



ORIGINAL ARTICLE

Use of machine-learning algorithms to determine features of systolic blood pressure variability that predict poor outcomes in hypertensive patients

Ronilda C. Lacson^{1,2}, Bowen Baker³, Harini Suresh³, Katherine Andriole^{1,2}, Peter Szolovits³ and Eduardo Lacson Jr^{4,5}

¹Brigham and Women's Hospital, Boston MA, USA, ²Harvard Medical School, Boston, MA, USA, ³Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, ⁴Dialysis Clinic, Inc., Nashville, TN, USA and ⁵Tufts Medical Center, Boston, MA, USA

Correspondence and offprint requests to: Ronilda C. Lacson; E-mail: rlacson@rics.bwh.harvard.edu

ABSTRACT

Background. We re-analyzed data from the Systolic Blood Pressure Intervention Trial (SPRINT) trial to identify features of systolic blood pressure (SBP) variability that portend poor cardiovascular outcomes using a nonlinear machine-learning algorithm.

Methods. We included all patients who completed 1 year of the study without reaching any primary endpoint during the first year, specifically: myocardial infarction, other acute coronary syndromes, stroke, heart failure or death from a cardiovascular event ($n = 8799$; 94%). In addition to clinical variables, features representing longitudinal SBP trends and variability were determined and combined in a random forest algorithm, optimized using cross-validation, using 70% of patients in the training set. Area under the curve (AUC) was measured using a 30% testing set. Finally, feature importance was determined by minimizing node impurity averaging over all trees in the forest for a specific feature.

Results. A total of 365 patients (4.1%) reached the combined primary outcome over 37 months of follow-up. The random forest classifier had an AUC of 0.71 on the testing set. The 10 most significant features selected in order of importance by the automated algorithm included the urine albumin/creatinine (CR) ratio, estimated glomerular filtration rate, age, serum CR, history of subclinical cardiovascular disease (CVD), cholesterol, a variable representing SBP signals using wavelet transformation, high-density lipoprotein, the 90th percentile of SBP and triglyceride level.

Conclusions. We successfully demonstrated use of random forest algorithm to define best prognostic longitudinal SBP representations. In addition to known risk factors for CVD, transformed variables for time series SBP measurements were found to be important in predicting poor cardiovascular outcomes and require further evaluation.

Keywords: blood pressure, cardiovascular diseases, heart disease, hypertension, machine learning

Received: 28.3.2018. Editorial decision: 14.5.2018

© The Author(s) 2018. Published by Oxford University Press on behalf of ERA-EDTA.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

INTRODUCTION

Hypertension management decreases cardiovascular morbidity and mortality [1–4]. Recently, the Systolic Blood Pressure Intervention Trial (SPRINT) reported that targeting a systolic blood pressure (SBP) <120 mmHg compared with <140 mmHg resulted in a lower rate of a combined primary outcome of fatal and nonfatal major cardiovascular events. The primary outcome occurred in 1.65% of the intensive-treatment group compared with 2.18% in the standard-treatment group [5]. When data from the SPRINT trial were made publicly available [6], we sought to identify additional features of the originally collected clinical variables that could further predict poor cardiovascular outcomes.

Serial standardized SBP measurements were performed in SPRINT, allowing for characterization of variability over time [5]. Since long term variability in SBP, a.k.a. visit-to-visit variability (VVV), has been shown in a recent meta-analysis and two subsequent studies to predict cardiovascular outcomes including death [7–9], we hypothesized that longitudinal changes and variability in blood pressure will have prognostic clinical significance. The need for parsimony in conventional statistical model building has created an arduous process of variable selection and limited use of conventional representations of trends and variability such as mean/median, standard deviation (SD) and coefficient of variation, and less commonly slope, variation independent of the mean and root mean squared error [7]. We previously demonstrated that incorporating multiple SBP trends in a machine-learning model significantly improved mortality prediction in patients on hemodialysis, compared with using only standard SBP representation including the mean and selected individual SBP values [10]. Therefore, we further hypothesized that we could overcome limitations of conventional statistical methodology by evaluating a wide range of representations of the features of longitudinal changes in SBP concurrently within a novel machine-learning framework.

