MACHINE LEARNING TO PREDICT CARDIOVASCULAR DISEASE WITH NUTRITION

APPLYING MACHINE LEARNING TO EXPLORE NUTRIENTS PREDICTIVE OF CARDIOVASCULAR DISEASE USING CANADIAN LINKED POPULATION-BASED DATA

By JASON D. MORGENSTERN, B.Sc., M.D.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Public Health

McMaster University MASTER OF PUBLIC HEALTH (2020) Hamilton, Ontario

(Health Research Methods, Evidence, and Impact)

TITLE: Applying Machine Learning to Determine Nutrients Predictive of Cardiovascular Disease Using Canadian Linked Population-Based Data AUTHOR: Jason D. Morgenstern, B.Sc. (University of Guelph), M.D. (Western University) SUPERVISOR: Professor L.N. Anderson, NUMBER OF PAGES: xv, 121

Lay Abstract

This work explores the potential for machine learning to improve the study of diet and disease. In chapter 2, opportunities are identified for big data to make diet easier to measure. Also, we highlight how machine learning could find new, complex relationships between diet and disease. In chapter 3, we apply a machine learning algorithm, called conditional inference forests, to a unique Canadian dataset to predict whether people developed strokes or heart attacks. This dataset included responses to a health survey conducted in 2004, where participants' responses have been linked to administrative databases that record when people go to hospital or die up until 2017. Using these techniques, we identified aspects of nutrition that predicted disease, including caffeine, alcohol, and supplement-use. This work suggests that machine learning may be helpful in our attempts to understand the relationships between diet and health.

Abstract

The use of big data and machine learning may help to address some challenges in nutritional epidemiology. The first objective of this thesis was to explore the use of machine learning prediction models in a hypothesis-generating approach to evaluate how detailed dietary features contribute to CVD risk prediction. The second objective was to assess the predictive performance of the models. A population-based retrospective cohort study was conducted using linked Canadian data from 2004 – 2018. Study participants were adults age 20 and older (n=12 130) who completed the 2004 Canadian Community Health Survey, Cycle 2.2, Nutrition (CCHS 2.2). Statistics Canada has linked the CCHS 2.2 data to the Discharge Abstracts Database and the Canadian Vital Statistics Death database, which were used to determine cardiovascular outcomes (stroke or ischemic heart disease events or deaths). Conditional inference forests were used to develop models. Then, permutation feature importance (PFI) and accumulated local effects (ALEs) were calculated to explore contributions of nutrients to predicted disease. Supplement-use (median PFI (M)=4.09×10^{-4}, IQR=8.25×10^{-7} – 1.11×10^{-3}) and caffeine (M=2.79×10^{-4}, IQR= -9.11×10^{-5} – 5.86×10^{-4}) had the highest median PFIs for nutrition-related features. Supplement-use was associated with decreased predicted risk of CVD (accumulated local effects range (ALER)= -3.02×10^{-4} – 2.76×10^{-4}) and caffeine was associated with increased predicted risk (ALER= -9.96×10^{-4} – 0.035). The best-performing model had a logarithmic loss of 0.248. Overall, many non-linear relationships were observed, including threshold, j-shaped, and u-shaped. The results of this exploratory study suggest that applying machine learning to the nutritional epidemiology

of CVD, particularly using big datasets, may help elucidate risks and improve predictive models. Given the limited application thus far, work such as this could lead to improvements in public health recommendations and policy related to dietary behaviours.

Acknowledgements

I would like to thank my supervisor, Dr. Laura N. Anderson, for providing exceptional guidance, support, and mentoring throughout this thesis work. Her expertise in the epidemiology of cardiometabolic diseases, prevention of chronic diseases, the design of cohort studies, and many other areas, as well as her enthusiasm for innovation, has been invaluable throughout this process. I would also like to thank my committee members, Dr. Laura Rosella and Dr. Andrew Costa, for their support, guidance, and expertise, including in applying machine learning to epidemiology and public health. Additionally, I would like to thank Dr. Russell de Souza for his guidance and support, particularly in sharing his expertise in nutritional epidemiology.

I am very grateful to the Statistics Canada Research Data Centre at McMaster (RDC) for reviewing my proposal and providing access to an exceptional data resource. I would like to thank Li Wang, Peter Kitchen, and Anna Kata for their immense support throughout all aspects of my work at the RDC and for creating a very collegial and productive research environment.

Throughout my thesis work I have been fortunate to learn from my peers in Dr. Laura N. Anderson's research group. This has been an extremely supportive and collegial research environment. I've also benefited greatly from the guidance and patience of the MPH program director, Dr. Emma Apatu; curriculum coordinator, Angie Donato; and program coordinator, Stephanie Seiler.

I am very grateful for the support provided by the Canadian Institutes of Health Research (CIHR) through the Canada Graduate Scholarships – Masters Award, the clinician investigator program at McMaster, and the MPH program at McMaster.

I completed and submitted this thesis during the COVID-19 pandemic, while also training as a medical resident in Public Health and Preventive Medicine. I am grateful for the outstanding support from my supervisor; committee; the RDC staff; my supervisors and colleagues in public health; and CIHR during this exceptional period that made the completion of this research possible.

Finally, I would like to thank my fiancée Hannah and other family and friends who have all been very understanding, caring, and supportive during my thesis work.

Preface

This thesis is formatted as a sandwich thesis and comprised of an introduction in chapter 1, the first manuscript in chapter 2, the second manuscript in chapter 3, and the conclusion in chapter 1.The first manuscript is currently under review. The second manuscript will be submitted shortly. All the work conducted for the two manuscripts was completed as part of and during my MPH thesis work, towards the objectives of this thesis. I led the development of the research question, study design, and analysis plan and conducted the analysis, writing of the first draft, and the review/editing of both included manuscripts. My supervisor, Dr. Laura N. Anderson, contributed to the development of the research question, study design, and analysis plan; supervised the analysis and writing; and reviewed/edited all included work. Dr. Laura Rosella and Dr. Andrew Costa also contributed to the conceptualization and review of both manuscripts. All co-authors reviewed and edited the included work.

Table of Contents

List of Figures and Tables

List of all Abbreviations and Symbols

95% CI	95% Confidence Interval
a	Year
ASA24	Automated Self-Administered 24-hour
ALE	Accumulated Local Effects
ALER	Accumulated Local Effects Range
AUROC	Area Under the Receiver Operator Curve
CCHS 2.2	2004 Canadian Community Health Survey, Cycle 2.2, Nutrition
CVD	Cardiovascular Disease
FAMD	Factorial Analysis of Mixed Data
g	Gram
h	Hour
ICD-9	International Classification of Diseases 9
ICD-10	International Classification of Diseases 10
IQR	Interquartile Range
IU	International Units
kcal	Kilocalories
kg	Kilogram
LASSO	Least Absolute Shrinkage and Selection Operator
LOESS	Locally Estimated Scatterplot Smoothing
M	Median
mg	milligrams
PCA	Principal Components Analysis
PFI	Permutation Feature Importance
PUFA	Polyunsaturated Fatty Acids
RAE	Retinol Activity Equivalents
RCT	Randomized Controlled Trial

TMLE	Targeted Maximum Likelihood Estimation
µg	micrograms

Declaration of Academic Achievement

I declare this work to be my own. Specifically, I led the development of the research question, perspectives, study design, and analysis plan for the thesis, including both manuscripts. I conducted the analysis, wrote the first draft, and reviewed/edited the thesis and both included manuscripts. My supervisor, Dr. Laura N. Anderson, contributed to the development of the research question, perspectives, study design, and analysis plan; supervised the analysis and writing; and reviewed/edited all included work. Dr. Laura Rosella and Dr. Andrew Costa also contributed to the conceptualization of both manuscripts. All co-authors reviewed and edited the included work.

CHAPTER 1: Introduction and Objectives

Suboptimal diet recently surpassed smoking as the leading risk factor for non-communicable disease morbidity and mortality in the Global Burden of Disease Study.[1] One of the major drivers of this diet-related burden is cardiovascular disease (CVD), which accounts for one third of deaths worldwide.[2] While nutritional risk factors for CVD have been identified, many aspects of the role of diet in CVD remain poorly understood.[3] There is some disagreement among experts about nutritional factors such as carbohydrates,[4,5] eggs,[6,7] and red meat.[8,9]

There are many reasons for this ongoing debate. Recent commentaries have highlighted the complexity of dietary exposures, difficulty in accurately measuring food consumption, and long latency until disease onset as contributing factors.[10,11] Also, it has been argued that tiny effect sizes, multicollinearity, residual confounding, and measurement error are major issues for observational studies, which form the bulk of research in nutritional epidemiology.[12,13] Furthermore, the common use of single micro- and macronutrients, or even simple food groups and dietary pattern scores, fails to consider the full richness of diet. Finally, the lack of transparency in analysis and reporting of these observational studies has been emphasized as a further issue leading to the lack of reproducibility in nutritional epidemiology.[12–14] These challenges have led some scientists to go so far as to suggest that much of current nutrition knowledge may be a reflection of cumulative biases rather than scientific facts.[15]

A machine learning approach may help to mitigate some of these concerns. Specifically, it may help to better incorporate non-linear and non-additive relationships;

incorporate more complex dietary exposures, and account for multicollinearity. Given the current controversy surrounding domain knowledge in nutritional epidemiology, it may be helpful to use empirical analytic approaches that traditionally have not been applied in the field. Feature selection methods in machine learning can explore the totality of diet to find factors most predictive of disease.[16] Additionally, some algorithms like random forests or conditional inference forests incorporate variable importance rankings into the modelling process and are relatively resistant to multicollinearity.[17] Furthermore, these methods could analyze multiple levels of nutrition characterization, such as micro- and macronutrients, individual foods, and food groups, simultaneously for their relationship to predicted disease. An additional benefit of looking at diet as a whole may be a lessened risk of analytical bias, as all factors could be included, in some types of studies. Machine learning models are also generally more adept than statistical methods at incorporating non-linear and non-additive relationships, of which many are already known in nutritional epidemiology, and many likely remain to be discovered.[4,18,19] Finally, the machine learning practice of using resampling techniques for validation may help elucidate the most locally relevant dietary drivers of disease. Overall, these approaches may suggest novel avenues of investigation and improve prediction models by making better use of dietary factors.

Thus far there has been limited application of machine learning in nutritional epidemiology, but initial studies appear promising.[20,21] In the first manuscript presented in this thesis, "Perspective: Big Data and Machine Learning Could Help to Advance Nutritional Epidemiology" we outline the full rationale for applying big data and machine

learning in nutritional epidemiology. In the second manuscript, we apply conditional inference forests, a machine learning method, to make predictive models for cardiovascular disease (CVD) using a Canadian health survey that has been linked to administrative health databases. Conditional inference forests are a variant of random forests, which is a supervised machine learning technique that incorporates a form of variable selection into the model-building process, making it well suited to situations with many covariates. Additionally, random forests require little pre-processing of included features, easily incorporating features with different units and both categorical and continuous features.[16] They have frequently been used to identify gene-disease associations.[22] More recently, epidemiologists have begun using random forests to both predict health outcomes of interest and identify potentially important risk factors.[23,24] Random forests have proven highly adaptable, flexible in modelling non-linear relationships, able to identify complex interactions between variables, and useful for ranking variable importance.[25] They have previously been successfully used to model the relationship between dietary patterns and cardiac risk factors;[26,27] however, they have not frequently been used to predict CVD directly. Additionally, no relevant previous studies were conducted in Canada.

There are, however, some limitations associated with applying machine learning to nutritional epidemiology. For example, the theoretical properties of these models are unlike traditional advanced or causal epidemiology methods, which makes it more difficult to control for traditional sources of bias (such as confounding).[28] Additionally, although machine learning models may better incorporate non-linearity and non-

additivity, this added flexibility also makes them more vulnerable to overfitting (i.e. modelling non-generlizable noise in the training data).[16] This problem can be mitigated through the use of resampling-based validation, but it is unknown whether the added model complexity will provide sufficient predictive benefits to warrant the risks of overfitting. As such, in this thesis we investigate the prevalence of non-linear relationships to help ascertain any value of flexible modelling when using complex dietary exposures to predict cardiovascular disease. We also compare our results to existing literature. Another disadvantage of machine learning is the loss of model interpretability, which further obfuscates attempts to make inferences and limits model usability.[29] Generally, simpler models are preferred when they have equivalent predictive performance. This disadvantage is partly mitigated through application of interpretable machine learning methods.[29] These approaches may help to further delineate any advantage of the complex modelling methods.

In summary, there are several reasons why research applying machine learning to predict CVD with nutrition is necessary and novel. Despite decades of intensive research, there remains considerable debate/disagreement regarding dietary recommendations for CVD. This is significant as CVD remains the second leading cause of death in Canada after cancer, having caused 66 922 deaths in 2017.[30] Also, historical progress in reducing incidence of the disease may have stalled.[31] There remains some debate among experts on the nutritional causes of CVD, which may be due in part to complexities that are not adequately captured by traditional nutrient categories, dietary patterns, and statistical techniques. Therefore, the flexibility of ML approaches may offer novel insights. These

methods may also facilitate better predictive models. Notwithstanding its potential, there has been very little application of ML to understanding diet and CVD, adding further value to this investigation. Thus, the objectives of this thesis were:

Primary Research Objective

- To explore the use of machine learning prediction models in a hypothesis-generating approach to evaluate how detailed dietary features contribute to CVD risk prediction with interpretable machine learning methods

Secondary Research Objective

- To evaluate the predictive performance of the machine learning models incorporating detailed dietary variables.

.

CHAPTER 2: Manuscript – Perspective: Big Data and Machine Learning Could Help to Advance Nutritional Epidemiology

Jason D. Morgenstern[1], MD; Laura C. Rosella[2,3], PhD; Andrew P. Costa[1], PhD; Russell J. de Souza[1,4], ScD, RD; Laura N. Anderson[1], PhD

Affiliations

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

2. Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

3. Vector Institute, Toronto, ON, Canada

4. Population Health Research Institute, Hamilton Health Sciences, Hamilton, ON, Canada

Currently Under Review

2.1 Abstract

The field of nutritional epidemiology has faced exceptional challenges in establishing consistent and uncontested conclusions regarding the effects of diet on disease. Two major issues that have slowed progress are foods' inherent complexity and difficulties in measuring diet precisely and accurately. Diet's complexity leads to difficulties in defining exposures, controlling for confounding, and modelling non-linear and non-additive relationships. Most findings in nutrition research are currently based on observational studies, so wider use of randomized controlled trials (RCTs) have been advocated as a means of achieving clarity. However, adequate RCTs would be cost-prohibitive for most questions in nutritional epidemiology and would still entail concerns regarding generalizability and intervention definitions. We propose that judicious application of big data and machine learning in nutrition science could help to enhance observational studies by offering new means of dietary measurement, more tools to model the complexity of diet and its relationships with disease, and additional potential ways of addressing confounding. These developments could help to improve the reliability of findings in nutritional epidemiology.

2.2 Introduction

Conservatively, it is estimated that 250 000 different foods are consumed globally with many differing compositions and trillions of resulting combinations.[13] Given this degree of complexity, it is often unclear how best to assess relationships between nutrition and disease. Nutritional epidemiologists are faced with a diversity of exposures rivaling the genome, while also being equipped with less accurate and precise means of measurement.[32] Moreover, diet is not a constant, compounding measurement error at one timepoint with changes over the life course. Among other issues, the complexity of diet and exceptional obstacles to measurement have likely contributed to observed inconsistencies in the nutritional epidemiologic literature.[13,14,32] Many of these issues stem from a strong reliance on observational studies and could be partially addressed by greater reliance on randomized controlled trials (RCTs). However, significant obstacles to applying trials in nutrition science will probably restrict exploration of most research questions to observational study designs.[32–34] As such, uncertainty stemming from the complexity of diet and high levels of measurement error must continue to be addressed. We propose that judicious application of big datasets and machine learning may mitigate some issues stymieing progress in nutritional epidemiology by enabling better use of observational data.

