PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the United States using nationally randomly sampled data
AUTHORS	Rigdon, Joseph; Basu, Sanjay

VERSION 1 - REVIEW

REVIEWER	Demosthenes Panagiotakos Harokopio University, Greece
REVIEW RETURNED	25-Jul-2019

GENERAL COMMENTS	Very well designed work and interpretation. Congratulations to the authors. I would suggest to further test for potential mediating and moderating effects by SES and other environmental factors, to further discuss the limitations of rapid dietary assessment and to more clearly interpret their findings in a clinical setting.

REVIEWER	Shameer Khader Advanced Analytics Center
REVIEW RETURNED	10-Aug-2019

GENERAL COMMENTS	Authors ask an interesting question -- whether adding nutritional values would improve the prediction of cardiovascular death or adding nutritional data in a machine learning performs better than the classical statistical model. The work is relevant in the current context of categorizing cardiovascular diseases as a lifestyle disease and data-driven medicine. However, the paper is difficult to follow as authors switch the narrative back and forth between the prediction vs. algorithmic performance. Careful restructuring and providing additional details on various methods would strengthen the manuscript. Major comments: Outcome: It will be useful if authors can stratify the analyses by cardiovascular death and cerebrovascular death separately. Then show the combined analyses. Please plot the outcome across the data and show as a figure. Features selection: It is not clear about the feature selection strategy used in the paper. Also, it will be interesting to stratify the models by age/gender.

	Imputation: It is not clear about the percentage of missing for the variables groups (nutrition vs. non-nutrition variables). The method used for imputation; what is the error rate for the imputation method? Machine learning algorithms: Please provide several metrics to assess the machine learning algorithm performances across training and testing data sets. Also, report AUC, MCC, etc. Authors used two ML algorithms, however, a work like this deserves a bit more attention beyond "commonly used" and "decision tree" algorithms. Encourage the authors to use one of each from different classes of ML (RF, SVM, ANN, DL, etc.) Authors should submit the code and model as part of the publication to ensure the reuse of the work. Please provide a GitHub repository with all data and code for further review.

REVIEWER	Collin Stultz Massachusetts Institute of Technology Massachusetts General Hospital
REVIEW RETURNED	20-Aug-2019

GENERAL COMMENTS	This is an interesting paper that purports to use nutritional data in conjunction with machine learning methods to predict cardiovascular mortality. The metrics of success are model calibration and discriminatory ability, and the main conclusions are that including machine learning coupled with the use of nutritional can significantly improve risk prediction. While there is much to like in this work, the data, as presented do not support the main conclusions. Additional data are needed to truly evaluate the potential improvement that nutritional data and machine learning (in particular gradient boosting machines and random forests) can provide. Detailed comments include: 1. Confidence intervals are calculated using Rubin's rules. The manuscript, however, is devoid of any details with respect to how this was implemented and the associated reference (#3) contains no additional information. This reviewer recognizes that Rubin's rules are not-infrequently used to calculate confidence intervals from imputed datasets, but the method is wrought with a number of challenges that make the interpretation of the resulting confidence intervals problematic. Firstly, the authors created 10 imputed training sets using an established boilerplate imputation method, however, no information is provided on how much data are missing and how many features need to be imputed. If, for example, a small fraction of the data missing, then the 10 imputed datasets will be very similar and the trained models will be very similar, leading to small confidence intervals. Additionally, if the imputed values are very similar this will also yield very similar models and small confidence intervals. As the authors use the calculated confidence intervals to determine what models are statistically superior, this is an important point. The authors should provide information on how different the various training datasets are to help the reader decide whether the resulting confidence intervals are truly trustworthy. 2. Confidence intervals can also be generated using bootstrapping (i.e., sampling with replacement) where imputation could be one on each bootstrapped training and testing sets. This method also has its challenges, but it is less likely to run in the problems

	mentioned above and it is more standard, at least in the machine learning literature. 3. The PIs often refer to models that use “machine learning” and those that are “standard” (Cox proportional hazard modeling). It should be noted that a great deal of regression models also aspire to “learn” from data and hence would be considered to be “machine learning” models by many purists. More importantly, the authors only examine two types of machine learning models and therefore it is not clear whether other methods would yield better results (e.g., artificial neural networks). The manuscript would be improved if the PIs would refrain from making statements about “machine learning” in general and focus their conclusions on the precise computational models that they examined in this work. 4. An annoying aspect of machine learning model is that there are rarely guiding principles to dictate what the optimal parameters for any given model should be; e.g., a systematic parameter search is often required to find optimal parameters. Was this done in this case? The methods section only contains a short paragraph that lists one set of parameters for the gradient boosting machine and the random forests. The reader is therefore left to wonder whether these parameters are truly optimal for the problem at hand. Consequently, it is not clear whether the use of “machine learning algorithms alone” are “not of substantial benefit” as the authors claim. 5. A minor point: The abstract lists this as a “Prospective study”, yet it uses an observational data set; i.e., all of the used are retrospective. This needs to be clarified.

