BMJ Open

Using machine learning to incorporate sparse nutrition data into cardiovascular mortality risk prediction

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-032703
Article Type:	Research
Date Submitted by the Author:	04-Jul-2019
Complete List of Authors:	Rigdon, Joseph ; Stanford University, Quantitative Sciences Unit Basu, Sanjay; Harvard Medical School
Keywords:	Cardiovascular disease, machine learning, Nutrition < TROPICAL MEDICINE, risk prediction

Using machine learning to incorporate sparse nutrition data into cardiovascular mortality risk prediction

Joseph Rigdon, PhD[*]

Quantitative Sciences Unit, Stanford University School of Medicine,

1070 Arastradero Road #3C3104, MC 5559

Palo Alto, California 94304

jrigdon@stanford.edu

Sanjay Basu, MD, PhD

Center for Primary Care, Harvard Medical School

Research and Analytics, Collective Health

School of Public Health, Imperial College London

635 Huntington Avenue, 2nd floor

Boston, MA 02115

sanjay_basu@hms.harvard.edu

[*]Denotes corresponding author

Abstract

Objectives: We aimed to test whether or not adding (i) nutrition predictor variables and/or (ii) using machine learning models improves cardiovascular death prediction versus standard Cox models without nutrition predictor variables

Design: Prospective study

Setting: Six waves of NHANES data collected from 1999-2011 linked to the National Death Index

Participants: 29,390 participants were included in the training set for model derivation and 12,600 were included in the test set for model evaluation. Our study sample was approximately 20% black race and 25% Hispanic ethnicity.

Primary and Secondary Outcome Measures: Time from NHANES interview until the minimum of time of cardiovascular death or censoring

Results: A standard risk model excluding nutrition data overestimated risk nearly two-fold [calibration slope of predicted versus true risk: 0.53 (95% CI: 0.49, 0.57)] with moderate discrimination [C-statistic: 0.87 (0.85, 0.88)]. Nutrition data alone, or machine learning alone, failed to improve performance, but both together improved calibration [slope: 1.08 (0.83, 1.33)] and discrimination [C-statistic: 0.93 (0.92, 0.94)].

Conclusions: Our results indicate that the inclusion of nutrition data with available machine learning algorithms can substantially improve cardiovascular risk prediction.

Keywords: Cardiovascular disease, machine learning, nutrition, risk prediction

Word Count: 3,167

Article Summary

Article focus

- Cardiovascular risk prediction models are commonplace in primary care medicine, and current models are built using Cox regression models with simple demographic and clinical variables
- Could using machine learning models and incorporating nutrition predictor variables improve cardiovascular risk prediction?

Key messages

- Use of survival random forest models with nutrition variables can yield well-calibrated models whereas standard models overestimate risk nearly two-fold and can improve model discrimination from 87% to 93%
- This study supports the clinical scenario where a patient fills out a 24-hour dietary recall in the waiting room prior to seeing the physician, and this nutrition data is used in concert with a machine learning model to more accurately predict CVD risk

Strengths and limitations of this study

- Nationally representative data with a comprehensive evaluation of nutrition, direct laboratory assessment of biomarkers, and direct examination of blood pressure
- Comprehensive follow-up with mortality adjudication by cause of death
- Limitations include the need to impute missing data, a short follow-up duration among individuals collected in the later waves of NHANES, and the lack of information about CVD events in addition to CVD mortality.

Introduction

Nutrition is thought to be a major contributor to cardiovascular disease mortality risk[1–4], but as yet is not explicitly incorporated into cardiovascular risk models that are used to guide clinical prescribing of statins and other preventive medications[5–9]. Nutrition is both imperfectly measured, typically through 24-hour dietary recalls, and nutrition data are sparse and multi-variable, with numerous metrics from individual kilocalorie intakes across a wide range of macro and micronutrients[10,11], making it difficult to determine how an overall nutritional profile might be incorporated into clinical practice. Several groups have offered composite nutrition quality scores (e.g., the Healthy Eating Index and alternatives)[12–14], which correlate to some degree with cardiovascular mortality [15–22] but have not yet been incorporated into common risk equations that use more traditional risk markers (e.g., systolic blood pressure)[5]. Optimizing cardiovascular disease risk prediction is important in clinical practice, because many modern clinical guidelines recommend that physicians prescribe therapies (such as statins, aspirin, and intensive blood pressure treatment) based in part on estimates of overall cardiovascular disease risk, not simply based on the levels of a single biomarker such as cholesterol or blood pressure levels, which fail to fully capture the influence of nutrition on risk [23–26].

With modern machine learning methods, it may be possible to avoid the problems of composite indices, such as reducing a large amount of sparse data to a rough composite that does not explain substantial variation in observed risk[27]. Machine learning approaches are particularly adept at capturing a complex array of large data represented by the sparse matrices of nutrition variables, and incorporating interactions among the data variables (such as between different types of nutrients, e.g., different fats, different carbohydrates, etc.), and identify nonlinear relationships between risk factors and outcomes (e.g., increasing carbohydrate to a very high level from a medium level may differ in impact than increasing from low to medium) that traditional regression models may not fully capture[28–31]. Additionally, with high-quality, more rapid 24-hour dietary recall techniques that can more comprehensively assess a person's dietary behaviors and link them to large nutritional databases, it is now possible to assess nutritional profiles in detail in the clinician's office or clinic waiting room[32–35]. It remains unclear, however, whether nutritional information from a 24-hour recall can add meaningful value to cardiovascular mortality risk prediction beyond biomarker values—such as lipid profile, blood pressure, and diabetes status—and whether using a machine learning approach can advance the predictive power of dietary recalls for cardiovascular risk assessment beyond composite indices already available.

Here, we use a 2-by-2 factorial experimental design to test two hypotheses using observational data: (i) that the data from a single 24-hour dietary recall can add substantial predictive value to cardiovascular mortality risk estimation beyond that afforded by standard biomarkers already included in traditional cardiovascular risk calculators; and (ii) that machine learning approaches to directly incorporate sparse matrices of nutrition data into risk estimates can be superior to standard regression models or the composite nutritional indices constructed through linear modeling methods in the past.

Methods

We conducted a 2-by-2 factorial experiment in which we compared the calibration and discrimination of cardiovascular disease mortality risk prediction models with and without data from a 24-hour dietary recall, and with and without a machine learning approach.

Data Source

Six waves of cross-sectional data from the National Health and Nutrition Examination Survey (NHANES, 1999-2000, 2001-2002, 2003-2004, 2005-2006, 2007-2008, and 2009-2010) were used to develop and validate the risk prediction models. The details of the NHANES sampling scheme are described elsewhere[36]. Briefly, NHANES is a survey including laboratory biomarkers and clinical examination, collected in two-year waves among children and adults, sampled to represent the non-institutionalized civilian U.S. population. Each observation within each wave was linked to the National Death Index (NDI, through 2011) by the Centers for Disease Control. The NDI provided data on the time of CVD death or censoring of follow-up, and additionally a variable attributing death to one of nine-cause specific categories (heart disease, cancer, chronic lower respiratory disease, cerebrovascular diseases, diabetes, pneumonia and influenza, Alzheimer's disease, kidney disease, and unintentional injuries).

The primary statistical outcome was defined as time from NHANES interview to the minimum of time of censoring or time of death from heart disease or cerebrovascular diseases, henceforth CVD mortality. Death from any other cause was treated as censored. Inclusion criteria were age 20-79 years old at time of interview with no prior CVD history. No actions were taken to blind assessment of predictors for the outcome and other predictors. No actions were taken to blind assessment of the outcome.

All potential predictors in the models were collected at time of NHANES interview to mimic a hypothetical scenario where a medical provider may want to conduct an in-clinic 24-hour dietary recall to improve prediction of CVD mortality. Demographic variables included age, sex, and race (Black race, Hispanic ethnicity), and currently-employed cardiovascular disease risk factors of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no)[5]. Nutrition variables included daily standardized intake of micronutrients (e.g., sodium, selenium) and macronutrients (e.g., fat, carbohydrates, protein) collected during a single 24-hour dietary recall following the NHANES interview (Supplementary Table A).

Patient and Public Involvement

No patient involved.

Model Development

Random samples of 70% of each NHANES wave were pooled to form the training sample from which the models were derived, with the remaining 30% prospectively held out to form the test set to assess performance of each model without refitting or recalibration. To train the models in the presence of missing data, 10 imputed data sets for the training sample were created using multiple imputation via chained equations[37,38].

In one arm of the 2-by-2 design, we tested whether or not switching from the standard Cox proportional hazards model to a machine learning algorithm could improve calibration and discrimination. The machine learning algorithms tested were those commonly used for clinical event risk prediction for censored time-to-event data: survival gradient boosted machines (GBMs)[39] and survival random forests (RFs)[40]. Both of these machine learning approaches construct decision trees from data. In a typical decision tree, each branch of the tree divides the sampled study population into increasingly-smaller subgroups that differ in their probability of the outcome. A good decision tree will separate the sampled population into groups that have low within-group variability and high between-group variability in the probability of the outcome. GBMs average many trees where errors made by the first tree contribute to learning of a less erroneous tree in the next iteration (a "boosting" strategy)[41,42]. RFs also build numerous decision trees, but average a forest composed of many trees, where each tree is independently fitted (a "bagging" strategy) with a random subset of covariates selected to be eligible to define the branches[42–45]. RFs use inverse probability of censoring weights to address censoring.

In the second arm of the 2-by-2 design, we tested whether or not adding nutrition variables, including all micro and macronutrients assessed in the NHANES dietary recall, to the standard demographic and biomarker variables could improve prediction. We additional compare incorporating all nutrition data versus using common existing composite nutrition indices: the Healthy Eating Index (HEI)[46], Alternate Healthy Eating Index (AHEI)[47], Mediterranean Diet Score (MDS)[48], and the Dietary Approaches to Stop Hypertension diet score (DASH)[49].

In total, our 2-by-2 design contained 18 models in four quadrants (Supplementary Table B). The no machine learning, no nutrition (standard model) quadrant included only one model: a Cox regression model with demographics and biomarker variables. The machine learning, no nutrition quadrant included two models: a gradient boosted machine and a random forest, both using only demographics and biomarker variables. The no machine learning, nutrition quadrant included five models: a Cox regression including demographics, biomarkers, and either HEI, AHEI, MDS, DASH, or all micro and macronutrients from NHANES. Finally, the machine learning, nutrition quadrant included 10 total models: gradient boosted machines or random forests including demographics, biomarkers, and either HEI, AHEI, MDS, DASH, or all micro and macronutrients from NHANES.

Cox regression models, a gradient boosted machine with 100 trees, a maximum tree depth of 1, and a learning rate of 0.1[50], and a survival random forest based on 20 conditional inference trees[51,52] were fit to each of the 10 imputed data sets. For the best performing model, we increased the number of trees from 20 to 500 to further improve model fit.

Outcome metrics

Model performance was assessed in terms of calibration (using the Greenwood-Nam-D'Agostino [GND] test) and discrimination (using the C-statistic). In the GND test, model predicted probability of 10-year CVD mortality risk was compared to actual death from CVD within 10 years after the NHANES interview by decile of predicted risk. A slope and intercept line were then drawn using these values across deciles of predicted risk, such that a calibration slope of 1 reflects perfect calibration (a perfect 45-degree line between predicted risk and actual event rates).

Model discrimination was assessed using the C-statistic (area under receiver operating characteristic [ROC] curve). Each point on the ROC curve was defined by the sensitivity (x-axis) and 1-specificity (y-axis) for a given cutpoint. The calculation of sensitivity and specificity followed from model predicted risk (above/below cutpoint) versus gold standard of outcome (whether or not CVD mortality happened within 10 years after NHANES interview). As with the GND statistics, C-statistics were calculated for each of the 10 imputed data sets and an overall C-statistic for each model was estimated by Rubin's rules.

Each model developed on imputed training data set k = 1, …, 10 was applied to imputed test set k=1, …, 10 to avoid overlap between training data model development and test set evaluation. Calibration and discrimination mean values and 95% confidence intervals for each model were calculated using Rubin's rules to combine the 10 calibration values[3] (one per imputed data set).

No model updating was done in this study, and no risk groups were created. There were no differences in setting, eligibility criteria, outcome, or predictors between the training (development) set and the test (validation) set. There was no need for participant consent or Ethical Review Board approval as the data are publicly available. All statistical analyses were carried out in Stata 15 software[53] and R version 3.5.1[54]. This manuscript was written in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) recommendations[55], summarized in Supplementary Table I. All data relevant to the study are included in the article or uploaded as supplementary information, and statistical code, and dataset (upon request) are available at https://github.com/joerigdon/CVD_Prediction.

Results

Descriptive statistics on the study sample

Distributions of demographics, covariates and outcome rates were nearly equivalent in training and test sets (Table 1). Of the n=29390 individuals in the training set, 1171/29390 (4.0%) experienced CVD mortality within the follow-up period; of the n=12600 in the test set, 515/12600 (4.1%) experienced CVD mortality. The median follow-up time was 79 months in both training and test sets, with a mean age of 50 years, and 47% of the population being male, 20% Black, 26% Hispanic, 16% with diabetes, and 19% actively smoking tobacco. Composite nutrition indices were identical to within rounding error between the train and test datasets, with a mean HEI score of 47 (out of 100[46]), AHEI score of 47 (out of 110[47]), MDS score of 5 (out of 10[48]), and DASH score of 47 (out of 80[49]); higher scores indicate better adherence to the recommended dietary guidelines for all four of the composite scores.

Compared to individuals without CVD mortality, individuals experiencing CVD mortality were older (74.3 vs. 49.0 years old), more likely to be male (55.0% vs. 46.9%), had higher systolic blood pressure (142.9 vs. 124.8 mmHg), were more likely to take blood pressure medications (74.2% vs. 30.8%), and were more likely to have diabetes (33.3% vs. 15.5%; Table 2). Regarding nutrition variables, those experiencing CVD mortality counter-intuitively had a higher HEI score (51.0 vs. 46.9), a higher AHEI score (48.0 vs. 47.1), and a higher DASH score (48.1 vs. 47.4; Table 2), and comparable MDS scores (5.1 vs. 5.1).

Calibration and discrimination of standard models with and without nutrition data

Using the standard approach to CVD risk prediction modeling[5], a Cox proportional hazards model with variables of age, sex, Black race, and Hispanic ethnicity, total cholesterol, HDL cholesterol, systolic blood pressure, blood pressure medication, diabetes, and tobacco use, yielded a GND calibration slope of 0.53 (95% CI: 0.49, 0.57), reflecting profound risk over-estimation consistent with prior estimates[56,57]. Adding HEI, AHEI, MDS, or DASH score to the model did not change the calibration slope of 0.53, however the addition of the raw (not composite) 24-hour recall data decreased the slope to 0.48 (0.44, 0.53), reflecting a worsening of over-estimation of risk (Figure 1, Supplementary Table E).

The exclusion or inclusion of nutrition data did not affect discrimination of the standard Cox risk models. The Cox model with the above-mentioned non-nutrition data had a C-statistic of 0.87 (0.85, 0.88) in the test set. Adding HEI, AHEI, MDS, DASH, or all raw 24-hour recall data left the C-statistic unchanged at 0.87 (0.85, 0.88) (Figure 2, Supplementary Table F).

Calibration and discrimination of machine learning models with and without nutrition data

When using a machine learning GBM approach instead of a Cox proportional hazards model, but still excluding nutrition data, model calibration improved to 0.54 (0.47, 0.61), and when using random forest in place of Cox, the calibration improved further to 0.58 (0.49, 0.67). Adding nutrition variables improved the machine learning models' calibration when raw 24-hour recall data were used, but not when composite dietary indices were used. Adding HEI, AHEI, MDS, or DASH left the calibration slope unchanged at 0.54 for the GBM models and minimally changed the calibration slope for the random forest models from 0.58 to 0.59 or 0.60. The GBM model had the best calibration when using all 24-hour recall data, producing a calibration slope of 0.56 (0.50, 0.62). The random forest model with raw 24-hour nutrition data was the only model for which the 95% confidence intervals included the ideal value of 1, with a calibration slope of 1.08 (0.83, 1.33) (Figure 1, Supplementary Table E).

Model discrimination also improved with use of machine learning. Using a GBM in place of a Cox model improved discrimination slightly, from C-statistics of 0.87 (0.85, 0.88) in Cox models to 0.88 (0.87, 0.89) for all GBM models without nutrition data and 0.91

(0.90, 0.93) for the random forest without nutrition data. The discrimination was not significantly different with the addition of composite nutritional indices, but did improve to 0.93 (0.92, 0.94) with the addition of raw nutrition data (Figure 2, Supplementary Table F).

As expected, model calibration values (Supplementary Figure A, Supplementary Table C), and model discrimination values (Supplementary Figure B, Supplementary Table D) were better in the training data sets versus the held-out test set.

Cox model coefficients are detailed in Supplementary Table G and gradient boosted machine model relative influences are detailed in Supplementary Table H. Notable associations with cardiovascular death included age (HR for 1-year increase in age of 1.1 [1.09, 1.1], female sex (HR vs. males of 0.62 [0.55, 0.71]), Hispanic ethnicity (HR vs. non-Hispanics of 0.72 [0.61, 0.86]), systolic BP (HR for 1-unit increase of 1.01 [1.01, 1.01]), blood pressure medications (HR for each additional med of 1.22 [1.11, 1.34]), type 2 diabetes (HR vs. non-diabetics of 1.46 [1.23, 1.73]), and tobacco use (HR vs. non-users 1.82 [1.53, 2.17]) (Supplementary Table G). No associations with cardiovascular death were found with HEI, AHEI, MDS, or DASH.

In the comprehensive evaluation of all 24-hour nutrition variables, protective associations were seen with fiber (HR 0.97 [0.96, 0.99] for 1-gram increase) and niacin (HR 0.97 [0.95, 0.99] for 1-milligram increase), and harmful association with vitamin B6 (HR 1.17 [1.02, 1.35] for 1-milligram increase). Relative influences in a GBM display how much of a 0-100 importance total is accounted for by each variable in the model (Supplementary Table H). Age consistently had relative influences of around 70/100, with the next most important variables being SBP (around 11), blood pressure medications (around 7), total cholesterol (around 3), diabetes (3), and sex (2). Of the composite indices, only HEI (1.92) exceeded a relative influence of 1. Of the 24-hour nutrition variables, only potassium (1.82) exceeded a relative influence of 1. Partial dependence plots for the random forest model with all nutrition variables reveal an exponential increase in 10-year probability of CVD death starting at about age 65, and an S-shaped risk curve for 10-year probability of CVD death with spike around 145 mmHg systolic blood pressure (Supplementary Figure C)

Discussion

We examined whether or not improvements in CVD mortality prediction could be achieved by including sparse nutrition data into models derived through machine learning algorithms. We observed that the addition of nutrition variables to a standard Cox proportional hazards model was not of substantial benefit alone, nor was the use of machine learning algorithms alone, but when both nutrition data and machine learning were combined, we could substantially improve risk prediction beyond the inclusion of standard demographics and biomarkers alone. Calibration particularly improved when both nutrition data and machine learning algorithms were used.

Our findings are of clinical relevance as more rapid, automated or mobile device-based 24-hour dietary recalls make it feasible to provide a nutrition profile for patients at or before visiting a doctor's office[1,2], and as automated cardiovascular disease risk prediction models become an increasingly-important part of precision medicine guidelines that aim to improve the ability of medical practitioners to prescribe preventive cardiovascular treatments to patients with the highest risk[6]. As standard biomarkers fail to explain the full extent to which nutrition relates to cardiovascular mortality[58,59], machine learning approaches that directly incorporate raw dietary data appear to have benefits over composite nutritional indices that may excessively reduce complexity in nutritional interactions and non-linear relationships that confer risk. Our study benefits from being conducted on a nationally representative sample of US adults, including a comprehensive evaluation of nutrition, direct laboratory assessment of biomarkers, direct examination of blood pressure, and comprehensive follow-up with mortality adjudication by cause of death. Nevertheless, our study has important limitations, including the need to impute missing data, a short follow-up duration among individuals collected in the later waves of NHANES, and the lack of information about CVD events in addition to CVD mortality.

In the future, further research can assess whether the performance of rapid dietary recalls and associated cardiovascular risk estimation can be implemented in practice, whether the level of improvements to calibration and discrimination observed in this assessment produce clinically-meaningful changes in the level of prescribing of key preventive therapies for patients, and whether the difficulties of interpreting machine learning models are compared to traditional Cox-type risk models poses challenges to the acceptability of these models in clinical practice.

At present, our results indicate that the inclusion of nutrition data with available machine learning algorithms can substantially improve cardiovascular risk prediction.

Author Contributions

SB conceptualized the study and design and contributed to data preparation and analysis. JR contributed to data preparation and analysis. Both authors contributed to writing and critically reviewing the manuscript.

Competing Interests statement

JR and SB have no competing interests to report.

Acknowledgements

The authors acknowledge two anonymous reviewers at the Stanford Quantitative Sciences Unit.

Funding

This work was supported by the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Number DP2MD010478. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Shivappa, N., Steck, S. E., Hussey, J. R., Ma, Y. & Hebert, J. R. Inflammatory potential of diet and all-cause, cardiovascular, and cancer mortality in National Health and Nutrition Examination Survey III Study. *Eur. J. Nutr.* **56**, 683–692 (2017).
2. Aune, D. *et al.* Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—a systematic review and dose-response meta-analysis of prospective studies. *Int. J. Epidemiol.* **46**, 1029–1056 (2017).
3. Wang, D. D. *et al.* Association of Specific Dietary Fats With Total and Cause-Specific Mortality. *JAMA Intern. Med.* **176**, 1134–1145 (2016).

4. Langley-Evans, S. C. Nutrition in early life and the programming of adult disease: a review. *J. Hum. Nutr. Diet.* **28**, 1–14 (2015).

5. Goff David C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129**, S49–S73 (2014).

6. Stone Neil J. *et al.* 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults. *Circulation* **129**, S1–S45 (2014).

7. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA J. Am. Med. Assoc.* **285**, 2486–2497 (2001).

8. Lloyd-Jones, D. M. *et al.* Prediction of Lifetime Risk for Cardiovascular Disease by Risk Factor Burden at 50 Years of Age. *Circulation* **113**, 791–798 (2006).

9. Yadlowsky, S. *et al.* Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann. Intern. Med.* **169**, 20 (2018).

10. Stumbo, P. Considerations for selecting a dietary assessment system. *J. Food Compos. Anal. Off. Publ. U. N. Univ. Int. Netw. Food Data Syst.* **21**, S13–S19 (2008).

11. Stewart, K. K. & Whitaker, J. R. *Modern Methods of Food Analysis*. (Springer Science & Business Media, 2012).

12. Kennedy, E. T., Ohls, J., Carlson, S. & Fleming, K. The Healthy Eating Index: Design and Applications. *J. Am. Diet. Assoc.* **95**, 1103–1108 (1995).

13. McCullough, M. L. & Willett, W. C. Evaluating adherence to recommended diets in adults: the Alternate Healthy Eating Index. *Public Health Nutr.* **9**, (2006).

14. Panagiotakos, D. B., Pitsavos, C. & Stefanadis, C. Dietary patterns: A Mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk. *Nutr. Metab. Cardiovasc. Dis.* **16**, 559–568 (2006).

15. Reedy, J. *et al.* Higher Diet Quality Is Associated with Decreased Risk of All-Cause, Cardiovascular Disease, and Cancer Mortality among Older Adults. *J. Nutr.* **144**, 881–889 (2014).

16. Onvani, S., Haghighatdoost, F., Surkan, P. J., Larijani, B. & Azadbakht, L. Adherence to the Healthy Eating Index and Alternative Healthy Eating Index dietary patterns and mortality from all causes, cardiovascular disease and cancer: a meta-analysis of observational studies. *J. Hum. Nutr. Diet.* **30**, 216–226 (2017).

17. Fung, T. T. *et al.* Mediterranean diet and incidence and mortality of coronary heart disease and stroke in women. *Circulation* **119**, 1093–1100 (2009).

18. Akbaraly, T. N. *et al.* Alternative Healthy Eating Index and mortality over 18 y of follow-up: results from the Whitehall II cohort. *Am. J. Clin. Nutr.* **94**, 247–253 (2011).

19. Schwingshackl, L. & Hoffmann, G. Diet Quality as Assessed by the Healthy Eating Index, the Alternate Healthy Eating Index, the Dietary Approaches to Stop Hypertension Score, and Health Outcomes: A Systematic Review and Meta-Analysis of Cohort Studies. *J. Acad. Nutr. Diet.* **115**, 780-800.e5 (2015).

20. Kant, A. K. Indexes of Overall Diet Quality: A Review. *J. Am. Diet. Assoc.* **96**, 785–791 (1996).

21. Folsom, A. R., Parker, E. D. & Harnack, L. J. Degree of Concordance With DASH Diet Guidelines and Incidence of Hypertension and Fatal Cardiovascular Disease. *Am. J. Hypertens.* **20**, 225–232 (2007).

22. Fung, T. T. *et al.* Adherence to a DASH-Style Diet and Risk of Coronary Heart Disease and Stroke in Women. *Arch. Intern. Med.* **168**, 713–720 (2008).

23. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* 25709 (2018). doi:10.1016/j.jacc.2018.11.003

24. Bibbins-Domingo, K. *et al.* Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: US Preventive Services Task Force Recommendation Statement. *JAMA* **316**, 1997–2007 (2016).

25. Bibbins-Domingo, K. & on behalf of the U.S. Preventive Services Task Force. Aspirin Use for the Primary Prevention of Cardiovascular Disease and Colorectal Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann. Intern. Med.* **164**, 836 (2016).

26. Whelton, P. K. *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. *J. Am. Coll. Cardiol.* **71**, e127–e248 (2018).

27. Suresh, S., Saraswathi, S. & Sundararajan, N. Performance enhancement of extreme learning machine for multi-category sparse data classification problems. *Eng. Appl. Artif. Intell.* **23**, 1149–1157 (2010).