2.2.1 Elaboration of Issues in Nutritional Epidemiology

Suboptimal diet recently surpassed smoking as the leading risk factor for non-communicable disease morbidity and mortality in the Global Burden of Disease Study.[1] This indicates that developing a detailed understanding of diet's effects on health must be

a top priority in efforts to improve public health. Unfortunately, diet and nutritional factors may be among the least understood risk factors.[35] After successful early years tackling micronutrient deficiencies, nutritional epidemiology largely shifted focus to understanding dietary risk factors for chronic diseases of aging.[36] A key focus of this endeavour has been the elaboration of the nutritional determinants of coronary artery disease, such as saturated fat. However, even on this long-studied topic, there remains some debate.[4,37–39] For example, numerous systematic reviews of the RCT evidence for saturated fats' effect have reached differing conclusions despite relying on the same evidence base.[37,38,40–45] Moreover, ongoing uncertainties range far beyond saturated fat and heart disease, such as the effects of carbohydrates,[4,5] eggs,[6,7] and red meat on all-cause mortality and/or cancer.[8,9] For instance, the International Agency for Research on Cancer classified red meat as "probably carcinogenic to humans" in 2015,[9] while a more recent systematic review recommended not reducing red meat consumption.[8] Despite decades of intensive study, substantial ambiguity remains regarding important health effects of many common nutrients and foods.

As revealed by these ongoing uncertainties, nutritional epidemiology has proven to be one of the most challenging areas of health research. A source of this uncertainty that has recently been criticized is a reliance on observational studies, wherein it is difficult to reconcile diet's complexity and the impacts of high levels of measurement error.[10,11] Given the vast space of potential dietary exposures and their combinations, it is difficult to know how to properly specify models. Frequently, only a subset of foods, food groups, or single nutrients are included. Even the growing practice of incorporating

dietary pattern scores fails to approach the full richness of diet. Moreover, changes in diet over the life course are often ignored.[46] Additional issues related to this complexity include strong correlations between diverse nutrients and also between dietary factors and other important determinants of health.[32] Together, these strong interdependencies can result in unstable coefficients in statistical models and residual confounding. A further result of the complexity of the exposures in nutritional epidemiology is the enablement of selective analyses, multiplicity issues in frequentist analyses, and biased reporting.[12–14] All of the issues related to dietary complexity are only exacerbated by difficulties in precisely and accurately measuring dietary intake.[12,13] Nondifferential measurement error can lead to both exaggeration and diminution of nutrients' effects, particularly in individual studies.[47,48] This problem has been highlighted as a contributor to reproducibility crises in scientific areas outside of nutritional epidemiology.[48] Furthermore, nondifferential measurement error makes it difficult to detect the small effect sizes that may be typical of individual nutrients.[13,32] It also exacerbates problems with multicollinearity and residual confounding.[47,49] Meanwhile, differential measurement error, when it exists, could compromise the internal validity of studies and further contribute to reproducibility issues. Overall, the challenging tasks of overcoming dietary complexity and measurement error in observational studies are probably major causes of inconsistency in nutritional epidemiology.

Proposed solutions to improve nutritional epidemiology include greater use of large RCTs,[12] improved dietary measurement with biomarkers,[50] and the use of Mendelian randomization.[51] It is true that large, long, and well-conducted RCTs may be

able to elucidate several well-characterized problems. However, there will be ongoing issues with blinding, ensuring adherence, and the long latency until disease onset.[34] It is likely possible to address these issues at great expense, but the cost will be prohibitive for most questions that are not considered of highest priority. Additionally, enduring questions about generalizability and the exact characterization of dietary interventions will remain. Therefore, observational studies are likely to remain the dominant, and perhaps even preferable, study design for many questions in nutritional epidemiology. The use of biomarkers to confirm and better characterize nutrition exposures is a promising method for reducing measurement error. However, there remain concerns regarding the specificity of such markers.[52] Additionally, these methods are unlikely to fully address issues stemming from dietary complexity. Mendelian randomization, which uses genetic polymorphisms as instrumental variables to infer causal effect estimates, is another promising avenue.[53] In addition to mimicking randomized interventions, Mendelian randomization studies assess lifelong exposures. As an example, they have provided further evidence that LDL is likely to be a causal factor in coronary heart disease.[54] However, it is difficult to find genetic factors for nutritional exposures that meet the necessary criteria to be an instrumental variable for Mendelian randomization studies. There have been attempts to use genes associated with cruciferous vegetable and dairy consumption,[55] but concerns have been raised about pleiotropy and population stratification. Mendelian randomization studies may well be a useful tool for advancing nutrition research methodology; however, they will likely not be able to address all questions.

Nutritional epidemiology has relied heavily on observational studies, which currently entails grappling with dietary complexity and high levels of measurement error. These issues notwithstanding, observational studies are likely to remain the dominant research modality in nutritional epidemiology given the barriers and fundamental issues involved in more widespread application of RCTs. Innovations in the use of dietary biomarkers and Mendelian randomization offer substantial promise but will probably have limited applicability. In the context of high dietary complexity and measurement error with no easy solutions, greater consideration of big data and machine learning could enhance the use of observational data to advance nutritional epidemiology.

2.2.2 Big Data and Machine Learning

"Big data" refers to datasets that usually include both many observations and variables, making the use of traditional software and statistical tools difficult.[56] As a result, there is often a need for more flexible modelling than provided for in classical statistical analysis. The specific size of datasets required to constitute big data varies depending on the context. Generally, it has been characterized by the 'three V's', which include the data's volume, velocity, and variety.[57] Big datasets are also often less structured than traditionally collected data, and may be a byproduct of something, rather than an intentionally collected sample.[58] Use and availability of big data has risen alongside exponential improvement and expansion of computing devices, data storage capacity, and the internet of things. The internet of things refers to networks of computerized objects that record and share data among themselves with no human

intervention.[59] In addition to many effluent data sources relevant to health (e.g. electronic health records and social media), researchers have recently begun to contend with big data arising from the investigation of complex biological systems such as the genome and microbiome.[60] While diet has usually been studied using simplified constructs, its complexity could become another source of big data. More recently, it has become possible to imagine capturing both the complexity in food itself, which would entail many variables, and to make many more observations than we have in traditional investigator-generated datasets.

Machine learning is a subfield of artificial intelligence, which encompasses a wide range of approaches that seek to provide computers with the ability to learn tasks without being explicitly programmed.[61] These approaches rely on algorithms that derive patterns from data with little human input.[62] This contrasts with statistical techniques that rely more on human knowledge for verification of model assumptions and variable selection.[63] Statistical techniques also emphasize a theoretical approach to hypothesis testing and uncertainty estimation, which is not common in machine learning. Finally, machine learning approaches tend to make greater use of cross-validation than statistical regression methodologies, although cross-validation can be used in both approaches. Cross-validation refers to randomly splitting a dataset into mutually exclusive components followed by iterative training and testing of a model on the different components to generate an average estimate of performance, which is more likely to apply out-of-sample. Machine learning is often applied to big data, where it is sometimes difficult to apply conventional statistical approaches.

Machine learning can be broadly classified into supervised and unsupervised approaches.[64] For supervised approaches, an example dataset, including complete label or outcome information, is used by a learning algorithm to identify patterns in the explanatory variables. The trained model is then applied to make predictions on new data. Common examples of supervised algorithms include neural networks, random forests, and support vector machines. In contrast, for unsupervised approaches, there are no human-supplied examples for the observations in a dataset and the algorithm searches for latent patterns or groupings.[65] Examples of unsupervised approaches include dimensionality reduction, such as principal components analysis (PCA) and autoencoders; and clustering approaches, such as k-means and k-medoids.[64] An additional subfield of machine learning is feature selection, which aims to remove variables that are irrelevant to the outcome in supervised problems,[66] thereby overcoming the curse of dimensionality. The curse of dimensionality refers to a phenomenon whereby prediction accuracy decreases if too many irrelevant variables are added to an analysis, especially in the context of a limited sample size.[67] Typical examples of feature selection algorithms are least absolute shrinkage and selection operator (LASSO), genetic algorithms, and recursive feature elimination. In health research, machine learning has been applied to the analysis of genome- and microbiome-derived data, where conventional analyses are limited by the curse of dimensionality[66] and there is limited mechanistic understanding or theory to guide analysis. Several comprehensive review articles relating big data and machine learning to epidemiology and public health provide greater detail on both topics, but nutritional epidemiology has not yet been discussed in detail.[68,69]

2.3 Current and Potential Applications of Big Data and Machine Learning in Nutritional Epidemiology

The objective of this perspective article is to highlight some of the ways that developments in big data and machine learning can address issues in the use of observational data in nutritional epidemiology. Their application could reduce measurement error with new tools, improve modelling of nutrition's non-linear and non-additive relationships with disease, and allow better characterization of the complexity of diet and its confounders. Such developments could improve predictive models for chronic diseases and enhance inferences about the relationship between diet and disease.

2.3.1 New Measurement Methods

Big data related to nutrition are now generated through multiple means. These data may lead to reduced measurement error in nutritional epidemiology through the provision of more objective, scalable, and affordable means of data collection. For example, web-based tools, such as the Automated Self-Administered 24-hour (ASA24) Dietary Assessment Tool, capture 24-hour recalls without the time and expense required by trained interviewers;[69] however, this does shift the burden of collection to the respondent, who may not be willing. Such online self-report modalities are readily accessible and could allow the recruitment of larger study populations, with more frequent, detailed, and longitudinal characterization of their diets. Less user-burdensome methods include the large and detailed grocery purchase habits of populations generated by consumer rewards programs, or the eating patterns recorded in smartphone tracking

apps, which could be used to develop cohorts with unprecedented dietary detail and sample sizes. An example of the latter type of effort is the Harvard Apple Women's Health study, the first long-term research study that aims to use health app data to advance the understanding of menstrual cycles and their relationship to various health conditions.[70] Further, machine learning is being used to find reliable and specific biomarkers to characterize dietary exposures,[71] which may improve measurement of some aspects of diet.

An additional means of collecting nutritional data for research may come from the use of machine learning models to classify pictures of food.[72–77] Given the ubiquity of smartphones, such techniques may facilitate less onerous and more regular diet records, reducing both differential and nondifferential measurement error. They could also enable practical, accurate, and detailed measurement of diet trajectories over longer periods of time. Additionally, machine learning-based food recognition could incorporate auditory and other contextual information to improve the accuracy of measurement. Considering the rapid growth of the internet of things industry, it is conceivable that such measurements could even be conducted passively. Home security systems, thermostats, fridges, voice assistants, and many other appliances are being equipped with cameras, microphones, and WIFI functionality. Thus, with the permission of their owners, such devices could be recruited for dietary measurement. Another approach is the digestion of the data produced on social media and web search platforms, which often includes integrated food and health-related information.[78] With these new data collection modalities, observational studies could be rapidly scaled either passively or through

dissemination of relevant apps, increasing study precision and potentially generalizability. However, a cautious consideration of the impact of selection bias and systematic measurement error in such large and often secondary datasets would require further investigation. Additionally, many of these approaches entail major privacy concerns. Careful, collaborative work will be needed to ensure research projects involving this data are ethical, collect only strictly necessary information, and that security is sufficiently robust to ensure that other parties (e.g. insurance companies) cannot access it.

2.3.2 Tools to Model the Complexity of Diet in Relation to Disease

Current approaches to modelling the relationship between nutritional exposures and disease typically focus on single nutrients or foods, though there has been a growth of studies including simplified, low-dimensional representations of overall diet, such as the nine-point Mediterranean Diet Score,[79] (alternative) Healthy Eating Index,[80,81] or DASH diet score.[82] Machine learning could afford inclusion of more dietary explanatory variables and help to identify the most predictive ones empirically.[83] Nutritional epidemiology may benefit from incorporating rich, *a posteriori*, dietary exposures with machine learning approaches. This is not completely novel, as some clustering, dimensionality reduction, and feature selection approaches have been applied to derive important aspects of diet. For example, PCA,[84] k-means clustering,[85] and LASSO[86] have been used to generate *a posteriori* dietary patterns and associate them with various disease outcomes. However, thus far mostly linear approaches have been used. Additionally, there has been no use of similar techniques to analyze multiple levels of food classification simultaneously. Machine learning could be used to incorporate these

multiple levels, such as micro- and macronutrient content, specific food types, and food groups, within the same analysis. As a result, the most important aspects of diet could be determined empirically for a given problem, which was called for in a recent commentary.[10] However, these techniques are not without caution. With no initial expert curation of variables and careful validation, important predictors could be missed and unimportant predictors incorrectly emphasized.

In addition to better capturing the richness of nutrition, machine learning can model non-linear and non-additive relationships more flexibly, particularly when they are unknown. Most existing models in nutritional epidemiology assume monotonic, linear, and additive relationships between diet and disease. However, there is emerging evidence that non-linear relationships may be more common than previously thought. For example, salt,[18] carbohydrate,[4] and fats[19] may all have u- or j-shaped relationships with cardiovascular diseases. Additionally, there is support for various interactions in nutritional epidemiology. For instance, the impact of salt on hypertension seems to be moderated by the potassium and simple carbohydrate content of the diet.[87–89] Machine learning models could incorporate both known and unknown non-linear and interactive relationships in models that include numerous predictors.

While limited, there are studies that have applied machine learning to incorporate greater dietary complexity and to more flexibly model health-related outcomes thus far. For example, a stochastic gradient boosting regression algorithm was used to accurately predict individual glycemic responses to food with detailed dietary, lifestyle, medical, laboratory, anthropometric, and microbiota data.[20] The model included thousands of

variables and used permutation feature importance and partial dependence plots to interpret their contributions to predictions. Unexpectedly, the model placed greater emphasis on microbiota-related variables. This study was unique among nutrition studies in using a surrogate outcome with low latency and having unusually precise dietary measurements. Another more typical nutritional epidemiologic cohort study found a 22% increase in the accuracy of cardiometabolic risk factor prediction when comparing random forest to linear regression.[90] This study incorporated rich dietary independent variables and used PCA for dimensionality reduction. Lastly, another recent cohort study compared the performance of random survival forests and gradient boosted machines using nutritional explanatory variables with a standard Cox proportional hazards model lacking nutritional data in predicting cardiovascular mortality.[91] The machine learning models outperformed the statistical model in both predictive discrimination and calibration. Interestingly, addition of nutrition data to the statistical model did not improve its predictive discrimination or calibration, but, when added to the machine learning models, both measures of prediction performance improved. This lends support to the proposition that machine learning models may better leverage the full richness of diet in modelling health outcomes.

2.3.3 New Means of Controlling for Confounding Variables

Incorporating data with both higher numbers of observations and more variables into nutritional epidemiologic studies, alongside machine learning analytical techniques, could possibly reduce residual confounding. Potential opportunities include a higher chance of avoiding unmeasured confounding with higher dimensionality, using machine

learning to include novel types of unstructured data, leveraging higher dimensionality for greater use of negative controls and instrumental variables, and using new machine learning approaches to controlling for confounding with high dimensional data when applied within causal frameworks. Specifically, big datasets including variables related to microbiota, genetics, metabolomics, behavioural factors, environment, and social determinants of health could enhance analyses by helping to avoid missing unmeasured confounders.[92] Furthermore, machine learning can make entirely new types of data available for inclusion in models. For example, deep learning has been used to derive variables describing the built environment from satellite images.[93] Further big data-types that could be considered include medical information from free-text clinical notes,[94] physiological data from wearable devices,[95] and populations' demographic, socioeconomic, and health records from linked government datasets.[96] Another potential advantage of incorporating big data is the greater availability of negative controls, which can help to ascertain the likelihood of residual confounding; and instrumental variables, which can allow observational studies to mimic randomized trials under certain assumptions.[97–101] Finally, new machine learning methods are being developed that may help to reduce residual confounding, including feature selection approaches[102,103] and methods of combining many weaker proxy variables for stronger but unobserved confounders into propensity scores.[104] These approaches have often performed comparably to or better than expert-based propensity scores.[102,104–112] However, they should be used with caution due to their early stage of development and their potential for worsening model instability.[103] Altogether, big data provides an opportunity to improve

measurement and representation of factors beyond diet, while machine learning could facilitate the analysis of these high-dimensional datasets.