VERSION 1 – AUTHOR RESPONSE

Editor	Response	Page number
Reviewer 1		
Very well designed work and interpretation. Congratulations to the authors. I would suggest to further test for potential mediating and moderating effects by SES and other environmental factors, to further discuss the limitations of rapid dietary assessment and to more clearly interpret their findings in a clinical setting.	We applied the best-performing algorithm, survival random forest with 100 trees and 10 randomly sampled variables at each node, inclusive of nutrition variables, to the data, with the inclusion of education level (NHANES variable DMDEDUC2) and ratio of family income to poverty (NHANES variable INDFMPIR). We were unable to adjust for environment variables at the zip code and census tract level as we don’t have zip code and census in NHANES.	14
Reviewer 2		
Authors ask an interesting question -- whether adding nutritional values would improve the prediction of cardiovascular death or adding nutritional data in a machine learning performs	Thank you. To improve readability, we have restructured the Results section into subsections of key metrics. Currently, the Results section is structured as “Descriptive statistics of the study sample”, “Model calibration performance”, “Model discrimination performance”, and “Important associations”.	10-14

better than the classical statistical model. The work is relevant in the current context of categorizing cardiovascular diseases as a lifestyle disease and data-driven medicine. However, the paper is difficult to follow as authors switch the narrative back and forth between the prediction vs. algorithmic performance. Careful restructuring and providing additional details on various methods would strengthen the manuscript.		
Outcome: It will be useful if authors can stratify the analyses by cardiovascular death and cerebrovascular death separately. Then show the combined analyses. Please plot the outcome across the data and show as a figure.	We applied the best-performing algorithm, survival random forest with 100 trees and 10 randomly sampled variables at each node, inclusive of nutrition variables, to the data, separately for the outcomes of heart disease and cerebrovascular disease (rather than combined outcome of either).	14
Features selection: It is not clear about the feature selection strategy used in the paper. Also, it will be interesting to stratify the models by age/gender.	Given the large volume of individual level data (42K observations), and relatively small number of features (approximately 200), we opted for no data-driven feature selection. Rather our features are selected using the AHA/ACC ASCVD risk estimation tool's predictor variables and inclusion of all nutrition related variables from a 24-hour dietary recall. We currently have a partial dependence plot for our best-performing algorithm (Supplementary Figure C), that shows model predictions over the whole range of age, indicating a spike in risk after age 65. Additionally, in Supplementary Table P (and mentioned in the Discussion), we document a protective association for female sex [HR vs. males of 0.65 (0.57, 0.73)].	6, Supplementary Table A
Imputation: It is not clear about the percentage of missing for the variables groups (nutrition vs. non-nutrition variables). The method used for imputation; what is the error rate for the imputation method?	We have added Supplementary Table B that displays the percentage of missing data in both training and held-out test sets.	Supplement

Machine learning algorithms: Please provide several metrics to assess the machine learning algorithm performances across training and testing data sets. Also, report AUC, MCC, etc. Authors used two ML algorithms, however, a work like this deserves a bit more attention beyond "commonly used" and "decision tree" algorithms. Encourage the authors to use one of each from different classes of ML (RF, SVM, ANN, DL, etc.)	We currently provide two key metrics across the train and test set: calibration as measured by the GND slope and discrimination as measured by the C-statistic (equivalent to AUC). We chose machine learning models that could handle a right-censored time-to-event outcome. We were able to find gradient boosted machines and also random forests, but had trouble implementing neural nets and deep learners, and furthermore NNs/DLs didn't serve our purpose of a theory-driven modeling approach.	9
Authors should submit the code and model as part of the publication to ensure the reuse of the work. Please provide a GitHub repository with all data and code for further review.	We provide a GitHub repository at the end of the Methods section: "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"	10
Reviewer 3		
This is an interesting paper that purports to use nutritional data in conjunction with machine learning methods to predict cardiovascular mortality. The metrics of success are model calibration and discriminatory ability, and the main conclusions are that including machine learning coupled with the use of nutritional can significantly improve risk prediction. While there is much to like in this work, the data, as presented do not support the main conclusions. Additional data are needed to truly evaluate the potential improvement that nutritional data and machine learning (in particular gradient boosting	Thank you for a thoughtful review. We defer to specific comments below.	