28. Messina, M., Lampe, J. W., Birt, D. F., Appel, L. J. & al, et. Reductionism and the narrowing nutrition perspective: Time for reevaluation and emphasis on food synergy. *Am. Diet. Assoc. J. Am. Diet. Assoc. Chic.* **101**, 1416–9 (2001).

29. Wang, J., Li, D., Dangott, L. J. & Wu, G. Proteomics and Its Role in Nutrition Research,. *J. Nutr.* **136**, 1759–1762 (2006).

30. Marcos, A., Nova, E. & Montero, A. Changes in the immune system are conditioned by nutrition. *Eur. J. Clin. Nutr.* **57**, S66–S69 (2003).

31. Zeisel, S. H. *et al.* Nutrition: A Reservoir for Integrative Science. *J. Nutr.* **131**, 1319–1321 (2001).

32. Subar, A. F. *et al.* The Automated Self-Administered 24-Hour Dietary Recall (ASA24): A Resource for Researchers, Clinicians, and Educators from the National Cancer Institute. *J. Acad. Nutr. Diet.* **112**, 1134–1137 (2012).

33. Vereecken, C. A., Covents, M., Matthys, C. & Maes, L. Young adolescents' nutrition assessment on computer (YANA-C). *Eur. J. Clin. Nutr.* **59**, 658–667 (2005).

34. Hongu, N. *et al.* Dietary Assessment Tools Using Mobile Technology. *Top. Clin. Nutr.* **26**, 300 (2011).

35. Thompson, F. E. *et al.* Comparison of Interviewer-Administered and Automated Self-Administered 24-Hour Dietary Recalls in 3 Diverse Integrated Health Systems. *Am. J. Epidemiol.* **181**, 970–978 (2015).

36. NHANES - About the National Health and Nutrition Examination Survey. (2017). Available at: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm. (Accessed: 11th March 2019)

37. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).

38. Vergouwe, Y., Royston, P., Moons, K. G. M. & Altman, D. G. Development and validation of a prediction model with missing predictor data: a practical approach. *J. Clin. Epidemiol.* **63**, 205–214 (2010).

39. Chen, Y., Jia, Z., Mercola, D. & Xie, X. A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. *Computational and Mathematical Methods in Medicine* (2013). doi:10.1155/2013/873595

40. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).

41. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

42. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189–1232 (2001).

43. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).

44. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

45. Friedman, J., Hastie, T. & Tibshirani, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann. Stat.* **28**, 337–407 (2000).

46. Guenther, P. M. *et al.* Update of the Healthy Eating Index: HEI-2010. *J. Acad. Nutr. Diet.* **113**, 569–580 (2013).

47. Chiuve, S. E. *et al.* Alternative Dietary Indices Both Strongly Predict Risk of Chronic Disease. *J. Nutr.* **142**, 1009–1018 (2012).

48. Trichopoulou, A., Costacou, T., Bamia, C. & Trichopoulos, D. Adherence to a Mediterranean Diet and Survival in a Greek Population. *N. Engl. J. Med.* **348**, 2599–2608 (2003).

49. Günther, A. L. B. *et al.* Association Between the Dietary Approaches to Hypertension Diet and Hypertension in Youth With Diabetes Mellitus. *Hypertension* **53**, 6–12 (2009).

50. Greenwell, B., Boehmke, B., Cunningham, J. & Developers (https://github.com/gbm-developers), G. B. M. *gbm: Generalized Boosted Regression Models*. (2019).

51. Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. & Van Der Laan, M. J. Survival ensembles. *Biostatistics* **7**, 355–373 (2006).

52. Hothorn, T., Hornik, K., Strobl, C. & Zeileis, A. *party: A Laboratory for Recursive Partytioning*. (2019).

53. StataCorp. *Stata Statistical Software: Release 15*. (StataCorp LLC, 2017).

54. R Core Team. *R: A language and environment for statistical computing*. (2018).

55. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann. Intern. Med.* **162**, 55 (2015).

56. Yadlowsky, S. *et al.* Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann. Intern. Med.* (2018). doi:10.7326/M17-3011

57. Ridker, P. M. & Cook, N. R. Statins: new American guidelines for prevention of cardiovascular disease. *The Lancet* **382**, 1762–1765 (2013).

58. Kant, A. K. Dietary patterns: biomarkers and chronic disease riskThis paper is one of a selection of papers published in the CSCN–CSNS 2009 Conference, entitled Are dietary patterns the best way to make nutrition recommendations for chronic disease prevention? *Appl. Physiol. Nutr. Metab.* **35**, 199–206 (2010).

59. Boushey, C. J., Coulston, A. M., Rock, C. L. & Monsen, E. *Nutrition in the Prevention and Treatment of Disease*. (Elsevier, 2001).

Figure Legends

Figure 1: Calibration slopes and confidence intervals of models in the hold-out test set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Figure 2: Model discrimination (C-statistic) in the hold-out test set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Tables

Table 1: Descriptive statistics on the study sample (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N=41990). Statistics are grouped to reflect participants in the training (n=29390/41990 = 70%) or test (n=12600/41990 = 30%) data subsets. CVD = cardiovascular disease, HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest. Mean (±standard deviation) reported for continuous variables and N (%) reported for categorical variables.

	Training data for model derivation n=29390	**Test data for model evaluation** n=12600	**P-value for difference**[1]
CVD death			
No	28,219 (96.0%)	12,085 (95.9%)	0.63
Yes	1,171 (4.0%)	515 (4.1%)	
Time since interview (months)	79.3 (±41.5)	79.5 (±41.4)	0.71
Wave			
99-00	3,810 (13.0%)	1,633 (13.0%)	1.00
01-02	8,853 (30.1%)	3,795 (30.1%)	
03-04	3,926 (13.4%)	1,684 (13.4%)	
05-06	3,891 (13.2%)	1,669 (13.2%)	
07-08	4,353 (14.8%)	1,866 (14.8%)	
09-10	4,557 (15.5%)	1,953 (15.5%)	
Age	50.1 (±20.4)	50.0 (±20.4)	0.55
Sex			
Male	13,870 (47.2%)	5,941 (47.2%)	0.94
Female	15,520 (52.8%)	6,659 (52.8%)	
Black			
No	14,826 (50.4%)	6,316 (50.1%)	0.35
Yes	5,839 (19.9%)	2,554 (20.3%)	
Missing	8,725 (29.7%)	3,730 (29.6%)	
Hispanic			
No	21,861 (74.4%)	9,369 (74.4%)	0.96

	Training data for model derivation	**Test data for model evaluation**	**P-value for difference**[1]
Yes	7,529 (25.6%)	3,231 (25.6%)	
Total chol	197.8 (±42.9)	198.5 (±44.3)	0.33
Missing	3,640 (12.4%)	1,485 (11.8%)	
HDL	45.6 (±23.0)	45.4 (±22.9)	0.63
Missing	3,641 (12.4%)	1,486 (11.8%)	
SBP	125.5 (±20.8)	125.4 (±20.7)	0.81
Missing	3,166 (10.8%)	1,357 (10.8%)	
DBP	69.8 (±12.7)	69.9 (±12.5)	0.58
Missing	3,377 (11.5%)	1,428 (11.3%)	
Number of blood pressure medications			
0	19,855 (67.6%)	8,473 (67.2%)	0.66
1	7,875 (26.8%)	3,428 (27.2%)	
2 or more	1,660 (5.6%)	699 (5.5%)	
T2DM			
No	10,560 (35.9%)	4,518 (35.9%)	0.18
Yes	4,695 (16.0%)	2,096 (16.6%)	
Missing	14,135 (48.1%)	5,986 (47.5%)	
Smoking			
No	23,713 (80.7%)	10,246 (81.3%)	0.14
Yes	5,675 (19.3%)	2,354 (18.7%)	
Missing	2 (0.0%)	0 (0.0%)	
HEI	47.0 (±11.0)	47.1 (±11.0)	0.58
Missing	3,274 (11.1%)	1,364 (10.8%)	
AHEI	47.2 (±11.0)	47.1 (±11.1)	0.59
Missing	3,258 (11.1%)	1,358 (10.8%)	
MDS	5.1 (±1.2)	5.1 (±1.2)	0.70
Missing	3,270 (11.1%)	1,368 (10.9%)	
DASH	47.4 (±9.3)	47.4 (±9.4)	0.77
Missing	8,700 (29.6%)	3,796 (30.1%)	

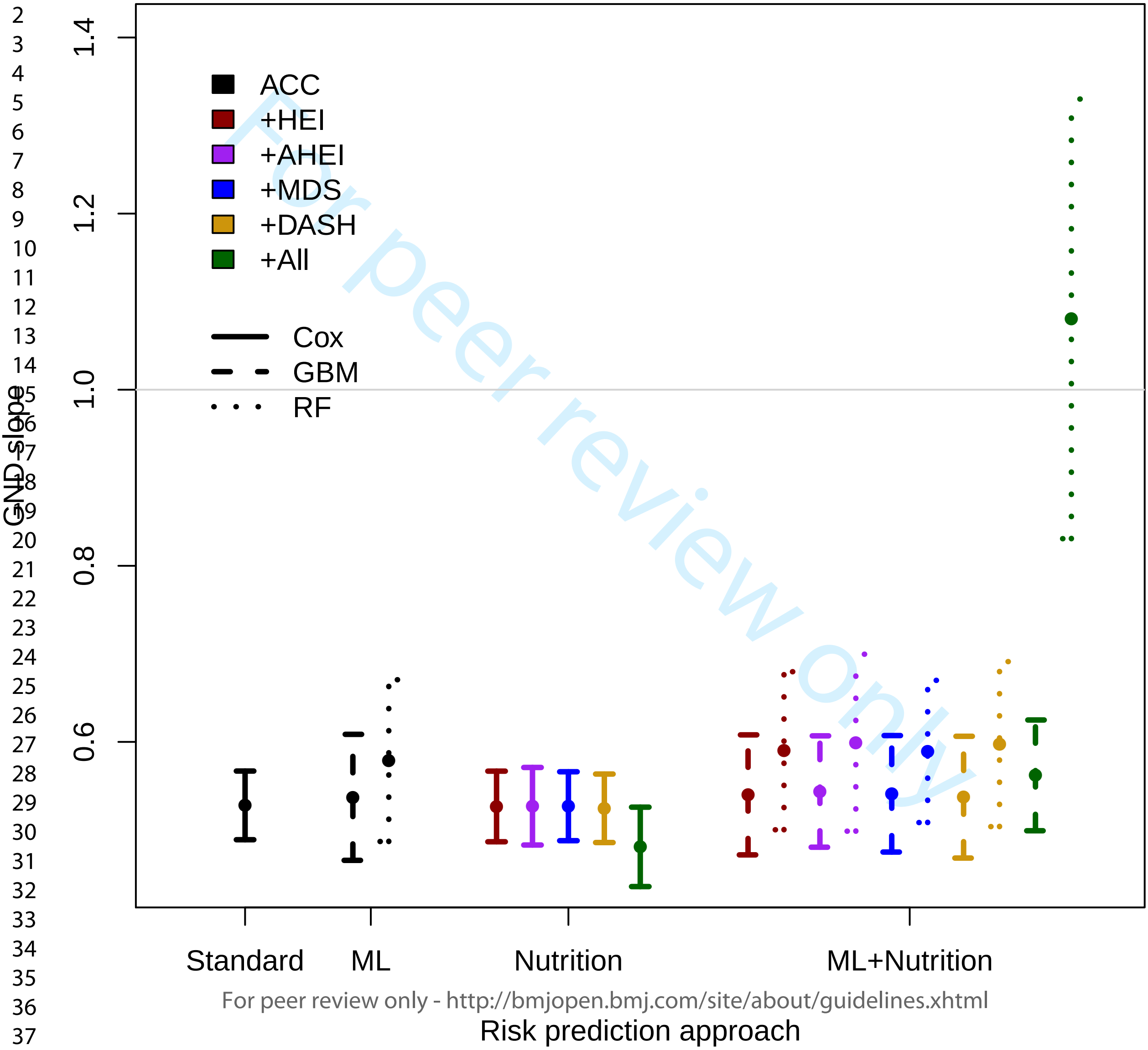
[1]Wilcoxon rank sum test for continuous variables, e.g., age, and Fisher's exact test for categorical variables, e.g., black race

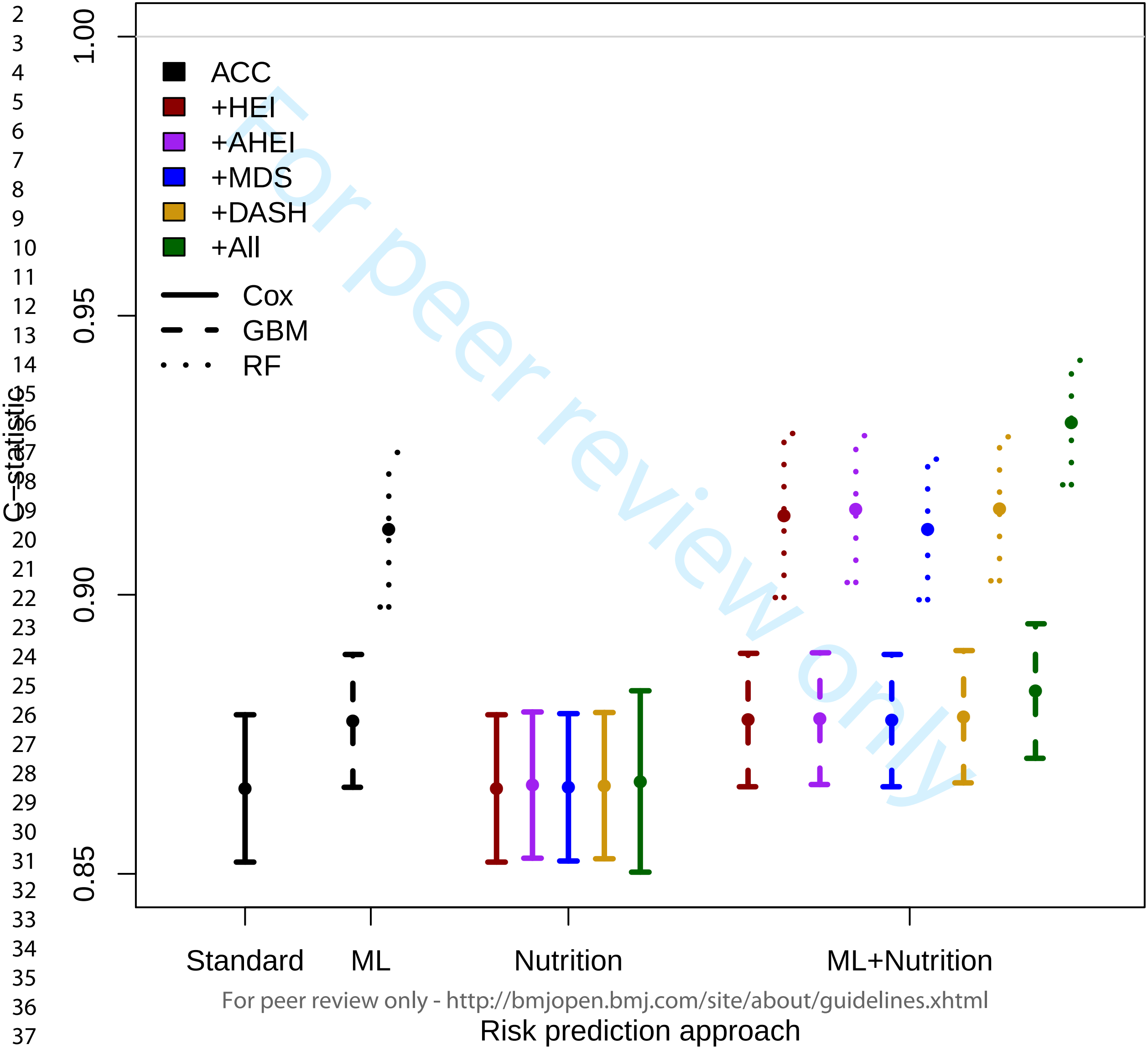
Table 2: Comparisons of participant characteristics by outcome (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N=41990). Descriptive summary of variables in those participants without CVD event (n=40304) vs. those with a CVD event (n=1686) during the follow-up period. Mean (±standard deviation) reported for continuous variables and N (%) reported for categorical variables.

	No CVD	CVD	P-value for difference[1]
	n=40304	n=1686	
Time since interview (months)	80.3 (±41.4)	55.7 (±34.9)	<0.0001
Wave			
99-00	5,168 (12.8%)	275 (16.3%)	<0.0001
01-02	11,681 (29.0%)	967 (57.4%)	
03-04	5,401 (13.4%)	209 (12.4%)	
05-06	5,451 (13.5%)	109 (6.5%)	
07-08	6,127 (15.2%)	92 (5.5%)	
09-10	6,476 (16.1%)	34 (2.0%)	
Age	49.0 (±20.1)	74.3 (±11.9)	<0.0001
Sex			
Male	18,883 (46.9%)	928 (55.0%)	<0.0001
Female	21,421 (53.1%)	758 (45.0%)	
Black			
No	20,005 (49.6%)	1,137 (67.4%)	<0.0001
Yes	8,110 (20.1%)	283 (16.8%)	
Missing	12,189 (30.2%)	266 (15.8%)	
Hispanic			
No	29,781 (73.9%)	1,449 (85.9%)	<0.0001
Yes	10,523 (26.1%)	237 (14.1%)	
Total chol	198.1 (±43.2)	196.2 (±47.0)	0.10
Missing	4,670 (11.6%)	455 (27.0%)	
HDL	45.5 (±23.0)	45.0 (±24.2)	0.002
Missing	4,672 (11.6%)	455 (27.0%)	
SBP	124.8 (±20.3)	142.9 (±26.8)	<0.0001
Missing	4,114 (10.2%)	409 (24.3%)	
DBP	70.0 (±12.5)	67.5 (±14.7)	<0.0001
Missing	4,359 (10.8%)	446 (26.5%)	
Number of blood pressure medications			
0	27,894 (69.2%)	434 (25.7%)	<0.0001
1	10,205 (25.3%)	1,098 (65.1%)	
2	2,205 (5.5%)	154 (9.1%)	
T2DM			
No	14,680 (36.4%)	398 (23.6%)	<0.0001
Yes	6,229 (15.5%)	562 (33.3%)	
Missing	19,395 (48.1%)	726 (43.1%)	
Smoking			
No	32,508 (80.7%)	1,451 (86.1%)	<0.0001
Yes	7,794 (19.3%)	235 (13.9%)	
Missing	2 (0.0%)	0 (0.0%)	
HEI	46.9 (±11.0)	51.0 (±10.3)	<0.0001
Missing	4,179 (10.4%)	459 (27.2%)	
AHEI	47.1 (±11.1)	48.0 (±10.9)	0.006
Missing	4,158 (10.3%)	458 (27.2%)	
MDS	5.1 (±1.2)	5.1 (±1.2)	0.10
Missing	4,472 (11.1%)	166 (9.8%)	

	No CVD	CVD	P-value for difference[1]
DASH	47.4 (±9.4)	48.1 (±9.2)	0.01
Missing	11,774 (29.2%)	722 (42.8%)	

[1] Wilcoxon rank sum test for continuous variables, e.g., age, and Fisher's exact test for categorical variables, e.g., black race


1.4
1.2
1.0
0.8
0.6
GND slope
ACC
+HEI
+AHEI
+MDS
+DASH
+All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach


C−statistic
1.00
0.95
0.90
0.85
ACC
+HEI
+AHEI
+MDS
+DASH
+All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach

Supplementary Appendix

Figure Legends

Supplementary Figure A: Calibration slopes and confidence intervals of models in training set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Supplementary Figure B: Model discrimination (C-statistic) in training set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Supplementary Figure C: Partial dependence plots for best model (500 trees using full data) for (a) age and (b) systolic blood pressure. Plots estimated by averaging model predictions for 1000 random samples from the training data at each decile of age or SBP.

Supplementary Table A: List of all predictor variables included in statistical models

Variable name	**Definition**
Demographic and risk factors (4)	
age	Age in years
sex	Sex (0 if male, 1 if female)
black	Black race (0 if no, 1 if yes)
hispanic	Hispanic ethnicity (0 if no, 1 if yes)
ACC covariates (7)	
total_chol	Total cholesterol (mg/dL)
hdl	HDL cholesterol (mg/dL)
sbp	Systolic blood pressure (mmHg)
dbp	Diastolic blood pressure (mmHg)
bpmeds	Number of blood pressure medications
dm	Type 2 diabetes (0 if no, 1 if yes)
tob	Current smoking (0 if no, 1 if yes)
Composite nutrition variables (4)	
hei	Healthy eating index (0-100)
ahei	Alternative healthy eating index (0-110)
mds	Mediterranean diet score (0-9)
dash	DASH diet score (0-80)
24-hour recall variables (103)	
milk_g	Milk and milk drinks (g)
cream_g	Creams and cream substitutes (g)
milk_dessert_g	Milk desserts, sauces, gravies (g)
cheese_g	Cheeses (g)
meat_ns_g	Meat, not specified as to type (g)
beef_g	Beef (g)
pork_g	Pork (g)
lamb_g	Lamb, veal, game, other carcass meat (g)
poultry_g	Poulty (g)
organ_meat_g	Organ meats, sausages, and lunchmeats, and meat spreads (g)
fish_g	Fish and shellfish (g)
meat_nonmeat_g	Meat, poultry, fish with nonmeat items (g)
protein_frozen_g	Proetin and shelf-stable plate meals, soups, and gravies with meat, poulty fish base; gelatin and gelatin-based drinks
eggs_g	Eggs (g)
egg_mixture_g	Egg mixtures (g)
egg_sub_g	Egg substitutes (g)
egg_frozen_g	Frozen plate meals with egg as major ingredient (g)
legumes_g	Legumes (g)
nuts_g	Nuts, nut butters, and nut mixtures (g)
seeds_g	Seeds and seed mixtures (g)
carob_g	Carob products (g)
flour_mix_g	Flour and dry mixes (g)
bread_yeast_g	Yeast breads, rolls (g)

bread_quick_g	Quick breads (g)
pastries_g	Cakes, cookies, pies, pastries, bars (g)
crackers_g	Crackers and salty snacks from grain products (g)
pancakes_g	Pancakes, waffles, French toast, other grain products (g)
pastas_g	Pastas, cooked cereals, rice (g)
cereals_g	Cereals, not cooked or not specified as to cooked (g)
grain_mix_g	Grain mixtures, frozen plate meals, soups (g)
meat_sub_g	Meat substitutes, mainly cereal protein (g)
citrus_g	Citrus fruits, juices (g)
fruit_dried_g	Dried fruits (g)
fruit_other_g	Other fruits (g)
fruit_juice_g	Fruit juices and nectars excluding citrus (g)
fruit_baby_g	Fruit and juices baby food (g)
potatoes_g	White potatoes and Puerto Rican starchy vegetables (g)
veg_darkgreen_g	Dark-green vegetables (g)
veg_deepyellow_g	Deep-yellow vegetables (g)
tomatoes_g	Tomatoes and tomato mixtures (g)
veg_other_g	Other vegetables (g)
veg_baby_g	Vegetables and mixtures mostly vegetables baby food (g)
veg_meat_g	Vegetables with meat, poultry, fish (g)
veg_mixture_g	Mixtures mostly vegetables without meat, poultry, fish (g)
fats_g	Fats (g)
oils_g	Oils (g)
salad_dressing_g	Salad dressings (g)
sweets_g	Sugars and sweets (g)
bev_nonalcohol_g	Nonalcoholic beverages (g)
bev_alcohol_g	Alcoholic beverages (g)
water_g	Water, noncarbonated (g)
bev_nutrition_g	Formulated nutrition beverages, energy drinks, sports drinks, functional beverages (g)
kcal	Energy (kcal)
protein_g	Protein (g)
carb_g	Carbohydrates (g)
fiber_g	Fiber (g)
fat_g	Fat (g)
fat_sat_g	Saturated fats (g)
fat_mono_g	Monounsaturated fats (g)
fat_poly_g	Polyunsaturated fats (g)
cholesterol_mg	Cholesterol (mg)
vite_mg	Vitamin-E as alpha-tocopherol (mg)
vita_mcg	Vitamin A, RAE (mcg)
betacaro_mcg	Beta-carotene (mcg)
vitb1_mg	Thiamin (Vitamin B1) (mg)

vitb2_mg	Riboflavin (Vitamin B2) (mg)
niacin_mg	Niacin (mg)
vitb6_mg	Vitamin B6 (mg)
folate_mcg	Total folate (mcg)
vitb12_mcg	Vitamin B12 (mcg)
vitc_mg	Vitamin C (mg)
calcium_mg	Calcium (mg)
phosphorus_mg	Phosphorus (mg)
magnesium_mg	Magnesium (mg)
iron_mg	Iron (mg)
zinc_mg	Zing (mg)
copper_mg	Copper (mg)
sodium_mg	Sodium (mg)
potassium_mg	Potassium (mg)
selenium_mcg	Selenium (mg)
caffeine_mg	Caffeine (mg)
theobromine_mg	Theobromine (mg)
alcohol_gm	Alcohol (gm)
sfa_40_gm	SFA 4:0 (Butanoic) (g)
sfa_60_gm	SFA 6:0 (Hexanoic) (g)
sfa_80_gm	SFA 8:0 (Octanoic) (g)
sfa_100_gm	SFA 10:0 (Decanoic) (g)
sfa_120_gm	SFA 12:0 (Dodecanoic) (g)
sfa_140_gm	SFA 14:0 (Tetradecanoic) (g)
sfa_160_gm	SFA 16:0 (Hexadecanoic) (g)
sfa_180_gm	SFA 18:0 (Octadecanoic) (g)
mfa_161h_gm	MFA 16:1 (Hexadecanoic) (g)
mfa_161o_gm	MFA 16:1 (Octadecanoic) (g)
mfa_201_gm	MFA 20:1 (Eicosenoic) (g)
mfa_221_gm	MFA 22:1 (Docosenoic) (g)
pfa_182_gm	PFA 18:2 (Octadecadienoic) (g)
pfa_183_gm	PFA 18:3 (Octadecatrienoic) (g)
pfa_184_gm	PFA 18:4 (Octadecatatraenoic) (g)
pfa_204_gm	PFA 20:4 (Eicosatetraenoic) (g)
pfa_205_gm	PFA 20:5 (Eicosapentaenoic) (g)
pfa_225_gm	PFA 22:5 (Docosapentaenoic) (g)
pfa_226_gm	PFA 22:6 (Docosahexaenoic) (g)
water_yesterday_gm	Total plain water drank yesterday (g)