To optimize analysis of the SPRINT data, we elected to utilize a nonlinear machine-learning algorithm, random forest search, to classify patients into a binary grouping of either event-free or having the SPRINT primary outcome. Random forest is a method for classification using multiple classification trees that are combined in an ensemble [11]. Multiple trees are trained and fitted to bootstrapped training data obtained through random sampling with replacement. The goal is to decrease the correlation between individual trees, which results in diminished variance when the trees are aggregated. Random forests accommodate sparsity [12], which is favorable in this case, due to a low percentage of patients who reached the primary outcome. The individual trees are designed to overfit on features (making very specific decisions that only account for part of the data set), whereas the voting strategy mitigates these effects by generalizing over the decisions of multiple trees. We sought to determine the best representation of time series SBP, with an assumption that one or more representations of SBP will be significant prognostic factors because the primary SPRINT trial result indicated that SBP targets were significantly related to outcome.

MATERIALS AND METHODS

Study population

We retrospectively analyzed data from SPRINT, which included 9361 patients with a SBP of 130 mmHg or higher and an increased cardiovascular risk, but without diabetes. *A priori*, we

limited this analysis to include all patients who completed 1 year of the study without reaching any primary endpoint during the first year, specifically myocardial infarction, other acute coronary syndromes, stroke, heart failure or death from a cardiovascular event ($n=8799$; 94%). This 1-year period represented our study baseline. Waiver of consent was granted for this de-identified data set (previously approved for release by the SPRINT committee) and the study was deemed 'exempt' by the Partners Healthcare Institutional Review Board.

Feature inclusion and calculation

All variables in the SPRINT baseline data were utilized in the analysis, including the binary variable indicating whether the patient was assigned to the intensive or usual care treatment group. Derived variables were computed using Tsfresh (Copyright 2016 Maximilian Christ, Blue Yonder GmbH), a Python package available through unrestricted software license and funded in part by the German Federal Ministry of Education and Research [13]. Tsfresh calculates and returns features from time series data, including mean, median and mode, SD, counts above mean for a time series (i.e. number of SBP values above mean for a patient over time), counts below mean, sum of absolute value of consecutive SBP change, variance, maximum and minimum values, SBP values above different percentiles (e.g. 90th percentile), linear regression slope, entropy and coefficients for continuous wavelet transform (CWT) of the SBP signal over time (Appendix Table S1). From these features, Tsfresh performs standard univariate analysis to predict the outcome, including Fisher's exact test, Kolmogorov-Smirnov test and the Kendal rank test. It takes into account multiple testing by using the Benjamini-Yekutieli correction. Figure 1a illustrates several variables that were automatically computed from time-series data. Significant features that predicted poor outcomes were all automatically identified within the advanced machine-learning framework. Figure 1b illustrates time series SBP data from two patients in the study: although the initial and last recorded SBP values were similar, variation in SBP, the maximum-recorded SBP and the number of SBP dips evident in the graph are not routinely used in conventional analyses.

Random forest algorithm

All features were normalized and included in a random forest algorithm implemented in Python (Scikit learn, BSD license). Random forest fitted several decision tree classifiers on subsamples of the data set and used averaging to measure the model's predictive accuracy. We varied the node splitting criterion for the decision tree classifiers using Gini index [14] and entropy [15], with entropy performing better. We then optimized hyperparameters of the random forest including number of trees and the minimum leaf size (i.e. the minimum number of training examples that may be at any leaf of a classification tree) by implementing hyperparameter grid search using 10-fold cross validation. A random forest with 50 trees and a minimum leaf size of 46 was the final result.

Feature importance determination

Finally, feature importance was determined using the Scikit algorithm for minimizing node impurity averaging over all trees in the forest for a specific feature. Specifically, every time a split of a node is made on a feature, the Gini impurity for the two descendant nodes is less than the parent node. That is, the

