Published in final edited form as:

J Vasc Surg. 2016 November ; 64(5): 1515–1522.e3. doi:10.1016/j.jvs.2016.04.026.

The use of machine learning for the identification of peripheral artery disease and future mortality risk

Elsie Gyang Ross, MD, MSc[1,*], **Nigam Shah, MBBS, PhD**[2,*], **Ronald Dalman, MD**[1], **Kevin Nead, MD, MPhil**[3], **John Cooke, MD, PhD**[4,5], and **Nicholas J. Leeper, MD**[1]

[1]Division of Vascular Surgery, Stanford Health Care, Stanford, CA, USA

[2]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA

[3]University of Pennsylvania, Philadelphia, PA, USA

[4]Department of Cardiovascular Sciences, Houston Methodist Research Institute, Houston, Texas, USA

[5]Center for Cardiovascular Regeneration, Houston Methodist DeBakey Heart and Vascular Center, Houston, Texas, USA

Abstract

Objective—A key aspect of the precision medicine effort is the development of informatics tools that can analyze and interpret 'big data' sets in an automated and adaptive fashion, while providing accurate and actionable clinical information. The aims of this study were to develop machine learning algorithms for the identification of disease and the prognostication of mortality risk, and to determine whether such models perform better than classical statistical analyses.

Methods—Focusing on peripheral artery disease (PAD), patient data were derived from a prospective, observational study of 1,755 patients who presented for elective coronary angiography. We employed multiple supervised machine learning algorithms and utilized diverse clinical, demographic, imaging and genomic information in a hypothesis-free manner to build models that could identify patients with PAD and predict future mortality. Comparison was made to standard stepwise linear regression models.

Results—Our machine-learned models outperformed stepwise logistic regression models both for the identification of patients with PAD (AUC 0.87 versus 0.76, respectively, P=0.03), and predicting future mortality (AUC 0.76 versus 0.65, respectively, P=0.10). Both machine-learned models were markedly better calibrated than the stepwise logistic regression models, thus providing more accurate disease and mortality risk estimates.

Corresponding Author: Nicholas Leeper, MD, Division of Vascular Surgery, 300 Pasteur Drive, Suite H3638, Stanford, CA 94305, Phone: (650)-723-3639, Fax: (650)-498-6044, nleeper@stanford.edu.

[*]Authors contributed equally

Conflict of interest: None of the authors involved in this work have a conflict of interest to report.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusions—Machine learning approaches can produce more accurate disease classification and prediction models. These tools may prove clinically useful for the automated identification of patients with highly morbid diseases for which aggressive risk factor management can improve outcomes.

Introduction

The promise of precision medicine is beginning to be realized in some areas of medicine. In oncology for example, genetic profiling is now being used to identify patients for whom tailored chemotherapy regimens - directed against the individual's personal cancer mutation - can be used to significantly improve outcomes relative to traditional therapy[1]. Rather than the current empiric approach to treatment, there is hope that with a deeper understanding of biology and pharmacogenomics that we may one day be able to guarantee that every patient receives the right dose of the right medicine at the right time[2].

In addition to efforts directed towards personalizing therapy, the precision medicine effort is also focused on refining our diagnostic and predictive capabilities. Indeed, it is becoming increasingly clear that traditional risk prediction models (commonly based on a handful of epidemiological factors) fail to capture important information about the individual's behavior, environment, comorbidities and personal biological makeup. Now, with an influx of detailed patient data residing in large, prospective datasets[3, 4] and the rapid adoption of electronic health records (EHRs), there is an increasing pool of information that can potentially be analyzed to refine our predictive capabilities to deliver better care[30].

Here, we evaluated the utility of using machine learning algorithms and their ability to integrate disparate data inputs to make predictions about presence or absence of disease and the risk of future events. Specifically we built several models to identify peripheral artery disease, a highly prevalent[5] and morbid disease[6] that frequently goes undiagnosed[7, 8], and predict risk of future mortality. We compared our models with standard logistic regression models to evaluate whether there were significant improvements in predictive ability from using modern machine learning techniques.

Methods

Study Population

Patient data were derived from the Genetic Determinants of Peripheral Artery Disease (GenePAD) study, a prospective observational study funded by the National Heart, Lung and Blood Institute with the goal of identifying key demographic, clinical and genomic factors that differentiate patients with PAD from those without disease[9, 10]. While PAD was the disease of interest in this observational study, it was not used as a factor in enrolling patients. Thus, we used this data set due to the depth and breadth of patient variables collected, its longitudinal nature and the potential impact on improving care for this subset of patients.

Data from the study includes a cohort of 1,755 patients who were enrolled at presentation to Stanford University Medical Center or Mount Sinai Medical Center for elective coronary angiography between January 2004 and March 2008. Patients included those referred for complaints of angina, dyspnea or had abnormal stress test results. Clinical data were collected at the time of enrollment and patients were prospectively followed for observation of any adverse events. PAD status was not known prior to enrollment nor was it used as inclusion or exclusion criteria. Patients were followed until the study completion in 2012.