Supplementary Table B: Outline of prediction models assessed

		Standard	Machine learning	
		A. Cox regression model	B. Gradient boosted machine	C. Survival random forest
Standard	1. Demographics, ACC	Model 1A	Model 1B	Model 1C
Add nutrition variables	2. Demographics, ACC, HEI	Model 2A	Model 2B	Model 2C
	3. Demographics, ACC, AHEI	Model 3A	Model 3B	Model 3C
	4. Demographics, ACC, Med diet score	Model 4A	Model 4B	Model 4C
	5. Demographics, ACC, DASH diet score	Model 5A	Model 5B	Model 5C
	6. Demographics, ACC, all 24-hour recall data	Model 6A	Model 6B	Model 6C

Supplementary Table C: Calibration slopes and confidence intervals on the training data

		Standard	Machine learning	
		Cox model	GBM	Random forest
Standard	Demographics, ACC	0.52 (0.50, 0.54)	0.55 (0.51, 0.60)	0.74 (0.52, 0.95)
Plus nutrition variables	Demographics, ACC, HEI	0.52 (0.50, 0.54)	0.55 (0.51, 0.60)	0.76 (0.52, 1.00)
	Demographics, ACC, AHEI	0.52 (0.50, 0.54)	0.56 (0.51, 0.60)	0.76 (0.53, 0.98)
	Demographics, ACC, Med diet score	0.51 (0.49, 0.54)	0.55 (0.51, 0.60)	0.75 (0.54, 0.97)
	Demographics, ACC, DASH diet score	0.52 (0.50, 0.53)	0.55 (0.50, 0.60)	0.76 (0.53, 1.00)
	Demographics, ACC, all 24-hour recall data	0.54 (0.51, 0.57)	0.57 (0.53, 0.62)	1.13 (0.73, 1.52)

Supplementary Table D: C-statistics on the training data

		Standard	Machine learning	
		Cox model	GBM	Random forest
Standard	Demographics, ACC	0.87 (0.86, 0.88)	0.88 (0.87, 0.89)	0.97 (0.96, 0.98)
Plus nutrition variables	Demographics, ACC, HEI	0.87 (0.86, 0.88)	0.88 (0.87, 0.89)	0.97 (0.97, 0.98)
	Demographics, ACC, AHEI	0.87 (0.86, 0.88)	0.88 (0.87, 0.89)	0.97 (0.97, 0.98)
	Demographics, ACC, Med diet score	0.87 (0.86, 0.88)	0.88 (0.87, 0.89)	0.97 (0.97, 0.98)
	Demographics, ACC, DASH diet score	0.87 (0.86, 0.88)	0.88 (0.87, 0.89)	0.97 (0.97, 0.98)
	Demographics, ACC, all 24-hour recall data	0.88 (0.88, 0.89)	0.88 (0.88, 0.89)	0.99 (0.99, 0.99)

Supplementary Table E: Calibration slopes and confidence intervals on the held-out test data

		Standard	Machine learning	
		Cox model	GBM	Random forest
Standard	Demographics, ACC	0.53 (0.49, 0.57)	0.54 (0.47, 0.61)	0.58 (0.49, 0.67)
Plus nutrition variables	Demographics, ACC, HEI	0.53 (0.49, 0.57)	0.54 (0.47, 0.61)	0.59 (0.50, 0.68)
	Demographics, ACC, AHEI	0.53 (0.48, 0.57)	0.54 (0.48, 0.61)	0.60 (0.50, 0.70)
	Demographics, ACC, Med diet score	0.53 (0.49, 0.57)	0.54 (0.47, 0.61)	0.59 (0.51, 0.67)
	Demographics, ACC, DASH diet score	0.52 (0.49, 0.56)	0.54 (0.47, 0.61)	0.60 (0.50, 0.69)
	Demographics, ACC, all 24-hour recall data	0.48 (0.44, 0.53)	0.56 (0.50, 0.62)	1.08 (0.83, 1.33)[1]

[1] Model built using 500 trees; 20-tree model had slope 0.88 (0.69, 1.07)

Supplementary Table F: C-statistics on the held out test data

		Standard	Machine learning	
		Cox model	GBM	Random forest
Standard	Demographics, ACC	0.87 (0.85, 0.88)	0.88 (0.87, 0.89)	0.91 (0.90, 0.93)
	Demographics, ACC, HEI	0.87 (0.85, 0.88)	0.88 (0.87, 0.89)	0.91 (0.90, 0.93)
	Demographics, ACC, AHEI	0.87 (0.85, 0.88)	0.88 (0.87, 0.89)	0.92 (0.90, 0.93)
Plus nutrition variables	Demographics, ACC, Med diet score	0.87 (0.85, 0.88)	0.88 (0.87, 0.89)	0.91 (0.90, 0.92)
	Demographics, ACC, DASH diet score	0.87 (0.85, 0.88)	0.88 (0.87, 0.89)	0.92 (0.90, 0.93)
	Demographics, ACC, all 24-hour recall data	0.87 (0.85, 0.88)	0.88 (0.87, 0.89)	0.93 (0.92, 0.94)[1]

[1]Model built using 500 trees; 20-tree model had C-statistic 0.90 (0.89, 0.92)

Supplementary Table G: Hazard ratios (95% CIs) from Cox models developed on training data. Estimates of hazard ratios and confidence intervals estimated using Rubin's rules, combining results from the 10 imputed training sets. See Supplementary Table A for variable definitions.

	Model 1 (ACC)	**Model 2 (+HEI)**	**Model 3 (+AHEI)**	**Model 4 (+MDS)**	**Model 5 (+DASH)**	**Model 6 (+All)**
age	1.1 (1.09, 1.1)	1.1 (1.09, 1.1)	1.1 (1.09, 1.1)	1.1 (1.09, 1.1)	1.1 (1.09, 1.1)	1.09 (1.09, 1.1)
sex	0.62 (0.55, 0.71)	0.62 (0.55, 0.71)	0.62 (0.55, 0.7)	0.62 (0.55, 0.71)	0.62 (0.55, 0.71)	0.56 (0.49, 0.64)
black	1.06 (0.9, 1.26)	1.07 (0.91, 1.26)	1.08 (0.91, 1.27)	1.07 (0.91, 1.26)	1.05 (0.89, 1.24)	1.03 (0.85, 1.23)
hispanic	0.72 (0.61, 0.86)	0.72 (0.61, 0.86)	0.73 (0.61, 0.86)	0.72 (0.61, 0.86)	0.73 (0.61, 0.86)	0.65 (0.54, 0.79)
total_chol	1 (0.99, 1)	1 (0.99, 1)	1 (0.99, 1)	1 (0.99, 1)	1 (0.99, 1)	1 (0.99, 1)
hdl	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)
sbp	1.01 (1.01, 1.01)	1.01 (1.01, 1.01)	1.01 (1.01, 1.01)	1.01 (1.01, 1.01)	1.01 (1.01, 1.01)	1.01 (1.01, 1.01)
bpmeds	1.22 (1.11, 1.34)	1.22 (1.11, 1.34)	1.22 (1.11, 1.34)	1.22 (1.11, 1.34)	1.21 (1.1, 1.33)	1.24 (1.12, 1.37)
dm	1.46 (1.23, 1.73)	1.48 (1.26, 1.74)	1.47 (1.25, 1.73)	1.48 (1.25, 1.74)	1.46 (1.24, 1.72)	1.38 (1.16, 1.63)
tob	1.82 (1.53, 2.17)	1.82 (1.52, 2.17)	1.8 (1.51, 2.14)	1.82 (1.53, 2.17)	1.78 (1.49, 2.13)	1.72 (1.42, 2.07)
hei		1 (0.99, 1.01)				
ahei			1 (0.99, 1)			
mds				1.02 (0.97, 1.08)		
dash					0.99 (0.98, 1)	
milk_g						1 (1, 1)
cream_g						1 (0.99, 1)
milk_dessert_g						1 (1, 1)
cheese_g						1 (1, 1)
meat_ns_g						1 (0.99, 1.02)
beef_g						1 (1, 1)
pork_g						1 (1, 1)
lamb_g						1 (1, 1)
poultry_g						1 (1, 1)
organ_meat_g						1 (1, 1)
fish_g						1 (0.99, 1)
meat_nonmeat_g						1 (1, 1)
protein_frozen_g						1 (1, 1)
eggs_g						1 (1, 1)
egg_mixture_g						1 (1, 1)
egg_sub_g						1 (0.99, 1)
legumes_g						1 (1, 1)
nuts_g						1 (1, 1)
seeds_g						1 (0.99, 1.01)

	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
carob_g						0.94 (0, ∞)
flour_mix_g						0.39 (0, ∞)
bread_yeast_g						1 (1, 1)
bread_quick_g						1 (1, 1)
pastries_g						1 (1, 1)
crackers_g						1 (1, 1)
pancakes_g						1 (1, 1)
pastas_g						1 (1, 1)
cereals_g						1 (1, 1)
grain_mix_g						1 (1, 1)
meat_sub_g						0.91 (0, ∞)
citrus_g						1 (1, 1)
fruit_dried_g						1 (1, 1.01)
fruit_other_g						1 (1, 1)
fruit_juice_g						1 (1, 1)
fruit_baby_g						1 (0.99, 1.02)
potatoes_g						1 (1, 1)
veg_darkgreen_g						1 (1, 1)
veg_deepyellow_g						1 (1, 1.01)
tomatoes_g						1 (1, 1)
veg_other_g						1 (1, 1)
veg_baby_g						0.8 (0, ∞)
veg_meat_g						1 (1, 1)
veg_mixture_g						1 (1, 1)
fats_g						1 (0.99, 1.01)
oils_g						1.01 (0.99, 1.03)
salad_dressing_g						1 (1, 1.01)
sweets_g						1 (1, 1)
bev_nonalcohol_g						1 (1, 1)
bev_alcohol_g						1 (1, 1)
water_g						1 (1, 1)
kcal						1 (1, 1)
protein_g						1.01 (1, 1.02)
carb_g						1 (1, 1.01)
fiber_g						0.97 (0.96, 0.99)
fat_g						1 (0.97, 1.03)
fat_sat_g						1.06 (0.91, 1.23)
fat_mono_g						1 (0.94, 1.07)
fat_poly_g						1 (0.96, 1.03)
cholesterol_mg						1 (1, 1)
vite_mg						0.99 (0.97, 1.01)
vita_mg						1 (1, 1)
betacaro_mcg						1 (1, 1)
vitb1_mg						1.05 (0.81, 1.35)
vitb2_mg						1.07 (0.85, 1.34)
niacin_mg						0.97 (0.95, 0.99)
vitb6_mg						1.17 (1.02, 1.35)
folate_mcg						1 (1, 1)
vitb12_mcg						1 (0.98, 1.02)
vitc_mg						1 (1, 1)
calcium_mg						1 (1, 1)
phosphorus_mg						1 (1, 1)
magnesium_mg						1 (1, 1)

	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
iron_mg						1 (0.98, 1.02)
zinc_mg						1.01 (1, 1.03)
copper_mg						0.86 (0.66, 1.11)
sodium_mg						1 (1, 1)
potassium_mg						1 (1, 1)
selenium_mcg						1 (1, 1)
caffeine_mg						1 (1, 1)
theobromine_mg						1 (1, 1)
alcohol_gm						1.01 (1, 1.02)
sfa_40_gm						1.4 (0.6, 3.27)
sfa_60_gm						0.58 (0.13, 2.64)
sfa_80_gm						1.2 (0.4, 3.59)
sfa_100_gm						0.75 (0.16, 3.51)
sfa_120_gm						1.01 (0.85, 1.2)
sfa_140_gm						0.9 (0.59, 1.37)
sfa_160_gm						0.95 (0.79, 1.14)
sfa_180_gm						0.96 (0.79, 1.17)
mfa_161h_gm						0.95 (0.71, 1.26)
mfa_161o_gm						1 (0.95, 1.06)
mfa_201_gm						1.12 (0.81, 1.54)
mfa_221_gm						0.67 (0.24, 1.87)
pfa_182_gm						1.04 (0.99, 1.09)
pfa_183_gm						0.84 (0.66, 1.07)
pfa_184_gm						0.05 (0, 39.37)
pfa_204_gm						0.28 (0.05, 1.61)
pfa_205_gm						0.34 (0.04, 2.66)
pfa_225_gm						27.42 (0.19, 3905.43)
pfa_226_gm						2.91 (0.52, 16.29)
water_yesterday_gm						1 (1, 1)

Supplementary Table H: Relative influences of variables in GBM models, averaged across the 10 imputed training sets. See Supplementary Table A for variable definitions.

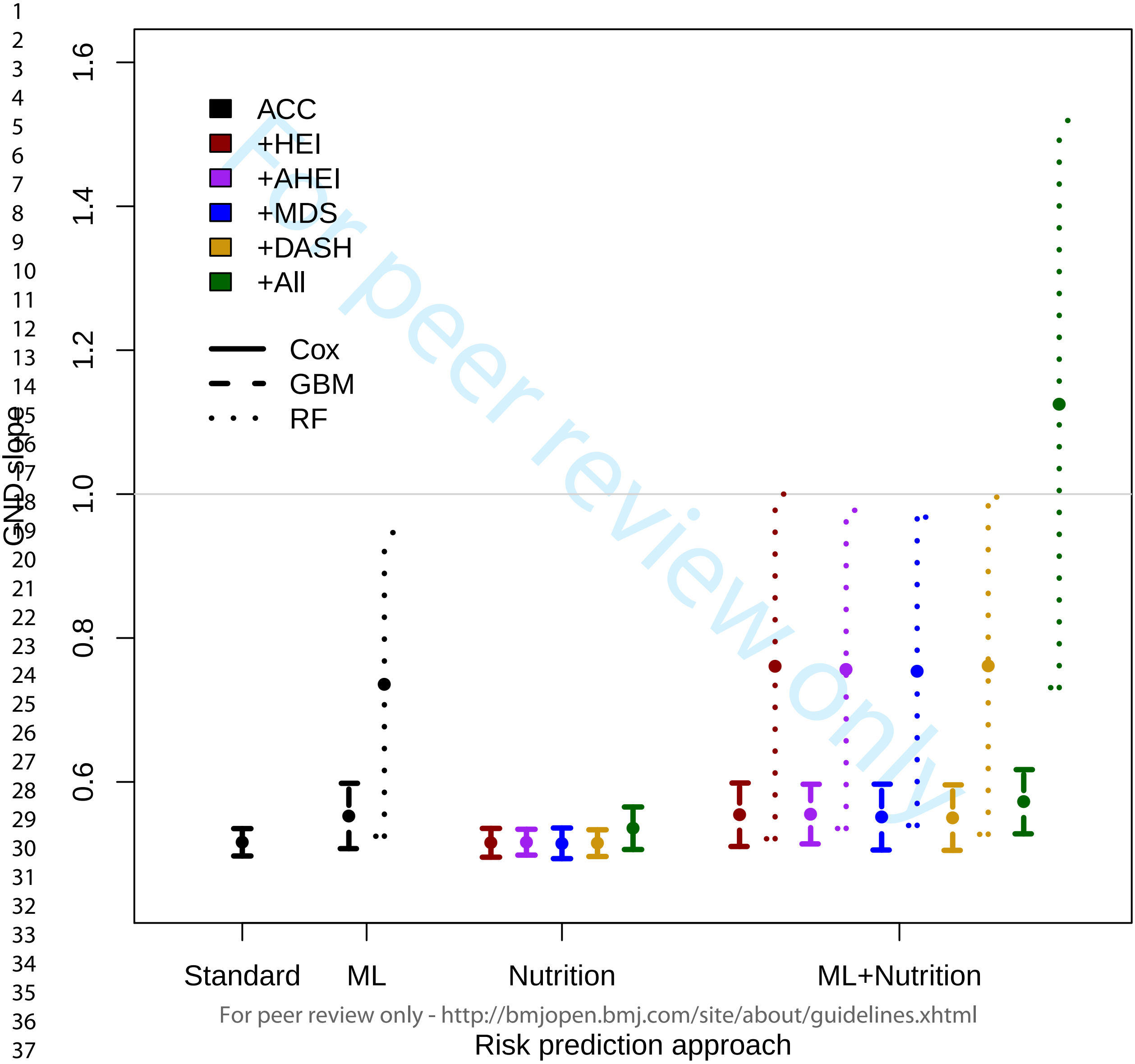
	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
age	70.98	70.79	70.84	71.41	71.02	66.58
sex	2.44	2.38	2.42	2.50	2.32	2.02
black	0.00	0.00	0.00	0.00	0.00	0.00
hispanic	0.01	0.02	0.01	0.00	0.01	0.00
total_chol	3.60	3.48	3.47	3.30	3.60	2.16
hdl	0.42	0.37	0.45	0.41	0.33	0.05
sbp	11.81	10.62	11.83	11.84	11.70	8.42
bpmeds	7.45	7.35	7.32	7.29	7.50	6.49
dm	3.06	2.85	3.11	2.99	2.90	2.61
tob	0.23	0.23	0.27	0.26	0.26	0.00
hei		1.92				
ahei			0.28			
mds				0.00		
dash					0.35	
milk_g						0.08
cream_g						0.09
milk_dessert_g						0.17

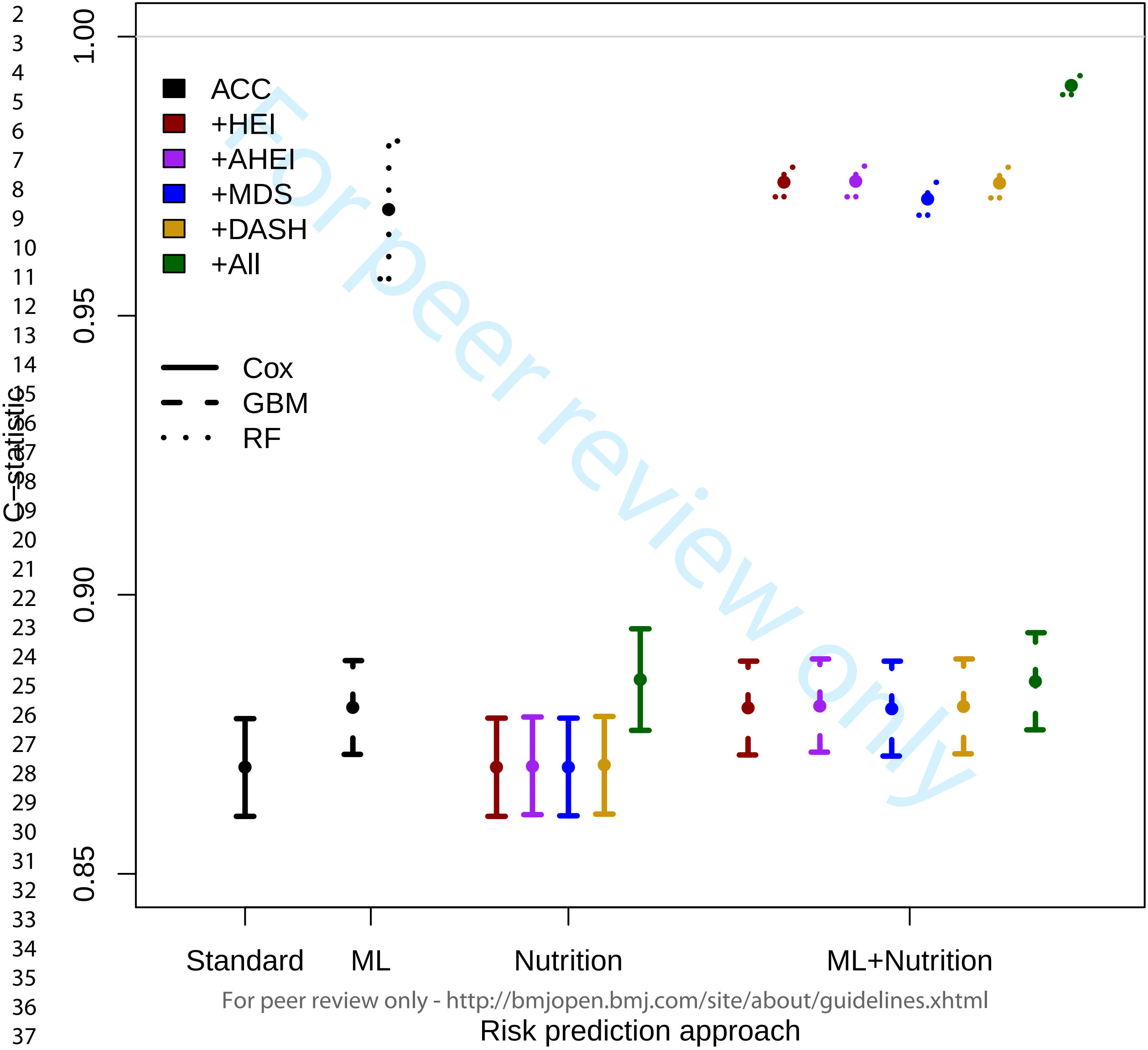
	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
cheese_g						0.00
meat_ns_g						0.29
beef_g						0.00
pork_g						0.14
lamb_g						0.08
poultry_g						0.00
organ_meat_g						0.00
fish_g						0.02
meat_nonmeat_g						0.00
protein_frozen_g						0.00
eggs_g						0.03
egg_mixture_g						0.00
egg_sub_g						0.23
legumes_g						0.12
nuts_g						0.09
seeds_g						0.34
carob_g						0.00
flour_mix_g						0.00
bread_yeast_g						0.16
bread_quick_g						0.03
pastries_g						0.08
crackers_g						0.06
pancakes_g						0.00
pastas_g						0.13
cereals_g						0.00
grain_mix_g						0.00
meat_sub_g						0.00
citrus_g						0.00
fruit_dried_g						0.00
fruit_other_g						0.00
fruit_juice_g						0.00
fruit_baby_g						0.00
potatoes_g						0.00
veg_darkgreen_g						0.02
veg_deepyellow_g						0.00
tomatoes_g						0.06
veg_other_g						0.12
veg_baby_g						0.00
veg_meat_g						0.06
veg_mixture_g						0.00
fats_g						0.15
oils_g						0.24
salad_dressing_g						0.06
sweets_g						0.07
bev_nonalcohol_g						0.00
bev_alcohol_g						0.00
water_g						0.00
kcal						0.29
protein_g						0.44
carb_g						0.55
fiber_g						1.69
fat_g						0.00
fat_sat_g						0.21
fat_mono_g						0.17

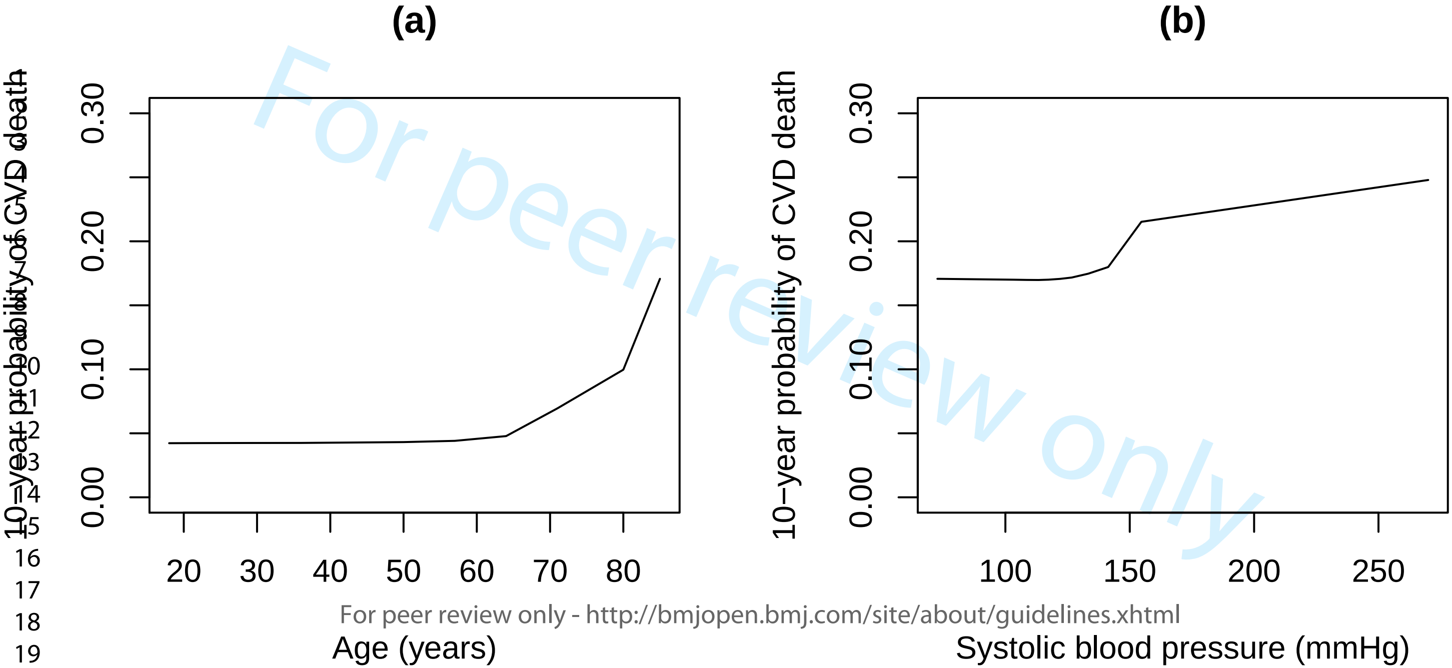
	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
fat_poly_g						0.00
cholesterol_mg						0.00
vite_mg						0.00
vita_mg						0.18
betacaro_mcg						0.19
vitb1_mg						0.05
vitb2_mg						0.02
niacin_mg						0.02
vitb6_mg						0.32
folate_mcg						0.11
vitb12_mcg						0.00
vitc_mg						0.00
calcium_mg						0.23
phosphorus_mg						0.13
magnesium_mg						0.47
iron_mg						0.11
zinc_mg						0.08
copper_mg						0.29
sodium_mg						0.02
potassium_mg						1.82
selenium_mcg						0.09
caffeine_mg						0.00
theobromine_mg						0.00
alcohol_gm						0.02
sfa_40_gm						0.10
sfa_60_gm						0.00
sfa_80_gm						0.07
sfa_100_gm						0.00
sfa_120_gm						0.14
sfa_140_gm						0.02
sfa_160_gm						0.00
sfa_180_gm						0.30
mfa_161h_gm						0.17
mfa_161o_gm						0.35
mfa_201_gm						0.00
mfa_221_gm						0.00
pfa_182_gm						0.00
pfa_183_gm						0.07
pfa_184_gm						0.02
pfa_204_gm						0.00
pfa_205_gm						0.00
pfa_225_gm						0.00
pfa_226_gm						0.04
water_yesterday_gm						0.00