machines and random forests) can provide. Detailed comments include:		
Confidence intervals are calculated using Rubin's rules. The manuscript, however, is devoid of any details with respect to how this was implemented and the associated reference (#3) contains no additional information. This reviewer recognizes that Rubin's rules are not-infrequently used to calculate confidence intervals from imputed datasets, but the method is wrought with a number of challenges that make the interpretation of the resulting confidence intervals problematic. Firstly, the authors created 10 imputed training sets using an established boilerplate imputation method, however, no information is provided on how much data are missing and how many features need to be imputed. If, for example, a small fraction of the data missing, then the 10 imputed datasets will be very similar and the trained models will be very similar, leading to small confidence intervals. Additionally, if the imputed values are very similar this will also yield very similar models and small confidence intervals. As the authors use the calculated confidence intervals to determine what models are statistically superior, this is an important point. The authors should provide information on how different the various training datasets	We have added a table to the supplement that outlines how much missing data is present in each variable in the training and test data set. For simplicity, we have opted for one imputation (rather than 10) for each of training and test to fill in the missing data. We are now evaluating C-statistics on the test set using the confidence interval results from DeLong's test.	Supplementary Table B 9

are to help the reader decide whether the resulting confidence intervals are truly trustworthy.		
Confidence intervals can also be generated using bootstrapping (i.e., sampling with replacement) where imputation could be one on each bootstrapped training and testing sets. This method also has its challenges, but it is less likely to run in the problems mentioned above and it is more standard, at least in the machine learning literature.	We have opted for a simpler approach given your comments. First, we impute missing data in training and test separately to yield one complete training and one complete test set. Then, we do a manual grid search across two key parameters in each of random forest and gradient boosted machines. Finally we choose the best RF and GBM by choosing the model that minimizes the $(\text{slope}-1)^2 + (\text{C-statistic}-1)^2$ in the test set.	8-9
The PIs often refer to models that use “machine learning” and those that are “standard” (Cox proportional hazard modeling). It should be noted that a great deal of regression models also aspire to “learn” from data and hence would be considered to be “machine learning” models by many purists. More importantly, the authors only examine two types of machine learning models and therefore it is not clear whether other methods would yield better results (e.g., artificial neural networks). The manuscript would be improved if the PIs would refrain from making statements about “machine learning” in general and focus their conclusions on the precise computational models that they examined in this work.	The two machine learning methods we chose represent the most common alternatives to Cox modeling we found in the literature for time-to-event prediction that can be classified as machine learners other than Cox regression, with one representing bagging (random forest) and the other boosting (GBM). We explored an artificial neural network based approach called deepSurv without success.	7
An annoying aspect of machine learning model is that there are rarely guiding principles to dictate what the optimal parameters for any given model should be; e.g.,	We have retrained the models, employing the following manual grid search approach. For survival random forest, we assessed internal and external calibration and discrimination for number of trees (100, 300, 500) by number of randomly sampled variables at each node (1, 5, 10). For	8

a systematic parameter search is often required to find optimal parameters. Was this done in this case? The methods section only contains a short paragraph that lists one set of parameters for the gradient boosting machine and the random forests. The reader is therefore left to wonder whether these parameters are truly optimal for the problem at hand. Consequently, it is not clear whether the use of “machine learning algorithms alone” are “not of substantial benefit” as the authors claim.	gradient boosted machine, we tuned manually over the grid of number of trees (100, 300, 500) by tree depth (1, 5, 10).	
A minor point: The abstract lists this as a “Prospective study”, yet it uses an observational data set; i.e., all of the used are retrospective. This needs to be clarified.	We have changed prospective to retrospective.	2

VERSION 2 – REVIEW

REVIEWER	Shameer Khader AstraZeneca, USA
REVIEW RETURNED	14-Oct-2019

GENERAL COMMENTS	The authors have successfully addressed my queries and concerns. I have no further questions at this time.