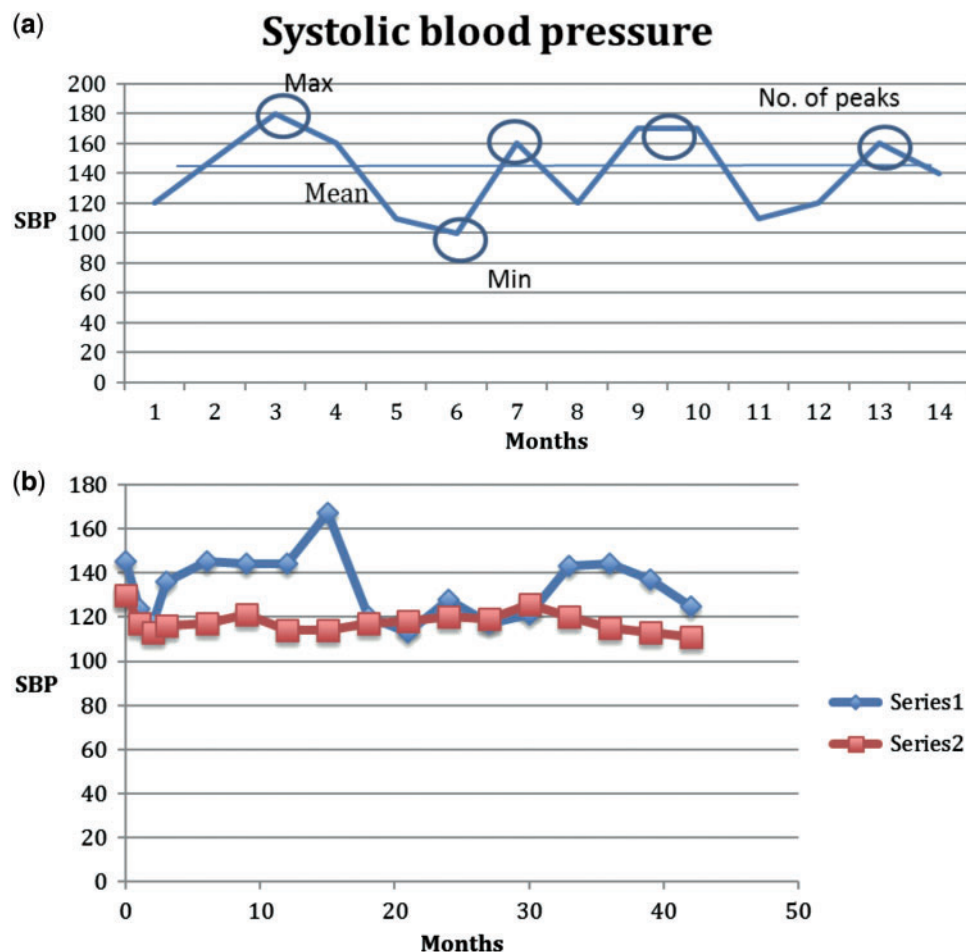


FIGURE 1: (a) Sample time-series data; (b) time series data for two patients.

descendant nodes contain similar values in terms of the study outcome. The formula for Gini impurity is as follows:

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i)$$

where n_c is the number of classes in the target feature and p_i is the ratio of this class. Adding up the decreases in Gini impurity for each individual variable over all trees in the forest gives an efficient measure of variable importance. Using this algorithm for minimizing node impurity for each feature allows automatic selection of important features to include in the model. This automatic feature selection precludes our having to manually choose features for modeling.

Evaluation

The data set was randomly divided into two sets—70% for training and 30% hold-out as the testing set. Using the 70% training set, an ensemble of random forests was created. For each random forest in the ensemble, the minority class (patients who reached the primary outcome) was repeatedly sampled so that there was equal number of samples in each class. Random forests were trained, each with their own subset, and their predictions were ensemble. Ten-fold cross-validation was performed on the training set to select hyperparameters of the model. Finally, the best hyperparameters were used on the entire

training set, and the final model was evaluated on the hold-out data in the testing set. Area under the curve (AUC) was measured for the final model.

RESULTS

Among 8799 patients included in the study, 365 patients (4.1%) reached the combined primary outcome over an average of 37 months of follow-up. Characteristics of patients who reached the endpoint relative to the rest of the cohort are shown in Table 1. These patients were older and had more smokers. In addition, there were more patients with history of cardiovascular disease (CVD) as well as chronic kidney disease (CKD). As expected, there were fewer patients on intensive management among patients who reached the primary outcome.

The best random forest classifier had an average AUC of 0.68 on the training data set, measured using 10-fold cross validation (95% confidence interval 0.57–0.79), as shown in Table 2. Using this model with 27 significant variables, the random forest classifier had an AUC of 0.71 on the test (hold-out) data set.