Supplementary Table I: TRIPOD checklist

Title and abstract			**Page number**
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions	2
Introduction			
Background and objectives	3a	Explain the medical context (including diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	4-5
	3b	Specify the objectives, including whether the study describes the development or validation of the model, or both	5
Methods			
Source of data	4a	Describe the study design or sources of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable	6
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up)	6
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centers	6
	5b	Describe eligibility criteria for participants	6
	5c	Give details of treatments received, if relevant	N/A
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	6
	6b	Report any actions to blind assessment of the outcome to be predicted	6
Predictors	7a	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured	6-7, Supp Table A
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors	6
Sample size	8	Explain how the study size was arrived at	7
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	7
Statistical analysis	10a	Describe how predictors were handled in the analysis (D)	6-7
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation (D)	7-8
	10c	For validation, describe how predictions were calculated (V)	7
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models	8-9
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done (V)	10
Risk groups	11	Provide details on how risk groups were created, if done	N/A
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors (V)	N/A
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	10
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including number of participants with missing data for predictors and outcome	10, Table 1
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome) (V)	10, Table 1
Model development	14a	Specify the number of participants and outcome events in each analysis (D)	10-11
	14b	If done, report the unadjusted association between each candidate predictor and outcome (D)	12-13, Supp Table G
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point) (D)	12-13, Supp Table G
	15b	Explain how to use the prediction model (D)	12-13
Model performance	16	Report performance measures (with CIs) for the prediction model	11-13
Model updating	17	If done, report the results from any model updating (i.e., model specification, model performance) (V)	N/A
Discussion			
Limitations	18	Discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data)	14
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data (V)	14
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence	15
Implications	20	Discuss the potential clinical use of the model and implications for future research	15
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets	25-37
Funding	22	Give the source of funding and the role of the funders for the present study	16


GND slope
1.6
1.4
1.2
1.0
0.8
0.6
ACC
+HEI
+AHEI
+MDS
+DASH
+All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach


1.00
0.95
0.90
0.85
C−statistic
ACC
+HEI
+AHEI
+MDS
+DASH
+All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach


(a)
10−year probability of CVD death
0.00
0.10
0.20
0.30
20
30
40
50
60
70
80
Age (years)
(b)
10−year probability of CVD death
0.00
0.10
0.20
0.30
100
150
200
250
Systolic blood pressure (mmHg)

BMJ Open

Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the United States using nationally randomly sampled data

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-032703.R1
Article Type:	Original research
Date Submitted by the Author:	04-Oct-2019
Complete List of Authors:	Rigdon, Joseph ; Wake Forest School of Medicine, Department of Biostatistics and Data Science Basu, Sanjay; Harvard Medical School
<b>Primary Subject Heading</b>:	Cardiovascular medicine
Secondary Subject Heading:	Research methods
Keywords:	Cardiovascular disease, machine learning, Nutrition < TROPICAL MEDICINE, risk prediction

Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the United States using nationally randomly sampled data

Joseph Rigdon, PhD*

Department of Biostatistics and Data Science, Wake Forest School of Medicine,

525@Vine #4403

Winston-Salem, North Carolina 27157

jrigdon@wakehealth.edu

Sanjay Basu, MD, PhD

Center for Primary Care, Harvard Medical School

Research and Analytics, Collective Health

School of Public Health, Imperial College London

635 Huntington Avenue, 2nd floor

Boston, MA 02115

sanjay_basu@hms.harvard.edu

*Denotes corresponding author

Abstract

Objectives: We aimed to test whether or not adding (i) nutrition predictor variables and/or (ii) using machine learning models improves cardiovascular death prediction versus standard Cox models without nutrition predictor variables

Design: Retrospective study

Setting: Six waves of National Health and Nutrition Examination Survey (NHANES) data collected from 1999-2011 linked to the National Death Index (NDI)

Participants: 29,390 participants were included in the training set for model derivation and 12,600 were included in the test set for model evaluation. Our study sample was approximately 20% black race and 25% Hispanic ethnicity.

Primary and Secondary Outcome Measures: Time from NHANES interview until the minimum of time of cardiovascular death or censoring

Results: A standard risk model excluding nutrition data overestimated risk nearly two-fold [calibration slope of predicted versus true risk: 0.53 (95% CI: 0.50, 0.55)] with moderate discrimination [C-statistic: 0.87 (0.86, 0.89)]. Nutrition data alone failed to improve performance while machine learning alone improved calibration to 1.18 (0.92, 1.44) and discrimination to 0.91 (0.90, 0.92). Both together substantially improved calibration [slope: 1.01 (0.76, 1.27)] and discrimination [C-statistic: 0.93 (0.92, 0.94)].

Conclusions: Our results indicate that the inclusion of nutrition data with available machine learning algorithms can substantially improve cardiovascular risk prediction.

Keywords: Cardiovascular disease, machine learning, nutrition, risk prediction

Word Count: 3,475

Strengths and limitations of this study

- Nationally representative data with a comprehensive evaluation of nutrition, direct laboratory assessment of biomarkers, and direct examination of blood pressure
- Comprehensive follow-up with mortality adjudication by cause of death
- Limitations include the need to impute missing data, a short follow-up duration among individuals collected in the later waves of NHANES, and the lack of information about cardiovascular disease (CVD) events in addition to CVD mortality.

Introduction

Nutrition is thought to be a major contributor to cardiovascular disease mortality risk[1–4], but as yet is not explicitly incorporated into cardiovascular risk models that are used to guide clinical prescribing of statins and other preventive medications[5–9]. Nutrition is both imperfectly measured, typically through 24-hour dietary recalls, and nutrition data are sparse and multi-variable, with numerous metrics from individual kilocalorie intakes across a wide range of macro and micronutrients[10,11], making it difficult to determine how an overall nutritional profile might be incorporated into clinical practice. Several groups have offered composite nutrition quality scores (e.g., the Healthy Eating Index and alternatives)[12–14], which correlate to some degree with cardiovascular mortality [15–22] but have not yet been incorporated into common risk equations that use more traditional risk markers (e.g., systolic blood pressure)[5]. Optimizing cardiovascular disease risk prediction is important in clinical practice, because many modern clinical guidelines recommend that physicians prescribe therapies (such as statins, aspirin, and intensive blood pressure treatment) based in part on estimates of overall cardiovascular disease risk, not simply based on the levels of a single biomarker such as cholesterol or blood pressure levels, which fail to fully capture the influence of nutrition on risk [23–26].

With modern machine learning methods, it may be possible to avoid the problems of composite indices, such as reducing a large amount of sparse data to a rough composite that does not explain substantial variation in observed risk[27]. Machine learning approaches are particularly adept at capturing a complex array of large data represented by the sparse matrices of nutrition variables, and incorporating interactions among the data variables (such as between different types of nutrients, e.g., different fats, different carbohydrates, etc.), and identify nonlinear relationships between risk factors and outcomes (e.g., increasing carbohydrate to a very high level from a medium level may differ in impact than increasing from low to medium) that traditional regression models may not fully capture[28–31]. Additionally, with high-quality, more rapid 24-hour dietary recall techniques that can more comprehensively assess a person's dietary behaviors and link them to large nutritional databases, it is now possible to assess nutritional profiles in detail in the clinician's office or clinic waiting room[32–35]. It remains unclear, however, whether nutritional information from a 24-hour recall can add meaningful value to cardiovascular mortality risk prediction beyond biomarker values—such as lipid profile, blood pressure, and diabetes status—and whether using a machine learning approach can advance the predictive power of dietary recalls for cardiovascular risk assessment beyond composite indices already available.

Here, we use a 2-by-2 factorial experimental design to test two hypotheses using observational data: (i) that the data from a single 24-hour dietary recall can add substantial predictive value to cardiovascular mortality risk estimation beyond that afforded by standard biomarkers already included in traditional cardiovascular risk calculators; and (ii) that machine learning approaches to directly incorporate sparse matrices of nutrition data into risk estimates can be superior to standard regression models or the composite nutritional indices constructed through linear modeling methods in the past.

Methods

We conducted a 2-by-2 factorial experiment in which we compared the calibration and discrimination of cardiovascular disease mortality risk prediction models with and without data from a 24-hour dietary recall, and with and without a machine learning approach.

Data Source

Six waves of cross-sectional data from the National Health and Nutrition Examination Survey (NHANES, 1999-2000, 2001-2002, 2003-2004, 2005-2006, 2007-2008, and 2009-2010) were used to develop and validate the risk prediction models. The details of the NHANES sampling scheme are described elsewhere[36]. Briefly, NHANES is a survey including laboratory biomarkers and clinical examination, collected in two-year waves among children and adults, sampled to represent the non-institutionalized civilian U.S. population. Each observation within each wave was linked to the National Death Index (NDI, through 2011) by the Centers for Disease Control. The NDI provided data on the time of CVD death or censoring of follow-up, and additionally a variable attributing death to one of nine-cause specific categories (heart disease, cancer, chronic lower respiratory disease, cerebrovascular diseases, diabetes, pneumonia and influenza, Alzheimer's disease, kidney disease, and unintentional injuries).

The primary statistical outcome was defined as time from NHANES interview to the minimum of time of censoring or time of death from heart disease or cerebrovascular diseases, henceforth CVD mortality. Death from any other cause was treated as censored. Inclusion criteria were age 20-79 years old at time of interview with no prior CVD history. No actions were taken to blind assessment of predictors for the outcome and other predictors. No actions were taken to blind assessment of the outcome.

All potential predictors in the models were collected at time of NHANES interview to mimic a hypothetical scenario where a medical provider may want to conduct an in-clinic 24-hour dietary recall to improve prediction of CVD mortality. Demographic variables included age, sex, and race (Black race, Hispanic ethnicity), and currently-employed cardiovascular disease risk factors of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no)[5]. Nutrition variables included daily standardized intake of micronutrients (e.g., sodium, selenium) and macronutrients (e.g., fat, carbohydrates, protein) collected during a single 24-hour dietary recall following the NHANES interview (Supplementary Table A).

Patient and Public Involvement

No patient involved.

Model Development

Random samples of 70% of each NHANES wave were pooled to form the training sample from which the models were derived, with the remaining 30% prospectively held out to form the test set to assess performance of each model without refitting or recalibration. To train the models in the presence of missing data, multiple imputation via chained equations[37,38] was employed to fill in missing values (Supplementary Table B) so that one complete data set was available.

In one arm of the 2-by-2 design, we tested whether or not switching from the standard Cox proportional hazards model to a machine learning algorithm could improve calibration and discrimination. The machine learning algorithms tested were those commonly used for clinical event risk prediction for censored time-to-event data: survival gradient boosted machines (GBMs)[39] and survival random forests (RFs)[40]. Both of these machine learning approaches construct decision trees from data. In a typical decision tree, each branch of the tree divides the sampled study population into increasingly-smaller subgroups that differ in their probability of the outcome. A good decision tree will separate the sampled population into groups that have low within-group variability and high between-group variability in the probability of the outcome. GBMs average many trees where errors made by the first tree contribute to learning of a less erroneous tree in the next iteration (a “boosting” strategy)[41,42]. RFs also build numerous decision trees, but average a forest composed of many trees, where each tree is independently fitted (a “bagging” strategy) with a random subset of covariates selected to be eligible to define the branches[42–45]. RFs use inverse probability of censoring weights to address censoring.

In the second arm of the 2-by-2 design, we tested whether or not adding nutrition variables, including all micro and macronutrients assessed in the NHANES dietary recall, to the standard demographic and biomarker variables could improve prediction. We additional compare incorporating all nutrition data versus using common existing composite nutrition indices: the Healthy Eating Index (HEI)[46], Alternate Healthy Eating

Index (AHEI)[47], Mediterranean Diet Score (MDS)[48], and the Dietary Approaches to Stop Hypertension diet score (DASH)[49].

In total, our 2-by-2 design contained 18 models in four quadrants. The no machine learning, no nutrition (standard model) quadrant included only one model: a Cox regression model with demographics and biomarker variables. The machine learning, no nutrition quadrant included two models: a gradient boosted machine and a random forest, both using only demographics and biomarker variables. The no machine learning, nutrition quadrant included five models: a Cox regression including demographics, biomarkers, and either HEI, AHEI, MDS, DASH, or all micro and macronutrients from NHANES. Finally, the machine learning, nutrition quadrant included 10 total models: gradient boosted machines or random forests including demographics, biomarkers, and either HEI, AHEI, MDS, DASH, or all micro and macronutrients from NHANES.

Cox regression models, GBM, and RF were fit to the 70% training data. GBMs were tuned via manual grid search over number of trees equal to 100, 300, or 500 and tree depth equal to 1, 5 or 10, with learning rate set to 0.1[50]. RFs based on conditional inference trees[51,52] were tuned via manual grid search over number of trees equal to 100, 300, or 500 and number of input variables randomly sampled at each node equal to 1, 5, or 10. The best performing GBM and RF models were those that minimized in the 30% held-out test set the sum of (i) the squared error between the calibration metric (described below) and the ideal target of 1 and (ii) the squared error between the discrimination metric (described below) and the ideal target of 1.

Outcome metrics

Model performance was assessed in terms of calibration (using the Greenwood-Nam-D'Agostino [GND] test) and discrimination (using the C-statistic). In the GND test, model predicted probability of 10-year CVD mortality risk was compared to observed rates of death from CVD within 10 years after the NHANES interview by decile of predicted risk. A slope and intercept line were then drawn using these values across deciles of predicted risk, such that a calibration slope of 1 reflects perfect calibration (a perfect 45-degree line between predicted and observed risk).

Model discrimination was assessed using the C-statistic (area under receiver operating characteristic [ROC] curve). Each point on the ROC curve was defined by the sensitivity (x-axis) and 1-specificity (y-axis) for a given cutpoint. The calculation of sensitivity and specificity followed from model predicted risk (above/below cutpoint) versus gold standard of outcome (whether or not CVD mortality happened within 10 years after NHANES interview). Confidence intervals for C-statistics were calculated using DeLong's test[53] as implemented in the R package 'pROC'[54].

Sensitivity analyses included (i) adding education and poverty to the best performing model and (ii) applying the best performing model to the component outcomes CVD mortality, heart disease and cerebrovascular diseases, separately. No model updating was done in this study, and no risk groups were created. There were no differences in setting, eligibility criteria, outcome, or predictors between the training (development) set and the test (validation) set. There was no need for participant consent or Ethical Review Board approval as the data are publicly available. All statistical analyses were carried out in Stata 15 software[55] and R version 3.6.1[56].

This manuscript was written in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) recommendations[57], summarized in Supplementary Table C.

Data Availability Statement

Statistical code used for data scraping (from NHANES and NDI websites, as specified in comments in the code), training and test data sets, data management, model fitting, and table and figure creation are available in the following public, open access repository: https://github.com/joerigdon/CVD_Prediction.

Results

Descriptive statistics on the study sample

Distributions of demographics, covariates and outcome rates were nearly equivalent in training and test sets (Table 1). Of the n=29390 individuals in the training set, 1179/29390 (4.0%) experienced CVD mortality within the follow-up period; of the n=12600 in the test set, 507/12600 (4.0%) experienced CVD mortality. The median follow-up time was 79 months in both training and test sets, with a mean age of 50 years, and 47% of the population being male, 20% Black, 26% Hispanic, 16% with diabetes, and 19% actively smoking tobacco. Composite nutrition indices were identical to within rounding error between the train and test datasets, with a mean HEI score of 47 (out of 100[46]), AHEI score of 47 (out of 110[47]), MDS score of 5 (out of 10[48]), and DASH score of 47 (out of 80[49]); higher scores indicate better adherence to the recommended dietary guidelines for all four of the composite scores.

Compared to individuals without CVD mortality, individuals experiencing CVD mortality were older (74.3 vs. 49.0 years old), more likely to be male (55.0% vs. 46.9%), had higher systolic blood pressure (142.9 vs. 124.8 mmHg), were more likely to take blood pressure medications (74.2% vs. 30.8%), and were more likely to have diabetes (33.3% vs. 15.5%; Table 2). Regarding nutrition variables, those experiencing CVD mortality counter-intuitively had a higher HEI score (51.0 vs. 46.9), a higher AHEI score (48.0 vs. 47.1), and a higher DASH score (48.1 vs. 47.4; Table 2), and comparable MDS scores (5.1 vs. 5.1).

Model calibration performance

As expected, model calibration values were better in the training (Supplementary Figure A, Supplementary Tables D, E, F, G, H, I) versus the held-out test set (Figure 1, Supplementary Tables J, K, L, M, N, O). Using the standard approach to CVD risk prediction modeling[5], a Cox proportional hazards model with variables of age, sex, Black race, and Hispanic ethnicity, total cholesterol, HDL cholesterol, systolic blood pressure, blood pressure medication, diabetes, and tobacco use, yielded a GND calibration slope of 0.53 (95% CI: 0.50, 0.55), reflecting profound risk over-estimation consistent with prior estimates[58,59]. Adding HEI, AHEI, MDS, or DASH score to the model did not change the calibration slope of 0.53, however the addition of the raw (not composite) 24-hour recall data decreased the slope to 0.46 (0.43, 0.50), reflecting a worsening of over-estimation of risk (Figure 1, Supplementary Tables J, K, L, M, N, O).

When using a machine learning GBM approach instead of a Cox proportional hazards model, but still excluding nutrition data, model calibration improved to 0.56 (0.51, 0.61), and when using random forest in place of Cox, the calibration improved further to 1.18 (0.92, 1.44). Adding nutrition variables improved the machine learning models' calibration when raw 24-hour recall data were used, but not when composite dietary indices were used. Adding HEI, AHEI, MDS, or DASH slightly improved calibration slope to 0.59 for the GBM models and improved calibration slope for the random forest models from 1.18 to 1.13. The GBM model had the best calibration when using all 24-hour recall data, producing a calibration slope of 0.83 (0.77, 0.89). The random forest model with raw 24-hour nutrition data was the closest to the ideal value of 1, with a calibration slope of 1.01 (0.76, 1.27) (Figure 1, Supplementary Table O).

Model discrimination performance

Model discrimination values were better in the training (Supplementary Figure B, Supplementary Tables D, E, F, G, H, I) versus the held-out test set (Figure 2, Supplementary Tables J, K, L, M, N, O). The exclusion or inclusion of nutrition data did not affect discrimination of the standard Cox risk models. The Cox model with the above-mentioned non-nutrition data had a C-statistic of 0.88 (0.87, 0.89) in the test set. Adding HEI, AHEI, MDS, DASH, or all raw 24-hour recall data left the C-statistic unchanged at 0.88 (Figure 2, Supplementary Tables J, K, L, M, N, O).

Model discrimination also improved with use of machine learning. Using a GBM in place of a Cox model improved discrimination slightly, from C-statistics of 0.88 in Cox models to 0.90 (0.89, 0.91) for all GBM models without nutrition data and 0.91 (0.90, 0.92) for the random forest without nutrition data. The discrimination was not significantly different with the addition of composite nutritional indices, but did improve to 0.93 (0.92, 0.94) with the addition of raw nutrition data (Figure 2, Supplementary Table O).

Important associations

Cox model coefficients are detailed in Supplementary Table P and gradient boosted machine model relative influences are detailed in Supplementary Table Q. Notable associations with cardiovascular death included age (HR for 1-year increase in age of 1.1 [1.09, 1.1], female sex (HR vs. males of 0.65 [0.57, 0.73]), Hispanic ethnicity (HR vs. non-Hispanics of 0.69 [0.58, 0.81]), systolic BP (HR for 1-unit increase of 1.0050 [1.0024, 1.0075]), blood pressure medications (HR for each additional med of 1.19 [1.08, 1.30]), type 2 diabetes (HR vs. non-diabetics of 1.46 [1.29, 1.65]), and tobacco use (HR vs. non-users 1.91 [1.61, 2.27]) (Supplementary Table P). No associations with cardiovascular death were found with HEI or AHEI. A one-unit increase of MDS slightly increased risk: 1.0481 (1.0004, 1.0980), and a one-unit increase in DASH score slightly reduced risk: 0.9870 (0.9806, 0.9935).

In the comprehensive evaluation of all 24-hour nutrition variables, protective associations were seen with fiber (HR 0.96 [0.95, 0.97] for 1-gram increase) and niacin (HR 0.98 [0.96, 0.99] for 1-milligram increase), and harmful association with saturated fat (HR 1.19 [1.07, 1.32] for 1-gram increase). Examining fat intake per one-gram increase more closely, SFA 16:0 intake was protective [0.85 (0.76, 0.94)], as was SFA 18:0 [0.85 (0.75, 0.98)]. MFA 16:1 [1.06 (1.02, 1.10)], and MFA 20:1 [1.32 (1.03, 1.69)] slightly increased risk, as did PFA 18:2 [1.07 (1.04, 1.11)]. MFA 22:1 [0.34 (0.13, 0.90)] and PFA 18:3 [0.80 (0.68, 0.95)] reduced risk.

Relative influences in a GBM display how much of a 0-100 importance total is accounted for by each variable in the model (Supplementary Table Q). Age consistently had relative influences of 20-30, with the exception of Model 3 with AHEI (relative influence 6), and Model 4 with MDS (relative influence 3). SBP had a relative influence of 19-41 in all models except Model 6 with all nutrition variables (relative influence 3). HDL ranged from 10-37 with the exception of Model 4 with AHEI (3) and Model 6 with all nutrition variables (3). Total cholesterol ranged from 13-24 with the exception of Model 6 (2). Tobacco use was unusually influential in Model 3 (46) while remaining below 4 in all other models. HEI was important in Model 1 (14) and DASH in Model 5 (17), whereas relative influences for AHEI and MDS failed to exceed 2. Of the 24-hour nutrition variables, iron, legumes, sweets, and pastries had relative influences of 5 or greater. Partial dependence plots for the random forest model with all nutrition variables reveal an exponential increase in 10-year probability of CVD death starting at about age 65, and a linear increase in risk for 10-year probability of CVD death after 120 mmHg systolic blood pressure (Supplementary Figure C).

Sensitivity Analyses

Adding education and poverty to the best performing model did not substantially improve calibration (1.0120 with vs. 1.0137 without), or discrimination (0.9336 with vs. 0.9320 without). Applying the best performing model separately to death from heart disease yielded calibration slope 0.9670 (0.7525, 1.1814) and discrimination C-statistic 0.9256 (0.9120, 0.9391). Applying the best performing model separately to death from cerebrovascular disease yielded calibration slope 0.7406 (0.5636, 0.9177) and discrimination C-statistic 0.9157 (0.8898, 0.9416).

Discussion

We examined whether or not improvements in CVD mortality prediction could be achieved by including sparse nutrition data into models derived through machine learning algorithms. We observed that the addition of nutrition variables to a standard Cox proportional hazards model was not of substantial benefit alone, machine learning alone improved calibration and moderately improved discrimination, and when both nutrition data and machine learning were combined, we could substantially improve risk prediction beyond the inclusion of standard demographics and biomarkers alone. Calibration particularly improved when both nutrition data and machine learning algorithms were used.

Our findings are of clinical relevance as more rapid, automated or mobile device-based 24-hour dietary recalls make it feasible to provide a nutrition profile for patients at or before visiting a doctor's office[1,2], and as automated cardiovascular disease risk prediction models become an increasingly-important part of precision medicine guidelines that aim to improve the ability of medical practitioners to prescribe preventive cardiovascular treatments to patients with the highest risk[6]. As standard biomarkers fail to explain the full extent to which nutrition relates to cardiovascular mortality[60,61], machine learning approaches that directly incorporate raw dietary data appear to have benefits over composite nutritional indices that may excessively reduce complexity in nutritional interactions and non-linear relationships that confer risk. Our study benefits from being conducted on a nationally representative sample of US adults, including a comprehensive evaluation of nutrition, direct laboratory assessment of biomarkers, direct examination of blood pressure, and comprehensive follow-up with mortality adjudication by cause of death.

Nevertheless, our study has important limitations, including the need to impute missing data, a short follow-up duration among individuals collected in the later waves of NHANES, the lack of information about CVD events in addition to CVD mortality, and the need to assess feasibility of model implementation in practice. In the future, further research can assess whether the performance of rapid dietary recalls and associated cardiovascular risk estimation can be implemented in practice, whether the level of improvements to calibration and discrimination observed in this assessment produce clinically-meaningful changes in the level of prescribing of key preventive therapies for patients, and whether the difficulties of interpreting machine learning models compared to traditional Cox-type risk models poses challenges to the acceptability of these models in clinical practice.

At present, our results indicate that the inclusion of nutrition data with available machine learning algorithms can substantially improve cardiovascular risk prediction.

Author Contributions

SB conceptualized the study and design and contributed to data preparation and analysis. JR contributed to data preparation and analysis. Both authors contributed to writing and critically reviewing the manuscript.

Competing Interests statement

JR and SB have no competing interests to report.

Acknowledgements

The authors acknowledge two anonymous reviewers at the Stanford Quantitative Sciences Unit.