Trained research assistants and nurse practitioners conducted extensive participant interviews and performed careful review of patient history. Variables collected included demographic variables, clinical comorbidities, medications, lab tests, physical exam variables, physical activity and smoking behaviors, socioeconomic variables, selected genomic markers associated with PAD as well as results of coronary angiograms (Table I). Ankle-brachial indices (ABIs) were measured for all patients and PAD was defined as an ABI of < 0.9. All patients provided written consent to participate in the GenePAD study, which was approved by the Stanford and Mount Sinai Institutional Review Boards.

Machine Learning

Machine learning refers to methods developed within the fields of statistics, computer science, and artificial intelligence that allow the creation of algorithms that can learn from and make predictions using data. Some commonly used algorithms, which we utilized in this study, include *elastic net*[11] and *random forest*[12].

Elastic net is a linear modeling technique similar to linear or logistic regression. The advantage of an elastic net is that in addition to fitting an optimized linear model, a penalty is applied to independent variables in the model such that variables that have little influence on the dependent variable are minimized or dropped from the final model. This has the effect of reducing model complexity while improving the generalizability of the model, which can improve predictive accuracy. Random forest is a “tree-based” algorithm whereby multiple decision trees are built using a random assortment of independent variables, which are used to predict an outcome label for a random subset of samples. Using a “majority vote” system, a new sample is predicted by the multiple decision trees in the random forest model and the ultimate classification of this new sample is based on the classification predicted by a majority of the decision trees.

Model Building

The goals of our predictive models were to predict which patients had PAD based on baseline demographic, clinical, and genomic factors and predict their future risk of mortality. In building our machine learning algorithms, we used a “hypothesis-free” approach to identifying which variables (e.g. clinical, demographic, or social) should be included in our models and elected to include any variable for which the majority of patients had a data value. That is, we did not include variables based on any *a priori* hypotheses of their ability to predict disease presence or absence or risk of future mortality. Variables were simply included if they were available. We then included patients with complete data across variables included in the model, then randomly splitting them into a 70% training set and a 30% test set.

For model training, data from 70% of patients were used to approximate model parameters. We used 10-fold cross-validation for training our model to decrease risk of model over-fitting[13]. Multiple machine learning algorithms were utilized to build predictive models including Elastic Net, a penalized regression model, and Random Forest using R version 3.2.1[14].

Model Performance

Each model's ability to discriminate between high and low risk patients was determined using the area under the receiver operator curve (AUC-ROC) metric[15]. We also evaluated model calibration (i.e. the model's ability to accurately predict observed absolute risk) using the Hosmer-Lemeshow test for goodness of fit, where a P-value < 0.05 would indicate poor calibration[16]. From our learned models we selected the model with the best discrimination and calibration. AUC and calibration are based on model performance on the test set of patients.

Identifying PAD

The ultimate goal of our learned model was to establish an accurate classification algorithm that could identify patients with PAD. To this end, cases of PAD were defined as patients with an ABI < 0.9 identified at the time of the GenePAD study, whether or not they had a diagnosis prior to enrollment. Controls were defined as patients with ABI > 0.9. We also had a secondary aim of evaluating whether we could build a model that could identify patients with undiagnosed PAD, given how frequently PAD goes undiagnosed. In the GenePAD data set for instance, of patients identified as having PAD by ABI at the time of enrollment, 68% had no prior diagnosis[8]. Thus for our "undiagnosed PAD" model we included patients with an ABI < 0.9 and excluded those who had a prior PAD diagnosis. Table I lists variables included in this model.

Predicting Mortality

The goal of this predictive model was to most accurately predict all-cause mortality by the end of the GenePAD study, which lasted a total of 8 years. Patient follow-up consisted of regular reviews of patient electronic health records and calls to patients and/or their families for regular updates. All patients included in our PAD prediction model were also included in the mortality prediction model.

Model comparison

To compare whether our models provided better predictive accuracy than standard methodology such as logistic regression, we also built stepwise logistic regression models for identification of PAD patients and prediction of mortality. Using the patient training set we used univariate analysis to identify which variables were significantly associated with presence of PAD or mortality using Chi-Square test for categorical variables and analysis of variance for continuous variables. We then performed forward stepwise logistic regression, including variables with a significance level of at least 10%. The final stepwise logistic regression models were selected based on optimizing the Akaike Information Content. These final models were then applied to the test patient set and AUC and calibration were calculated accordingly. To compare differences in discriminatory performance of the machine learned models to the stepwise logistic regression models, we used the bootstrap test for unpaired ROC curves[17] using 2000 bootstraped samples. P-values < 0.05 were considered significant.

Variable importance

An important aspect of model building is identifying which variables may be of greatest importance in contributing to the accuracy of a model. We thus identified the most heavily weighted variables for each machine learned model and compared differences in algorithm discrimination performance with and without these variables using integrated discrimination improvement index (IDI)[15] and compared calibration performance using the category-free net reclassification improvement index (NRI)[18], which assessed whether patient risk reclassification was appropriate (i.e. higher risk assignment for cases and lower risk assignment for non-cases).