2.3.4 Improving Disease Prediction

Prediction models for cardiovascular disease, one of the major focuses of the science of diet, have been extensively studied for the past five decades. Risk prediction tools, such as the one originally developed from the Framingham study in 1967, are still commonly used in clinical practice to determine the need for hypertensive and cholesterol medications.[113] More recently, population-level models have been developed that can be used to guide the implementation of public health preventive interventions, inform policymakers about future disease burden, and assess the impact of public health actions.[114–117] Typically, prediction models have included very few dietary components and,[113] when included, greatly simplified dietary factors are typically used (e.g. only a small number of foods or nutrient ratios[114,118]). Their absence in prediction models is likely related to commonly being omitted from the data sources used to generate prediction models, as well as the collection of dietary data being relatively arduous. Additionally, oversimplification of dietary variables, when they are included, may result in a lack of added predictive performance, further disincentivizing their inclusion by later researchers. Inclusion of rich dietary data in predictive models could be an important and largely untapped avenue for improved performance. Such data is more likely to be of benefit if new data sources permit a reduction in nondifferential measurement error, allowing models to take advantage of ensembles of relatively small effect sizes. Additionally, prediction applications are where machine learning models have historically

excelled. Therefore, the use of better data alongside machine learning models, with their ability to incorporate richer dietary variables, more comprehensive covariates, and higher complexity of relationships, offers additional opportunities to improve prediction models. A recent cohort study supports this idea, as it demonstrated synergistic prediction performance improvements for cardiovascular mortality when combining rich dietary data with machine learning methods.[91] A further advantage of applying the machine learning paradigm is that cross-validation makes many algorithms largely resistant to the effects of multicollinearity in the context of prediction. Furthermore, this internal validation could permit the identification of dietary patterns and factors that are most relevant in specific populations for prediction of specific diseases. Overall, both novel data sources and machine learning methods offer opportunities to improve chronic disease prediction models through incorporation of rich dietary data.

Notwithstanding the potential positive impacts on predictive modelling, the application of big data and machine learning has several potential pitfalls. The impact of selection bias and systematic measurement error in novel data sources has already been described. If excluded from training datasets, vulnerable populations could be further marginalized by predictive algorithms that are inaccurate for them. Additionally, given that machine learning methods are usually atheoretical and sometimes inscrutable, they are vulnerable should some aspect of the underlying data generating process change. In that case they may unexpectedly become inaccurate, so researchers should take steps to safeguard against this eventuality. Another important consideration is that complex machine learning models do not always improve prediction. They are more flexible than

most parametric regression models; however, this makes them more susceptible to overfitting.[64] Their relative advantage depends on the importance of interactions and non-linearity for a given problem. Ideally, many machine learning and statistical models should be trialed and evaluated using cross-validation for a given prediction problem. Non-linear parametric statistical models such as fractional polynomials and restricted cubic splines should also be considered.[119,120] A related issue with most machine learning approaches is that they typically require more observations per variable to make robust predictions.[64] Therefore, it may often not be appropriate to apply machine learning techniques in smaller datasets. Alternatively, the numerous feature selection and dimensionality reduction techniques in the machine learning corpus can be used, alongside domain knowledge, to reduce the number of included variables. Also, some supervised machine learning algorithms, such as random forest, are relatively robust in the presence of uninformative variables. In general, statistical techniques will perform better and be more generalizable in situations where only a small dataset is available and both non-linear and non-additive relationships are not very influential. Finally, it is important to note that modelling health outcomes is distinct from the application domains in which machine learning was originally developed.[121] For example, in most computer vision contexts there is a very high signal-to-noise ratio. Meanwhile, in medical domains, a significant proportion of prediction error likely comes from unmodifiable stochasticity, posing a lower ceiling on possible prediction accuracy. So, in health research, uncertainty estimates and probability predictions are more important than they often have been in machine learning. While not often done, uncertainty estimates can be derived for machine

learning analyses using resampling and Bayesian approaches. Finally, in the health research context it is important to focus primarily on calibration as a predictive performance metric, which entails the concordance between predicted and observed absolute probabilities across the full spectrum of risk.[122,123] This is in contrast to the more frequent use of discriminative performance metrics such as area under the receiver operator curve in machine learning research.

2.3.5 Informing Causal Studies

While most machine learning and big data research has focused on prediction or classification, it could also help to inform inferential studies in nutritional epidemiology. First, by reducing nondifferential measurement error and increasing sample sizes, big data could allow detection of smaller effect sizes and reduce the effects of multicollinearity on coefficient stability.[32,47] Further, application of machine learning could help with hypothesis generation, particularly as methods for interpreting complex algorithms improve. Already, current techniques such as permutation feature importance, accumulated local effects, partial dependence plots, Shapley values, local interpretable model-agnostic explanations, and interaction h-statistics can be used with almost any machine learning model to reveal the shape of relationships between predictors and outcomes, as well as important interactions.[29] Additionally, dimensionality reduction and feature selection techniques can be used to derive empirical dietary patterns and predictive dietary factors for further study. Given nutrition's high level of complexity, these exploratory approaches may be particularly helpful. Also, an advantage of data-driven dietary patterns and variable selection is that they may be more reflective of

relevant dietary variation in a local population than *a priori* scores developed elsewhere.[124] Furthermore, if the totality of dietary exposure data is incorporated into an analysis with machine learning techniques, including multiple food/nutrient classification levels, there may be less temptation or possible explanations for conducting selective analyses. An additional consideration is that big data and machine learning may enable more comprehensive and precise incorporation of confounders into analysis, possibly reducing residual confounding. Finally, greater availability of big data might allow more study of meta-dietary factors such as timing of meals, preparation and cooking methods, social aspects of dining, the location of eating, and additional contextual factors (e.g. eating while watching TV).

Machine learning could also enhance formal causal inference studies in nutritional epidemiology within a potential outcomes framework. New ways of using machine learning to automatically generate propensity scores and select confounders from high dimensional data have already been described.[102,103] Additionally, targeted maximum likelihood estimation (TMLE) can serve as an alternative to propensity score- and G-computation-based causal effect estimation while incorporating ensemble machine learning methods, such as Super Learner.[125] In concert with Super Learner, TMLE has demonstrated less biased estimation of causal effects than traditional approaches. The advantage seems to stem from using the machine learning ensemble during a secondary targeting phase to better balance the bias-variance tradeoff in estimation of causal effects.[125]

While big data and machine learning may be helpful for informing causal studies through both hypothesis generation and application within causal inference frameworks, they are not enough for causal inference on their own. For any experimental data, many causal models typically exist that could explain observed relationships.[126] Therefore, experts' domain knowledge is essential for informing *a priori* causal models, interpreting results generated by algorithms, and putting findings into the wider context of evidence. In particular, while big data may provide additional opportunities to control for unmeasured confounders, use negative controls, and find instrumental variables; without adequate forethought it also poses a higher risk of biasing effect estimates and masking direct effects through unintended inclusion of collider and mediator variables in models.[103] Further issues when using big data and machine learning to inform causal studies are selection bias and systematic measurement error. Both must be better understood to ensure valid and generalizable results. Lastly, feature selection techniques should be used in this context with caution. If these techniques are used to specify a final model, particularly if the outcome variable was used during feature selection, there is a high risk of inaccurate inferences.

2.4 Conclusion

Overall, greater use of big data and machine learning could improve the reliability and validity of nutritional epidemiologic findings, while still using primarily observational evidence. Specifically, the incorporation of big data and machine learning into epidemiologic analyses could enable reduced measurement error, better representation of the complexity of diet and its confounders, and improved consideration

of intricate relationships between diet and disease. In turn, such improvements could help to improve both predictions and inferences regarding the relationships between diet and disease. With increased use of big data and machine learning, many of the challenges and criticisms of nutritional epidemiology could potentially be addressed.

CHAPTER 3: Manuscript - Development of Machine Learning Prediction Models to Explore Nutrients Predictive of Cardiovascular Disease Using Canadian Linked Population-Based Data

Jason D. Morgenstern[1], Laura Rosella[2,3,4,5], Andrew Costa[1,3,6], Laura N. Anderson[1,7]

Affiliations

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
2. Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
3. Institute for Clinical Evaluative Sciences (ICES), Toronto, Ontario, Canada
4. Public Health Ontario (PHO), Toronto, Ontario, Canada
5. Vector Institute, Toronto, Ontario, Canada
6. Department of Medicine, McMaster University, Hamilton, Ontario, Canada
7. Population Health Research Institute, Hamilton Health Sciences, Hamilton, ON, Canada

Manuscript Under Preparation for Submission: Planned for submission soon.

3.1 Abstract

Importance: Machine learning may improve use of observational data to understand the nutritional epidemiology of cardiovascular diseases (CVD) through better modelling of non-linearity, non-additivity, and dietary complexity.

Objective: To develop machine learning prediction models for exploring how detailed dietary features are related to CVD risk prediction and evaluating their predictive performance.

Design: A retrospective cohort study from 2004 – 2018 with a maximum follow-up of 14 years.

Setting: A Canadian population-based study.

Participants: A total of 35 107 adults who completed the 2004 Canadian Community Health Survey, Cycle 2.2, Nutrition (CCHS 2.2) were considered for inclusion. This is a national, population-based survey with prospective linkage to administrative data until 2018. The Survey is representative of 98% of the population and had a 76.5% participation rate. Individuals who were less than 20 years old, lived in Quebec, were pregnant, did not provide a dietary recall, or who had a known history of CVD were excluded.

Exposures: Sixty-one nutrition-related features measured from the 24-hour dietary recall and general health components of the CCHS 2.2.

Main Outcome and Measure: Ischemic heart disease or stroke event or death defined as International Classification of Diseases (ICD)-9 codes 410-414 and 430-438 or ICD-10 codes I20-25 and I60-69 used in linked administrative databases of hospital discharges and national deaths.

Results: 12 130 individuals were included in our study with a median age of 50.0 (IQR=34.0 – 65.0) and 6850 were female (56.5%). 1120 (9.2%) individuals developed ischemic heart disease or stroke. Twenty-three (37.7%) nutrition features had a positive median permutation feature importance (PFI). Supplement-use (median PFI (M)=4.09 x 10^{-4}, IQR=8.25 x 10^{-7} – 1.11 x 10^{-3}), caffeine (M=2.79 x 10^{-4}, IQR= -9.11 x 10^{-5} – 5.86 x 10^{-4}), and alcohol (M=1.52 x 10^{-4}, IQR=1.99 x 10^{-5} – 5.02 x 10^{-4}) had the highest median PFIs for nutrition-related features. Supplement-use was related to decreased predicted risk of CVD (accumulated local effects range (ALER)= -3.02 x 10^{-4} – 2.76 x 10^{-4}), caffeine was related to increased predicted risk (ALER= -9.96 x 10^{-4} – 0.035), and frequency of alcohol-use had a u-shaped relationship with predicted risk (ALER= -8.38 x 10^{-4} – 0.002) . A diverse mixture of non-linear dose-response curves was observed, such as threshold, j-shaped, and u-shaped relationships. The model with the best prediction performance during training had a test logarithmic loss of 0.248.

Conclusions and Relevance: Our study is one of the first to apply machine learning techniques to the prediction of CVD using detailed population-based dietary data and showed competitive prediction performance. Machine learning models identified numerous nutrition features important for prediction of CVD risk in exploratory analyses, which demonstrated a mix of linear and non-linear relationships. More research applying machine learning to the nutritional epidemiology of CVD, particularly using big datasets, may help elucidate risks and improve predictive models.

3.2 Introduction

Suboptimal diet is the leading risk factor for deaths globally.[1] Cardiovascular diseases (CVD), which cause one-third of deaths worldwide, are responsible for much of the morbidity and mortality stemming from suboptimal diet. Age-standardized mortality from CVD decreased from 1990 to 2015 in high-income countries, but this trend may now be slowing or reversing.[127,128] Randomized controlled trial (RCT) evidence supports a protective effect of Mediterranean dietary patterns, and observational evidence implicates multiple dietary factors such as vegetables, fruits, and trans-fat in CVD.[38] However, there remain significant inconsistencies in the evidence and recommendations regarding the impacts of common nutrients and foods on CVD, such as red meat,[8,9] carbohydrates,[4,5] and eggs.[6,7] The lack of randomized controlled trials, challenges addressing measurement error, long latencies until disease onset, multicollinearity, and the complexity of foods/diets are some of the factors that make the dietary determinants of CVD difficult to study.[10,11]

Machine learning methods may address some issues in the study of dietary risks for CVD through the incorporation of non-linear effects, interactions, and high-dimensional sets of covariates. Many potential non-linear, including u- or j-shaped, and non-additive relationships between dietary risk factors and CVD have been identified.[18,19,87–89,129] Additionally, at least 250 000 distinct foods are consumed globally, making consideration of large numbers of covariates important.[15] Therefore, more complex modelling approaches may be advantageous when studying the nutritional epidemiology of CVD. There has been some use of machine learning to assess the relationship between dietary factors and cardiometabolic risk or glycemic responses, but

none have studied CVD directly.[20,21] Finally, few existing CVD predictive models incorporate dietary features.[113]

The primary objective of this study was to develop machine learning prediction models to explore how detailed dietary features are related to CVD risk prediction with interpretable machine learning methods in a Canadian population-based cohort. The secondary objective was to evaluate the predictive performance of the models.

3.3 Methods

3.3.1 Study Design and Data Sources

We conducted a retrospective cohort study and developed a prediction model for cardiovascular disease with detailed dietary data. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) reporting guidelines were followed where relevant.[130] The study sample was comprised of participants in the Canadian Community Health Survey, Cycle 2.2, Nutrition (CCHS 2.2), which was administered from 2004 – 2005 by Statistics Canada.[131] This survey had a multistage, stratified cluster design representative of 98% of Canadians of any age. The survey had a response rate of 76.5% and a sample size of 35 107. For our study, we excluded individuals if they were younger than 20 years of age, resided in the province of Quebec (because linked outcome data was not available for these residents), were pregnant at the baseline survey interview, did not provide a dietary recall, or had pre-existing CVD. Pre-existing CVD was identified if participants reported heart disease in the CCHS 2.2 general health component or had a CVD-related hospital discharge recorded in linked

administrative databases anytime from April 1st, 1999 (the earliest available linked date) until one year after completing the survey.

Individual-level data from the CCHS 2.2 was linked to the Discharge Abstracts Database and the Canadian Vital Statistics Death Database by Statistics Canada. The Discharge Abstracts Database is maintained by the Canadian Institute for Health Information and includes International Classification of Diseases-9 and -10 (ICD-9 and -10) codes for diagnoses relating to all hospital discharges of Canadian citizens in every province except for Quebec.[132] The Discharge Abstracts Database has been linked to the CCHS 2.2 from April 1st, 1999 to March 31st, 2018. The Canadian Vital Statistics Death Database is a census of all deaths occurring in Canada, including the related ICD-9 and -10 codes. This database was linked to CCHS 2.2 from January 1st, 2000 until December 31st, 2018 (see supplementary figure 1 for a graphical representation of the study design, including exclusion criteria). Both linked datasets are de-identified and stored in the Statistics Canada Research Data Centres, which are secure facilities restricted to vetted employees.