Funding

This work was supported by the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Number DP2MD010478. The

content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Shivappa, N., Steck, S. E., Hussey, J. R., Ma, Y. & Hebert, J. R. Inflammatory potential of diet and all-cause, cardiovascular, and cancer mortality in National Health and Nutrition Examination Survey III Study. *Eur. J. Nutr.* **56**, 683–692 (2017).
2. Aune, D. *et al.* Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—a systematic review and dose-response meta-analysis of prospective studies. *Int. J. Epidemiol.* **46**, 1029–1056 (2017).
3. Wang, D. D. *et al.* Association of Specific Dietary Fats With Total and Cause-Specific Mortality. *JAMA Intern. Med.* **176**, 1134–1145 (2016).
4. Langley‐Evans, S. C. Nutrition in early life and the programming of adult disease: a review. *J. Hum. Nutr. Diet.* **28**, 1–14 (2015).
5. Goff David C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129**, S49–S73 (2014).
6. Stone Neil J. *et al.* 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults. *Circulation* **129**, S1–S45 (2014).
7. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA J. Am. Med. Assoc.* **285**, 2486–2497 (2001).

8. Lloyd-Jones, D. M. *et al.* Prediction of Lifetime Risk for Cardiovascular Disease by Risk Factor Burden at 50 Years of Age. *Circulation* **113**, 791–798 (2006).

9. Yadlowsky, S. *et al.* Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann. Intern. Med.* **169**, 20 (2018).

10. Stumbo, P. Considerations for selecting a dietary assessment system. *J. Food Compos. Anal. Off. Publ. U. N. Univ. Int. Netw. Food Data Syst.* **21**, S13–S19 (2008).

11. Stewart, K. K. & Whitaker, J. R. *Modern Methods of Food Analysis*. (Springer Science & Business Media, 2012).

12. Kennedy, E. T., Ohls, J., Carlson, S. & Fleming, K. The Healthy Eating Index: Design and Applications. *J. Am. Diet. Assoc.* **95**, 1103–1108 (1995).

13. McCullough, M. L. & Willett, W. C. Evaluating adherence to recommended diets in adults: the Alternate Healthy Eating Index. *Public Health Nutr.* **9**, (2006).

14. Panagiotakos, D. B., Pitsavos, C. & Stefanadis, C. Dietary patterns: A Mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk. *Nutr. Metab. Cardiovasc. Dis.* **16**, 559–568 (2006).

15. Reedy, J. *et al.* Higher Diet Quality Is Associated with Decreased Risk of All-Cause, Cardiovascular Disease, and Cancer Mortality among Older Adults. *J. Nutr.* **144**, 881–889 (2014).

16. Onvani, S., Haghighatdoost, F., Surkan, P. J., Larijani, B. & Azadbakht, L. Adherence to the Healthy Eating Index and Alternative Healthy Eating Index dietary patterns and mortality from all causes, cardiovascular disease and cancer: a meta-analysis of observational studies. *J. Hum. Nutr. Diet.* **30**, 216–226 (2017).

17. Fung, T. T. *et al.* Mediterranean diet and incidence and mortality of coronary heart disease and stroke in women. *Circulation* **119**, 1093–1100 (2009).

18. Akbaraly, T. N. *et al.* Alternative Healthy Eating Index and mortality over 18 y of follow-up: results from the Whitehall II cohort. *Am. J. Clin. Nutr.* **94**, 247–253 (2011).

19. Schwingshackl, L. & Hoffmann, G. Diet Quality as Assessed by the Healthy Eating Index, the Alternate Healthy Eating Index, the Dietary Approaches to Stop Hypertension Score, and Health Outcomes: A Systematic Review and Meta-Analysis of Cohort Studies. *J. Acad. Nutr. Diet.* **115**, 780-800.e5 (2015).

20. Kant, A. K. Indexes of Overall Diet Quality: A Review. *J. Am. Diet. Assoc.* **96**, 785–791 (1996).

21. Folsom, A. R., Parker, E. D. & Harnack, L. J. Degree of Concordance With DASH Diet Guidelines and Incidence of Hypertension and Fatal Cardiovascular Disease. *Am. J. Hypertens.* **20**, 225–232 (2007).

22. Fung, T. T. *et al.* Adherence to a DASH-Style Diet and Risk of Coronary Heart Disease and Stroke in Women. *Arch. Intern. Med.* **168**, 713–720 (2008).

23. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* 25709 (2018). doi:10.1016/j.jacc.2018.11.003

24. Bibbins-Domingo, K. *et al.* Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: US Preventive Services Task Force Recommendation Statement. *JAMA* **316**, 1997–2007 (2016).

25. Bibbins-Domingo, K. & on behalf of the U.S. Preventive Services Task Force. Aspirin Use for the Primary Prevention of Cardiovascular Disease and Colorectal Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann. Intern. Med.* **164**, 836 (2016).

26. Whelton, P. K. *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. *J. Am. Coll. Cardiol.* **71**, e127–e248 (2018).

27. Suresh, S., Saraswathi, S. & Sundararajan, N. Performance enhancement of extreme learning machine for multi-category sparse data classification problems. *Eng. Appl. Artif. Intell.* **23**, 1149–1157 (2010).

28. Messina, M., Lampe, J. W., Birt, D. F., Appel, L. J. & al, et. Reductionism and the narrowing nutrition perspective: Time for reevaluation and emphasis on food synergy. *Am. Diet. Assoc. J. Am. Diet. Assoc. Chic.* **101**, 1416–9 (2001).

29. Wang, J., Li, D., Dangott, L. J. & Wu, G. Proteomics and Its Role in Nutrition Research,. *J. Nutr.* **136**, 1759–1762 (2006).

30. Marcos, A., Nova, E. & Montero, A. Changes in the immune system are conditioned by nutrition. *Eur. J. Clin. Nutr.* **57**, S66–S69 (2003).

31. Zeisel, S. H. *et al.* Nutrition: A Reservoir for Integrative Science. *J. Nutr.* **131**, 1319–1321 (2001).

32. Subar, A. F. *et al.* The Automated Self-Administered 24-Hour Dietary Recall (ASA24): A Resource for Researchers, Clinicians, and Educators from the National Cancer Institute. *J. Acad. Nutr. Diet.* **112**, 1134–1137 (2012).

33. Vereecken, C. A., Covents, M., Matthys, C. & Maes, L. Young adolescents' nutrition assessment on computer (YANA-C). *Eur. J. Clin. Nutr.* **59**, 658–667 (2005).

34. Hongu, N. *et al.* Dietary Assessment Tools Using Mobile Technology. *Top. Clin. Nutr.* **26**, 300 (2011).

35. Thompson, F. E. *et al.* Comparison of Interviewer-Administered and Automated Self-Administered 24-Hour Dietary Recalls in 3 Diverse Integrated Health Systems. *Am. J. Epidemiol.* **181**, 970–978 (2015).

36. NHANES - About the National Health and Nutrition Examination Survey. (2017). Available at: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm. (Accessed: 11th March 2019)

37. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).

38. Vergouwe, Y., Royston, P., Moons, K. G. M. & Altman, D. G. Development and validation of a prediction model with missing predictor data: a practical approach. *J. Clin. Epidemiol.* **63**, 205–214 (2010).

39. Chen, Y., Jia, Z., Mercola, D. & Xie, X. A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. *Computational and Mathematical Methods in Medicine* (2013). doi:10.1155/2013/873595

40. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).

41. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

42. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189–1232 (2001).

43. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).

44. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

45. Friedman, J., Hastie, T. & Tibshirani, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann. Stat.* **28**, 337–407 (2000).

46. Guenther, P. M. *et al.* Update of the Healthy Eating Index: HEI-2010. *J. Acad. Nutr. Diet.* **113**, 569–580 (2013).

47. Chiuve, S. E. *et al.* Alternative Dietary Indices Both Strongly Predict Risk of Chronic Disease. *J. Nutr.* **142**, 1009–1018 (2012).

48. Trichopoulou, A., Costacou, T., Bamia, C. & Trichopoulos, D. Adherence to a Mediterranean Diet and Survival in a Greek Population. *N. Engl. J. Med.* **348**, 2599–2608 (2003).

49. Günther, A. L. B. *et al.* Association Between the Dietary Approaches to Hypertension Diet and Hypertension in Youth With Diabetes Mellitus. *Hypertension* **53**, 6–12 (2009).

50. Greenwell, B., Boehmke, B., Cunningham, J. & Developers (https://github.com/gbm-developers), G. B. M. *gbm: Generalized Boosted Regression Models*. (2019).

51. Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. & Van Der Laan, M. J. Survival ensembles. *Biostatistics* **7**, 355–373 (2006).

52. Hothorn, T., Hornik, K., Strobl, C. & Zeileis, A. *party: A Laboratory for Recursive Partytioning*. (2019).

53. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).

54. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

55. StataCorp. *Stata Statistical Software: Release 15*. (StataCorp LLC, 2017).

56. R Core Team. *R: A language and environment for statistical computing*. (2018).

57. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann. Intern. Med.* **162**, 55 (2015).

58. Yadlowsky, S. *et al.* Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann. Intern. Med.* (2018). doi:10.7326/M17-3011

59. Ridker, P. M. & Cook, N. R. Statins: new American guidelines for prevention of cardiovascular disease. *The Lancet* **382**, 1762–1765 (2013).

60. Kant, A. K. Dietary patterns: biomarkers and chronic disease riskThis paper is one of a selection of papers published in the CSCN–CSNS 2009 Conference, entitled Are dietary patterns the best way to make nutrition recommendations for chronic disease prevention? *Appl. Physiol. Nutr. Metab.* **35**, 199–206 (2010).

61. Boushey, C. J., Coulston, A. M., Rock, C. L. & Monsen, E. *Nutrition in the Prevention and Treatment of Disease*. (Elsevier, 2001).

Figure Legends

Figure 1: Calibration slopes and confidence intervals of models in the hold-out test set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Figure 2: Model discrimination (C-statistic) in the hold-out test set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Tables

Table 1: Descriptive statistics on the study sample (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N=41990). Statistics are grouped to reflect participants in the training (n=29390/41990 = 70%) or test (n=12600/41990 = 30%) data subsets. CVD = cardiovascular disease, HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest. Mean (±standard deviation) reported for continuous variables and N (%) reported for categorical variables.

	Training data for model derivation n=29390	Test data for model evaluation n=12600	P-value for difference[1]
CVD death			
No	28,211 (96.0%)	12,093 (96.0%)	0.96
Yes	1,179 (4.0%)	507 (4.0%)	
Heart disease death			
No	28,507 (97.0%)	12,214 (96.9%)	0.76
Yes	883 (3.0%)	386 (3.1%)	
Cerebrovascular death			
No	29,094 (99.0%)	12,479 (99.0%)	0.71
Yes	296 (1.0%)	121 (1.0%)	
Time since interview (months)	79.3 (±41.4)	79.4 (±41.6)	0.84
Wave			
99-00	3,810 (13.0%)	1,633 (13.0%)	1.0
01-02	8,853 (30.1%)	3,795 (30.1%)	
03-04	3,926 (13.4%)	1,684 (13.4%)	
05-06	3,891 (13.2%)	1,669 (13.2%)	
07-08	4,353 (14.8%)	1,866 (14.8%)	
09-10	4,557 (15.5%)	1,953 (15.5%)	
Age	50.0 (±20.4)	50.1 (±20.6)	0.60
Sex			
Male	13,924 (47.4%)	5,887 (46.7%)	0.22
Female	15,466 (52.6%)	6,713 (53.3%)	
Black			
No	14,807 (50.4%)	6,335 (50.3%)	0.94
Yes	5,882 (20.0%)	2,511 (19.9%)	
Missing	8,701 (29.6%)	3,754 (29.8%)	
Hispanic			
No	21,871 (74.4%)	9,359 (74.3%)	0.77
Yes	7,519 (25.6%)	3,241 (25.7%)	
Education level			
<9th	3,942 (13.4%)	1,756 (13.9%)	0.087
9-11	4,538 (15.4%)	1,954 (15.5%)	
HS degree	6,543 (22.3%)	2,716 (21.6%)	
Some college or Associate's	7,138 (24.3%)	2,986 (23.7%)	
College degree	5,061 (17.2%)	2,268 (18.0%)	
Missing	2,168 (7.4%)	920 (7.3%)	
Ratio of family income to poverty threshold	2.5 (±1.6)	2.5 (±1.6)	0.59
Missing	2,655 (9.0%)	1,109 (8.8%)	
Total chol	198.0 (±43.1)	198.0 (±43.9)	0.86
Missing	3,641 (12.4%)	1,484 (11.8%)	

HDL	45.5 (±23.0)	45.6 (±23.0)	0.36
Missing	3,643 (12.4%)	1,484 (11.8%)	
SBP	125.4 (±20.6)	125.6 (±21.1)	0.38
Missing	3,175 (10.8%)	1,348 (10.7%)	
DBP	69.9 (±12.6)	69.8 (±12.7)	0.50
Missing	3,374 (11.5%)	1,431 (11.4%)	
Number of blood pressure medications			
0	19,892 (67.7%)	8,436 (67.0%)	0.32
1	7,851 (26.7%)	3,452 (27.4%)	
2 or more	1,647 (5.6%)	712 (5.7%)	
Type 2 diabetes			
No	10,537 (35.9%)	4,541 (36.0%)	0.42
Yes	4,783 (16.3%)	2,008 (15.9%)	
Missing	14,070 (47.9%)	6,051 (48.0%)	
Smoking			
No	23,774 (80.9%)	10,185 (80.8%)	0.90
Yes	5,615 (19.1%)	2,414 (19.2%)	
Missing	1 (0.0%)	1 (0.0%)	
HEI	47.0 (±11.0)	47.2 (±11.0)	0.28
Missing	3,277 (11.2%)	1,361 (10.8%)	
AHEI	47.1 (±11.1)	47.1 (±11.0)	0.76
Missing	3,263 (11.1%)	1,353 (10.7%)	
MDS	5.1 (±1.2)	5.1 (±1.2)	0.095
Missing	3,270 (11.1%)	1,368 (10.9%)	
DASH	47.4 (±9.3)	47.4 (±9.4)	0.75
Missing	8,835 (30.1%)	3,661 (29.1%)	

[1] Wilcoxon rank sum test for continuous variables, e.g., age, and Fisher's exact test for categorical variables, e.g., black race

Table 2: Comparisons of participant characteristics by outcome (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N=41990). Descriptive summary of variables in those participants without CVD event (n=40304) vs. those with a CVD event (n=1686) during the follow-up period. Mean (±standard deviation) reported for continuous variables and N (%) reported for categorical variables.

	No CVD	**CVD**	**P-value for difference**[1]
	n=40304	n=1686	
Time since interview (months)	80.3 (±41.4)	55.7 (±34.9)	<0.0001
Wave			
99-00	5,168 (12.8%)	275 (16.3%)	<0.0001
01-02	11,681 (29.0%)	967 (57.4%)	
03-04	5,401 (13.4%)	209 (12.4%)	

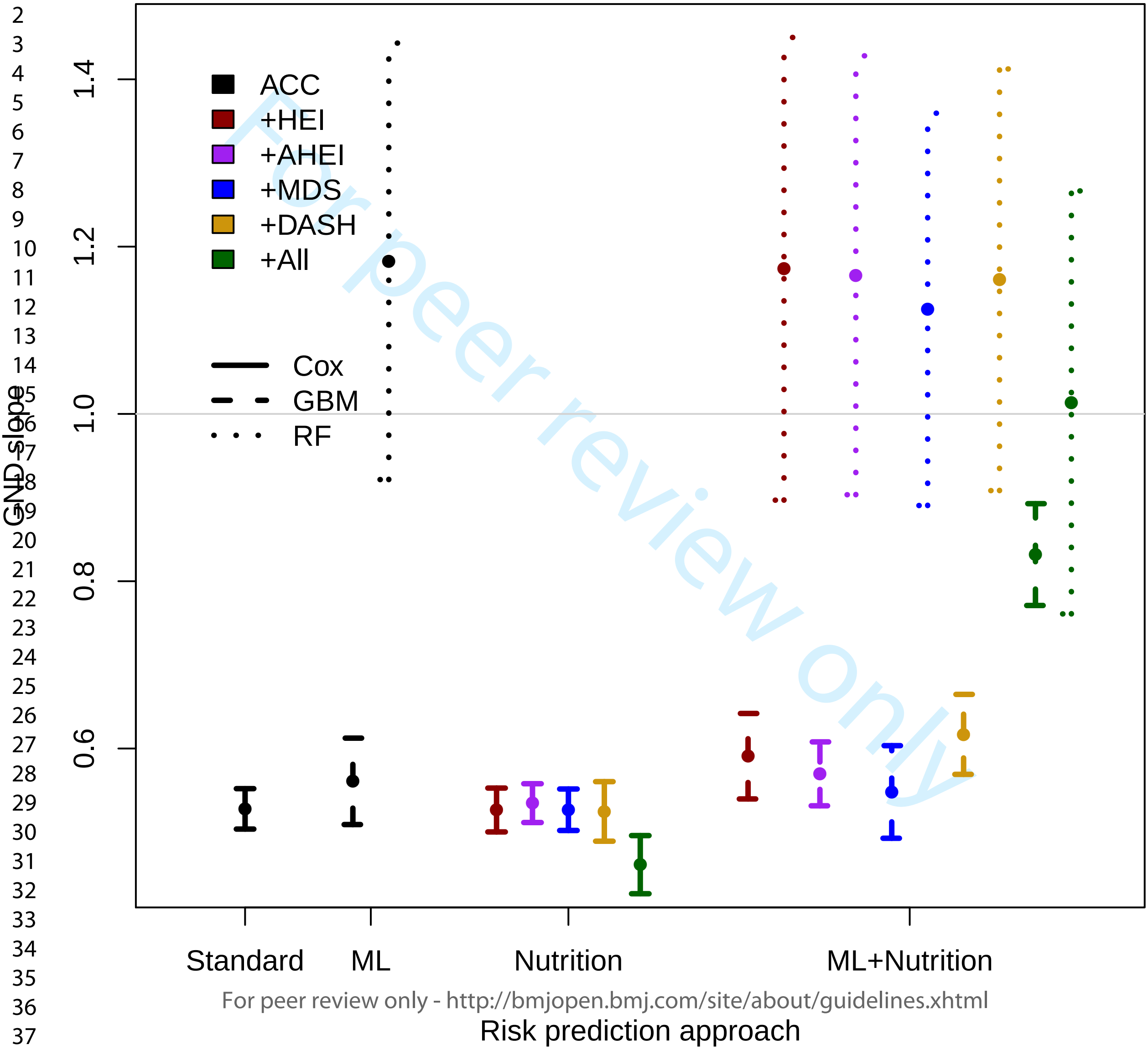
	No CVD	CVD	P-value for difference[1]
05-06	5,451 (13.5%)	109 (6.5%)	
07-08	6,127 (15.2%)	92 (5.5%)	
09-10	6,476 (16.1%)	34 (2.0%)	
Age	49.0 (±20.1)	74.3 (±11.9)	<0.0001
Sex			
Male	18,883 (46.9%)	928 (55.0%)	<0.0001
Female	21,421 (53.1%)	758 (45.0%)	
Black			
No	20,005 (49.6%)	1,137 (67.4%)	<0.0001
Yes	8,110 (20.1%)	283 (16.8%)	
Missing	12,189 (30.2%)	266 (15.8%)	
Hispanic			
No	29,781 (73.9%)	1,449 (85.9%)	<0.0001
Yes	10,523 (26.1%)	237 (14.1%)	
Education level			
<9th	5,223 (13.0%)	475 (28.2%)	<0.0001
9-11	6,201 (15.4%)	291 (17.3%)	
HS degree	8,923 (22.1%)	336 (19.9%)	
Some college or Associate's	9,776 (24.3%)	348 (20.6%)	
College degree	7,111 (17.6%)	218 (12.9%)	
Missing	3,070 (7.6%)	18 (1.1%)	
Ratio of family income to poverty threshold	2.5 (±1.6)	2.1 (±1.4)	<0.0001
Missing	3,565 (8.8%)	199 (11.8%)	
Total chol	198.1 (±43.2)	196.2 (±47.0)	0.10
Missing	4,670 (11.6%)	455 (27.0%)	
HDL	45.5 (±23.0)	45.0 (±24.2)	0.002
Missing	4,672 (11.6%)	455 (27.0%)	
SBP	124.8 (±20.3)	142.9 (±26.8)	<0.0001
Missing	4,114 (10.2%)	409 (24.3%)	
DBP	70.0 (±12.5)	67.5 (±14.7)	<0.0001
Missing	4,359 (10.8%)	446 (26.5%)	
Number of blood pressure medications			
0	27,894 (69.2%)	434 (25.7%)	<0.0001
1	10,205 (25.3%)	1,098 (65.1%)	
2 or more	2,205 (5.5%)	154 (9.1%)	
Type 2 diabetes			
No	14,680 (36.4%)	398 (23.6%)	<0.0001
Yes	6,229 (15.5%)	562 (33.3%)	
Missing	19,395 (48.1%)	726 (43.1%)	

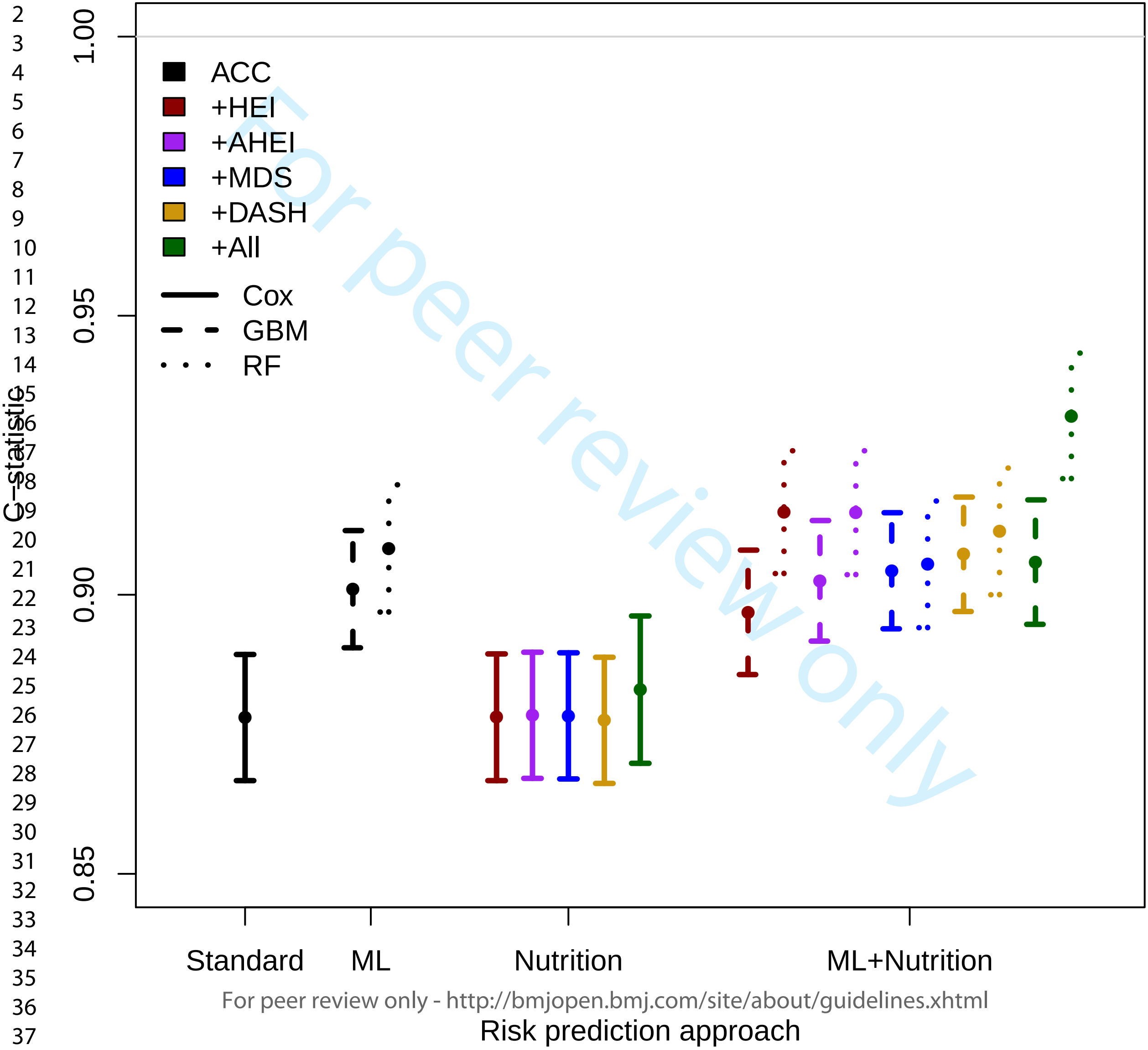
	No CVD	CVD	P-value for difference[1]
Smoking			
No	32,508 (80.7%)	1,451 (86.1%)	<0.0001
Yes	7,794 (19.3%)	235 (13.9%)	
Missing	2 (0.0%)	0 (0.0%)	
HEI	46.9 (±11.0)	51.0 (±10.3)	<0.0001
Missing	4,179 (10.4%)	459 (27.2%)	
AHEI	47.1 (±11.1)	48.0 (±10.9)	0.006
Missing	4,158 (10.3%)	458 (27.2%)	
MDS	5.1 (±1.2)	5.1 (±1.2)	0.10
Missing	4,472 (11.1%)	166 (9.8%)	
DASH	47.4 (±9.4)	48.1 (±9.2)	0.01
Missing	11,774 (29.2%)	722 (42.8%)	