The 10 most significant features selected in order of importance by the automated algorithm included the urine albumin/creatinine (CR) ratio, estimated glomerular filtration rate, age, serum CR, history of subclinical CVD, total cholesterol, a variable representing time-series SBP signals using wavelet transformation, high-density lipoprotein (HDL), the 90th percentile SBP and triglyceride. The SPRINT treatment assignment labeled

Table 1. Characteristics of patients who reached and did not reach the primary outcome

Characteristics	Patients who reached primary outcome	Patients who did not reach primary outcome
Demographics		
Mean age, years (SD)	71 (10)	68 (9)
Female sex, n (%)	104/365 (28)	2943/8296 (35)
Race, n (%)		
White	245 (67)	4782 (58)
Black	91 (25)	2480 (30)
Hispanic	23 (6)	886 (11)
Others	6 (2)	148 (2)
Other patient characteristics		
Mean baseline SBP (mmHg)	141	140
Mean baseline DBP (mmHg)	76	78
Mean BMI	29	30
Smoking status, n (%)		
Never	134 (37)	3700 (45)
Former	173 (47)	3514 (42)
Current	58 (16)	1074 (13)
Subgroup with history of clinical/subclinical CVD, n (%)	128 (35)	1578 (19)
Subgroup with CKD (eGFR <60 mL/min/1.73 m ²), n (%)	152 (42)	2251 (27)
Medications		
Mean number of medications prescribed	2	2
Participants on no antihypertensive agents, n (%)	22 (6)	805 (10)
Aspirin use, n (%)	223 (61)	4199 (51)
Statin use, n (%)	125 (34)	3587 (43)
Mean laboratory parameters		
eGFR (mL/min/1.73 m ²)	66.5	72.2
Serum CR (mg/dL)	1.2	1.1
Total cholesterol (mg/dL)	185.8	190.2
Glucose (mg/dL)	99.7	98.8
HDL direct (mg/dL)	50.8	52.9
Triglycerides (mg/dL)	131.3	125.8
Urine albumin (mg/g CR)	96.3	38.1
Intervention		
Intensive, n (%)	152 (42)	4187 (50)

DBP, diastolic blood pressure; BMI, body mass index

Table 2. AUC measured using 10-fold cross-validation on the training data set and on the test data set for the best model

Data set	AUC
Training set (10-fold cross-validation)	
1	0.60
2	0.62
3	0.70
4	0.75
5	0.71
6	0.68
7	0.74
8	0.69
9	0.71
10	0.59
Mean (95% confidence interval)	0.68 (0.57–0.79)
Test set	0.71

‘intensive’ (a yes/no binary variable referring to the SBP goal) was significant albeit was nearly ranked the lowest (26th) among the 27 significant variables identified by the model. Additional features of the SBP time series data and other clinical variables were also included. A ranked list of all significant variables in the final model is shown in Figure 2.

Nine of the 27 significant variables (33%) that were automatically selected by the algorithm were related to SBP. The SBP feature labeled `SBP_cwt_coefficient_1` represents amplitude and duration of a particular SBP wave pattern that was reflected as a significant wavelet coefficient. In particular, four wavelet scales were used (2, 5, 10 and 20), and a coefficient for the wavelet with a scale of 2 was selected as a feature. Two additional SBP coefficients were also included in the model, representing other discernable wave patterns in the data. Furthermore, the 90th percentile of SBP values was ranked more significant, although the median and 60th to 80th percentiles as well as the last baseline SBP were also significant in the model.

DISCUSSION

This study successfully combined next generation automated machine-learning algorithms selected for optimal handling of time series data with low event outcomes to demonstrate that features of time series variables that are unaccounted for in conventional statistical models have prognostic significance. The identification of several significant features of time series of SBP also highlight the nonlinear nature of the random forest algorithm utilized herein. This study extends beyond central tendency and variability information previously used to