3.3.2 Ethics and consent

The data linkage used for this study was approved by Statistics Canada's Executive Management Board, and access to the data is governed by Statistics Canada's directive on record linkage.[133] Under applicable ethical standards, this governance structure allows analyses to occur without research ethics board approval.[134]

3.3.3 Outcome

We defined the incidence of CVD as hospital discharges or deaths where the primary associated diagnosis was ischemic heart disease or stroke, as identified in the Discharge Abstracts Database and Canadian Vital Statistics Death Database. We used ICD-9 codes 410-414 or 430-438 and ICD-10 codes I20-25 or I60-69 to identify ischemic heart disease or stroke.[135,136] We categorized the outcome as a binary feature (yes/no). Available data allowed relevant hospital discharges to be detected anytime from the completion of the CCHS 2.2 survey in 2004 or 2005 until March 31st, 2018, while deaths could be detected up until December 31st, 2018. Therefore, there was a follow-up time of up to 14 years.

3.3.4 Features

All features included in the models were measured during participants' baseline survey interviews from 2004-2005. Participants in the CCHS 2.2 completed at least one 24-hour dietary recall conducted by trained interviewers using a computer-based application called the Automated Multiple-Pass Method.[131] The general health component of the survey was collected using a standardized questionnaire. Most interviews for both the 24-hour dietary recall and general health component were conducted in-person. When this was not possible, they were conducted over the telephone.

All features included in the models for this study originated from the "General Health, vitamin and mineral supplements and 24-Hour Dietary Recall - HS.txt" file, which contains a combination of answers to the general health questionnaire and nutrition-related features derived from responses to the 24-hour recall component of the

CCHS 2.2.[131] This file included the intake of macro- and micronutrients derived from the 24-hour dietary recall using the Canadian Nutrient File as a reference.[131] We included most nutrients derived from the 24-hour dietary recall as features in our models. Macronutrients were provided as both absolute grams consumed and percent of daily energy intake. We only included the percent of daily energy intake features in our models. Several different derivations of folate/folic acid were available. We only included one feature for both natural folate and added folic acid. We included fruit and vegetable consumption (average number of times eaten daily over the previous 30 days), frequency of alcohol-use, and supplement-use from the general health component of the CCHS 2.2. Overall, sixty-one nutrition-related features were included (see supplementary table 1 for more details about each feature). We also included fourteen socioeconomic, demographic, psychological, and behavioural features from the general health component that are well-established predictors of CVD (see supplementary table 1). Examples include age, sex, marital status, stress, physical activity, household income, and smoking status. Given our interest in exploring the total direct contribution of nutrition features to predicted CVD risk, metabolic risk factors that may be on the causal path between nutrition and CVD outcomes were not included (e.g. body mass index, hypertension, diabetes).

3.3.5 Data Pre-processing

After applying the exclusion criteria and creating a binary feature for the primary outcome, we examined summary statistics. We avoided including categorical feature levels with less than 20 participants to avoid the perfect separation problem.[137] As a result, one dichotomous feature was dropped (fibre supplement-use). Additionally, some

of the levels of two other categorical features were combined (cultural or racial origin and household income; see supplementary table 1). A feature for the percent of life lived in Canada was derived from participants' immigrant status, age, and years since immigration. Additionally, a more detailed smoking feature (see table 1) was derived from smoking status, the number of cigarettes smoked among daily smokers, and the number of years since quitting among former smokers.

9.32% percent of participants had missing data for at least one feature. The median percent missing data per feature was 0.02% (IQR=0.06%). See supplementary table 2 for the number of missing values for each feature included in the models. We conducted single imputation using factorial analysis of mixed data (FAMD), a principal components analysis method, with the "missMDA" R package.[138] First, we estimated the number of dimensions using 100-fold cross-validation by minimizing the mean squared error of predictions, considering one to five dimensions. Then, we computed FAMD with the best performing number of dimensions to estimate missing values.

3.3.6 Statistical Analysis

Descriptive statistics stratified by outcome were evaluated. We then developed conditional inference forest models[25] to predict CVD-status using the R package "mlr."[139] Conditional inference forests are like random forests but use a non-parametric significance test to reduce bias in feature importance calculations and lessen overfitting.[25]

We randomly assigned training and testing datasets with 70% and 30% of the total observations, respectively, stratified by the outcome.[140] Hyperparameters of the conditional inference forests, including the number of features randomly sampled for each

node of the base trees (mtry) and the minimum significance-level needed to perform splits in the base trees (mincriterion), were tuned on the training dataset with 10-fold cross-validation.[16] See supplementary table 3 for the full grid of hyperparameters constructed for testing and their associated performance values. We considered tuning the number of base trees used for each model, maximum depth, minimum number of observations at a split, and minimum number of observations at terminal nodes, but these had no impact on testing performance or overfitting tendency in initial tests; thus, they were not tuned or restricted. Regarding the number of base trees used for each model, it should be noted that only 750 and 1000 base trees were considered. Given that there was no identified performance difference, 750 trees were used in all models to reduce computational burden. Median logarithmic loss on out-of-fold data was used to select models, based on minimizing prediction error.

We selected the four hyperparameter sets with the best prediction performance to generate four different models on both the training and testing datasets for computation of permutation-based feature importance (PFI) and accumulated local effects (ALEs). Therefore, eight models in total were used to compute eight sets of PFIs and ALEs for each feature (four on the training dataset and four on the testing dataset). This was done to sample some of the variability in these estimates from several models with similarly high predictive performance. Prediction performance overall was evaluated using only the single best-performing hyperparameter set.

PFIs were generated using the "party" R package.[29,141] We used logarithmic loss to evaluate importance. Median PFI and interquartile ranges (IQRs) across all selected

models were calculated. Median PFIs can be interpreted as the median increase in logarithmic loss of model predictions after randomly permuting a specific feature.

We computed ALE plots across a grid of at most 20 values for each feature using the "iml" R package.[142] The number and position of the feature-levels at which accumulated local effects were calculated depended on the feature itself (i.e. whether ordinal or continuous) and density of the distribution of participants along the feature. Locally estimated scatterplot smoothing (LOESS) was used to fit lines to each model's ALEs trajectory in the plots. ALEs can be interpreted as the average main effects of a given feature-level on the risk of the outcome relative to average risk across all observations, independent of other features.[20,29]

Lastly, we predicted the probability of developing CVD. Prediction performance was evaluated by applying the selected model to the held-out test dataset. Predictive performance was determined using calibration plots, logarithmic loss, and area under the receiver operator curve (AUROC) with AUROC 95% confidence intervals (95% CIs) computed using 1000 bootstrapped samples. AUROCs and CIs were evaluated using the "pROC" package in R.[143]

3.4 Results

3.4.1 Descriptive Statistics

12 130 individuals were included in the final analyses (table 1). There were 6850 females and 5280 males included. 1120 instances of the primary outcome were observed (9.2%). 560 (8.2%) females and 560 (10.6%) males had a cardiovascular event or death.

The median age of participants was 50.0 (IQR=34.0-65.0). Median carbohydrate consumption was 49.3% of total calories (IQR=41.6 – 56.8%), median fat consumption was 31.5% of total calories (IQR=25.1-37.8%), median protein consumption was 15.7% of total calories (IQR=12.5 – 19.6%), and median reported calories were 1805 kCal (IQR=1318-2442 kCal).

Table 1. Descriptive statistics at baseline of selected features included in model.

Feature[a]	**Overall n = 12 130**	**Developed CVD n = 1120**	**Did Not Develop CVD n = 11 010**
Age, median (IQR), y	50.0 (34.0 - 65.0)	71.0 (59.0 - 79.0)	48.0 (33.0 - 62.0)
Sex, n (%)			
Female	5280 (43.5)	560 (50.0)	4720 (42.9)
Male	6850 (56.5)	560 (50.0)	6290 (57.1)
Marital Status, n (%)			
Married	5555 (45.8)	535 (47.8)	5020 (45.6)
Common-law	865 (7.1)	35 (3.1)	835 (7.6)
Widowed	1490 (12.3)	340 (30.4)	1150 (10.4)
Separated	435 (3.6)	40 (3.6)	400 (3.6)
Divorced	945 (7.8)	90 (8.0)	855 (7.8)
Single, Never Married	2835 (23.4)	85 (7.6)	2750 (25.0)
Educational Attainment, n (%)			
Grade 8 or lower	1210 (10.0)	245 (21.9)	970 (8.8)

Grade 9 – 10	1060 (8.7)	40 (3.6)	880 (8.0)
Grade 11 – 13	635 (5.2)	185 (16.5)	570 (5.2)
Secondary school, no post-secondary	2250 (18.5)	65 (5.8)	2085 (18.9)
Some post-secondary	1135 (9.4)	165 (14.7)	1065 (9.7)
Trades certificate or diploma	1525 (12.6)	65 (5.8)	1395 (2.7)
Diploma/certificate – college	2090 (17.2)	135 (12.1)	1950 (17.7)
University certificate below bachelor's	270 (2.2)	145 (12.9)	240 (2.2)
Bachelor's degree	1350 (11.1)	25 (2.2)	1295 (11.8)
University degree above bachelor's	600 (4.9)	55 (4.9)	565 (5.1)
Household Income, n (%)[b]			
0 to $9000	455 (3.8)	40 (3.6)	415 (3.8)
$10 000 to $14 999	805 (6.6)	130 (11.6)	675 (6.1)
$15 000 to $19 999	770 (6.3)	120 (10.7)	650 (5.9)
$20 000 to $29 999	2035 (16.8)	300 (26.8)	1740 (15.8)
$30 000 to $39 999	1495 (12.3)	150 (13.4)	1345 (12.2)
$40 000 to $49 999	1190 (9.8)	100 (8.9)	1095 (9.9)
$50 000 to $59 999	1120 (9.2)	70 (6.3)	1045 (9.5)
$60 000 to $79 999	1600 (13.2)	95 (8.5)	1505 (13.7)
$80 000 or more	2660 (21.9)	115 (10.3)	2545 (23.1)
Smoking Status, n (%)			
Daily smoker, more than 20 cigarettes	605 (5.0)	80 (7.1)	530 (4.8)
Daily smoker, 16 to 20 cigarettes	465 (3.8)	50 (4.5)	415 (3.8)

Daily smoker, 11 to 15 cigarettes	675 (5.6)	45 (4.0)	625 (5.7)
Daily smoker, 6 to 10 cigarettes	630 (5.2)	35 (3.1)	595 (5.4)
Daily smoker, 5 or less cigarettes	305 (2.5)	20 (1.8)	280 (2.5)
Occasional smoker	450 (3.7)	20 (1.8)	430 (3.9)
Former daily smoker, quit less than 1 year ago	250 (2.1)	15 (1.3)	235 (2.1)
Former daily smoker, quit 1 to 3 years ago	360 (3.0)	25 (2.2)	335 (3.0)
Former daily smoker, quit 3 or more years ago	2705 (22.3)	350 (31.3)	2350 (21.3)
Former occasional smoker	290 (2.4)	35 (3.1)	255 (2.3)
Never smoked	5395 (44.5)	435 (38.8)	4960 (45.0)
Frequency of Alcohol Consumption, n (%)			
None	2560 (21.1)	365 (32.6)	2200 (20.0)
Less than once a month	2405 (19.8)	240 (21.4)	2165 (19.7)
Once a month	1020 (8.4)	70 (6.3)	945 (8.6)
2 to 3 times a month	1430 (11.8)	80 (7.1)	1345 (12.2)
Once a week	1570 (12.9)	90 (8.0)	1480 (13.4)
2 to 3 times a week	1745 (14.4)	100 (8.9)	1650 (15.0)
4 to 6 times a week	530 (4.4)	50 (4.5)	480 (4.4)
Every day	870 (7.2)	125 (11.2)	740 (6.7)
Physical activity, median (IQR), kcal/kg/h	2.2 (1.4 - 3.5)	1.9 (1.2 - 3.2)	2.2 (1.4 - 3.5)

Caffeine, median (IQR), g	168.7 (52.9 - 326.7)	177.9 (76.6 - 322.1)	167.4 (51.6 - 327.5)
Energy intake, median (IQR), kCal	1805 (1318 - 2442)	1674 (1237 - 2250)	1825 (1329 - 2459)
Carbohydrate, median (IQR), % of calories	49.3 (41.6 - 56.8)	49.6 (41.9 - 56.7)	49.2 (41.6 - 56.8)
Fat, median (IQR), % of calories	31.47 (25.1 - 37.8)	31.5 (25.5 - 37.1)	31.5 (25.0 - 37.8)
Protein, median (IQR), % of calories	15.7 (12.5 - 19.6)	16.0 (12.9 - 19.8)	15.6 (12.5 - 19.6)
Saturated fat, median (IQR), % of calories	9.6 (7.2 - 12.5)	9.6 (7.4 - 12.7)	9.6 (7.2 - 12.5)
Monounsaturated fat, median (IQR), % of calories	12. 3 (9.4 - 15.4)	12.2 (9.4 - 15.1)	12.3 (9.4 - 15.5)
Polyunsaturated fat, median (IQR), % of calories	5.2 (3.8 - 7.0)	5.3 (3.8 - 7.0)	5.2 (3.8 - 7.0)
Total sugars, median (IQR), g	84.7 (52.8 - 129.6)	78.0 (48.8 - 119.8)	85.5 (53.1 - 130.9)
Total fibre, median (IQR) g	14.6 (9.7 - 21.2)	14.6 (10.1 - 21.1)	14.6 (9.6 - 21.2)
Daily fruit, median (IQR), number of times eaten per day	1.0 (0.4 - 2.0)	0.6 (0.1 - 1.0)	1.0 (0.4 - 2.0)
Daily other vegetables, median (IQR), number of times eaten per day	1.0 (0.4 - 1.0)	1.0 (0.4 - 1.0)	1.0 (0.4 - 1.0)
Folic acid, median (IQR), µg	76.2 (37.8 - 133.9)	67.8 (36.5 - 113.3)	77.3 (38.0 - 135.7)
Folate, median (IQR), µg	188.4 (127.8 - 274.5)	182.2 (134.2 - 252.4)	189.5 (127.1 - 276.8)

Vitamin B12, median (IQR), µg	3.0 (1.7 - 5.0)	3.0 (1.8 - 5.0)	3.0 (1.7 - 5.0)
Vitamin D, median (IQR), µg	4.0 (2.1 - 6.8)	4.3 (2.3 - 7.0)	4.0 (2.1 - 6.8)
Sodium, median (IQR), µg	2634 (1793 - 3750)	2566.88 (1796.13 - 3558.71)	2644 (1793 - 3773)
Potassium, median (IQR), µg	2819 (2011 - 3777)	2800 (2021 - 3647)	2821 (2010 - 3792)
Takes Supplements, n (%)			
Yes	5850 (48.2)	600 (53.6)	5255 (47.7)
No	6275 (51.7)	520 (46.4)	5755 (52.3)

CVD: cardiovascular disease; MET: metabolic equivalent of task.
[a]Descriptive statistics are unweighted. Counts are rounded to the nearest 5. Medians are rounded to the nearest 0.1. Cells containing less than 15 participants were combined.
[b]Measured in Canadian dollars