	No CVD	CVD	P-value for difference[1]
	n=40304	n=1686	
Time since interview (months)	80.3 (±41.4)	55.7 (±34.9)	<0.0001
Wave			
99-00	5,168 (12.8%)	275 (16.3%)	<0.0001
01-02	11,681 (29.0%)	967 (57.4%)	
03-04	5,401 (13.4%)	209 (12.4%)	
05-06	5,451 (13.5%)	109 (6.5%)	
07-08	6,127 (15.2%)	92 (5.5%)	
09-10	6,476 (16.1%)	34 (2.0%)	
Age	49.0 (±20.1)	74.3 (±11.9)	<0.0001
Sex			
Male	18,883 (46.9%)	928 (55.0%)	<0.0001
Female	21,421 (53.1%)	758 (45.0%)	
Black			
No	20,005 (49.6%)	1,137 (67.4%)	<0.0001
Yes	8,110 (20.1%)	283 (16.8%)	
Missing	12,189 (30.2%)	266 (15.8%)	
Hispanic			
No	29,781 (73.9%)	1,449 (85.9%)	<0.0001
Yes	10,523 (26.1%)	237 (14.1%)	
Education level			
<9th	5,223 (13.0%)	475 (28.2%)	<0.0001
9-11	6,201 (15.4%)	291 (17.3%)	
HS degree	8,923 (22.1%)	336 (19.9%)	
Some college or Associate’s	9,776 (24.3%)	348 (20.6%)	

	No CVD	CVD	P-value for difference[1]
College degree	7,111 (17.6%)	218 (12.9%)	
Missing	3,070 (7.6%)	18 (1.1%)	
Ratio of family income to poverty threshold	2.5 (±1.6)	2.1 (±1.4)	<0.0001
Missing	3,565 (8.8%)	199 (11.8%)	
Total chol	198.1 (±43.2)	196.2 (±47.0)	0.10
Missing	4,670 (11.6%)	455 (27.0%)	
HDL	45.5 (±23.0)	45.0 (±24.2)	0.002
Missing	4,672 (11.6%)	455 (27.0%)	
SBP	124.8 (±20.3)	142.9 (±26.8)	<0.0001
Missing	4,114 (10.2%)	409 (24.3%)	
DBP	70.0 (±12.5)	67.5 (±14.7)	<0.0001
Missing	4,359 (10.8%)	446 (26.5%)	
Number of blood pressure medications			
0	27,894 (69.2%)	434 (25.7%)	<0.0001
1	10,205 (25.3%)	1,098 (65.1%)	
2 or more	2,205 (5.5%)	154 (9.1%)	
Type 2 diabetes			
No	14,680 (36.4%)	398 (23.6%)	<0.0001
Yes	6,229 (15.5%)	562 (33.3%)	
Missing	19,395 (48.1%)	726 (43.1%)	
Smoking			
No	32,508 (80.7%)	1,451 (86.1%)	<0.0001
Yes	7,794 (19.3%)	235 (13.9%)	
Missing	2 (0.0%)	0 (0.0%)	
HEI	46.9 (±11.0)	51.0 (±10.3)	<0.0001
Missing	4,179 (10.4%)	459 (27.2%)	
AHEI	47.1 (±11.1)	48.0 (±10.9)	0.006
Missing	4,158 (10.3%)	458 (27.2%)	
MDS	5.1 (±1.2)	5.1 (±1.2)	0.10
Missing	4,472 (11.1%)	166 (9.8%)	
DASH	47.4 (±9.4)	48.1 (±9.2)	0.01
Missing	11,774 (29.2%)	722 (42.8%)	

[1] Wilcoxon rank sum test for continuous variables, e.g., age, and Fisher's exact test for categorical variables, e.g., black race


GND slope
1.4
1.2
1.0
0.8
0.6
■ ACC
■ +HEI
■ +AHEI
■ +MDS
■ +DASH
■ +All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach


1.00
0.95
0.90
0.85
C−statistic
ACC
+HEI
+AHEI
+MDS
+DASH
+All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach

Supplementary Appendix

Figure Legends

Supplementary Figure A: Calibration slopes and confidence intervals of models in training set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Supplementary Figure B: Model discrimination (C-statistic) in training set (National Health and Nutrition Examination Survey, 1999-2010 linked to the 2011 National Death Index, N= 12600). All models included demographic variables age, sex, and race (Black race, Hispanic ethnicity). ACC=American College of Cardiology covariates of total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL; mg/dL), systolic blood pressure (mmHg), blood pressure treatment status (yes/no), diabetes status (yes/no), and current smoking status (yes/no), HEI=healthy eating index, AHEI=alternative healthy eating index, MDS=Mediterranean diet score, DASH=dietary approaches to stop hypertension diet score, GBM=gradient boosted machine, RF=random forest

Supplementary Figure C: Partial dependence plots for best model (100 trees, interaction depth 5 using demographics, ACC variables, and full nutrition profile) for (a) age and (b) systolic blood pressure. Plots estimated by averaging model predictions for by decile of age or SBP.

Supplementary Table A: List of all predictor variables included in statistical models

Variable name	**Definition**
Demographic and risk factors (4)	
age	Age in years
sex	Sex (0 if male, 1 if female)
black	Black race (0 if no, 1 if yes)
hispanic	Hispanic ethnicity (0 if no, 1 if yes)
ACC covariates (7)	
total_chol	Total cholesterol (mg/dL)
hdl	HDL cholesterol (mg/dL)
sbp	Systolic blood pressure (mmHg)
dbp	Diastolic blood pressure (mmHg)
bpmeds	Number of blood pressure medications
dm	Type 2 diabetes (0 if no, 1 if yes)
tob	Current smoking (0 if no, 1 if yes)
Composite nutrition variables (4)	
hei	Healthy eating index (0-100)
ahei	Alternative healthy eating index (0-110)
mds	Mediterranean diet score (0-9)
dash	DASH diet score (0-80)
24-hour recall variables (103)	
milk_g	Milk and milk drinks (g)
cream_g	Creams and cream substitutes (g)
milk_dessert_g	Milk desserts, sauces, gravies (g)
cheese_g	Cheeses (g)
meat_ns_g	Meat, not specified as to type (g)
beef_g	Beef (g)
pork_g	Pork (g)
lamb_g	Lamb, veal, game, other carcass meat (g)
poultry_g	Poulty (g)
organ_meat_g	Organ meats, sausages, and lunchmeats, and meat spreads (g)
fish_g	Fish and shellfish (g)
meat_nonmeat_g	Meat, poultry, fish with nonmeat items (g)
protein_frozen_g	Proetin and shelf-stable plate meals, soups, and gravies with meat, poulty fish base; gelatin and gelatin-based drinks
eggs_g	Eggs (g)
egg_mixture_g	Egg mixtures (g)
egg_sub_g	Egg substitutes (g)
egg_frozen_g	Frozen plate meals with egg as major ingredient (g)
legumes_g	Legumes (g)
nuts_g	Nuts, nut butters, and nut mixtures (g)
seeds_g	Seeds and seed mixtures (g)
carob_g	Carob products (g)
flour_mix_g	Flour and dry mixes (g)
bread_yeast_g	Yeast breads, rolls (g)
bread_quick_g	Quick breads (g)
pastries_g	Cakes, cookies, pies, pastries, bars (g)

crackers_g	Crackers and salty snacks from grain products (g)
pancakes_g	Pancakes, waffles, French toast, other grain products (g)
pastas_g	Pastas, cooked cereals, rice (g)
cereals_g	Cereals, not cooked or not specified as to cooked (g)
grain_mix_g	Grain mixtures, frozen plate meals, soups (g)
meat_sub_g	Meat substitutes, mainly cereal protein (g)
citrus_g	Citrus fruits, juices (g)
fruit_dried_g	Dried fruits (g)
fruit_other_g	Other fruits (g)
fruit_juice_g	Fruit juices and nectars excluding citrus (g)
fruit_baby_g	Fruit and juices baby food (g)
potatoes_g	White potatoes and Puerto Rican starchy vegetables (g)
veg_darkgreen_g	Dark-green vegetables (g)
veg_deepyellow_g	Deep-yellow vegetables (g)
tomatoes_g	Tomatoes and tomato mixtures (g)
veg_other_g	Other vegetables (g)
veg_baby_g	Vegetables and mixtures mostly vegetables baby food (g)
veg_meat_g	Vegetables with meat, poultry, fish (g)
veg_mixture_g	Mixtures mostly vegetables without meat, poultry, fish (g)
fats_g	Fats (g)
oils_g	Oils (g)
salad_dressing_g	Salad dressings (g)
sweets_g	Sugars and sweets (g)
bev_nonalcohol_g	Nonalcoholic beverages (g)
bev_alcohol_g	Alcoholic beverages (g)
water_g	Water, noncarbonated (g)
bev_nutrition_g	Formulated nutrition beverages, energy drinks, sports drinks, functional beverages (g)
kcal	Energy (kcal)
protein_g	Protein (g)
carb_g	Carbohydrates (g)
fiber_g	Fiber (g)
fat_g	Fat (g)
fat_sat_g	Saturated fats (g)
fat_mono_g	Monounsaturated fats (g)
fat_poly_g	Polyunsaturated fats (g)
cholesterol_mg	Cholesterol (mg)
vite_mg	Vitamin-E as alpha-tocopherol (mg)
vita_mcg	Vitamin A, RAE (mcg)
betacaro_mcg	Beta-carotene (mcg)
vitb1_mg	Thiamin (Vitamin B1) (mg)
vitb2_mg	Riboflavin (Vitamin B2) (mg)

niacin_mg	Niacin (mg)
vitb6_mg	Vitamin B6 (mg)
folate_mcg	Total folate (mcg)
vitb12_mcg	Vitamin B12 (mcg)
vitc_mg	Vitamin C (mg)
calcium_mg	Calcium (mg)
phosphorus_mg	Phosphorus (mg)
magnesium_mg	Magnesium (mg)
iron_mg	Iron (mg)
zinc_mg	Zing (mg)
copper_mg	Copper (mg)
sodium_mg	Sodium (mg)
potassium_mg	Potassium (mg)
selenium_mcg	Selenium (mg)
caffeine_mg	Caffeine (mg)
theobromine_mg	Theobromine (mg)
alcohol_gm	Alcohol (gm)
sfa_40_gm	SFA 4:0 (Butanoic) (g)
sfa_60_gm	SFA 6:0 (Hexanoic) (g)
sfa_80_gm	SFA 8:0 (Octanoic) (g)
sfa_100_gm	SFA 10:0 (Decanoic) (g)
sfa_120_gm	SFA 12:0 (Dodecanoic) (g)
sfa_140_gm	SFA 14:0 (Tetradecanoic) (g)
sfa_160_gm	SFA 16:0 (Hexadecanoic) (g)
sfa_180_gm	SFA 18:0 (Octadecanoic) (g)
mfa_161h_gm	MFA 16:1 (Hexadecanoic) (g)
mfa_161o_gm	MFA 16:1 (Octadecanoic) (g)
mfa_201_gm	MFA 20:1 (Eicosenoic) (g)
mfa_221_gm	MFA 22:1 (Docosenoic) (g)
pfa_182_gm	PFA 18:2 (Octadecadienoic) (g)
pfa_183_gm	PFA 18:3 (Octadecatrienoic) (g)
pfa_184_gm	PFA 18:4 (Octadecatatraenoic) (g)
pfa_204_gm	PFA 20:4 (Eicosatetraenoic) (g)
pfa_205_gm	PFA 20:5 (Eicosapentaenoic) (g)
pfa_225_gm	PFA 22:5 (Docosapentaenoic) (g)
pfa_226_gm	PFA 22:6 (Docosahexaenoic) (g)
water_yesterday_gm	Total plain water drank yesterday (g)

Supplementary Table B: Percentage of missing data for variables included in analysis

Variable	Percentage missing
milk_g	10.99
cream_g	10.99
milk_dessert_g	10.99
cheese_g	10.99
meat_ns_g	10.99
beef_g	10.99
pork_g	10.99
lamb_g	10.99
poultry_g	10.99
organ_meat_g	10.99
fish_g	10.99
meat_nonmeat_g	10.99
protein_frozen_g	10.99
eggs_g	10.99
egg_mixture_g	10.99
egg_sub_g	10.99
egg_frozen_g	10.99
legumes_g	10.99
nuts_g	10.99
seeds_g	10.99
carob_g	10.99
flour_mix_g	10.99
bread_yeast_g	10.99
bread_quick_g	10.99
pastries_g	10.99
crackers_g	10.99
pancakes_g	10.99
pastas_g	10.99
cereals_g	10.99
grain_mix_g	10.99
meat_sub_g	10.99
citrus_g	10.99
fruit_dried_g	10.99
fruit_other_g	10.99
fruit_juice_g	10.99
fruit_baby_g	10.99
potatoes_g	10.99
veg_darkgreen_g	10.99
veg_deepyellow_g	10.99
tomatoes_g	10.99
veg_other_g	10.99
veg_baby_g	10.99
veg_meat_g	10.99
veg_mixture_g	10.99

Variable	Percentage missing
fats_g	10.99
oils_g	10.99
salad_dressing_g	10.99
sweets_g	10.99
bev_nonalcohol_g	10.99
bev_alcohol_g	10.99
water_g	10.99
bev_nutrition_g	10.99
permth_int	0.00
bpmeds	0.00
kcal	10.98
protein_g	10.98
carb_g	10.98
fiber_g	10.98
fat_g	10.98
fat_sat_g	10.98
fat_mono_g	10.98
fat_poly_g	10.98
cholesterol_mg	10.98
vite_mg	10.98
vita_mg	10.98
betacaro_mcg	10.98
vitb1_mg	10.98
vitb2_mg	10.98
niacin_mg	10.98
vitb6_mg	10.98
folate_mcg	10.98
vitb12_mcg	10.98
vitc_mg	10.98
calcium_mg	10.98
phosphorus_mg	10.98
magnesium_mg	10.98
iron_mg	10.98
zinc_mg	10.98
copper_mg	10.98
sodium_mg	10.98
potassium_mg	10.98
selenium_mcg	10.98
caffeine_mg	10.98
theobromine_mg	10.98
alcohol_gm	10.98
sfa_40_gm	10.98
sfa_60_gm	10.98
sfa_80_gm	10.98
sfa_100_gm	10.98

Variable	Percentage missing
sfa_120_gm	10.98
sfa_140_gm	10.98
sfa_160_gm	10.98
sfa_180_gm	10.98
mfa_161h_gm	10.98
mfa_161o_gm	10.98
mfa_201_gm	10.98
mfa_221_gm	10.98
pfa_182_gm	10.98
pfa_183_gm	10.98
pfa_184_gm	10.98
pfa_204_gm	10.98
pfa_205_gm	10.98
pfa_225_gm	10.98
pfa_226_gm	10.98
water_yesterday_gm	10.82
age	0.00
sex	0.00
black	29.66
hispanic	0.00
sbp	10.77
tob	0.00
hdl	12.21
total_chol	12.21
pov	8.96
dm	47.92
cvdevent	0.00
hd	0.00
cereb	0.00
educ2	7.35
hei	11.05
ahei	10.99
mds	11.05
dash	29.76

Supplementary Table C: TRIPOD checklist

Title and abstract			**Page number**
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions	2
Introduction			
Background and objectives	3a	Explain the medical context (including diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	4-5
	3b	Specify the objectives, including whether the study describes the development or validation of the model, or both	4-5
Methods			
Source of data	4a	Describe the study design or sources of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable	5
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up)	5
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centers	5
	5b	Describe eligibility criteria for participants	6
	5c	Give details of treatments received, if relevant	N/A
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	6
	6b	Report any actions to blind assessment of the outcome to be predicted	6
Predictors	7a	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured	6, Supp Table A
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors	6
Sample size	8	Explain how the study size was arrived at	7
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	7
Statistical analysis	10a	Describe how predictors were handled in the analysis (D)	6-7
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation (D)	7-8
	10c	For validation, describe how predictions were calculated (V)	9
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models	8-9
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done (V)	9
Risk groups	11	Provide details on how risk groups were created, if done	N/A
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors (V)	N/A
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	10
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including number of participants with missing data for predictors and outcome	10, Table 1
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome) (V)	10, Table 1
Model development	14a	Specify the number of participants and outcome events in each analysis (D)	10-11
	14b	If done, report the unadjusted association between each candidate predictor and outcome (D)	12-13, Supp Table P
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point) (D)	12-13, Supp Table P, GitHub repository
	15b	Explain how to use the prediction model (D)	12-13
Model performance	16	Report performance measures (with CIs) for the prediction model	11-13
Model updating	17	If done, report the results from any model updating (i.e., model specification, model performance) (V)	N/A
Discussion			
Limitations	18	Discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data)	15
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data (V)	14-15
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence	15-16
Implications	20	Discuss the potential clinical use of the model and implications for future research	15-16
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets	10
Funding	22	Give the source of funding and the role of the funders for the present study	16

Supplementary Table D: *Internal* validation results from models including demographic and ACC variables only. Criteria is equal to $(\text{slope}-1)^2 + (\text{C-statistic}-1)^2$.

	Intercept 95% CI	**Slope 95% CI**	**C-Statistic 95% CI**	**Criteria**
Cox	0.0011	0.5144	0.8607	0.2552
	-0.0016	0.4941	0.8517	
	0.0038	0.5348	0.8698	
GBM: 100, 1	-0.0004	0.5415	0.8761	0.2256
	-0.0070	0.4919	0.8680	
	0.0061	0.5910	0.8842	
GBM: 100, 5	-0.0022	0.5550	0.8990	0.2082
	-0.0044	0.5399	0.8912	
	0.0000	0.5702	0.9068	
GBM: 100, 10	-0.0039	0.5678	0.9163	0.1938
	-0.0106	0.5237	0.9088	
	0.0029	0.6118	0.9238	
GBM: 300, 1	0.0005	0.5388	0.8747	0.2284
	-0.0070	0.4847	0.8664	
	0.0079	0.5930	0.8831	
GBM: 300, 5	-0.0014	0.5436	0.8963	0.2191
	-0.0050	0.5186	0.8884	
	0.0023	0.5687	0.9042	
GBM: 300, 10	-0.0038	0.5719	0.9140	0.1907
	-0.0068	0.5514	0.9065	
	-0.0007	0.5924	0.9215	
GBM: 500, 1	-0.0004	0.5401	0.8767	0.2267
	-0.0070	0.4908	0.8685	
	0.0062	0.5894	0.8849	
GBM: 500, 5	-0.0014	0.5493	0.8985	0.2134
	-0.0042	0.5295	0.8907	
	0.0015	0.5691	0.9063	
GBM: 500, 10	-0.0020	0.5488	0.9113	0.2114
	-0.0052	0.5279	0.9037	
	0.0012	0.5696	0.9189	
RF: 100, 1	-0.0462	1.3190	0.9210	0.1080
	-0.0824	0.8935	0.9140	
	-0.0101	1.7445	0.9279	
RF: 100, 5	-0.0185	0.7434	0.9728	0.0666
	-0.0489	0.5668	0.9705	
	0.0118	0.9199	0.9751	
RF: 100, 10	-0.0191	0.7191	0.9720	0.0797
	-0.0526	0.5421	0.9696	
	0.0144	0.8961	0.9744	
RF: 300, 1	-0.0442	1.2884	0.9210	0.0894
	-0.0750	0.9315	0.9140	
	-0.0135	1.6454	0.9279	

RF: 300, 5	-0.0156	0.7380	0.9731	0.0694
	-0.0409	0.5808	0.9708	
	0.0096	0.8951	0.9755	
RF: 300, 10	-0.0194	0.7222	0.9724	0.0779
	-0.0535	0.5423	0.9701	
	0.0147	0.9021	0.9747	
RF: 500, 1	-0.0475	1.3431	0.9272	0.1230
	-0.0805	0.9557	0.9206	
	-0.0145	1.7304	0.9337	
RF: 500, 5	-0.0198	0.7633	0.9763	0.0566
	-0.0524	0.5706	0.9741	
	0.0128	0.9560	0.9784	
RF: 500, 10	-0.0219	0.7462	0.9758	0.0650
	-0.0610	0.5376	0.9736	
	0.0172	0.9549	0.9780	

Supplementary Table E: *Internal* validation results from models including demographic, ACC variables, and HEI. Criteria is equal to $(\text{slope-1})^2 + (\text{C-statistic-1})^2$.

	Intercept 95% CI	**Slope 95% CI**	**C-Statistic 95% CI**	**Criteria**
Cox	0.0009	0.5165	0.8608	0.2531
	-0.0018	0.4962	0.8517	
	0.0036	0.5368	0.8699	
GBM: 100, 1	-0.0006	0.5595	0.8762	0.2094
	-0.0065	0.5159	0.8679	
	0.0054	0.6031	0.8845	
GBM: 100, 5	-0.0018	0.5513	0.8992	0.2115
	-0.0041	0.5348	0.8914	
	0.0006	0.5678	0.9070	
GBM: 100, 10	-0.0043	0.5829	0.9107	0.1819
	-0.0113	0.5354	0.9027	
	0.0028	0.6305	0.9187	
GBM: 300, 1	-0.0015	0.5601	0.8752	0.2091
	-0.0068	0.5200	0.8668	
	0.0037	0.6003	0.8837	
GBM: 300, 5	-0.0032	0.5638	0.9027	0.1997
	-0.0071	0.5366	0.8950	
	0.0008	0.5910	0.9105	
GBM: 300, 10	-0.0049	0.5859	0.9191	0.1780
	-0.0106	0.5482	0.9118	
	0.0008	0.6236	0.9264	

GBM: 500, 1	-0.0007	0.5485	0.8754	0.2194
	-0.0076	0.4959	0.8671	
	0.0062	0.6011	0.8836	
GBM: 500, 5	-0.0030	0.5680	0.9009	0.1964
	-0.0063	0.5456	0.8931	
	0.0002	0.5904	0.9088	
GBM: 500, 10	-0.0035	0.5777	0.9144	0.1857
	-0.0086	0.5437	0.9068	
	0.0016	0.6117	0.9219	
RF: 100, 1	-0.0463	1.3193	0.9302	0.1068
	-0.0772	0.9646	0.9239	
	-0.0154	1.6740	0.9365	
RF: 100, 5	-0.0193	0.7561	0.9759	0.0601
	-0.0512	0.5684	0.9737	
	0.0125	0.9439	0.9782	
RF: 100, 10	-0.0207	0.7366	0.9757	0.0700
	-0.0575	0.5408	0.9735	
	0.0160	0.9325	0.9779	
RF: 300, 1	-0.0448	1.2936	0.9345	0.0905
	-0.0793	0.9023	0.9285	
	-0.0102	1.6848	0.9405	
RF: 300, 5	-0.0199	0.7645	0.9764	0.0560
	-0.0523	0.5724	0.9742	
	0.0125	0.9566	0.9785	
RF: 300, 10	-0.0213	0.7440	0.9762	0.0661
	-0.0591	0.5423	0.9740	
	0.0164	0.9457	0.9783	
RF: 500, 1	-0.0454	1.3038	0.9336	0.0967
	-0.0815	0.8937	0.9275	
	-0.0094	1.7139	0.9397	
RF: 500, 5	-0.0174	0.7627	0.9768	0.0568
	-0.0459	0.5824	0.9746	
	0.0112	0.9429	0.9789	
RF: 500, 10	-0.0182	0.7384	0.9766	0.0690
	-0.0500	0.5556	0.9744	
	0.0137	0.9212	0.9787	

Supplementary Table F: *Internal* validation results from models including demographic, ACC variables, and AHEI. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2.

	Intercept **95% CI**	**Slope** **95% CI**	**C-Statistic** **95% CI**	**Criteria**

Cox	0.0011	0.5142	0.8610	0.2553
	-0.0009	0.4993	0.8520	
	0.0031	0.5292	0.8701	
GBM: 100, 1	-0.0012	0.5533	0.8761	0.2149
	-0.0075	0.5057	0.8678	
	0.0050	0.6008	0.8844	
GBM: 100, 5	-0.0020	0.5502	0.8991	0.2125
	-0.0060	0.5231	0.8912	
	0.0019	0.5773	0.9071	
GBM: 100, 10	-0.0049	0.5887	0.9147	0.1764
	-0.0116	0.5440	0.9070	
	0.0017	0.6334	0.9225	
GBM: 300, 1	-0.0004	0.5399	0.8760	0.2271
	-0.0059	0.4989	0.8677	0.2271
	0.0051	0.5808	0.8842	0.2271
GBM: 300, 5	-0.0024	0.5586	0.8977	0.2053
	-0.0050	0.5407	0.8897	
	0.0001	0.5764	0.9057	
GBM: 300, 10	-0.0020	0.5685	0.9159	0.1933
	-0.0066	0.5385	0.9081	
	0.0026	0.5985	0.9237	
GBM: 500, 1	-0.0005	0.5416	0.8762	0.2255
	-0.0072	0.4909	0.8679	
	0.0063	0.5922	0.8844	
GBM: 500, 5	-0.0021	0.5564	0.8993	0.2069
	-0.0055	0.5328	0.8916	
	0.0013	0.5800	0.9071	
GBM: 500, 10	-0.0037	0.5697	0.9165	0.1921
	-0.0110	0.5227	0.9089	
	0.0035	0.6167	0.9242	
RF: 100, 1	-0.0481	1.3493	0.9317	0.1267
	-0.0844	0.9270	0.9255	
	-0.0118	1.7717	0.9379	
RF: 100, 5	-0.0202	0.7717	0.9770	0.0526
	-0.0539	0.5712	0.9749	
	0.0135	0.9722	0.9791	
RF: 100, 10	-0.0214	0.7427	0.9760	0.0668
	-0.0596	0.5396	0.9739	
	0.0168	0.9458	0.9782	
RF: 300, 1	-0.0438	1.2788	0.9327	0.0823
	-0.0756	0.9201	0.9267	
	-0.0120	1.6374	0.9387	
RF: 300, 5	-0.0171	0.7559	0.9766	0.0601
	-0.0450	0.5808	0.9745	
	0.0109	0.9311	0.9788	

RF: 300, 10	-0.0220	0.7478	0.9766	0.0642
	-0.0613	0.5385	0.9745	
	0.0173	0.9571	0.9787	
RF: 500, 1	-0.0498	1.3774	0.9330	0.1469
	-0.0862	0.9518	0.9269	
	-0.0135	1.8029	0.9391	
RF: 500, 5	-0.0176	0.7642	0.9772	0.0561
	-0.0467	0.5813	0.9750	
	0.0115	0.9471	0.9793	
RF: 500, 10	-0.0183	0.7369	0.9768	0.0698
	-0.0505	0.5538	0.9747	
	0.0138	0.9200	0.9789	

Supplementary Table G: *Internal* validation results from models including demographic, ACC variables, and MDS. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2.