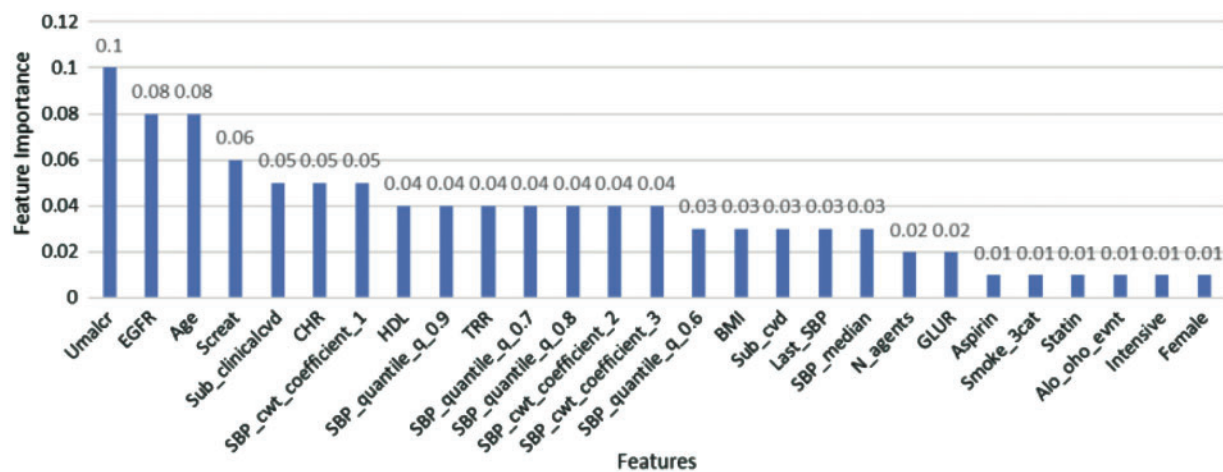


FIGURE 2: Feature importance for all features included in the model.

characterize long term fluctuations in SBP readings in studies reporting VVV. The constraints posed by linear models include collinearity, which leads to inflation of the variance as variables that overlap in distribution relative to each other and to the outcome is input into the same model [16]. We have overcome such constraints, allowing for a greater use of the different features of time series data that we believe will open the door to an entire spectrum of time series predictive analytics using clinical data usually available in electronic health records. The random forest method, which relies on multiple classification trees, uses nonparametric classification for recursively partitioning data per the value of a predictor variable until observations in a partition become increasingly homogenous. The random forest model is protected from highly collinear variables because even if two of the variables in a tree provide the same child node homogeneity, it simply selects one without affecting model quality. In other words, while some level of collinearity may be present, the random forest method ignores overlap between potentially collinear variables, leading to an overall increase in discrimination. This is a key advantage over generalized linear models in conventional statistics, which calculate a marginal contribution for each variable in the model.

Therefore, we are able to include several representations of the time trend of SBP data including transformations (e.g. Fourier and wavelet) [10, 17, 18] that have resulted in novel features of the SBP time series emerging as potentially useful predictors of outcome. The current study results are proof of concept for utilizing SBP because it was the only time series data available in SPRINT, but we believe that we are now able to use multiple variables available longitudinally in future analyses, which will further expand capabilities of clinical prognostic models and predictive analytics.

We utilized data from the SPRINT study, which was made publicly available for a SPRINT Data Challenge, as our vehicle to model and identify features that are associated with poor cardiovascular outcomes in hypertensive adults. A nonlinear classifier was utilized to model the primary outcome to deal with possible nonlinearity in predicting outcomes, as exemplified by the 'J-curve' often reported in hypertension outcome studies [19]. In addition, the random forest classifier has the capacity to account for data sparsity associated with the low number of patients who achieved the primary outcome. Several machine learning algorithms have been utilized to predict poor outcomes

in medicine [20–26], including for predicting mortality using blood pressure trends as features [10, 18].

To our knowledge, a unique advance of the current study which has not been reported in any previous clinical predictive study that incorporated time-series data with multiple time duration granularity (e.g. 1 month intervals, 3 month intervals) was: (i) models that simultaneously represented multiple time durations in continuous wavelet transformation, as well as (ii) variables that reflected multiple aspects of time-series data with features that included SBP values for specific deciles, sum of consecutive SBP changes and number of SBP values above mean. Indeed, automatic data transformation for time-series data has been used in other domains and has been shown to predict hemodynamic deterioration in Intensive Care Unit (ICU) patients, when used up to 2 h prior to deterioration [18]. In addition, it has been used for visual representation and analysis of time-series Electroencephalography (EEG) data [17, 27, 28]. Wavelets, in particular, have been used for signal de-noising and data compression algorithms [29–33]. However, there are no previous studies that have combined various transformations of time-series data for predicting clinical outcomes. Our method that incorporated these techniques into an algorithm fit into the random forest classifier achieved reasonable discrimination for patients who were observed to have poor cardiovascular outcomes on the hold-out data set. We believe that the model could have been optimized further if the data set had contained more time series predictor variables.