3.4.2 Hyperparameter Tuning

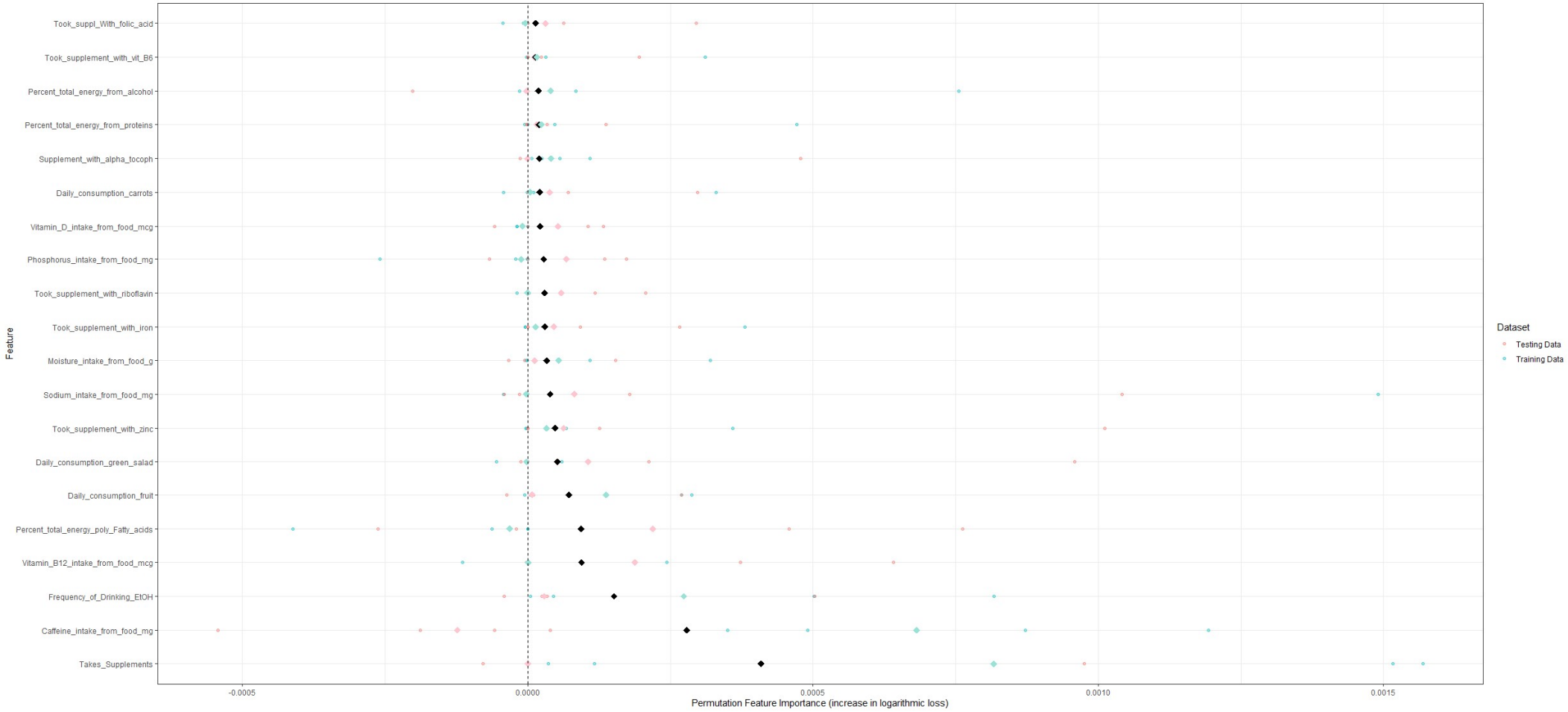
During cross-validation, the model with the lowest logarithmic loss for predictions on out-of-fold data had a median prediction error of 0.261. This model had an mtry value of 69 and a mincriterion value of 0.999. The three models with the next best performance had median prediction errors on out-of-fold data of 0.261, 0.261, and 0.261; mtry values of 39, 74, and 74 respectively; and mincriterion values of 0.95, 0.99, and 0.999 respectively. The range of median prediction error among all models tested was 0.261 – 0.285. See supplementary table 3 and supplementary figure 2 for more details regarding hyperparameter tuning. Overall, prediction error during cross-validation was highest with models using low mtry values. In models with a high mincriterion, prediction error was

lowest with high mtry values, while in models with a lower mincriterion, prediction error was lowest with a mid-range mtry values.

3.4.3 Permutation feature Importance

Overall, 23 nutrition features (figure 1 includes the 20 highest) (median PFI (M)=2.95 x 10^{-5}, IQR=2.00 x 10^{-5} – 6.21 x 10^{-5}) and 10 non-nutrition-related features (supplementary figure 3) had a positive median PFI. Age (M=0.213, IQR=0.154 - 0.296), sex (M=0.014, IQR=6.98 x 10^{-3} – 0.024), and smoking status (M=7.98 x 10^{-4}, IQR=3.24 x 10^{-4} – 1.87 x 10^{-3}) had the highest median PFIs among all included features (see supplementary table 5 for all PFIs). Vitamin or mineral supplement-use (M=4.09 x 10^{-4}, IQR=8.25 x 10^{-7} – 1.11 x 10^{-3}), caffeine intake (M=2.79 x 10^{-4}, IQR= -9.11 x 10^{-5} – 5.86 x 10^{-4}), and frequency of drinking alcohol (M=1.52 x 10^{-4}, IQR= 1.99 x 10^{-5} – 5.02 x 10^{-4}) were the nutrition-related features with the highest median PFIs.

Figure 1. The 20 nutrition-related features with the highest median permutation feature importance. The blue, red, and black diamonds are the median training, testing, and overall importance levels, respectively. Each circle is the permutation feature importance of an individual model developed using either the training or testing data. The vertical dashed line is zero.



3.4.4 Accumulated Local Effects

We evaluated the ALE plots of the twenty nutrition-related features and four non-nutrition features with the highest median PFIs. ALEs for all features can be found in supplementary table 5.

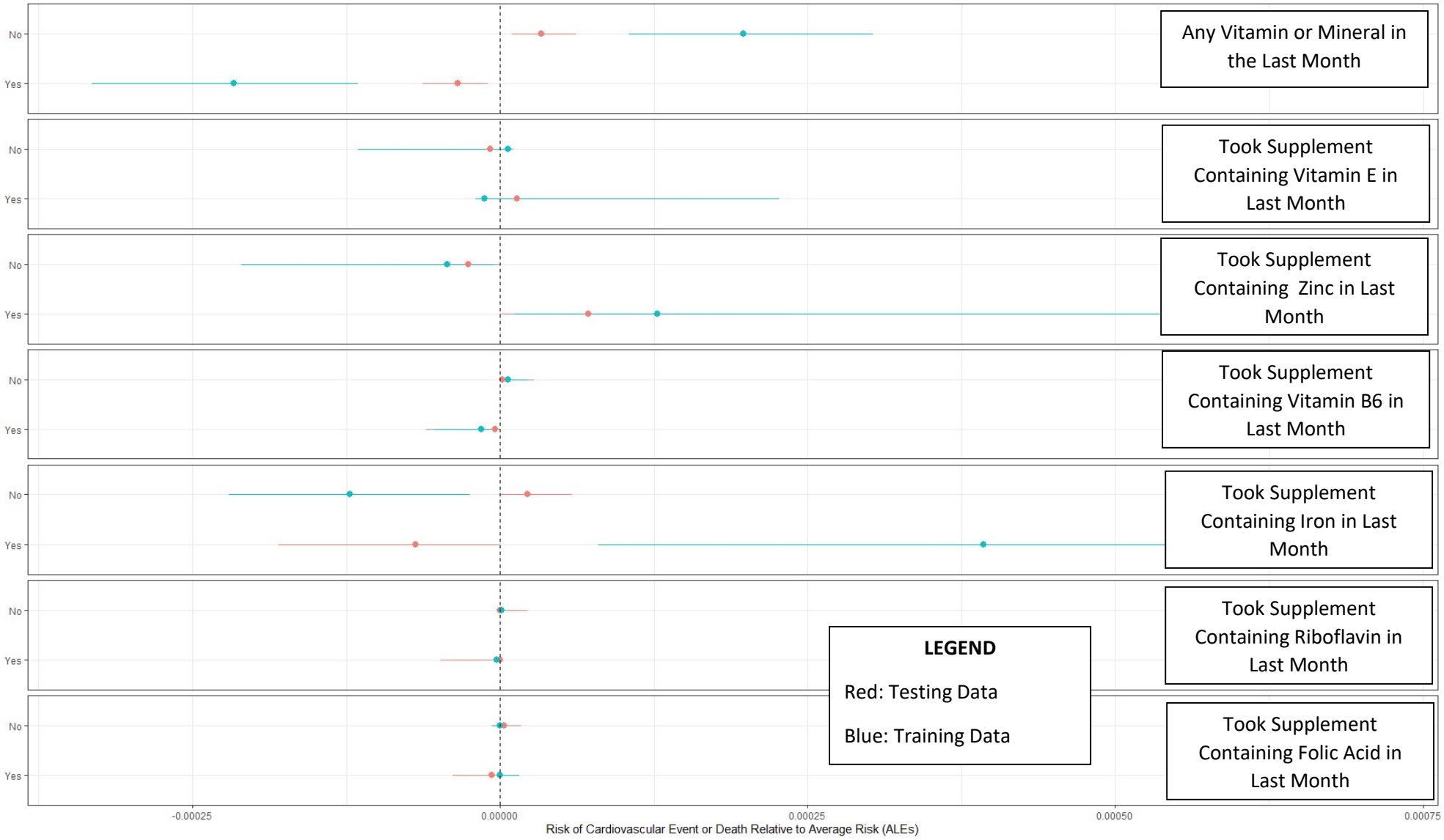
3.4.4.1 Supplements and Substances

The ALEs plots of any vitamin or mineral supplement-use in the last month (which had the highest PFI among nutrition-related features) and all other supplements with a positive median PFI are included in figure 2. The use of any supplement in the last month was related to decreased predicted risk in both training and testing dataset models (accumulated local effects range (ALER)= -4.21×10^{-4} – 3.84×10^{-4}). Zinc supplement-use (ALER= -5.97×10^{-4} – 0.002) was related to increased predicted risk in both training and testing models. As this result was consistent between training and testing datasets, the ALEs of zinc from food (ALER= -0.004 – 0.007) were also examined (see supplementary figure 4). Zinc from food appeared to have a threshold relationship with predicted risk, which started to increase near 10 – 15 mg of intake. Vitamin B6 supplement-use (ALER= -2.18×10^{-4} – 1.00×10^{-4}) was associated with decreased predicted risk in both training and testing models. Vitamin B6 intake from food (ALER= -3.39×10^{-4} – 6.71×10^{-4}) demonstrated a j-shaped relationship with predicted risk of CVD in both testing training models (when it had a non-zero impact on predictions) (see supplementary figure 4). The LOESS curves are not visible for all eight models, as in some models this feature had no impact on predictions; in this case the curves run along the x-axis. Predicted risk was lowest near 1.2 – 2.2 mg of vitamin B6 intake from food. Riboflavin (ALER= -1.92×10^{-4} – 2.35×10^{-6}) and folic acid (ALER= -1.14×10^{-4} - 6.31×10^{-5}) supplement-use had no impact on predicted risk among training models and were both associated with decreased predicted risk in testing models.

Vitamin E supplement-use (ALER= $-4.73 \times 10^{-4} – 9.27 \times 10^{-4}$) was related to decreased predicted risk in training models and increased predicted risk in testing models. Iron supplement-use (ALER= $-3.07 \times 10^{-4} – 7.65 \times 10^{-4}$) also had inconsistent ALEs between datasets, with increased predicted risk in training models and decreased predicted risk in testing models.

The ALE plots of caffeine intake from food/drink and percent of energy intake from alcohol are included in supplementary figure 5. Caffeine intake (ALER= $-0.002 \times 10^{-4} – 0.035$) had a threshold dose-response curve in most training and testing models, with predicted risk beginning to increase near 400 – 700 mg of daily intake. Percent of energy intake from alcohol (ALER= $-4.54 \times 10^{-4} – 0.008$) was related to increasing predicted risk across training and testing models. Some models demonstrated a threshold dose-response, with predicted risk increasing after 15% of energy intake, while others had a j-shaped dose response curve, with predicted risk near its lowest at 5% of energy intake. The ALE plot of frequency of alcohol consumption (ALER= $-0.002 \times 10^{-4} – 0.004$) (supplementary figure 6) was also examined. This feature reflects how often alcohol was consumed, but does not reflect the amount that was consumed on these occasions. There was a mostly u-shaped relationship with predicted risk in both training and testing models, with there being relatively low predicted risk associated with drinking anywhere from once a month to 2-3 times per week. Never drinking and drinking 4-6 times per week or daily were all associated with relatively higher predicted risk.

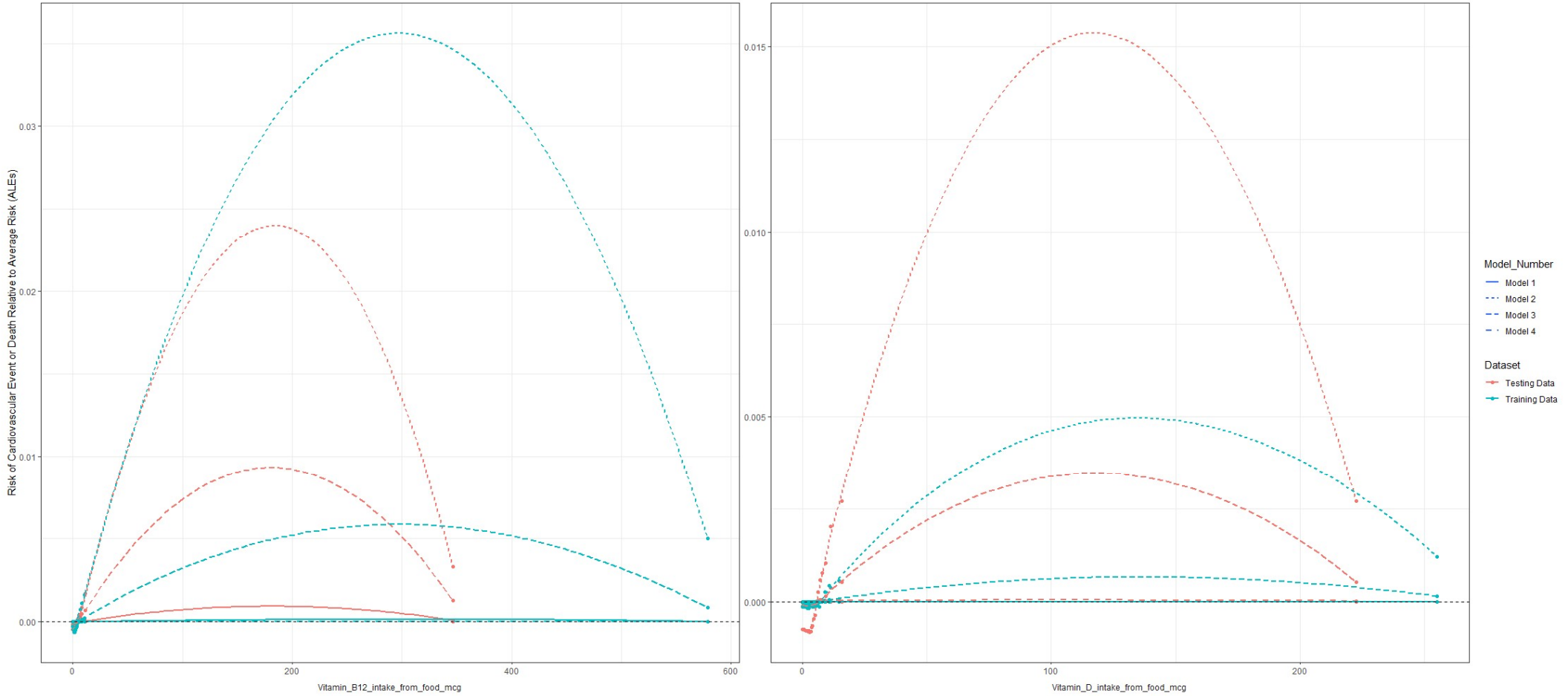
Figure 2. Accumulated local effects of the supplement features with the highest median PFI. Accumulated local effects, representing the average influences of specific feature levels on predicted risks of CVD relative to average predicted risk, are shown on the x-axis. Zero (or average predicted risk) is represented by a dashed line. Error bars demonstrate the interquartile range of estimates within testing and training datasets.



3.4.4.2 Vitamins from Food Sources

Vitamin B12 from food sources (ALER= -6.29×10^{-4} – 0.005), which had the fourth-highest PFI among nutrition-related features, was related to increasing predicted risk in both training and testing models (figure 3). Vitamin B12 supplementation (ALER= -4.04×10^{-5} – 9.27×10^{-5}) was also related to increased predicted risk in both training and testing models (supplementary figure 7). Vitamin D from food sources (ALER= -8.11×10^{-4} – 0.003) had a threshold dose-response relationship with predicted risk in the training and testing models that showed effects, with predicted rates of CVD increasing after approximately 2.9 – 6.2 µg (or 116 – 248 international units (IU)) of intake. Vitamin D supplementation (ALER= -3.73×10^{-4} – 2.11×10^{-4}), in turn, was associated with decreased predicted risk in both training and testing models (see supplementary figure 7).