	Intercept	**Slope**	**C-Statistic**	**Criteria**
	95% CI	**95% CI**	**95% CI**	
Cox	0.0009	0.5172	0.8609	0.2524
	-0.0015	0.4991	0.8518	
	0.0033	0.5352	0.8700	
GBM: 100, 1	-0.0017	0.5647	0.8763	0.2048
	-0.0064	0.5281	0.8680	
	0. 0031	0.6012	0.8847	
GBM: 100, 5	-0.0010	0.5495	0.8973	0.2135
	-0.0041	0.5284	0.8891	
	0.0020	0.5705	0.9055	
GBM: 100, 10	-0.0043	0.5771	0.9166	0.1858
	-0.0079	0.5530	0.9091	
	-0.0007	0.6011	0.9241	
GBM: 300, 1	-0.0006	0.5417	0.8760	0.2254
	-0.0075	0.4895	0.8677	
	0.0063	0.5939	0.8843	
GBM: 300, 5	-0.0020	0.5547	0.8997	0.2084
	-0.0046	0.5367	0.8920	
	0.0005	0.5727	0.9073	
GBM: 300, 10	-0.0037	0.5752	0.9151	0.1877
	-0.0091	0.5395	0.9075	
	0.0017	0.6109	0.9227	
GBM: 500, 1	-0.0011	0.5551	0.8769	0.2131
	-0.0074	0.5072	0.8687	
	0.0051	0.6029	0.8851	
GBM: 500, 5	-0.0019	0.5575	0.8984	0.2061
	-0.0056	0.5317	0.8905	

	0.0018	0.5832	0.9063	
GBM: 500, 10	-0.0047	0.5814	0.9167	0.1822
	-0.0115	0.5366	0.9092	
	0.0021	0.6263	0.9242	
RF: 100, 1	-0.0405	1.2255	0.9238	0.0567
	-0.0689	0.9059	0.9175	
	-0.0121	1.5451	0.9302	
RF: 100, 5	-0.0228	0.7646	0.9724	0.0562
	-0.0598	0.5597	0.9701	
	0.0142	0.9695	0.9748	
RF: 100, 10	-0.0207	0.7390	0.9731	0.0688
	-0.0569	0.5445	0.9707	
	0.0155	0.9336	0.9754	
RF: 300, 1	-0.0460	1.318	0.9262	0.1066
	-0.0788	0.935	0.9197	
	-0.0132	1.701	0.9326	
RF: 300, 5	-0.0169	0.7560	0.9733	0.0602
	-0.0442	0.5829	0.9709	
	0.0105	0.9291	0.9756	
RF: 300, 10	-0.0209	0.7435	0.9734	0.0665
	-0.0568	0.5489	0.9711	
	0.0151	0.9380	0.9757	
RF: 500, 1	-0.0457	1.3123	0.9274	0.1028
	-0.0790	0.9259	0.9211	
	-0.0125	1.6988	0.9338	
RF: 500, 5	-0.0168	0.7556	0.9734	0.0604
	-0.0440	0.5833	0.9711	
	0.0104	0.9280	0.9757	
RF: 500, 10	-0.0178	0.7375	0.9737	0.0696
	-0.0484	0.5601	0.9714	
	0.0128	0.9149	0.9760	

Supplementary Table H: *Internal* validation results from models including demographic, ACC variables, and DASH. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2.

	Intercept 95% CI	**Slope 95% CI**	**C-Statistic 95% CI**	**Criteria**
Cox	0.0009	0.5165	0.8615	0.2530
	-0.0027	0.4896	0.8525	
	0.0045	0.5434	0.8706	
GBM: 100, 1	-0.0006	0.5456	0.8769	0.2216
	-0.0073	0.4949	0.8687	
	0.0061	0.5964	0.8851	

GBM: 100, 5	-0.0032	0.5684	0.9018	0.1959
	-0.0074	0.5391	0.8940	
	0.0010	0.5977	0.9097	
GBM: 100, 10	-0.0048	0.5825	0.9183	0.1810
	-0.0099	0.5494	0.9108	
	0.0002	0.6157	0.9258	
GBM: 300, 1	-0.0006	0.5553	0.8766	0.2130
	-0.0075	0.5052	0.8683	
	0.0063	0.6054	0.8848	
GBM: 300, 5	-0.0022	0.5545	0.8990	0.2087
	-0.0064	0.5255	0.8910	
	0.0020	0.5836	0.9069	
GBM: 300, 10	-0.0041	0.5727	0.9172	0.1894
	-0.0105	0.5307	0.9098	
	0.0023	0.6146	0.9245	
GBM: 500, 1	-0.0004	0.5423	0.8772	0.2246
	-0.0076	0.4880	0.8690	
	0.0068	0.5965	0.8853	
GBM: 500, 5	-0.0033	0.5719	0.9016	0.1930
	-0.0078	0.5403	0.8938	
	0.0013	0.6035	0.9094	
GBM: 500, 10	-0.0029	0.5674	0.9064	0.1959
	-0.0083	0.5306	0.8986	
	0.0025	0.6043	0.9141	
RF: 100, 1	-0.0475	1.3431	0.9272	0.1230
	-0.0805	0.9557	0.9206	
	-0.0145	1.7304	0.9337	
RF: 100, 5	-0.0198	0.7633	0.9763	0.0566
	-0.0524	0.5706	0.9741	
	0.0128	0.9560	0.9784	
RF: 100, 10	-0.0219	0.7462	0.9758	0.0650
	-0.0610	0.5376	0.9736	
	0.0172	0.9549	0.9780	
RF: 300, 1	-0.0469	1.3320	0.9311	0.1150
	-0.0817	0.9285	0.9249	
	-0.0121	1.7354	0.9372	
RF: 300, 5	-0.0171	0.7578	0.9767	0.0592
	-0.0451	0.5818	0.9746	
	0.0108	0.9339	0.9789	
RF: 300, 10	-0.0225	0.7558	0.9767	0.0602
	-0.0630	0.5384	0.9746	
	0.0179	0.9731	0.9788	
RF: 500, 1	-0.0439	1.2784	0.9309	0.0823
	-0.0757	0.9184	0.9247	
	-0.0121	1.6383	0.9370	

RF: 500, 5	-0.0176	0.7640	0.9766	0.0562
	-0.0467	0.5804	0.9745	
	0.0115	0.9476	0.9788	
RF: 500, 10	-0.0184	0.7408	0.9766	0.0677
	-0.0506	0.5556	0.9745	
	0.0138	0.9260	0.9787	

Supplementary Table I: *Internal* validation results from models including demographic, ACC variables, and nutrition variables. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2.

	Intercept	**Slope**	**C-Statistic**	**Criteria**
	95% CI	**95% CI**	**95% CI**	
Cox	0.0007	0.5156	0.8750	0.2503
	-0.0016	0.4991	0.8661	
	0.0031	0.5321	0.8838	
GBM: 100, 1	-0.0027	0.5748	0.8811	0.1949
	-0.0075	0.5386	0.8729	
	0.0021	0.6111	0.8894	
GBM: 100, 5	-0.0063	0.6183	0.9169	0.1526
	-0.0121	0.5778	0.9092	
	-0.0004	0.6589	0.9246	
GBM: 100, 10	-0.0088	0.6767	0.9377	0.1084
	-0.0203	0.5990	0.9309	
	0.0026	0.7545	0.9445	
GBM: 300, 1	-0.0024	0.5723	0.8793	0.1975
	-0.0071	0.5354	0.8707	
	0.0024	0.6091	0.8878	
GBM: 300, 5	-0.0066	0.6294	0.9135	0.1448
	-0.0140	0.5778	0.9059	
	0.0007	0.6811	0.9211	
GBM: 300, 10	-0.0061	0.6427	0.9228	0.1336
	-0.0152	0.5795	0.9152	
	0.0029	0.7060	0.9303	
GBM: 500, 1	-0.0020	0.5616	0.8785	0.2070
	-0.0077	0.5188	0.8700	
	0.0036	0.6044	0.8870	
GBM: 500, 5	-0.0073	0.6395	0.9160	0.1370
	-0.0161	0.5770	0.9082	
	0.0016	0.7020	0.9239	
GBM: 500, 10	-0.0083	0.6644	0.9314	0.1173
	-0.0183	0.5961	0.9242	
	0.0016	0.7327	0.9386	

RF: 100, 1	-0.1754	3.3994	0.9874	5.7573
	-0.2884	1.7584	0.9853	
	-0.0624	5.0405	0.9895	
RF: 100, 5	-0.0427	1.2353	0.9967	0.0554
	-0.0884	0.8154	0.9960	
	0.0029	1.6552	0.9973	
RF: 100, 10	-0.0328	1.0458	0.9942	0.0021
	-0.0743	0.7056	0.9932	
	0.0087	1.3860	0.9952	
RF: 300, 1	-0.1742	3.3849	0.9919	5.6878
	-0.2843	1.7938	0.9903	
	-0.0642	4.9760	0.9934	
RF: 300, 5	-0.0432	1.2387	0.9969	0.0570
	-0.0884	0.8230	0.9963	
	0.0021	1.6544	0.9975	
RF: 300, 10	-0.0333	1.0426	0.9943	0.0018
	-0.0739	0.7138	0.9934	
	0.0072	1.3713	0.9953	
RF: 500, 1	-0.1813	3.4987	0.9921	6.2436
	-0.2962	1.8260	0.9907	
	-0.0664	5.1713	0.9935	
RF: 500, 5	-0.0436	1.2453	0.9970	0.0602
	-0.0885	0.8311	0.9964	
	0.0013	1.6596	0.9976	
RF: 500, 10	-0.0337	1.0453	0.9944	0.0021
	-0.0743	0.7155	0.9934	
	0.0069	1.3751	0.9953	

Table J: *External* validation results from models including demographic and ACC variables only. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2. Best performing GBM and RF are italicized.

	Intercept 95% CI	**Slope 95% CI**	**C-Statistic 95% CI**	**Criteria**
Cox	-0.0004	0.5278	0.8780	0.2379
	-0.0038	0.5037	0.8667	
	0.0029	0.5520	0.8893	
GBM: 100, 1	-0.0004	0.5276	0.8846	0.2365
	-0.0096	0.4621	0.8737	
	0.0088	0.5931	0.8956	
GBM: 100, 5	0.0004	0.5294	0.8948	0.2325
	-0.0064	0.4828	0.8840	
	0.0072	0.5761	0.9056	

GBM: 100, 10	0.0020	0.5358	0.9020	0.2251
	-0.0050	0.4875	0.8914	
	0.0090	0.5841	0.9126	
GBM: 300, 1	0.0004	0.5250	0.8838	0.2391
	-0.0101	0.4532	0.8728	
	0.0108	0.5968	0.8948	
GBM: 300, 5	0.0017	0.5254	0.8919	0.2369
	-0.0063	0.4696	0.8810	
	0.0097	0.5813	0.9027	
GBM: 300, 10	0.0004	0.5342	0.9022	0.2265
	-0.0058	0.4932	0.8917	
	0.0065	0.5751	0.9128	
GBM: 500, 1	0.0005	0.5173	0.8843	0.2464
	-0.0102	0.4408	0.8733	
	0.0113	0.5939	0.8952	
GBM: 500, 5	0.0011	0.5306	0.8944	0.2315
	-0.0052	0.4869	0.8837	
	0.0074	0.5743	0.9052	
GBM: 500, 10	*0.0030*	*0.5608*	*0.9010*	*0.2027*
	-0.0042	*0.5091*	*0.8905*	
	0.0102	*0.6124*	*0.9115*	
RF: 100, 1	-0.0427	1.2546	0.9097	0.0730
	-0.0744	0.8887	0.8982	
	-0.0109	1.6204	0.9213	
RF: 100, 5	-0.0077	0.6025	0.9273	0.1633
	-0.0224	0.5196	0.9167	
	0.0070	0.6853	0.9379	
RF: 100, 10	-0.0051	0.5591	0.9260	0.1999
	-0.0176	0.4954	0.9157	
	0.0075	0.6228	0.9363	
RF: 300, 1	*-0.0380*	*1.1824*	*0.9083*	*0.0417*
	-0.0609	*0.9215*	*0.8969*	
	-0.0150	*1.4433*	*0.9197*	
RF: 300, 5	-0.0058	0.5959	0.9281	0.1685
	-0.0171	0.5279	0.9180	
	0.0055	0.6639	0.9383	
RF: 300, 10	-0.0046	0.5559	0.9269	0.2026
	-0.0163	0.4970	0.9167	
	0.0070	0.6149	0.9371	
RF: 500, 1	-0.0410	1.2346	0.9079	0.0635
	-0.0659	0.9484	0.8963	
	-0.0162	1.5207	0.9195	
RF: 500, 5	-0.0066	0.5966	0.9281	0.1679
	-0.0186	0.5278	0.9182	
	0.0053	0.6654	0.9381	

RF: 500, 10	-0.0060	0.5671	0.9274	0.1927
	-0.0201	0.4952	0.9173	
	0.0080	0.6390	0.9375	

Supplementary Table K: *External* validation results from models including demographic, ACC variables, and HEI. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2. Best performing GBM and RF are italicized.

	Intercept	**Slope**	**C-Statistic**	**Criteria**
	95% CI	**95% CI**	**95% CI**	
Cox	-0.0003	0.5265	0.8781	0.2391
	-0.0040	0.5003	0.8667	
	0.0033	0.5527	0.8894	
GBM: 100, 1	0.0005	0.5395	0.8846	0.2254
	-0.0110	0.4587	0.8734	
	0.0120	0.6204	0.8958	
GBM: 100, 5	0.0012	0.5513	0.8943	0.2125
	-0.0071	0.4910	0.8834	
	0.0096	0.6116	0.9051	
GBM: 100, 10	*0.0020*	*0.5908*	*0.8968*	*0.1781*
	-0.0048	*0.5397*	*0.8857*	
	0.0088	*0.6419*	*0.9080*	
GBM: 300, 1	-0.0006	0.5416	0.8843	0.2235
	-0.0110	0.4644	0.8731	
	0.0098	0.6187	0.8955	
GBM: 300, 5	0.0007	0.5469	0.8963	0.2161
	-0.0062	0.4975	0.8855	
	0.0077	0.5963	0.9070	
GBM: 300, 10	0.0012	0.5769	0.9035	0.1883
	-0.0063	0.5229	0.8929	
	0.0087	0.6309	0.9142	
GBM: 500, 1	-0.0003	0.5362	0.8843	0.2285
	-0.0097	0.4677	0.8733	
	0.0091	0.6047	0.8954	
GBM: 500, 5	0.0012	0.5594	0.8969	0.2048
	-0.0068	0.5011	0.8858	
	0.0092	0.6177	0.9081	
GBM: 500, 10	0.0009	0.5699	0.9047	0.1941
	-0.0037	0.5371	0.8942	
	0.0056	0.6026	0.9152	
RF: 100, 1	-0.0395	1.2045	0.9127	0.0494
	-0.0619	0.9521	0.9015	
	-0.0171	1.4570	0.9239	

RF: 100, 5	-0.0076	0.6063	0.9309	0.1598
	-0.0212	0.5282	0.9213	
	0.0060	0.6844	0.9406	
RF: 100, 10	-0.0078	0.5851	0.9304	0.1770
	-0.0257	0.4934	0.9204	
	0.0101	0.6768	0.9403	
RF: 300, 1	-0.0378	1.1752	0.9154	0.0379
	-0.0633	0.8938	0.9043	
	-0.0124	1.4566	0.9264	
RF: 300, 5	-0.0084	0.6177	0.9314	0.1509
	-0.0241	0.5266	0.9216	
	0.0074	0.7088	0.9411	
RF: 300, 10	-0.0078	0.5867	0.9309	0.1756
	-0.0233	0.5065	0.9212	
	0.0078	0.6669	0.9406	
RF: 500, 1	*-0.0377*	*1.1735*	*0.9148*	*0.0374*
	-0.0625	*0.8969*	*0.9038*	
	-0.0129	*1.4501*	*0.9258*	
RF: 500, 5	-0.0077	0.6221	0.9318	0.1475
	-0.0222	0.5329	0.9222	
	0.0068	0.7112	0.9415	
RF: 500, 10	-0.0066	0.5851	0.9308	0.1769
	-0.0209	0.5060	0.9212	
	0.0078	0.6641	0.9403	

Supplementary Table L: *External* validation results from models including demographic, ACC variables, and AHEI. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2. Best performing GBM and RF are italicized.

	Intercept 95% CI	**Slope 95% CI**	**C-Statistic 95% CI**	**Criteria**
Cox	-0.0009	0.5347	0.8784	0.2313
	-0.0041	0.5115	0.8671	
	0.0023	0.5579	0.8897	
GBM: 100, 1	-0.0009	0.5326	0.8839	0.2319
	-0.0106	0.4627	0.8728	
	0.0088	0.6025	0.8951	
GBM: 100, 5	0.0005	0.5312	0.8964	0.2305
	-0.0052	0.4924	0.8857	
	0.0061	0.5700	0.9071	
GBM: 100, 10	*0.0009*	*0.5697*	*0.9025*	*0.1947*
	-0.0044	*0.5315*	*0.8917*	
	0.0063	*0.6079*	*0.9133*	
GBM: 300, 1	0.0001	0.5197	0.8852	0.2439
	-0.0088	0.4561	0.8741	

	0.0089	0.5833	0.8963	
GBM: 300, 5	0.0002	0.5223	0.8957	0.2391
	-0.0092	0.4583	0.8852	
	0.0097	0.5864	0.9062	
GBM: 300, 10	0.0030	0.5638	0.9061	0.1991
	-0.0034	0.5179	0.8954	
	0.0095	0.6096	0.9168	
GBM: 500, 1	-0.0004	0.5284	0.8848	0.2357
	-0.0097	0.4612	0.8737	
	0.0090	0.5955	0.8960	
GBM: 500, 5	0.0018	0.5348	0.8942	0.2276
	-0.0063	0.4780	0.8836	
	0.0098	0.5916	0.9047	
GBM: 500, 10	0.0011	0.5511	0.9054	0.2105
	-0.0038	0.5176	0.8948	
	0.0060	0.5846	0.9161	
RF: 100, 1	-0.0416	1.2373	0.9141	0.0637
	-0.0695	0.9188	0.9028	
	-0.0137	1.5558	0.9255	
RF: 100, 5	-0.0081	0.6211	0.9296	0.1485
	-0.0243	0.5268	0.9196	
	0.0080	0.7154	0.9395	
RF: 100, 10	-0.0064	0.5761	0.9288	0.1848
	-0.0200	0.5061	0.9191	
	0.0071	0.6460	0.9386	
RF: 300, 1	*-0.0372*	*1.1657*	*0.9147*	*0.0347*
	-0.0610	*0.9034*	*0.9036*	
	-0.0134	*1.4281*	*0.9258*	
RF: 300, 5	-0.0066	0.6066	0.9309	0.1595
	-0.0184	0.5344	0.9212	
	0.0053	0.6788	0.9406	
RF: 300, 10	-0.0067	0.5774	0.9299	0.1835
	-0.0206	0.5058	0.9201	
	0.0073	0.6491	0.9396	
RF: 500, 1	-0.0429	1.2622	0.9137	0.0762
	-0.0699	0.9513	0.9024	
	-0.0159	1.5731	0.9249	
RF: 500, 5	-0.0074	0.6195	0.9307	0.1496
	-0.0215	0.5326	0.9208	
	0.0068	0.7063	0.9407	
RF: 500, 10	-0.0055	0.5733	0.9295	0.1870
	-0.0175	0.5070	0.9196	
	0.0066	0.6396	0.9394	

Supplementary Table M: *External* validation results from models including demographic, ACC variables, and MDS. Criteria is equal to $(\text{slope-1})^2 + (\text{C-statistic-1})^2$. Best performing GBM and RF are italicized.

	Intercept **95% CI**	**Slope** **95% CI**	**C-Statistic** **95% CI**	**Criteria**
Cox	-0.0003	0.5268	0.8783	0.2387
	-0.0037	0.5020	0.8670	
	0.0032	0.5516	0.8896	
GBM: 100, 1	-0.0009	0.5401	0.8860	0.2245
	-0.0099	0.4738	0.8749	
	0.0081	0.6064	0.8972	
GBM: 100, 5	0.0012	0.5358	0.8960	0.2263
	-0.0047	0.4945	0.8846	
	0.0072	0.5770	0.9075	
GBM: 100, 10	*0.0015*	*0.5480*	*0.9043*	*0.2135*
	-0.0064	*0.4927*	*0.8939*	
	0.0094	*0.6034*	*0.9147*	
GBM: 300, 1	-0.0005	0.5253	0.8853	0.2385
	-0.0100	0.4578	0.8743	
	0.0090	0.5927	0.8963	
GBM: 300, 5	0.0009	0.5382	0.8930	0.2247
	-0.0066	0.4851	0.8823	
	0.0084	0.5914	0.9037	
GBM: 300, 10	0.0024	0.5390	0.9036	0.2218
	-0.0053	0.4860	0.8931	
	0.0100	0.5919	0.9141	
GBM: 500, 1	-0.0003	0.5304	0.8856	0.2336
	-0.0110	0.4526	0.8745	
	0.0103	0.6083	0.8966	
GBM: 500, 5	0.0011	0.5551	0.8974	0.2085
	-0.0067	0.4986	0.8867	
	0.0090	0.6116	0.9082	
GBM: 500, 10	0.0014	0.5220	0.9035	0.2378
	-0.0056	0.4750	0.8931	
	0.0085	0.5690	0.9139	
RF: 100, 1	*-0.0345*	*1.1250*	*0.9055*	*0.0246*
	-0.0557	*0.8905*	*0.8941*	
	-0.0133	*1.3595*	*0.9168*	
RF: 100, 5	-0.0084	0.6085	0.9275	0.1585
	-0.0232	0.5282	0.9178	
	0.0064	0.6887	0.9371	
RF: 100, 10	-0.0054	0.5666	0.9249	0.1935
	-0.0171	0.5063	0.9148	
	0.0062	0.6269	0.9351	

RF: 300, 1	-0.0404 -0.0659 -0.0150	1.2231 0.9316 1.5146	0.9094 0.8981 0.9207	0.0580
RF: 300, 5	-0.0066 -0.0190 0.0058	0.6099 0.5332 0.6866	0.9269 0.9168 0.9371	0.1575
RF: 300, 10	-0.0064 -0.0217 0.0090	0.5802 0.5000 0.6605	0.9254 0.9154 0.9354	0.1818
RF: 500, 1	-0.0388 -0.0632 -0.0145	1.1954 0.9179 1.4728	0.9094 0.8983 0.9206	0.0464
RF: 500, 5	-0.0060 -0.0169 0.0050	0.6030 0.5352 0.6708	0.9275 0.9177 0.9373	0.1629
RF: 500, 10	-0.0052 -0.0171 0.0066	0.5782 0.5118 0.6446	0.9267 0.9169 0.9364	0.1833

Supplementary Table N: *External* validation results from models including demographic, ACC variables, and DASH. Criteria is equal to $(\text{slope}-1)^2 + (\text{C-statistic}-1)^2$. Best performing GBM and RF are italicized.

	Intercept 95% CI	**Slope 95% CI**	**C-Statistic 95% CI**	**Criteria**
Cox	-0.0001 -0.0050 0.0048	0.5248 0.4892 0.5604	0.8775 0.8662 0.8888	0.2408
GBM: 100, 1	-0.0004 -0.0099 0.0091	0.5277 0.4598 0.5956	0.8847 0.8735 0.8959	0.2364
GBM: 100, 5	0.0008 -0.0056 0.0073	0.5548 0.5080 0.6015	0.8959 0.8851 0.9067	0.2090
GBM: 100, 10	*0.0002* *-0.0062* *0.0066*	*0.6169* *0.5691* *0.6647*	*0.9073* *0.8970* *0.9175*	*0.1554*
GBM: 300, 1	-0.0003 -0.0109 0.0103	0.5352 0.4618 0.6085	0.8849 0.8737 0.8961	0.2293
GBM: 300, 5	0.0010 -0.0059 0.0080	0.5268 0.4785 0.5750	0.8925 0.8812 0.9037	0.2355
GBM: 300, 10	0.0022 -0.0048 0.0092	0.5366 0.4889 0.5843	0.9015 0.8911 0.9120	0.2244

GBM: 500, 1	-0.0003	0.5276	0.8853	0.2363
	-0.0101	0.4577	0.8742	
	0.0094	0.5974	0.8964	
GBM: 500, 5	0.0006	0.5344	0.8963	0.2275
	-0.0074	0.4796	0.8851	
	0.0085	0.5892	0.9074	
GBM: 500, 10	0.0003	0.5544	0.8973	0.2091
	-0.0034	0.5286	0.8860	
	0.0039	0.5803	0.9086	
RF: 100, 1	-0.0410	1.2346	0.9079	0.0635
	-0.0659	0.9484	0.8963	
	-0.0162	1.5207	0.9195	
RF: 100, 5	-0.0066	0.5966	0.9281	0.1679
	-0.0186	0.5278	0.9182	
	0.0053	0.6654	0.9381	
RF: 100, 10	-0.0060	0.5671	0.9274	0.1927
	-0.0201	0.4952	0.9173	
	0.0080	0.6390	0.9375	
RF: 300, 1	-0.0393	1.2049	0.9104	0.0500
	-0.0636	0.9279	0.8988	
	-0.0149	1.4819	0.9219	
RF: 300, 5	-0.0062	0.6025	0.9289	0.1631
	-0.0178	0.5313	0.9189	
	0.0054	0.6738	0.9389	
RF: 300, 10	-0.0070	0.5789	0.9279	0.1825
	-0.0214	0.5044	0.9179	
	0.0074	0.6533	0.9379	
RF: 500, 1	*-0.0369*	*1.1604*	*0.9114*	*0.0336*
	-0.0597	*0.9083*	*0.9000*	
	-0.0142	*1.4124*	*0.9227*	
RF: 500, 5	-0.0053	0.5905	0.9300	0.1726
	-0.0142	0.5364	0.9205	
	0.0035	0.6446	0.9396	
RF: 500, 10	-0.0057	0.5756	0.9284	0.1852
	-0.0181	0.5073	0.9185	
	0.0067	0.6440	0.9383	

Supplementary Table O: *External* validation results from models including demographic, ACC variables, and nutrition variables. Criteria is equal to (slope-1)^2 + (C-statistic-1)^2. Best performing GBM and RF are italicized.