Results

Patient characteristics for those included in our predictive models are described in Table II. After paring down variables that were incomplete or redundant, we were left with 130 different variables that included genomic markers, physical activity assessments, demographics, clinical assessments, socioeconomic status, and family history. A total of 1,047 patients had complete data across these domains and were included in our PAD and mortality prediction models. For the PAD model this included 183 patients with PAD and 864 controls. For mortality prediction there were 129 patients with a mortality event and 918 controls. Median follow-up time for included patients was 5.3 years (interquartile range 4.3–6.2 years).

Identifying PAD

Balancing model accuracy and calibration, the best model that could identify patients with PAD was a penalized linear regression model, which achieved an AUC of 0.87 (95% Confidence Interval (CI), 0.81–0.92) (Figure 1, Supplemental Table 1). The stepwise linear regression model did not perform as well (AUC 0.76, 95% CI, 0.68–0.85), P = 0.03. The penalized regression model also demonstrated good calibration while the stepwise linear regression model had very poor calibration performance using Hosmer-Lemeshow goodness of fit tests (P=0.7, P < 0.0001, respectively) (Supplemental Figure 1, 2). Figure 2 illustrates the most important variables used in the penalized regression model. Removal of these variables from our penalized regression model resulted in significant changes in net reclassification and discriminatory power of the model (Net reclassification index (NRI) = −0.52, 95% CI, −0.23 to −0.80, P = 0.0003; Integrated Discrimination Improvement (IDI) = −0.05, 95% CI, −0.08 to −0.03), P < 0.0001).

Identifying Undiagnosed PAD

Given how frequently PAD goes undiagnosed, we were also interested in building a predictive model that could potentially identify patients with PAD who did not yet carry a diagnosis. Our training set included a total sample of 993 patients, of which 103 were found to have an ABI < 0.9 and had no prior diagnosis of PAD. Our best performing classification model for undiagnosed PAD was a random forest model, which achieved an AUC of 0.84

(95% Confidence Interval (CI), 0.77–0.91) (Supplemental Figure 3). The classical stepwise logistic regression model performed particularly poorly in attempting to identify cases of undiagnosed disease in comparison (AUC 0.60, CI 0.5–0.70), P = 0.0001. The random forest model also demonstrated good calibration while the linear regression model had poor calibration performance using Hosmer-Lemeshow goodness of fit tests (P = 0.06, P < 0.0001, respectively). The most important variables used in the identification of patients with undiagnosed PAD used in the random forest model are illustrated in Supplemental Figure 4.

Predicting Mortality

The best learned model for predicting mortality was a random forest model. The model AUC is 0.76 (95% CI, 0.68–0.84) (Figure 3) while the stepwise logistic regression model had poorer performance by AUC (0.65, 95% CI 0.61–0.78) (Supplemental Table II). This difference approached significance, P = 0.10. The learned model demonstrated much better calibration than the stepwise logistic regression model (P = 0.62, P < 0.0001, respectively) (Supplemental Figures 5, 6). Figure 4 illustrates the most important variables used in the random forest model. Removing these variables from the predictive model significantly reduced model discrimination (IDI −0.07, 95% CI, −0.13 to −0.01, P = 0.01), but did not significantly affect NRI (−0.28, 95% CI, −0.61 to 0.05, P = 0.1).

Discussion

We have described an approach leveraging machine learning to created predictive models to identify patients with peripheral artery disease and predict mortality in a high-risk population for which good “off the shelf” risk prediction models are lacking. Our predictive models, trained in a “hypothesis-free” fashion on a variety of patient data including genomic, imaging, and socioeconomic variables outperform standard logistic regression models. Ultimately, such tools may enable the automated detection of individuals with vascular disease and allow for personalized, preventative interventions prior to development of end-stage atherothrombosis.

Generally, identification of “at risk” patients and subsequent risk factor management are integral parts of preventative medical care. Risk stratification methods such as risk prediction scores that are highly utilized in medicine are often derived from epidemiologic studies and typically describe a handful of risk factors that predict future disease risk. Such risk scores can be of great clinical utility as they help clinicians identify patients for which further screening or intervention may change health trajectory [19–21]. There are limitations to such classical risk prediction scores, based exclusively on linear modeling, however. In cardiology, for example, the Framingham Risk Score (FRS) is widely used to identify patients at risk of future major adverse cardiovascular events (MACE). However, the FRS often fails to adequately risk stratify patients in more ethnically, socially and medically diverse populations [22–24]. Investigators have attempted to address these limitations over the last decade by providing “improved” risk scores that incorporate either new lab, biomarker or imaging data [25–27], however, such model adjustments typically have been incremental and confer relatively small or negligible improvements in risk prediction performance [28]. In the specific case of peripheral vascular disease, Duvall and colleagues have developed a PAD risk score based on a small number of factors derived from population studies[29], however the performance of this tool has been modest.