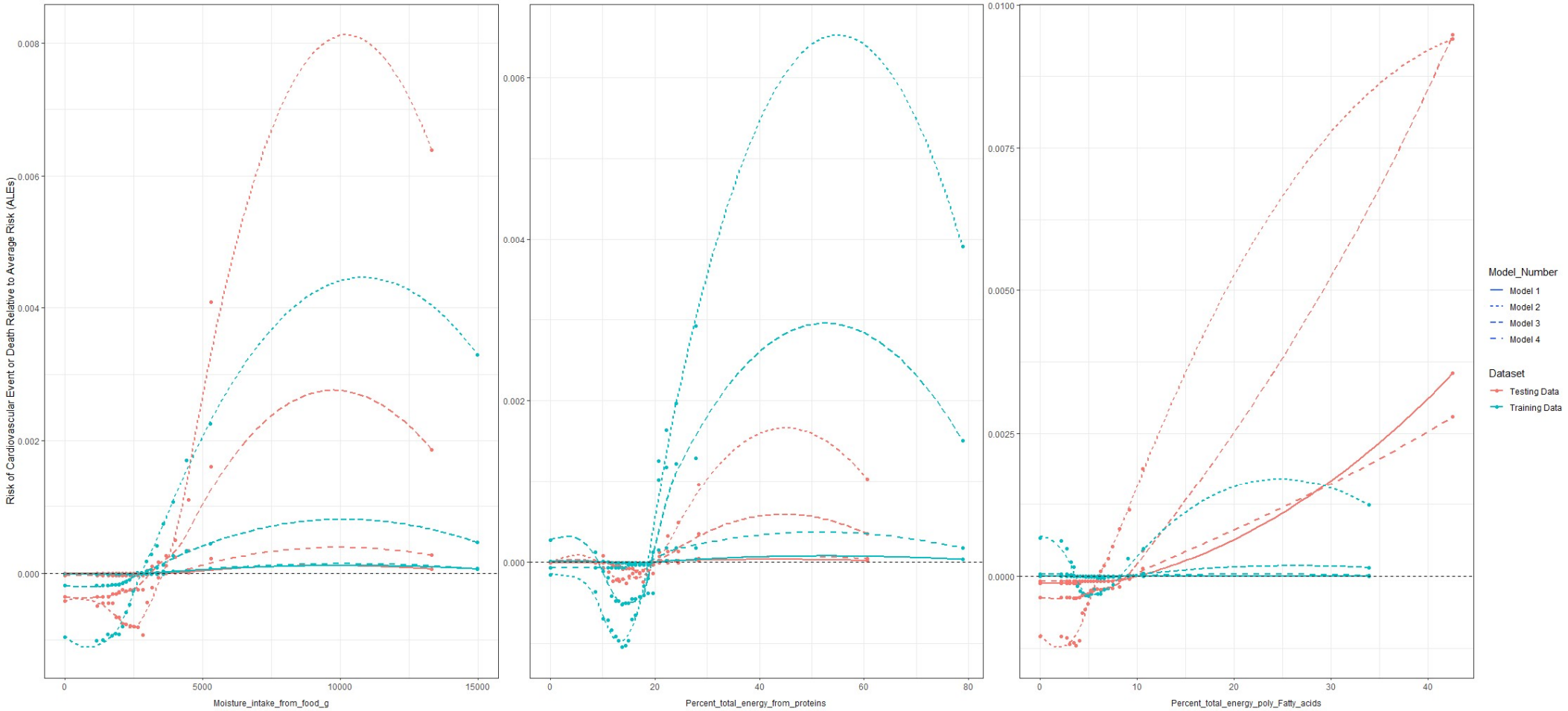
Figure 3. Accumulated local effects of vitamins from food sources with the greatest median PFI. The y-axis is accumulated local effects, representing the average influence of a given feature level on predicted risk of CVD relative to average predicted risk.



3.4.4.3 Macronutrients and Moisture

Percent of energy from polyunsaturated fatty acids (PUFAs) (ALER= -0.001 – 0.009) showed both threshold and j-shaped relationships with predicted risk in testing models, and little connection to predicted risk among training models (see figure 4). In the models demonstrating a j-shaped dose-response curve, risk was lowest at 3.4 – 5.9 % of energy intake from PUFAs. In the models with a threshold-type relationship, predicted risk increased after approximately 8.2% of energy from PUFAs. Moisture intake from food and beverages (ALER= -0.001 – 0.006) had threshold and j-shaped dose-response curves, with predicted risk increasing after approximately 2400 – 2800 g in threshold models and predicted risk being lowest at approximately 2800 g in the j-shaped model. Finally, percent of energy intake from protein (ALER= -0.001 – 0.004) mostly showed a j-shaped relationship with predicted risk in both testing and training models. Predicted risk was lowest when protein comprised 13.7 – 17.1% of energy intake.

Figure 4. Accumulated local effects of the macronutrients with the greatest median PFI. The y-axis is accumulated local effects, representing the average influence of a given feature level on predicted risk of CVD relative to average predicted risk.



3.4.4.4 Food Categories

The ALEs for daily consumption of fruit (ALER= $-8.70 \times 10^{-4} – 7.12 \times 10^{-4}$), which had the sixth highest PFI among nutrition-related features, are presented in supplementary figure 8. The daily fruit consumption feature, and all the other food category features, refer to the average number of times that the food was consumed per day over the previous 30 days (it does not reflect the specific quantity). Some testing models showed a u-shaped relationship between predicted risk and fruit consumption, while other training and testing models showed decreasing predicted risk with increasing fruit consumption. The u-shaped models had the lowest predicted risk when consuming fruit an average of 0.9 times per day. The models demonstrating decreasing predicted risk achieved most of this predicted risk reduction when consuming fruit approximately 1.25 – 2.5 times per day. Daily consumption of green salad (ALER= -0.002 – 0.002) was associated with decreasing predicted risk in testing models and had a u-shaped relationship with predicted risk in one of the training models. In the testing models low predicted risk was achieved when consuming green salad on average approximately 1 time daily and in the u-shaped model lowest predicted risk was seen with eating green salad on average 0.7 times per day. Lastly, daily consumption of carrots (ALER= $-6.77 \times 10^{-4} – 0.003$) did not show consistent relationships with predicted risk among models. The two models with the largest variation in predicted risk (training models 2/3 and testing model 2) had monotonically increasing predicted risk and an inverted j-shaped relationship with predicted risk. The model with increasing predicted risk achieved the highest predicted risk-level when consuming carrots one time daily on average and the u-

shaped model had the lowest predicted risk when eating carrots approximately 0.6 times daily.

3.4.4.5 Minerals

The ALEs for intake of sodium from food (ALER: -0.003 – 0.007), which had the eighth highest median PFI among nutrition-related features, are displayed in supplementary figure 9. Most models had a threshold relationship with predicted risk and one testing model had a j-shaped relationship with predicted risk. The threshold models had increasing predicted risk after approximately 2500 – 4000 mg of sodium intake and the j-shaped model had the lowest predicted risk at 2000 mg of intake. Sodium supplementation (ALER= -2.06×10^{-4} – 0.005) was also related to increased predicted risk in training and testing models (see supplementary figure 10). Phosphorous intake from food (ALER= -2.37×10^{-4} – 3.71×10^{-4}) had a u-shaped relationship with predicted risk in testing models, with the lowest risk seen at 667 – 1550 mg of intake. Threshold relationships were observed in the training models, which also had much less variation in predicted risk associated with phosphorous intake. These models showed increasing predicted risk after approximately 1100 mg of consumption. Phosphorous supplementation (ALER= -3.54×10^{-4} – 4.88×10^{-4}) had inconsistent ALEs between training and testing models.

3.4.4.6 Non-nutrition-related Features

Age (ALER= -0.083 – 0.240) had the highest median PFI overall (see supplementary figure 11). In all models, age displayed a threshold relationship with

predicted risk, with predicted risk starting to increase at approximately 35 years of age. Women had a lower predicted risk of CVD than men (ALER of sex= -0.014 – 0.018) in all models (see supplementary figure 12 for sex and all other remaining ALEs). Overall, smoking status (ALER= -0.002 – 0.023) was related to increasing predicted risk with higher frequency and recency across models. This relationship was most apparent in training models, whereas in testing models there was more variation in the influence on predicted risk. Among training models, predicted risk was similarly low among people who never smoked, former occasional smokers, and former daily smokers who had quit more than three years prior. Predicted risk was highest across models in those who smoked 16 or more cigarettes per day. Lastly, greater household food security (ALER= -0.001 – 0.035) was associated with less predicted risk in both training and testing models.

3.4.5 Prediction Performance

The model with the best prediction performance during training had a test dataset logarithmic loss of 0.248 and an AUROC of 0.821 (95% CI: 0.801 – 0.842) (see supplementary figure 13 for ROC plot). Absolute risk of CVD was modestly underpredicted from 0.000 – 0.300 predicted risk and correctly predicted in the 0.300 – 0.350 predicted risk decile (see supplementary figure 14 for calibration plot).

3.5 Discussion

We used conditional inference forests, a machine learning method, to build predictive models for CVD using 61 nutrition-related features and 14 socioeconomic, behavioural, psychological, and demographic covariates from a Canadian population-

based health survey linked to administrative health databases. Permutation feature importance and accumulated local effects were used to determine the contribution of features to predicted risk of CVD in the predictive models. We found many nutrition-related features with positive median PFIs. Accumulated local effects plots demonstrated a diverse mixture of linear, threshold, u-shaped, j-shaped, and other non-linear relationships.

Many of the nutrition-related features that we identified with the highest PFIs have been linked to CVD in the past, including alcohol, sodium, fruits, vegetables, supplement-use, caffeine, B vitamins, protein, PUFAs, and zinc.[38] Interestingly, many of the nutrition-related features with the highest PFIs were related to supplementation, substances, or food categories. These features were developed from survey questions about average consumption over the previous month, which may have reduced measurement error relative to other nutrients derived from the 24-hour dietary recall and therefore increased the strength of observed relationships. Additionally, food categories combine diverse mixtures of many nutrients and therefore, may have stronger impacts on risk of CVD than individual, isolated nutrients.

The nature of identified relationships between important nutrition-related features and predicted CVD risk were often broadly consistent with previous epidemiologic literature, as in the case of alcohol, sodium, fruit, and salad. We found a u-shaped relationship between frequency of drinking alcohol and CVD, which is similar to other available evidence.[144] Our results were less consistent regarding percent of energy of from alcohol; however, in some models we did observe a j-shaped dose-response curve. It

may be that the relationship stemming from percent of energy from alcohol was less precise, as it was derived from the 24-hour dietary recall rather than a question regarding average consumption. We also found that high levels of sodium were associated with higher predicted risk, which is consistent with the literature.[145–149] However, we found a threshold or j-shaped relationship, and there is not consensus on this in the literature.[35] Predicted risk in our models increased after 2000 – 4000 mg of intake, which at the lower end would be consistent with recommendations for sodium targets. Some systematic reviews of observational studies have not found evidence for non-linear relationships[145–147] while other studies and reviews have supported u- or threshold effects.[148,149] Fruits had u-shaped and decreasing relationships with predicted risk in our study. Overall, the literature suggests that fruits are protective; however, these systematic reviews have also found evidence of non-linear relationships.[150–153] Green salad also had mostly decreasing relationships with predicted risk, which is consistent with the general literature on vegetable intake and CVD.[150–153]

Other identified important features have less consistent relationships with CVD in previous literature or have demonstrated associations that differ from our findings, including those for supplement-use, vitamin B6, vitamin B12, caffeine, PUFAs, and protein. We found a link between use of any supplement and reduced predicted risk of CVD, while systematic reviews on this topic have concluded that there is not a relationship.[154,155] One review did find an initial protective effect against coronary heart disease incidence,[155] but this dissipated during a subgroup analysis restricted to randomized controlled trials. Overall, our results for supplementation and predicted risk

should not be interpreted independently as it is the context of other features in the model. We make no assumption regarding unmeasured features. We also found a relationship between vitamin B6 supplementation and lower predicted risk, as well as a j-shaped relationship between dietary vitamin B6 and predicted risk. Systematic reviews and recent studies have also found a protective effect of dietary/supplementary vitamin B6, but did not find evidence of non-linearity.[154,156–158] Our results showed an association between vitamin B12 intake from food/supplementation and increasing predicted risk. In contrast, reviews have found no relationship between vitamin B12 and CVD. Again, our results with vitamin B12 and predicted risk should not be interpreted independently as it is the context of other features in the model. We make no assumption regarding unmeasured features, such as meat consumption. An additional finding was that caffeine was related to increasing predicted risk of CVD, while most recent meta-analyses have found that moderate coffee or caffeine consumption is neutral or protective.[159–161] Again, our results for caffeine and its relationship to predicted risk may emerge due to unmeasured features associated with both caffeine and CVD, which we make no assumptions about.

Regarding PUFAs, there were j-shaped and threshold relationships found with predicted risk. The health effects of PUFAs remain controversial, with some systematic reviews finding no relationship[153,162] and others a protective effect.[45,163,164] Two reviews have reported evidence for non-linear relationships.[153,163] It has also been argued that omega-6 fatty acids in particular, may be harmful.[37,165] Lastly, the effect of protein intake on CVD remains uncertain. We found a u-shaped relationship with predicted risk. Systematic reviews have found null or harmful effects of total protein intake on CVD, beneficial

effects of plant protein, and harmful effects of animal protein.[166,167] Relationships between total and animal protein and risk were u- or j-shaped and similar to those found in our study.[167]

Other identified nutrients which were identified as predictors in our model have been less frequently discussed in the CVD literature, including zinc, and moisture intake from food and beverages. Both zinc supplementation and intake from food were associated with increased predicted risk of CVD. Studies have found null, harmful, and beneficial relationships between zinc intake and CVD.[168,169] One has reported that zinc from meat was associated with increased CVD, while zinc from other sources had no association.[170] We found that moisture had threshold or j-shaped dose-response curves. To our knowledge, no study has evaluated the relationship between moisture intake and CVD; however, moisture intake from beverage sources has been linked to increased body mass index.[171]

Overall, while our results are suggestive of future areas for investigation, they do not stem from a causal model and cannot be interpreted as causal effects. It must be interpreted in the context of other features in the model and we have not made any assumptions regarding unmeasured features. Given the absence of most whole foods in our analysis, it is impossible to determine whether associations arise from nutrients themselves or other aspects of their major food sources.

Age, sex, smoking status, and food security had the highest median PFIs among non-nutrition-related features. This is unsurprising considering that all have previously been linked to CVD and also would be expected to have less measurement error than the

nutrition-related features in our data.[172] Furthermore, the increasing predicted risk with age, increased predicted risk among males, increased predicted risk with higher levels of smoking, and increased predicted risk with greater food insecurity that we found were all consistent with what is known.[173,174]

During hyperparameter tuning, we noted that models with a higher mincriterion parameter (and therefore a higher required significance value to use features in making predictions) performed best with higher mtry values. Meanwhile, models with lower mincriterion values performed best with mid-range mtry values. Both types of models achieved nearly equivalent prediction performance. This may be significant in health research settings when interpretation of the contributions of features to predictions is of interest. This is because it has been reported that higher mtry values reduce the PFI of irrelevant features that are correlated with truly important features and also make feature contributions to model predictions more analogous to regression models.[175,176] Therefore, models with higher mtry values may be more amenable to interpretation of feature effects. Use of the mincriterion parameter, which does not exist in random forests, was important to achieve high prediction performance alongside high mtry values. As a result, conditional inference forests and attention to mincriterion/mtry during tuning may be important when model interpretation is needed.