	Intercept	**Slope**	**C-Statistic**	**Criteria**
	95% CI	**95% CI**	**95% CI**	

Cox	0.0010	0.4611	0.8830	0.3041
	-0.0034	0.4264	0.8698	
	0.0054	0.4959	0.8962	
GBM: 100, 1	-0.0030	0.5674	0.8896	0.1993
	-0.0092	0.5227	0.8784	
	0.0031	0.6120	0.9007	
GBM: 100, 5	-0.0016	0.5621	0.9072	0.2004
	-0.0073	0.5227	0.8966	
	0.0041	0.6015	0.9178	
GBM: 100, 10	*0.0027*	*0.6518*	*0.9090*	*0.1295*
	-0.0049	*0.5906*	*0.8981*	
	0.0103	*0.7131*	*0.9200*	
GBM: 300, 1	-0.0026	0.5681	0.8886	0.1989
	-0.0103	0.5108	0.8772	
	0.0051	0.6254	0.9000	
GBM: 300, 5	-0.0009	0.6548	0.9022	0.1287
	-0.0062	0.6121	0.8902	
	0.0044	0.6975	0.9143	
GBM: 300, 10	*0.0021*	*0.8318*	*0.9058*	*0.0372*
	-0.0039	*0.7710*	*0.8947*	
	0.0081	*0.8927*	*0.9170*	
GBM: 500, 1	-0.0026	0.5545	0.8894	0.2107
	-0.0101	0.5000	0.8781	
	0.0050	0.6090	0.9008	
GBM: 500, 5	-0.0029	0.5980	0.9030	0.1710
	-0.0060	0.5759	0.8912	
	0.0002	0.6202	0.9148	
GBM: 500, 10	0.0003	0.7133	0.9098	0.0903
	-0.0057	0.6624	0.8990	
	0.0063	0.7642	0.9206	
RF: 100, 1	-0.1254	2.5742	0.8937	2.4894
	-0.1941	1.5825	0.8781	
	-0.0567	3.5659	0.9093	
RF: 100, 5	*-0.0299*	*1.0137*	*0.9320*	*0.0048*
	-0.0567	*0.7609*	*0.9208*	
	-0.0031	*1.2666*	*0.9433*	
RF: 100, 10	-0.0201	0.8447	0.9336	0.0285
	-0.0412	0.6690	0.9226	
	0.0010	1.0204	0.9445	
RF: 300, 1	-0.1293	2.6387	0.9059	2.6942
	-0.1973	1.6579	0.8914	
	-0.0613	3.6195	0.9203	
RF: 300, 5	-0.0314	1.0368	0.9371	0.0053
	-0.0583	0.7826	0.9262	
	-0.0046	1.2909	0.9481	

RF: 300, 10	-0.0204	0.8343	0.9367	0.0315
	-0.0395	0.6773	0.9263	
	-0.0012	0.9913	0.9470	
RF: 500, 1	-0.1401	2.8162	0.9129	3.3062
	-0.2170	1.6982	0.8993	
	-0.0632	3.9342	0.9266	
RF: 500, 5	-0.0304	1.0242	0.9348	0.0048
	-0.0552	0.7896	0.9238	
	-0.0057	1.2588	0.9459	
RF: 500, 10	-0.0215	0.8494	0.9379	0.0265
	-0.0419	0.6824	0.9277	
	-0.0012	1.0165	0.9481	

Supplementary Table P: Hazard ratios (95% CIs) from Cox models developed on training data. See Supplementary Table A for variable definitions.

	Model 1 (ACC)	**Model 2 (+HEI)**	**Model 3 (+AHEI)**	**Model 4 (+MDS)**	**Model 5 (+DASH)**	**Model 6 (+All)**
age	1.10 (1.09, 1.10)	1.10 (1.09, 1.11)	1.10 (1.09, 1.11)	1.10 (1.09, 1.10)	1.10 (1.09, 1.11)	1.10 (1.09, 1.10)
sex	0.65 (0.57, 0.73)	0.65 (0.58, 0.74)	0.65 (0.58, 0.73	0.65 (0.57, 0.73)	0.65 (0.58, 0.74)	0.61 (0.54, 0.70)
black	1.14 (0.99, 1.32)	1.14 (0.99, 1.32)	1.15 (0.99, 1.33)	1.14 (0.99, 1.32)	1.11 (0.97, 1.29)	1.10 (0.99, 1.29)
hispanic	0.69 (0.58, 0.81)	0.69 (0.58, 0.82)	0.69 (0.58, 0.82)	0.69 (0.58, 0.82)	0.70 (0.59, 0.83)	0.64 (0.58, 0.77)
total_chol	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
hdl	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (1.00, 1.01)	1.00 (1.00, 1.00)
sbp	1.00 (1.00, 1.01)	1.00 (1.00, 1.01)	1.00 (1.00, 1.01)	1.00 (1.00, 1.01)	1.01 (1.00, 1.01)	1.00 (1.00, 1.01)
bpmeds	1.19 (1.08, 1.30)	1.19 (1.09, 1.30)	1.19 (1.09, 1.30)	1.19 (1.09, 1.31)	1.18 (1.07, 1.29)	1.21 (1.09, 1.33)
dm	1.46 (1.29, 1.65)	1.46 (1.29, 1.65)	1.45 (1.29, 1.64)	1.46 (1.29, 1.65)	1.45 (1.28, 1.63)	1.40 (1.29, 1.59)
tob	1.91 (1.61, 2.27)	1.89 (1.59, 2.25)	1.88 (1.59, 2.23)	1.91 (1.61, 2.26)	1.84 (1.55, 2.18)	1.84 (1.59, 2.19)
hei		1.00 (0.99, 1.01)				
ahei			1.00 (0.99, 1.00)			
mds				1.05 (1.00, 1.10)		
dash					0.99 (0.98, 0.99)	
milk_g						1 (1, 1)
cream_g						1 (0.99, 1)
milk_dessert_g						1 (1, 1)
cheese_g						1 (1, 1)
meat_ns_g						1 (0.99, 1.01)
beef_g						1 (1, 1)
pork_g						1 (1, 1)
lamb_g						1 (1, 1)
poultry_g						1 (1, 1)
organ_meat_g						1 (1, 1)
fish_g						1 (0.99, 1)
meat_nonmeat_g						1 (1, 1)
protein_frozen_g						1 (1, 1)
eggs_g						1 (1, 1)
egg_mixture_g						1 (1, 1)
egg_sub_g						0.99 (0.99, 1)
legumes_g						1 (1, 1)
nuts_g						1 (1, 1)
seeds_g						1 (0.99, 1.01)
flour_mix_g						0.22 (0, ∞)
bread_yeast_g						1 (1, 1)
bread_quick_g						1 (1, 1)
pastries_g						1 (1, 1)
crackers_g						1 (1, 1)

	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
pancakes_g						1 (1, 1)
pastas_g						1 (1, 1)
cereals_g						1 (1, 1)
grain_mix_g						1 (1, 1)
meat_sub_g						0.78 (0, ∞)
citrus_g						1 (1, 1)
fruit_dried_g						1 (1, 1.01)
fruit_other_g						1 (1, 1)
fruit_juice_g						1 (1, 1)
fruit_baby_g						0.84 (0, ∞)
potatoes_g						1 (1, 1)
veg_darkgreen_g						1 (1, 1)
veg_deepyellow_g						1 (1, 1.01)
tomatoes_g						1 (1, 1)
veg_other_g						1 (1, 1)
veg_meat_g						1 (1, 1)
veg_mixture_g						1 (1, 1)
fats_g						1 (1, 1.01)
oils_g						1 (0.98, 1.01)
salad_dressing_g						1 (1, 1.01)
sweets_g						1 (1, 1)
bev_nonalcohol_g						1 (1, 1)
bev_alcohol_g						1 (1, 1)
water_g						1 (1, 1)
kcal						1 (1, 1)
protein_g						1.01 (1, 1.02)
carb_g						1 (1, 1.01)
fiber_g						0.96 (0.95, 0.97)
fat_g						0.99 (0.97, 1.01)
fat_sat_g						1.19 (1.07, 1.32)
fat_mono_g						0.96 (0.93, 1)
fat_poly_g						0.97 (0.94, 0.99)
cholesterol_mg						1 (1, 1)
vite_mg						0.99 (0.98, 1.01)
vita_mg						1 (1, 1)
betacaro_mcg						1 (1, 1)
vitb1_mg						0.92 (0.78, 1.10)
vitb2_mg						1.02 (0.87, 1.19)
niacin_mg						0.98 (0.96, 0.99)
vitb6_mg						1.11 (0.98, 1.25)
folate_mcg						1 (1, 1)
vitb12_mcg						1 (0.99, 1.02)
vitc_mg						1 (1, 1)
calcium_mg						1 (1, 1)
phosphorus_mg						1 (1, 1)
magnesium_mg						1 (1, 1)
iron_mg						1.01 (1, 1.03)
zinc_mg						1.01 (1, 1.01)
copper_mg						0.93 (0.84, 1.03)
sodium_mg						1 (1, 1)
potassium_mg						1 (1, 1)
selenium_mcg						1 (0.99, 1)
caffeine_mg						1 (1, 1)

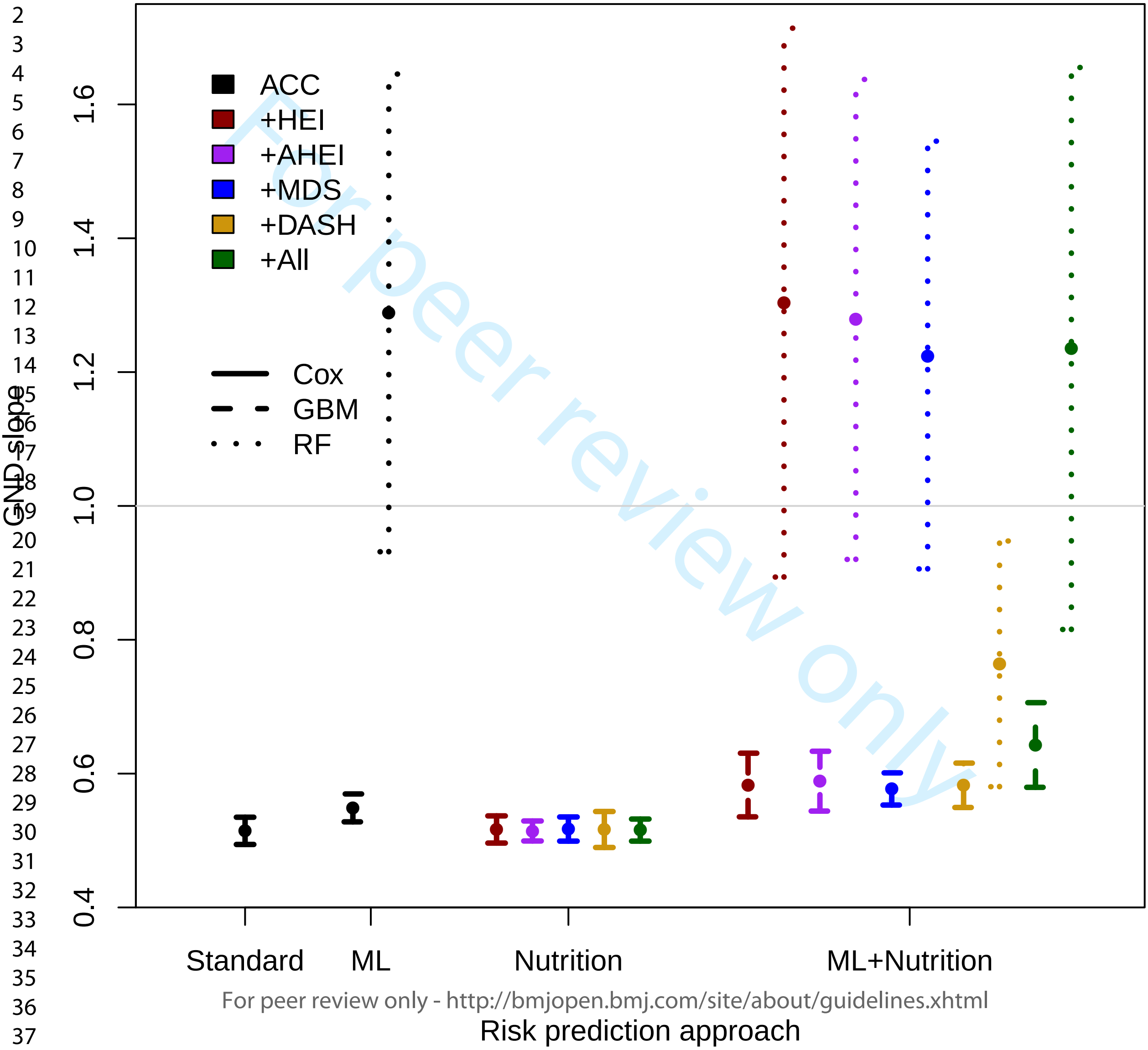
	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
theobromine_mg						1 (1, 1)
alcohol_gm						1.01 (1, 1.01)
sfa_40_gm						1.31 (0.69, 2.47)
sfa_60_gm						0.67 (0.24, 1.81)
sfa_80_gm						1.17 (0.53, 2.60)
sfa_100_gm						0.67 (0.22, 2.05)
sfa_120_gm						0.88 (0.77, 1.01)
sfa_140_gm						0.76 (0.57, 1.01)
sfa_160_gm						0.85 (0.76, 0.94)
sfa_180_gm						0.86 (0.75, 0.98)
mfa_161h_gm						0.85 (0.66, 1.09)
mfa_161o_gm						1.06 (1.02, 1.10)
mfa_201_gm						1.32 (1.03, 1.69)
mfa_221_gm						0.34 (0.13, 0.90)
pfa_182_gm						1.07 (1.04, 1.11)
pfa_183_gm						0.80 (0.68, 0.95)
pfa_184_gm						5.67 (0.15, 211.03)
pfa_204_gm						1.02 (0.29, 3.64)
pfa_205_gm						0.99 (0.21, 4.69)
pfa_225_gm						0.63 (0.01, 55.24)
pfa_226_gm						1.45 (0.40, 5.24)
water_yesterday_gm						1 (1, 1)

Supplementary Table Q: Relative influences of variables in best performing GBM models in training set from each modeling approach. See Supplementary Table A for variable definitions.

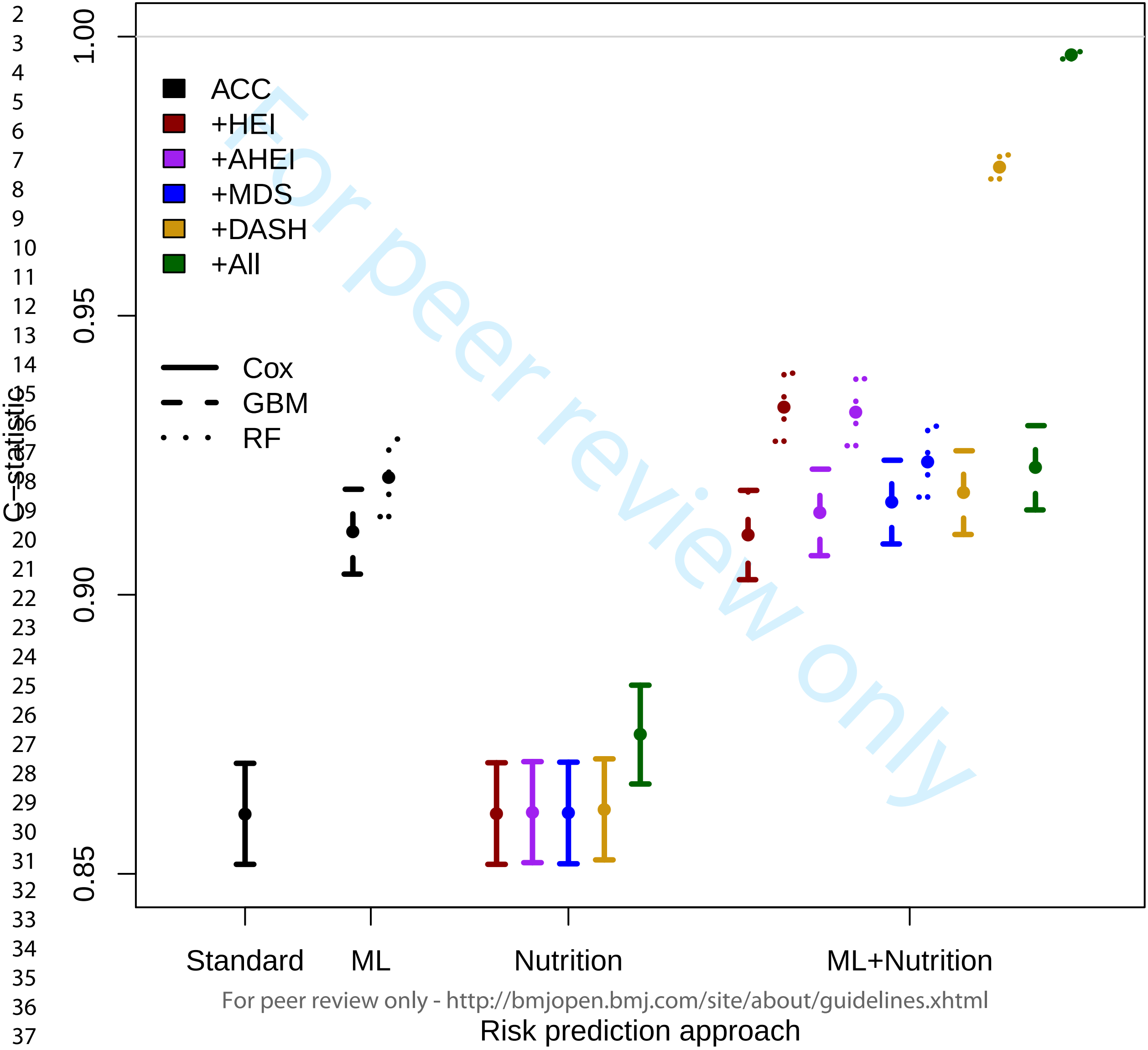
	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
age	19.89	30.33	5.59	2.93	29.70	19.25
sex	2.26	1.81	0.28	0.50	1.43	0.17
black	2.13	0.61	0.02	0.02	0.70	0.01
hispanic	0.98	0.68	0.05	0.02	0.71	0.01
total_chol	23.61	15.16	17.43	16.56	13.43	2.14
hdl	18.18	11.00	2.62	36.47	12.00	2.80
sbp	24.06	20.79	23.02	41.44	19.09	2.56
bpmeds	3.47	3.11	3.11	0.12	3.94	0.49
dm	2.08	1.53	0.12	0.05	1.64	0.27
tob	3.32	0.68	45.83	0.26	0.81	0.02
hei		14.30				
ahei			1.92			
mds				1.63		
dash					16.54	
iron_mg						10.86
legumes_g						8.42
sweets_g						6.55
pastries_g						5.75
pork_g						4.33
vita_mg						3.86
sfa_80_gm						2.99
cholesterol_mg						1.95
water_yesterday_gm						1.22
copper_mg						1.00

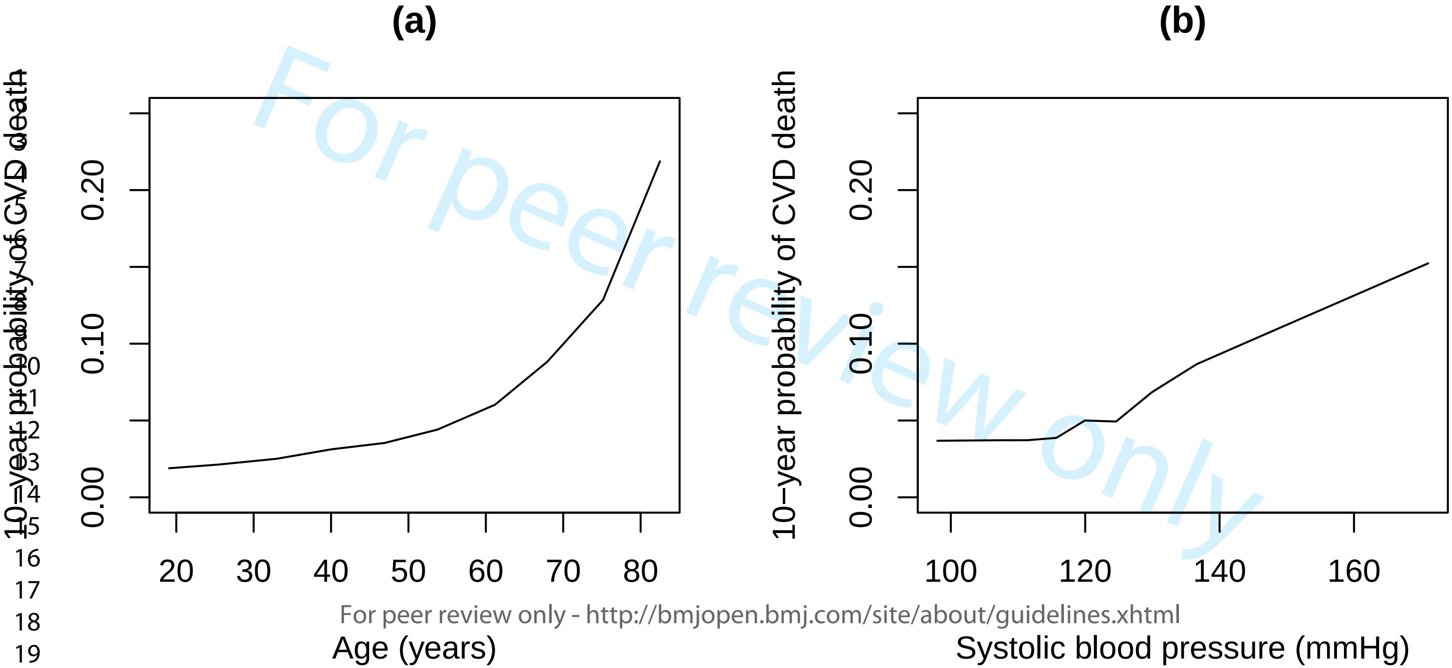
	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
fats_g						0.97
beef_g						0.92
vite_mg						0.76
bread_quick_g						0.70
calcium_mg						0.67
mfa_201_gm						0.66
vitb12_mcg						0.65
sfa_140_gm						0.65
betacaro_mcg						0.61
mfa_161o_gm						0.56
carb_g						0.54
kcal						0.51
mfa_161h_gm						0.50
caffeine_mg						0.47
veg_other_g						0.46
selenium_mcg						0.45
zinc_mg						0.44
vitb1_mg						0.43
pfa_183_gm						0.41
sfa_180_gm						0.39
sfa_120_gm						0.39
magnesium_mg						0.38
alcohol_gm						0.38
nuts_g						0.38
vitc_mg						0.37
fiber_g						0.37
phosphorus_mg						0.37
fat_poly_g						0.35
potassium_mg						0.35
salad_dressing_g						0.34
vitb6_mg						0.34
fat_g						0.33
bev_nonalcohol_g						0.33
fruit_other_g						0.32
sodium_mg						0.32
pancakes_g						0.31
protein_g						0.30
pfa_205_gm						0.30
poultry_g						0.29
sfa_160_gm						0.29
pfa_182_gm						0.28
milk_g						0.28
folate_mcg						0.28
fat_mono_g						0.28
cheese_g						0.26
milk_dessert_g						0.26
pfa_204_gm						0.26
niacin_mg						0.24
theobromine_mg						0.21
pastas_g						0.20

	Model 1 (ACC)	Model 2 (+HEI)	Model 3 (+AHEI)	Model 4 (+MDS)	Model 5 (+DASH)	Model 6 (+All)
pfa_226_gm						0.20
veg_darkgreen_g						0.19
bev_alcohol_g						0.19
tomatoes_g						0.18
fat_sat_g						0.16
crackers_g						0.16
vitb2_mg						0.16
sfa_100_gm						0.15
sfa_60_gm						0.14
pfa_225_gm						0.14
mfa_221_gm						0.14
egg_mixture_g						0.14
fruit_juice_g						0.14
citrus_g						0.12
veg_deepyellow_g						0.12
cream_g						0.12
organ_meat_g						0.11
potatoes_g						0.11
cereals_g						0.10
meat_nonmeat_g						0.09
seeds_g						0.08
water_g						0.06
fish_g						0.06
grain_mix_g						0.05
lamb_g						0.05
pfa_184_gm						0.04
meat_ns_g						0.03
eggs_g						0.03
protein_frozen_g						0.02
oils_g						0.02
fruit_dried_g						0.02
egg_sub_g						0.01
flour_mix_g						0.00
meat_sub_g						0.00
fruit_baby_g						0.00
veg_meat_g						0.00
veg_mixture_g						0.00



GND slope
0.4
0.6
0.8
1.0
1.2
1.4
1.6
ACC
+HEI
+AHEI
+MDS
+DASH
+All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach


1.00
0.95
0.90
0.85
C−statistic
ACC
+HEI
+AHEI
+MDS
+DASH
+All
Cox
GBM
RF
Standard
ML
Nutrition
ML+Nutrition
Risk prediction approach


(a)
(b)
10−year probability of CVD death
0.00
0.10
0.20
20
30
40
50
60
70
80
Age (years)
10−year probability of CVD death
0.00
0.10
0.20
100
120
140
160
Systolic blood pressure (mmHg)