Our PAD prediction score had excellent discrimination and calibration performance (AUC 0.87, Hosmer-Lemeshow P value = 0.7) and significantly outperformed a standard stepwise logistic regression model, especially when attempting to identify patients with undiagnosed PAD. While model AUC is an oft-cited metric, calibration is probably a more important aspect of model performance. In addition to providing more accurate classification (i.e. disease present or not present), well-calibrated models can also provide meaningful risk estimates for the classification task at hand. For instance, with a well-calibrated PAD model, physicians can be provided with a reliable estimate of the probability of PAD in a patient that might otherwise be missed, and potentially prioritize subjects for further screening.

In the current study, our models were generated from information captured in a relatively small clinical database. However, we are entering an era in which patient data is becoming ever more ubiquitous, and we will soon be able to apply predictive analytics to exponentially larger datasets which should generate even more meaningful and actionable insights[30]. For example, machine learning algorithms are now being designed to continuously surveil the electronic medical record in an automated fashion – *providing real-time predictive analytics for patient care.* Indeed, it is increasingly possible to extract even more nuanced patient data from the EHR with rule-based and natural language processing techniques[31, 32], and to use these data to accurately predict patient outcomes across a broad range of conditions[33–35]. Using newer analytic approaches, these models will not be static and will one day learn to better predict an individual's clinical trajectory over time[36, 37]. Further, such models can be "re-trained" locally at different institutions to maximize accuracy in different patient populations with different clinical and demographic profiles[38].

We chose to develop algorithms for the detection of PAD for several reasons. First, PAD is a disease that is both highly prevalent and notoriously difficult to diagnose, with over 50% of patients in general medical practice going undiagnosed in the U.S.[39]. Indeed, several studies have shown that PAD frequently goes unidentified, indicating that currently available diagnostic algorithms and screening efforts are insufficient[7, 8, 40]. Second, PAD is a highly morbid condition known to have worse outcomes than coronary or cerebrovascular disease, but is undertreated compared to other conditions[39, 40]. Third, underlying PAD increases the risk of complications from procedures for which this patient population frequently undergo, including percutaneous coronary intervention (PCI). PAD patients undergoing PCI have higher risk of vascular access failure, groin and retroperitoneal hematomas, limb ischemia and need for blood transfusions [41, 42]. Knowledge of a PAD diagnosis before these procedures can help reduce complication rates. Fourth, because intervention for PAD can preserve life and limb[43, 44], it could be useful to have an automated algorithm that can both identify latent disease, and single out those individuals at higher risk for adverse outcomes. Lastly, while there are currently no guidelines recommending routine ABI assessment, predictive algorithms have the potential benefit of providing a cost-effective way of identifying high-risk patients for which screening and early intervention can significantly impact disease trajectory and decrease health system costs[45].

A limitation of our study is that we only used patients who had complete data to build our models. In clinical practice patient data is frequently missing, which may reduce predictive accuracy. One way to address this limitation involves the imputation of missing data, which we found can actually improve predictive accuracy (data not shown). Such methods can be used for real-world data, balancing the limitations of different imputation techniques. Another limitation is that our models may not generalize well to other populations given that they were trained on data from high-risk patients enrolled at the time of coronary angiography. Though patients within the GenePAD database were not selectively enrolled for their likelihood of having PAD, there may be inherent bias in our data set. As previously mentioned, however, re-training machine learning algorithms for specific populations is recommended and can greatly improve predictive accuracy. Furthermore, our mortality prediction model had an AUC of 0.76, lower than our PAD classification model of 0.87. Even so, compared to the stepwise regression model for mortality risk, our machine-learned model performed 11 points higher.

Conclusion

Machine learning algorithms can produce accurate disease classification and prediction models, which outperform standard logistic regression models. Such methodology can be employed as potentially more accurate and easily automated ways of identifying at risk patients for conditions where good risk prediction methods do not exist or are found to have relatively poor performance. Future studies should attempt to test, automate and prospectively validate local models, with the aim of reducing the prevalence of undiagnosed diseases and the burden of adverse clinical outcomes related to delays in preventative interventions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This work was supported by 1T32HL098049-06 (EGR) and 5R01HL12522402 (NJL).

References

1. Garraway LA, Verweij J, Ballman KV. Precision oncology: An overview. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013; 31:1803–1805. [PubMed: 23589545]
2. Bielinski SJ, Olson JE, Pathak J, Weinshilboum RM, Wang L, Lyke KJ, et al. Preemptive genotyping for personalized medicine: Design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. Mayo Clinic proceedings. 2014; 89:25–33. [PubMed: 24388019]
3. Ingelsson EG. A. Ubble - uk longevity explorer. 2015
4. Ganna A, Ingelsson E. 5 year mortality predictors in 498 103 uk biobank participants: A prospective population-based study. The Lancet. 2015; 386:533–540.
5. Fowkes FG, Rudan D, Rudan I, Aboyans V, Denenberg JO, McDermott MM, et al. Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: A systematic review and analysis. Lancet. 2013; 382:1329–1340. [PubMed: 23915883]