The predictive discrimination and calibration of our models were comparable to or better than many existing CVD risk prediction tools.[177,178] This is despite the absence of any clinical, laboratory, or anthropometric features in our models, which suggests that rich nutritional features may add valuable information. Also, the features used in this

predictive model could potentially be collected relatively quickly and from home, which could be valuable. Lastly, machine learning prediction models have been criticized for poor calibration, but our results suggest that this is not always the case.

3.5.1 Strengths

Our study is one of the first to apply machine learning techniques to the prediction of CVD using detailed population-based dietary data. Furthermore, few existing models have used linked, population-based data with prospective follow-up of 14 years. An additional advantage of our study is the use of conditional inference forests rather than random forests. Given the use of non-parametric significance tests, these algorithms reduce feature importance bias in favour of continuous features and may permit more generalizable interpretations.[179] We applied comprehensive hyperparameter tuning to optimize the performance of our models, including mincriterion, which is infrequently considered. We are also one of the first groups to apply accumulated local effects to CVD epidemiology, which permits better isolation of feature effects in the presence of multicollinearity than partial dependence plots. Furthermore, effects were evaluated in both training and testing sets, and with multiple different models. Additionally, throughout development of our models we used logarithmic loss to measure prediction error rather than accuracy. Logarithmic loss is a proper scoring rule, which performs better in settings with more stochasticity (such as health research).[122] Finally, prediction performance was evaluated in terms of both discrimination and calibration, which is important in health research settings.[180]

3.5.2 Limitations

The major limitation of our study is that most nutrition features were gathered using a single 24-hour dietary recall. This is likely to introduce significant random measurement error, and likely reduced our ability to identify important features, while possibly resulting in some spurious findings.[172] Also, random measurement error can reduce both predictive discrimination and calibration of prediction models.[181] Additionally, many included nutrient features are highly correlated. Therefore, some of the features that we identified as important may only have been correlated with other truly important features. However, the use of a higher mtry value in most models should have mitigated this effect. A further issue is that multicollinearity and the use of higher mtry values results in greater PFI variance, which may have also obfuscated truly important and unimportant features. In the future, use of a conditional variant of permutation feature importance could mitigate some of these issues.[179]

We considered one machine learning model and future research may want to evaluate and compare more models to determine optimal modelling complexity. Friedman h-statistics will also help to determine the importance and nature of any interactions.[29] Both approaches will help to better understand whether and when machine learning offers advantages over traditional modelling approaches in the nutritional epidemiology of CVD.

3.6 Conclusions

Using machine learning methods, we were able to identify many nutrients important for prediction of CVD, with a mix of non-linear and linear relationships. Using

interpretable machine learning methods, we replicated many established relationships between nutrition and CVD, while highlighting potential novel areas of inquiry. Methods such as permutation feature importance and accumulated local effects challenge the common conception that machine learning algorithms are black boxes, offering an opportunity to improve the yield and usability of these methods in nutritional epidemiology and health research more broadly. Additionally, our models' predictive performances were comparable to existing tools despite lacking any laboratory or anthropometric features. Our results suggest that machine learning techniques warrant further investigation as an analytic tool in the nutritional epidemiology of CVD. Future work is needed to determine if the observed associations are causal. With growing recognition of the complexity of nutrition's relationship with disease, applying machine learning may be even more fruitful here than in other fields of health research. Also, incorporating both more machine learning algorithms and detailed dietary data into CVD predictive models may improve their performance. Finally, application of these methods to large cohort studies or other large data sources with repeated dietary recalls or food frequency questionnaires is likely to enhance the value of machine learning methods.

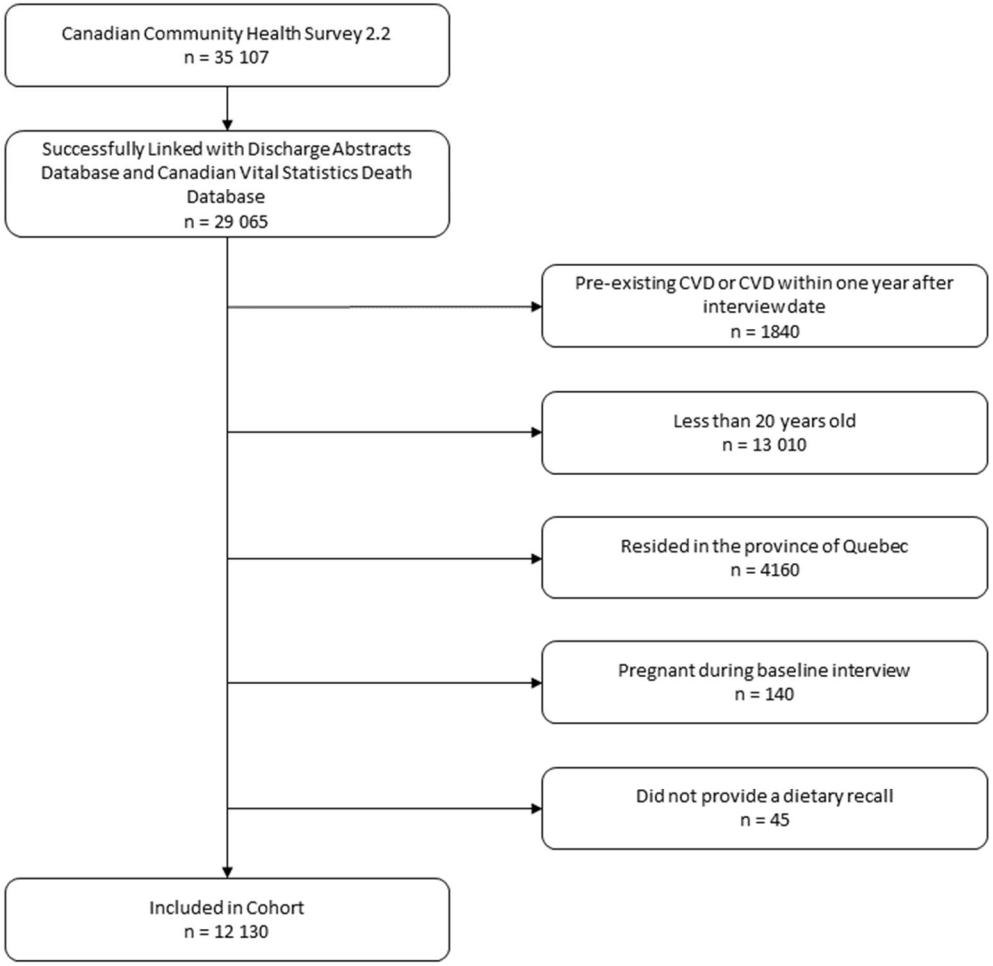
3.7 Acknowledgements

The analysis presented in this manuscript was conducted at the Statistics Canada Research Data Centre at McMaster (RDC) which is part of the Canadian Research Data Centre Network (CRDCN). The services and activities provided by the RDC are made possible by the financial or in-kind support of the SSHRC, the CIHR, the CFI, Statistics

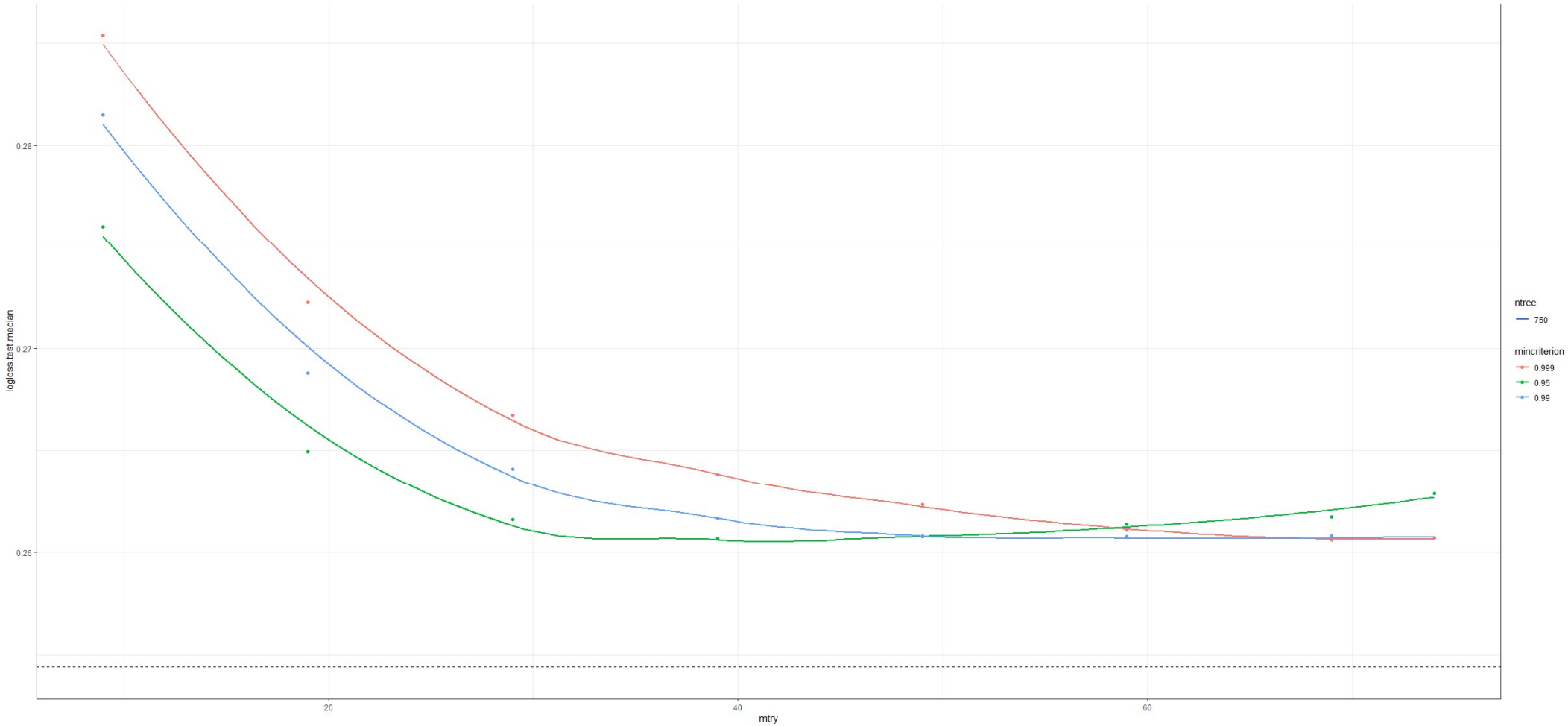
Canada, and McMaster University. The views expressed in this paper do not necessarily represent the CRDCN's or that of its partners.

3.9 Supplementary Figures

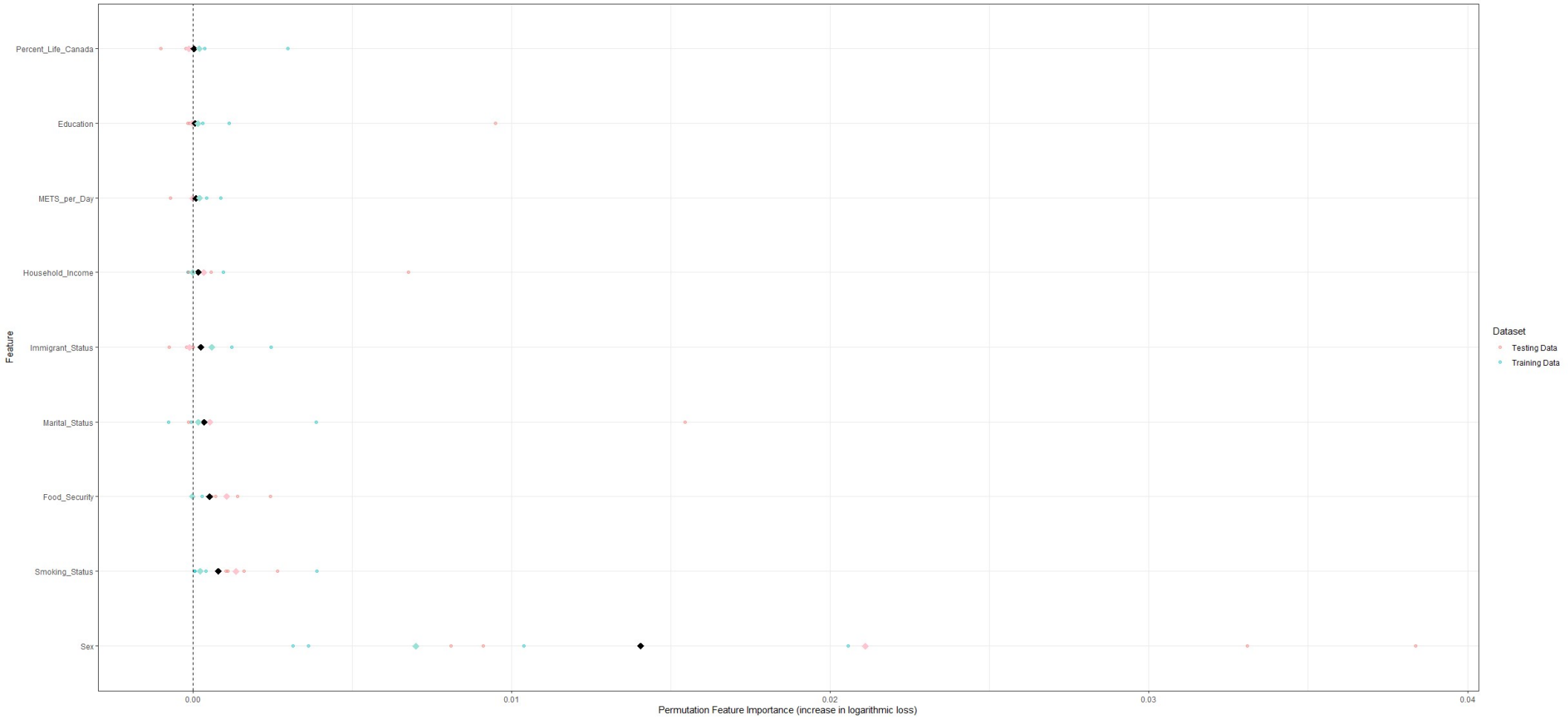
Supplementary figure 1: Study design



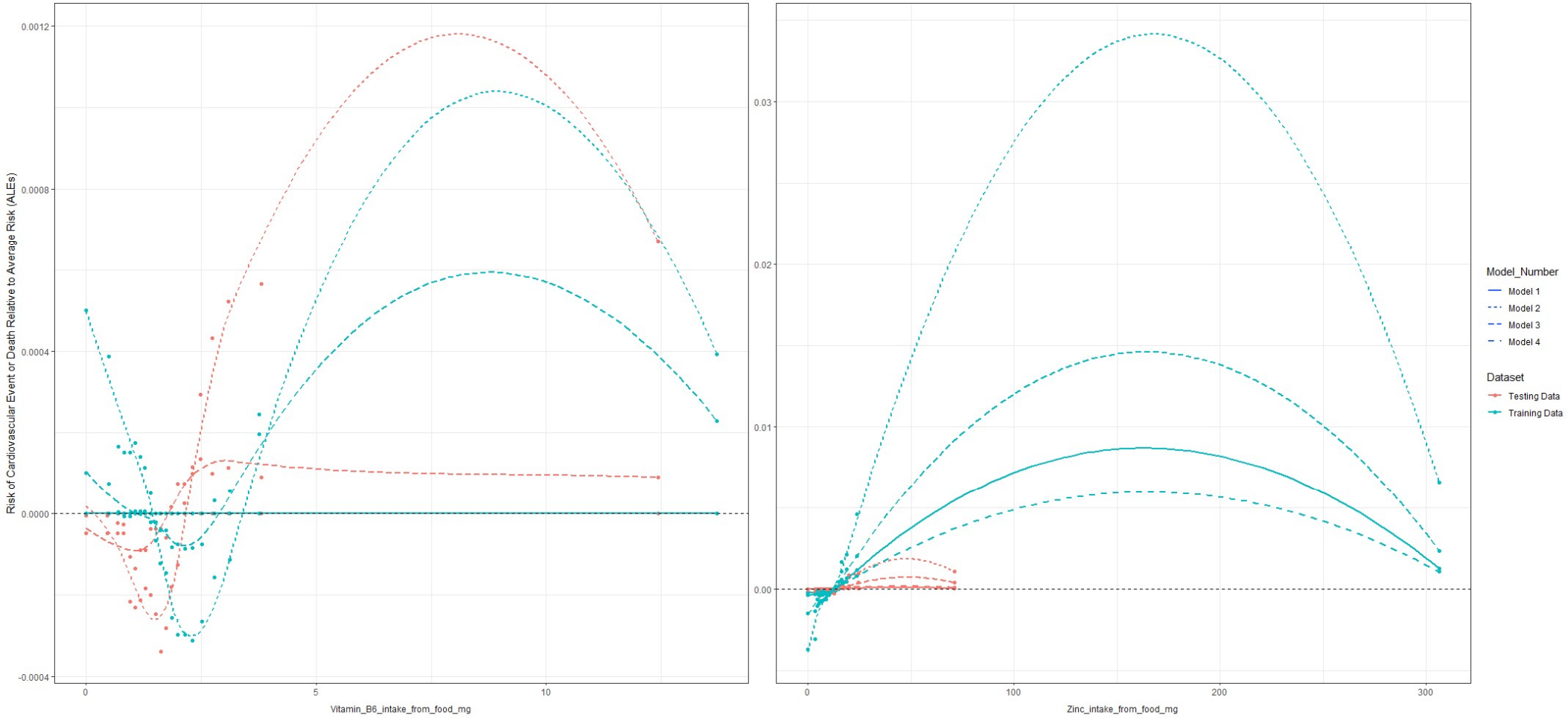
Supplementary figure 2: Prediction performance of the hyperparameter sets tested during cross-validation



