive features. Integration of socio-behavioral factors brought statistically significant improvement of the all three mortality risk predictions, including the risk of mortality due to cardiovascular diseases, coronary heart disease, and mortality resulting from all causes. In the overall analysis, age, race, and blood pressure were found to be relatively more important for the prediction. While in the analysis for odds ratio of socio-behavioral factors, race, educational level, smoking, and diet were found to have significant impact on cardiovascular-related outcomes.

Given that socio-behavioral factors manifested long-term impacts on the health outcome disparities, we expect more exciting and promising developments of robust lifetime risk prediction algorithms with tighter integration of socio-behavioral factors to inform policy making and improve the population health and longevity.

Table 4 – Odds ratio and significance of socio-behavioral factors

Features	OR for CVD_DTH	P-value	OR for CHD_DTH	P-value	OR for TOT_DTH	P-value
RACE_ASIAN			(reference group)			
RACE_BLACK	1.7074	0.001***	1.7074	0.001***	2.6543	0.000***
RACE_HISPANIC	0.7668	0.173	0.7668	0.173	0.8067	0.059
RACE_OTHER	0.5646	0.476	0.5646	0.476	1.6610	0.399
RACE_WHITE	1.9269	0.000***	1.9269	0.000***	2.7296	0.000***
EDU < HIGH SCHOOL			(reference group)			
EDU > HIGH SCHOOL	0.6655	0.000***	0.6655	0.000***	0.5489	0.000***
EDU_HIGH SHCOOL	0.7089	0.000***	0.7089	0.000***	0.6733	0.000***
SMOKER_0			(reference group)			
SMOKER_1	1.6952	0.000***	1.6952	0.000***	2.6555	0.000***
FORSMOKER_0			(reference group)			
FORSMOKER_1	1.0318	0.549	1.0318	0.549	1.3092	0.000***
DRINKER_0			(reference group)			
DRINKER_1	0.8913	0.503	0.8913	0.503	1.1589	0.037*
ALCO_ML_DAY	1.0036	0.235	1.0036	0.235	1.0027	0.043
DASH	1.0651	0.000***	1.0651	0.000***	1.0657	0.000***
aHEI	0.9540	0.000***	0.9540	0.000***	0.9531	0.000***
aMed	1.0326	0.098	1.0326	0.098	1.0590	0.000***
primaryPA	0.9526	0.109	0.9526	0.109	0.8895	0.000***
MVPA	1.0416	0.164	1.0416	0.164	1.0130	0.495

** : <0.05; ** : <0.01; *** : <0.001*

Acknowledgements

This work was supported by NIH Grant 1R21LM012618. We thank Ying Liu at UT Dallas for helpful comment.