Several variables in the top 10 list sorted by model importance that were automatically selected by the feature importance classifier were known risk factors for poor outcomes—measures for kidney dysfunction, baseline cholesterol profiles, age and prior CVD. It automatically selected estimated glomerular filtration rate (eGFR) and serum CR as the 2nd and 4th most important variables—both measures of kidney dysfunction, which are known predictors of cardiovascular outcomes in hypertensive patients. Other variables included values for CWT coefficients of time-series SBP measurements and the SBP value above the 90th percentile for each patient. The last feature highlights the importance of individual patients' high SBP measurements, corresponding to the value above the 90th percentile, as a predictor of poor cardiovascular outcomes. The CWT coefficients reflect changes in the SBP signal over several time durations, explained further below.

CWT uses an analyzing function called wavelets to decompose time-series signals into coefficients at several different time durations, commonly referred to as scales. Wavelets are localized in time, and can thus be used to model localized changes in time-series signals. To illustrate, a smaller-scale wavelet corresponds to a more compressed wavelet, whereas a larger-scale wavelet corresponds to a more stretched wavelet (see Figure 3). These various wavelets can then be shifted in time to model signal changes. An acute change over a short time will thus be reflected as a larger coefficient in a smaller-scale wavelet. The specific CWT coefficient features in the model corresponding to coefficients at different scales and shifting of a wavelet. Thus, it is important to understand how transformation of time-series data can best inform decisions regarding patient management, especially when they play an important role in predicting poor outcomes.

A model able to reliably predict and identify patients at risk for poor cardiovascular outcomes can be utilized for prognostication, informed decision making, triage, adjusting case-mix,

projecting resource utilization and public policy. More importantly, understanding specific features that are associated with poor outcomes allow clinicians to intervene and personalize patient management, in the context of hypertension and CVD. The current approach fully incorporated different representations of the same clinical variable to extract the most highly predictive combination(s) to optimize the model for predicting poor cardiovascular outcomes. This model can include multiple representations of other clinical variables that are already collected in most Electronic Medical Records (EMRs). The model optimizes the predictive contribution from data we currently collect, potentially alleviating the current drive to keep collecting more clinical variables (a heavy administrative burden for overworked clinicians) to improve predictive models' discrimination ability.

The SPRINT investigators concluded that intensive blood pressure treatment achieved improved cardiovascular outcomes, compared with standard blood pressure management. In contrast, another recent trial from the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Study Group demonstrated that intensive blood pressure control in patients at high risk for CVD did not reduce the rate of a composite outcome of fatal and nonfatal major cardiovascular events [34, 35]. Several more studies raise concerns that blood pressure reduction below a certain threshold may pose dangers, the so-called 'J-curve', which affects groups of individuals who are older and with comorbidities [19, 36, 37]. All these highlight the importance of personalized management, considering individual risks and benefits in managing hypertension. Identifying which features portend poor cardiovascular outcomes is a critical step in individualized management. Finally, incorporating changes in management and adherence as part of the time series data in the model may further provide insight into individualized optimal therapy.

Limitations

This work highlights the utility of modeling multiple time duration granularities. This study, however, did not consider a longer time forecasting horizon and was limited to 1 year from the beginning of SPRINT. Moreover, using time series transformation for other time-series data (e.g. laboratory values) was not applied in conjunction with SBP measurements because those data were not available for analysis. Changes in kidney function, some of the features deemed important in modeling outcomes, might be useful in further increasing the model's accuracy. Other serial/episodic clinical patient data would also be valuable. Finally, prospective validation of the model in a separate data set would be prudent to further assess the model's generalizability.

CONCLUSION

Features that predict poor cardiovascular outcomes were identified using an automated random forest algorithm. In addition to known risk factors for CVD, transformed variables for time series SBP measurements were found to be important in predicting poor cardiovascular outcomes and require further evaluation. Serial measurements of more clinical variables in addition to blood pressure, such as medication use, body weight, laboratory results and other relevant factors in the electronic health record, may now be studied concurrently utilizing this novel algorithm utilizing next generation automated machine-learning and signal-processing algorithms.