6. Alberts MJ, Bhatt DL, Mas JL, Ohman EM, Hirsch AT, Röther J, et al. Three-year follow-up and event rates in the international reduction of atherothrombosis for continued health registry. European Heart Journal. 2009; 30:2318–2326. [PubMed: 19720633]
7. Hirsch AT, Criqui MH, Treat-Jacobson D, Regensteiner JG, Creager MA, Olin JW, et al. Peripheral arterial disease detection, awareness, and treatment in primary care. Jama. 2001; 286:1317–1324. [PubMed: 11560536]
8. Chang P, Nead KT, Olin JW, Cooke JP, Leeper NJ. Clinical and socioeconomic factors associated with unrecognized peripheral artery disease. Vascular medicine (London, England). 2014; 19:289–296.
9. Sadrzadeh Rafie AH, Stefanick ML, Sims ST, Phan T, Higgins M, Gabriel A, Assimes T, et al. Sex differences in the prevalence of peripheral artery disease in patients undergoing coronary catheterization. Vascular medicine (London, England). 2010; 15:443–450.
10. Nead KT, Cooke JP, Olin JW, Leeper NJ. Alternative ankle-brachial index method identifies additional at-risk individuals. J Am Coll Cardiol. 2013; 62:553–559. [PubMed: 23707317]
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67:301–320.
12. Breiman L. Random forests. Machine Learning. 2001; 45:5–32.
13. Kuhn, M.; Johnson, K. Applied predictive modeling. New York: Springer; 2013.
14. Team RC. R: A language and environment for statistical computing. 2015.
15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for some traditional and novel measures. Epidemiology (Cambridge, Mass). 2010; 21:128–138.
16. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods. 1980; 9:1043–1069.
17. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. Proc: An open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics. 2011; 12:77. [PubMed: 21414208]
18. Pencina MJ, Steyerberg EW, D’Agostino RB. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in medicine. 2011; 30:11–21. [PubMed: 21204120]
19. Lindstrom J, Tuomilehto J. The diabetes risk score: A practical tool to predict type 2 diabetes risk. Diabetes care. 2003; 26:725–731. [PubMed: 12610029]
20. Wilson PW, D’Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837–1847. [PubMed: 9603539]
21. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute. 1989; 81:1879–1886. [PubMed: 2593165]
22. Ramsay SE, Morris RW, Whincup PH, Papacosta AO, Thomas MC, Wannamethee SG. Prediction of coronary heart disease risk by framingham and score risk assessments varies by socioeconomic position: Results from a study in british men. European journal of cardiovascular prevention and rehabilitation : official journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology. 2011; 18:186–193.
23. Arts EE, Popa C, Den Broeder AA, Semb AG, Toms T, Kitas GD, et al. Performance of four current risk algorithms in predicting cardiovascular events in patients with early rheumatoid arthritis. Annals of the rheumatic diseases. 2015; 74:668–674. [PubMed: 24389293]
24. Tillin T, Hughes AD, Whincup P, Mayet J, Sattar N, McKeigue PM, et al. Ethnicity and prediction of cardiovascular disease: Performance of qrisk2 and framingham scores in a u.K. Tri-ethnic prospective cohort study (sabre--southall and brent revisited). Heart (British Cardiac Society). 2014; 100:60–67. [PubMed: 24186564]
25. Murphy TP, Dhangana R, Pencina MJ, D’Agostino RB Sr. Ankle-brachial index and cardiovascular risk prediction: An analysis of 11,594 individuals with 10-year follow-up. Atherosclerosis. 2012; 220:160–167. [PubMed: 22099055]

26. Liabeuf S, Desjardins L, Diouf M, Temmar M, Renard C, Choukroun G, et al. The addition of vascular calcification scores to traditional risk factors improves cardiovascular risk assessment in patients with chronic kidney disease. PloS one. 2015; 10:e0131707. [PubMed: 26181592]

27. Kadowaki S, Shishido T, Honda Y, Narumi T, Otaki Y, Kinoshita D, et al. Additive clinical value of serum brain-derived neurotrophic factor for prediction of chronic heart failure outcome. Heart and vessels. 2015

28. Wang TJ. Assessing the role of circulating, genetic, and imaging biomarkers in cardiovascular risk prediction. Circulation. 2011; 123:551–565. [PubMed: 21300963]

29. Duval S, Massaro JM, Jaff MR, Boden WE, Alberts MJ, Califf RM, et al. An evidence-based score to detect prevalent peripheral artery disease (pad). Vascular medicine (London, England). 2012; 17:342–351.

30. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: The dawn of precision delivery. Jama. 2016; 315:651–652. [PubMed: 26881365]

31. Jonnagaddala J, Liaw ST, Ray P, Kumar M, Chang NW, Dai HJ. Coronary artery disease risk assessment from unstructured electronic health records using text mining. Journal of biomedical informatics. 2015

32. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: Profiling the safety of cilostazol by text-mining of clinical notes. PloS one. 2013; 8:e63499. [PubMed: 23717437]

33. Finlay GD, Rothman MJ, Smith RA. Measuring the modified early warning score and the rothman index: Advantages of utilizing the electronic medical record in an early warning system. Journal of hospital medicine. 2014; 9:116–119. [PubMed: 24357519]

34. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (trewscore) for septic shock. Science Translational Medicine. 2015; 7:299ra122–299ra122.

35. Jung K, Sheppard D, Covington S, Chandan SK, Januszyk M, Kirsner R, et al. Rapid identification of slow healing wounds. Wound Repair and Regeneration. In Press.

36. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D. Machine learning for personalized medicine: Predicting primary myocardial infarction from electronic health records. AI Magazine. 2012; 33:33.

37. Jung KS, Nigam. Implications of non-stationarity on predictive modeling using ehrs. Journal of biomedical informatics. In Press.

38. Celi LAG, Tang RJ, Villarroel MC, Davidzon GA, Lester WT, Chueh HC. A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury. 2011; 2:97–110.

39. Criqui MH, Aboyans V. Epidemiology of peripheral artery disease. Circulation research. 2015; 116:1509–1526. [PubMed: 25908725]

40. McDermott MM, Kerwin DR, Liu K, Martin GJ, O'Brien E, Kaplan H, et al. Prevalence and significance of unrecognized lower extremity peripheral arterial disease in general medicine practice*. Journal of general internal medicine. 2001; 16:384–390. [PubMed: 11422635]

41. Nikolsky E, Mehran R, Mintz GS, Dangas GD, Lansky AJ, Aymong ED, et al. Impact of symptomatic peripheral arterial disease on 1-year mortality in patients undergoing percutaneous coronary interventions. Journal of endovascular therapy : an official journal of the International Society of Endovascular Specialists. 2004; 11:60–70. [PubMed: 14748627]

42. Hildick-Smith DJ, Walsh JT, Lowe MD, Stone DL, Schofield PM, Shapiro LM, et al. Coronary angiography in the presence of peripheral vascular disease: Femoral or brachial/radial approach? Catheterization and cardiovascular interventions : official journal of the Society for Cardiac Angiography & Interventions. 2000; 49:32–37. [PubMed: 10627362]

43. Olin JW, Sealove BA. Peripheral artery disease: Current insight into the disease and its diagnosis and management. Mayo Clinic proceedings. 2010; 85:678–692. [PubMed: 20592174]

44. Bonaca MP, Creager MA. Pharmacological treatment and current management of peripheral artery disease. Circulation research. 2015; 116:1579–1598. [PubMed: 25908730]

45. Kent KC, Zwolak RM, Egorova NN, Riles TS, Manganaro A, Moskowitz AJ, et al. Analysis of risk factors for abdominal aortic aneurysm in a cohort of more than 3 million individuals. J Vasc Surg. 2010; 52:539–548. [PubMed: 20630687]

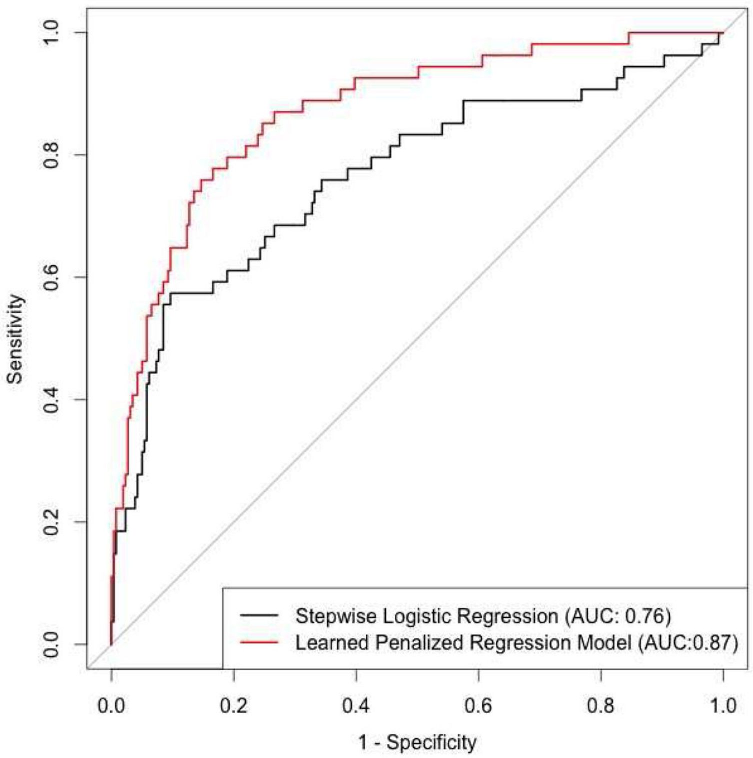

Figure 1.

Area Under the Curve for stepwise logistic regression and machine learned penalized regression model for identification of PAD patients.