References

[1] S. Mendis, S. Davis, and B. Norrving, "Organizational update: the world health organization global status report on noncommunicable diseases 2014; one more landmark step in the combat against stroke and vascular disease," (in eng), no. 1524-4628 (Electronic).
[2] W. H. Orgnization, "Cardiovascular diseases (CVDs)," 2017.
[3] M. J. Benjamin Ej Fau - Blaha *et al.*, "Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association," (in eng), no. 1524-4539 (Electronic).
[4] S. S. Khan, H. Ning, J. T. Wilkins, and et al., "Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity," *JAMA Cardiology,* **3** (2018), 280-287.
[5] D. E. Weiner *et al.*, "Chronic kidney disease as a risk factor for cardiovascular disease and all-cause mortality: a pooled analysis of community-based studies," **15** (2004), 1307-1315.
[6] F. S. Ahmad, H. Ning, J. D. Rich, C. W. Yancy, D. M. Lloyd-Jones, and J. T. Wilkins, "Hypertension, Obesity, Diabetes, and Heart Failure–Free Survival: The Cardiovascular Disease Lifetime Risk Pooling Project," *JACC: Heart Failure,* **4** (2016), 911-919.
[7] M. J. Sarnak *et al.*, "Anemia as a risk factor for cardiovascular disease in The Atherosclerosis Risk in Communities (ARIC) study," **40** (2002), 27-33.
[8] B. E. Garrett, S. R. Dube, S. Babb, and T. McAfee, "Addressing the Social Determinants of Health to Reduce Tobacco-Related Disparities," (in eng), no. 1469-994X (Electronic).
[9] E. P. Havranek *et al.*, "Social Determinants of Risk and Outcomes for Cardiovascular Disease," **132** (2015), 873-898.
[10] N. E. Adler, M. M. Glymour, and J. Fielding, "Addressing Social Determinants of Health and Health Inequalities," (in eng), no. 1538-3598 (Electronic).
[11] E. Blas and A. S. Kurup, *Equity, social determinants and public health programmes*. World Health Organization, 2010.
[12] C. Kreatsoulas and S. S. J. C. J. o. C. Anand, "The impact of social determinants on cardiovascular disease," **26** (2010), 8C-13C.
[13] E. K. Chung *et al.*, "Screening for Social Determinants of Health Among Children and Families Living in Poverty: A Guide for Clinicians," (in eng), no. 1538-3199 (Electronic).
[14] J.-P. Després and B. J. N. r. r. Lamarche, "Effects of diet and physical activity on adiposity and body fat distribution: Implications for the prevention of cardiovascular disease," 6 (1993), 137-159.
[15] R. Estruch *et al.*, "Primary prevention of cardiovascular disease with a Mediterranean diet," 368 (2013), 1279-1290.
[16] J. H. O'Keefe Jr and L. Cordain, "Cardiovascular disease resulting from a diet and lifestyle at odds with our Paleolithic genome: how to become a 21st-century hunter-gatherer," in *Mayo Clinic Proceedings*, **79** (2004), 101-108: Elsevier.
[17] H. J. M. Kohl 3rd, s. i. sports, and exercise, "Physical activity and cardiovascular disease: evidence for a dose response," **6** (2001), S472.
[18] P. D. Thompson *et al.*, "Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease: a statement from the Council on Clinical Cardiology (Subcommittee on Exercise, Rehabilitation, and Prevention) and the Council on Nutrition, Physical Activity, and Metabolism (Subcommittee on Physical Activity)," 107 (2003), 3109-3116.
[19] J. M. McGinnis, J. R. Williams-Russo P Fau - Knickman, and J. R. Knickman, "The case for more active policy attention to health promotion," (in eng), no. 0278-2715 (Print).
[20] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care," (in eng), no. 1573-689X (Electronic).
[21] R. C. Deo, "Machine Learning in Medicine," (in eng), no. 1524-4539 (Electronic).
[22] T. Zheng *et al.*, "A machine learning-based framework to identify type 2 diabetes through electronic health records," (in eng), no. 1872-8243 (Electronic).
[23] J. T. Wilkins *et al.*, "Data Resource Profile: The Cardiovascular Disease Lifetime Risk Pooling Project," (in eng), no. 1464-3685 (Electronic).
[24] E. W. Noreen, *Computer-intensive methods for testing hypotheses: an introduction* (Computer). 1989.
[25] R. K. Upadhyay, "Emerging risk biomarkers in cardiovascular diseases and disorders," (in eng), *Journal of lipids,* (2015), 971453-971453.
[26] S. L. Stevens *et al.*, "Blood pressure variability and cardiovascular disease: systematic review and meta-analysis," (in eng), no. 1756-1833 (Electronic).
[27] X. Xie *et al.*, "Effects of intensive blood pressure lowering on cardiovascular and renal outcomes: updated systematic review and meta-analysis," (in eng), no. 1474-547X (Electronic).
[28] S. Karr, "Epidemiology and management of hyperlipidemia," (in eng), no. 1936-2692 (Electronic).
[29] S. Yusuf *et al.*, "Cholesterol Lowering in Intermediate-Risk Persons without Cardiovascular Disease," (in eng), no. 1533-4406 (Electronic).
[30] A. Rawshani *et al.*, "Mortality and Cardiovascular Disease in Type 1 and Type 2 Diabetes," (in eng), no. 1533-4406 (Electronic).
[31] S. N. Kasthurirathne, J. R. Vest, N. Menachemi, P. K. Halverson, and S. J. Grannis, "Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services," (in eng), no. 1527-974X (Electronic).
[32] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron, "Using Machine Learning to Predict Laboratory Test Results," *American Journal of Clinical Pathology,* **6** (2016), 778-788.
[33] S. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *Journal of statistical software,* 45 (2011).
[34] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron, "3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data," *Journal of the American Medical Informatics Association (JAMIA),* **6** (2017), 645-653.

Address for correspondence
yuan.luo@northwestern.edu