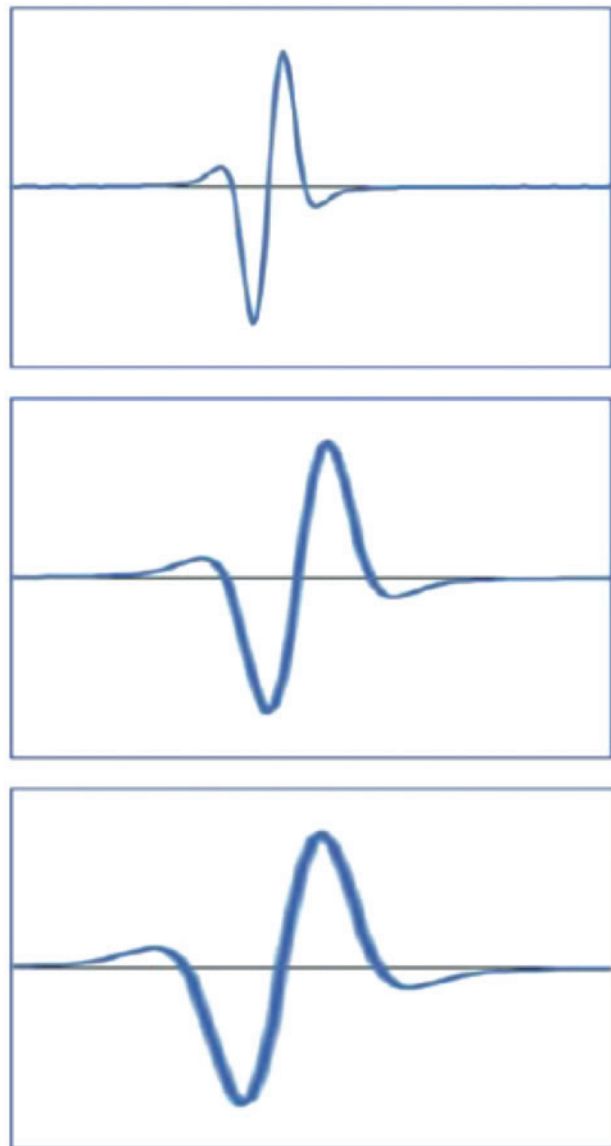


FIGURE 3: Ricker wavelets at increasing scales.

SUPPLEMENTARY DATA

Supplementary data are available at ckj online

CONFLICT OF INTEREST STATEMENT

None declared. The results presented in this paper have not been published previously, except in abstract format.