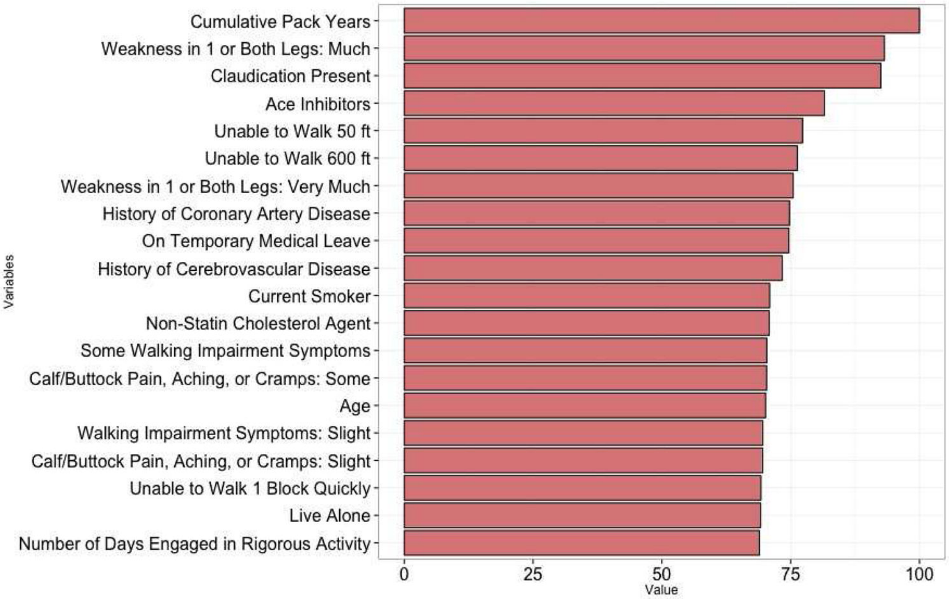

Figure 2.

Top 20 weighted variables used in the machine learned penalized regression model for identification of PAD patients.

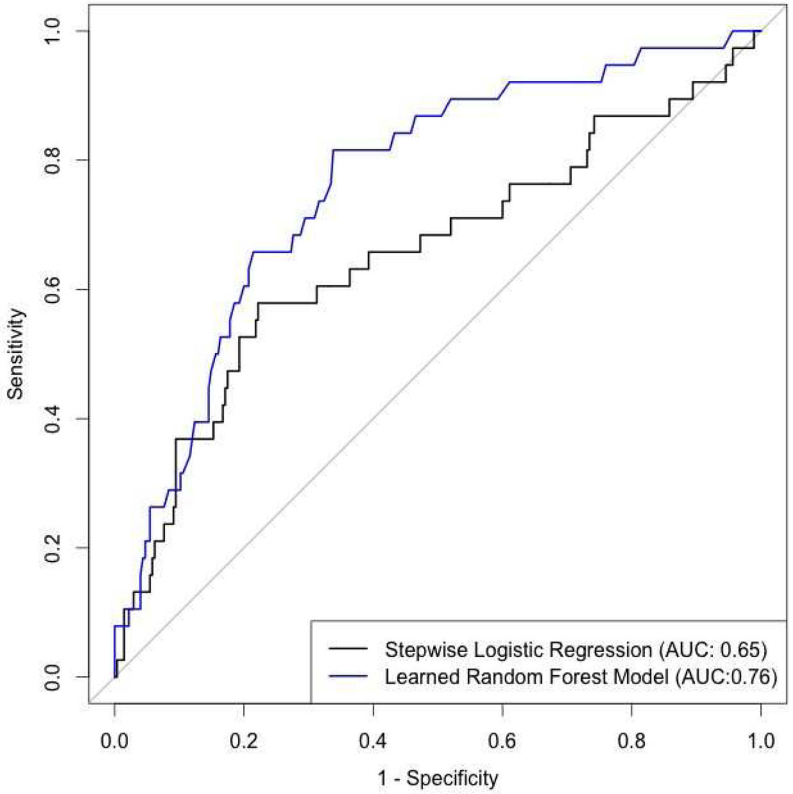

Figure 3.

Area Under the Curve for stepwise logistic regression and machine learned random forest model for prediction of mortality.