REFERENCES

- Chobanian AV, Bakris GL, Black HR et al. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. *JAMA* 2003; 289: 2560–2572
- Kannel WB, Castelli WP, McNamara PM et al. Role of blood pressure in the development of congestive heart failure. The Framingham study. *N Engl J Med* 1972; 287: 781–787
- Kannel WB, Sorlie P, Castelli WP et al. Blood pressure and survival after myocardial infarction: the Framingham study. *Am J Cardiol* 1980; 45: 326–330
- Kannel WB, Wolf PA, McGee DL et al. Systolic blood pressure, arterial rigidity, and risk of stroke. The Framingham study. *JAMA* 1981; 245: 1225–1229
- Group SR, Wright JT Jr, Williamson JD et al. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med* 2015; 373: 2103–2016
- Drazen JM, Morrissey S, Campion EW et al. A SPRINT to the Finish. *N Engl J Med* 2015; 373: 2174–2175
- Stevens SL, Wood S, Koshiaris C et al. Blood pressure variability and cardiovascular disease: systematic review and meta-analysis. *BMJ* 2016; 354: i4098
- Harden RM, Stevenson M, Downie WW et al. Assessment of clinical competence using objective structured examination. *BMJ* 1975; 1: 447–451
- Ohkuma T, Woodward M, Jun M et al. Prognostic value of variability in systolic blood pressure related to vascular events and premature death in type 2 diabetes mellitus: the ADVANCE-ON study. *Hypertension* 2017; 70: 461–468
- Lacson R. Predicting hemodialysis mortality utilizing blood pressure trends. *AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium*, Washington, DC, November 8–12, 2008: 369–373
- Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32
- Scheurwegs E, Sushil M, Tulkens S et al. Counting trees in random forests: predicting symptom severity in psychiatric intake reports. *J Biomed Inform* 2017; 75S: S112–S119
- Christ M. Tsfresh. Maximilian Christ, Blue Yonder GmbH 2016. <https://github.com/blue-yonder/tsfresh> (4 June 2018, date last accessed)
- Bellu L, Liberati P. Inequality Analysis: The Gini Index. In: Food and Agriculture Organization of the United Nations, editor. http://www.fao.org/docs/up/easypol/329/gini_index_040en.pdf (4 June 2018, date last accessed)
- Yentes JM, Hunt N, Schmid KK et al. The appropriate use of approximate entropy and sample entropy with short data sets. *Ann Biomed Eng* 2013; 41: 349–365
- Nathanson BH, Higgins TL. An introduction to statistical methods used in binary outcome modeling. *Semin Cardiothorac Vasc Anesth* 2008; 12: 153–166
- Quiroga RQ, Schurmann M. Functions and sources of event-related EEG alpha oscillations studied with the Wavelet Transform. *Clin Neurophysiol* 1999; 110: 643–654
- Saeed M, Mark RG. Efficient hemodynamic event detection utilizing relational databases and wavelet analysis. *Comput Cardiol* 2001; 28: 153–156
- Rahman F, McEvoy JW. The J-shaped curve for blood pressure and cardiovascular disease risk: historical context and recent updates. *Curr Atheroscler Rep* 2017; 19: 34
- Firoozbakht F, Rezaeian I, D'Agnillo M et al. An integrative approach for identifying network biomarkers of breast cancer subtypes using genomic, interactomic, and transcriptomic data. *J Comput Biol* 2017; 24: 756–766
- Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N Engl J Med* 2017; 376: 2507–2509
- Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med* 2001; 23: 89–109
- Tan L, Holland SK, Deshpande AK et al. A semi-supervised Support Vector Machine model for predicting the language outcomes following cochlear implantation based on pre-implant brain fMRI imaging. *Brain Behav* 2015; 5: e00391
- Ghassemi M, Pimentel MAF, Naumann T et al. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. *Proc Conf AAAI Artif Intell* 2015; 2015: 446–453
- Liu NT, Salinas J. Machine learning for predicting outcomes in trauma. *Shock* 2017; 48: 504–510
- Krittanawong C, Zhang H, Wang Z et al. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol* 2017; 69: 2657–2664
- Djemal R, AlSharabi K, Ibrahim S et al. EEG-based computer aided diagnosis of autism spectrum disorder using wavelet, entropy, and ANN. *Biomed Res Int* 2017; 2017: 9816591
- Jrad N, Kachenoura A, Merlet I et al. Classification of high frequency oscillations in epileptic intracerebral EEG. *Conf Proc IEEE Eng Med Biol Soc* 2015; 2015: 574–577
- Jobert M, Tismer C, Poiseau E et al. Wavelets-a new tool in sleep biosignal analysis. *J Sleep Res* 1994; 3: 223–232
- Kotter E, Roesner A, Torsten Winterer J et al. Evaluation of lossy data compression of chest X-rays: a receiver operating characteristic study. *Invest Radiol* 2003; 38: 243–249
- Zeng L, Jansen CP, Marsch S et al. Four-dimensional wavelet compression of arbitrarily sized echocardiographic data. *IEEE Trans Med Imaging* 2002; 21: 1179–1187
- Noubari HA, Fayazi A, Babapour F. De-noising of SPECT images via optimal thresholding by wavelets. *Conf Proc IEEE Eng Med Biol Soc* 2009; 2009: 352–355
- Mo F, Mo Q, Chen Y et al. WaveletQuant, an improved quantification software based on wavelet signal threshold denoising for labeled quantitative proteomic analysis. *BMC Bioinformatics* 2010; 11: 219
- Mancia G. Effects of intensive blood pressure control in the management of patients with type 2 diabetes mellitus in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial. *Circulation* 2010; 122: 847–849
- Cushman WC, Evans GW, Byington RP et al. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *N Engl J Med* 2010; 362: 1575–1585
- Frontoni S, Solini A, Fioretto P et al. The ideal blood pressure target to prevent cardiovascular disease in type 2 diabetes: a neutral viewpoint. *Nutr Metab Cardiovasc Dis* 2014; 24: 577–584
- Li Z, Lacson E, Lowrie EG et al. The epidemiology of systolic blood pressure and death risk in hemodialysis patients. *Am J Kidney Dis* 2006; 48: 606–615