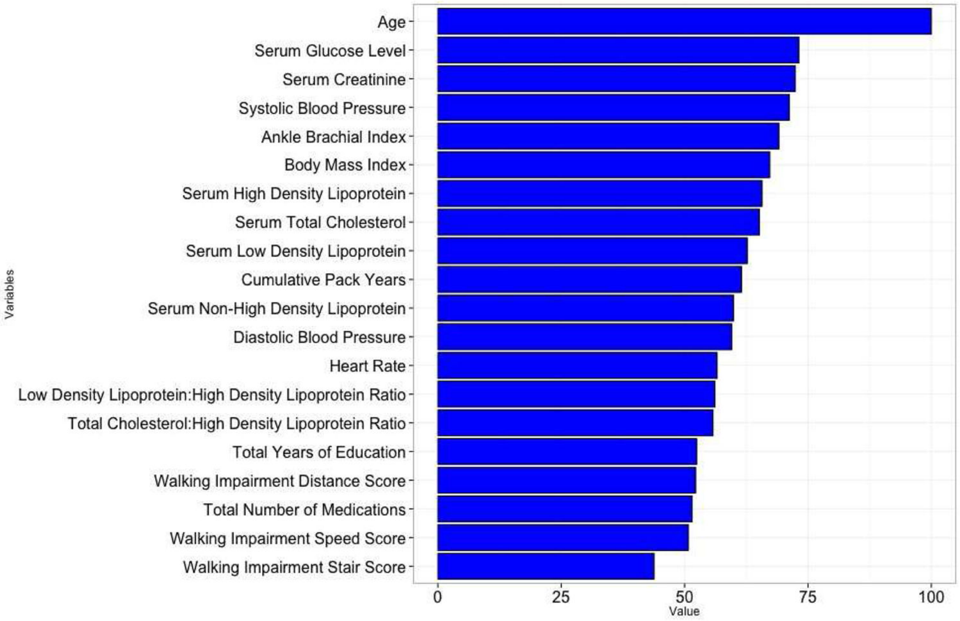

Figure 4.

Top 20 weighted variables used in the machine learned random forest model for mortality prediction.

Table I

Variables used in predictive models for identification of peripheral artery disease and future mortality

Categories	Variables
Demographics	Age, gender, self-reported race/ethnicity
Anthropomorphic measures	Height, weight, ankle brachial indices[a]
Physical exam	Blood pressure, heart rate
Laboratory results	Total Cholesterol, LDL, HDL, non-HDL, total cholesterol to HDL ratio, LDL to HDL ratio, serum creatinine, serum glucose
Imaging[a]	Coronary angiography results
Medical history	Coronary artery disease, peripheral artery disease[a], congestive heart failure, cerebrovascular disease, diabetes, cardiac arrhythmias, menopause
Clinical events during follow-up[a]	Stroke, myocardial infarction, coronary revascularization, new heart failure
Reported symptoms	Joint pain, claudication
Family history	Presence or absence of cardiovascular diseases in parents
Medications[b]	Aspirin, Plavix, statins, anti-hypertensives, diuretics, beta-blockers, anti-glycemic agents
Physical activity assessment	Walking Impairment Questionnaire [1,2], rigorous activity engagement
Genomic markers[c]	rs290481, rs819750, rs7100623, rs7003385, rs94855286, rs46599965, rs3745274, rs2171209, rs16824978, rs107572696
Social factors	Ever married, living situation, total education, current income, employment status, alcohol consumption
Smoking behavior	Ever smoked, current smoker, cumulative pack years

[a] Not used in building peripheral artery disease classification model.

[b] Also assessed medication compliance – "How many times in a month do you forget to take your medications?" and total number of medications;

[c] Selected single-nucleotide polymorphisms found to be associated with peripheral artery disease; HDL - highdensity lipoprotein; LDL - low-density lipoprotein;

Table II

Patient characteristics for all patients included in predictive models

Patient characteristics	All	PAD	No PAD
N	1,047	183	864
Age, mean (SD)	65.6 (10.7)	69.4 (9.4)	64.8 (10.8)
Male Gender, No. (%)	692 (66)	104 (57)	588 (68)
Race/Ethnicity, No., (%)			
Caucasian	581 (55)	102 (56)	479 (55)
African-American	141 (13)	41 (22)	100 (11)
Asian	72 (7)	7 (4)	65 (7.5)
Hispanic	106 (10)	19 (10)	87 (10)
Body mass index (kg/m^2), mean (SD)	29 (6)	28 (5)	29 (6)
Ankle Brachial Index, mean (SD)	1.0 (0.2)	0.7 (0.2)	1.1 (0.1)
Comorbidities, No. (%)			
Cardiac arrhythmias	215 (20)	42 (23)	173 (20)
Coronary Artery Disease	795 (76)	165 (90)	630 (73)
Congestive Heart Failure	63 (6)	17 (9)	46 (5)
Cerebrovascular Disease	49 (5)	21 (11)	28 (3)
Major adverse cardiovascular events *	250 (24)	61 (33)	189 (22)
Mortality	129 (11)	34 (18)	95 (11)
Medications, No. (%)			
Aspirin	787 (75)	138 (75)	649 (75)
Anti-hypertensives	919 (88)	173 (94)	746 (86)
Beta-blockers	617 (59)	112 (61)	505 (58)
Insulin	82 (8)	30 (16)	52 (6)
Clopidogrel	390 (37)	88 (48)	302 (35)
Statins	741 (71)	127 (69)	641 (74)
Smoking, No. (%)			
Current Smoker	104 (10)	34 (18)	70 (8)
Previous Smoker	420 (40)	61 (33)	359 (41)

*Major adverse cardiovascular events includes – stroke, myocardial infarction, coronary revascularization and new heart failure