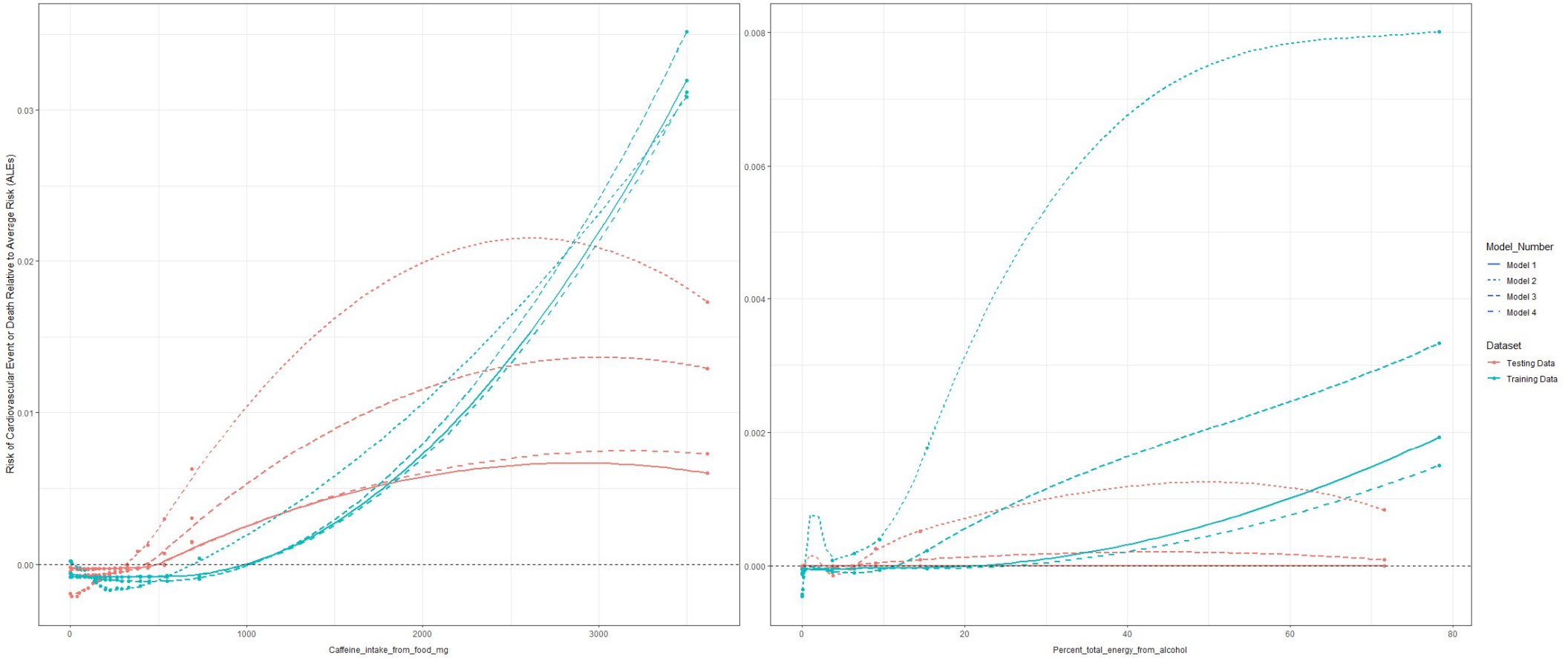
Supplementary figure 3: Permutation feature importance of the features not related to nutrition



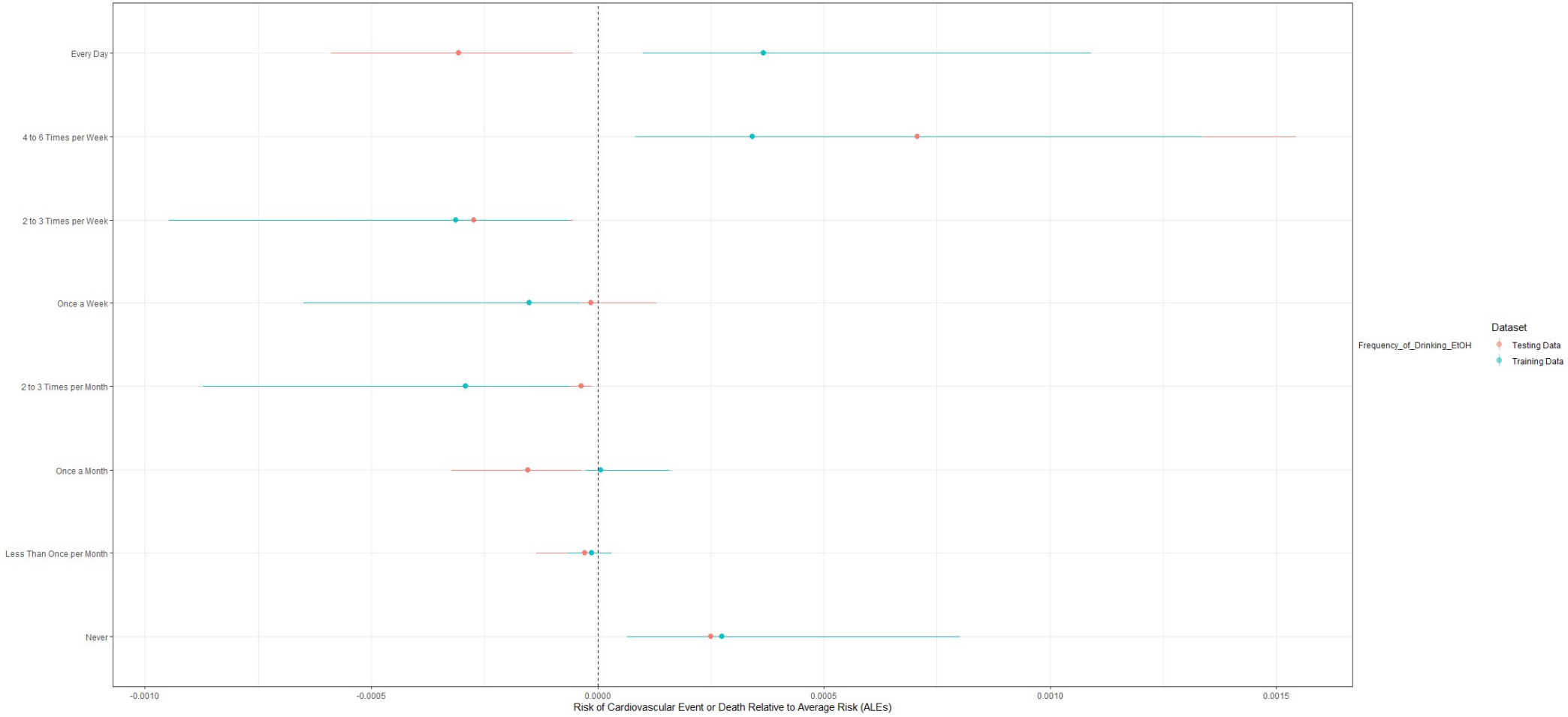
Supplementary figure 4: Accumulated local effects of zinc and vitamin B6 from food sources



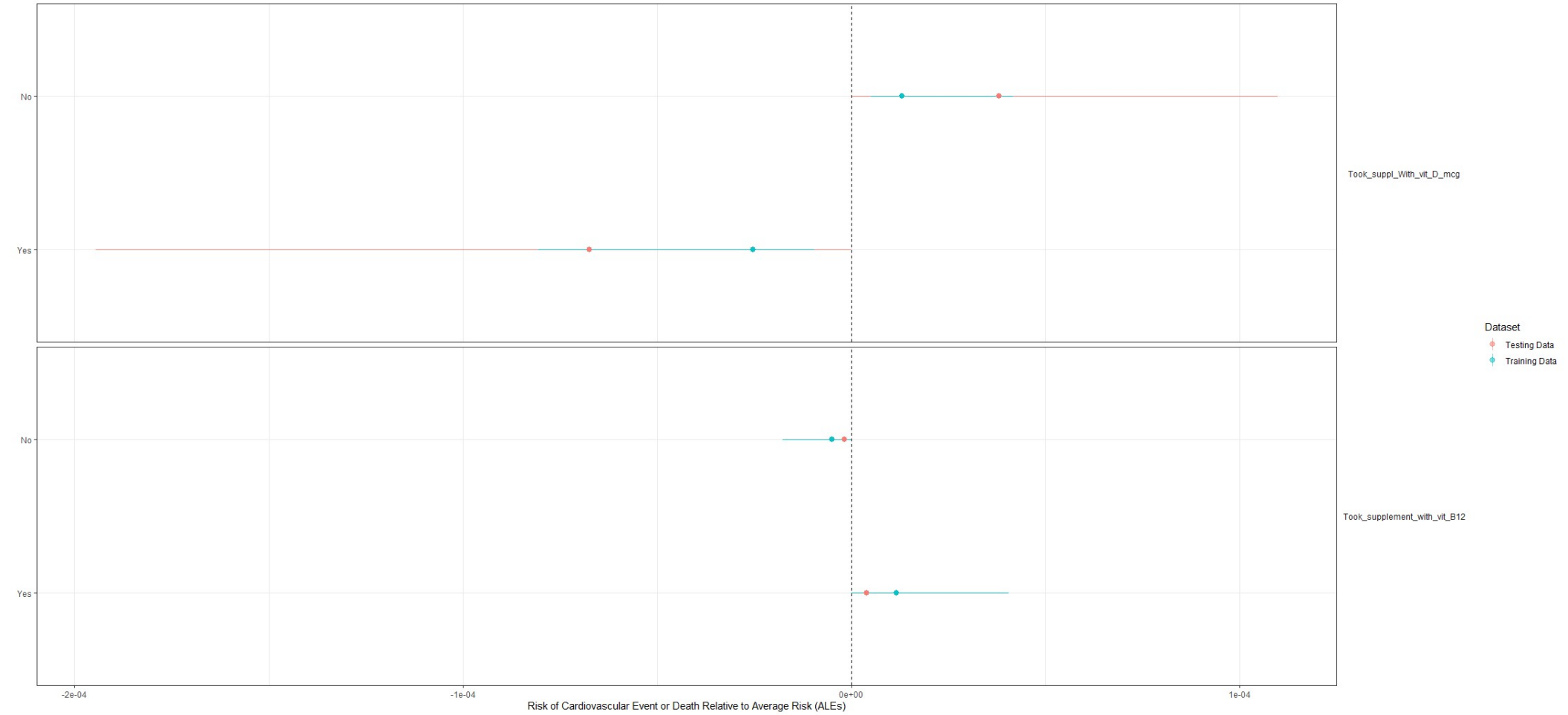
Supplementary figure 5: Accumulated local effects of caffeine and percent of daily energy from alcohol



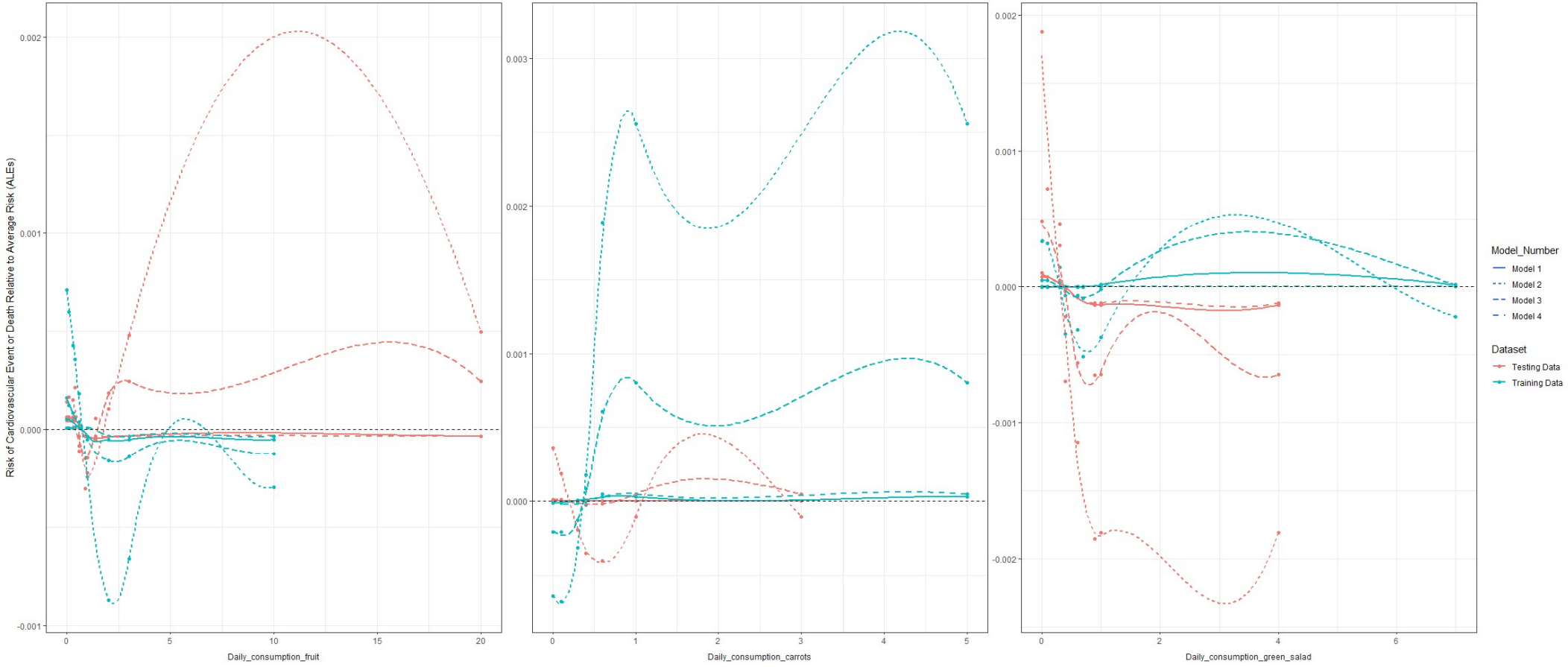
Supplementary figure 6: Accumulated local effects of frequency of drinking alcohol



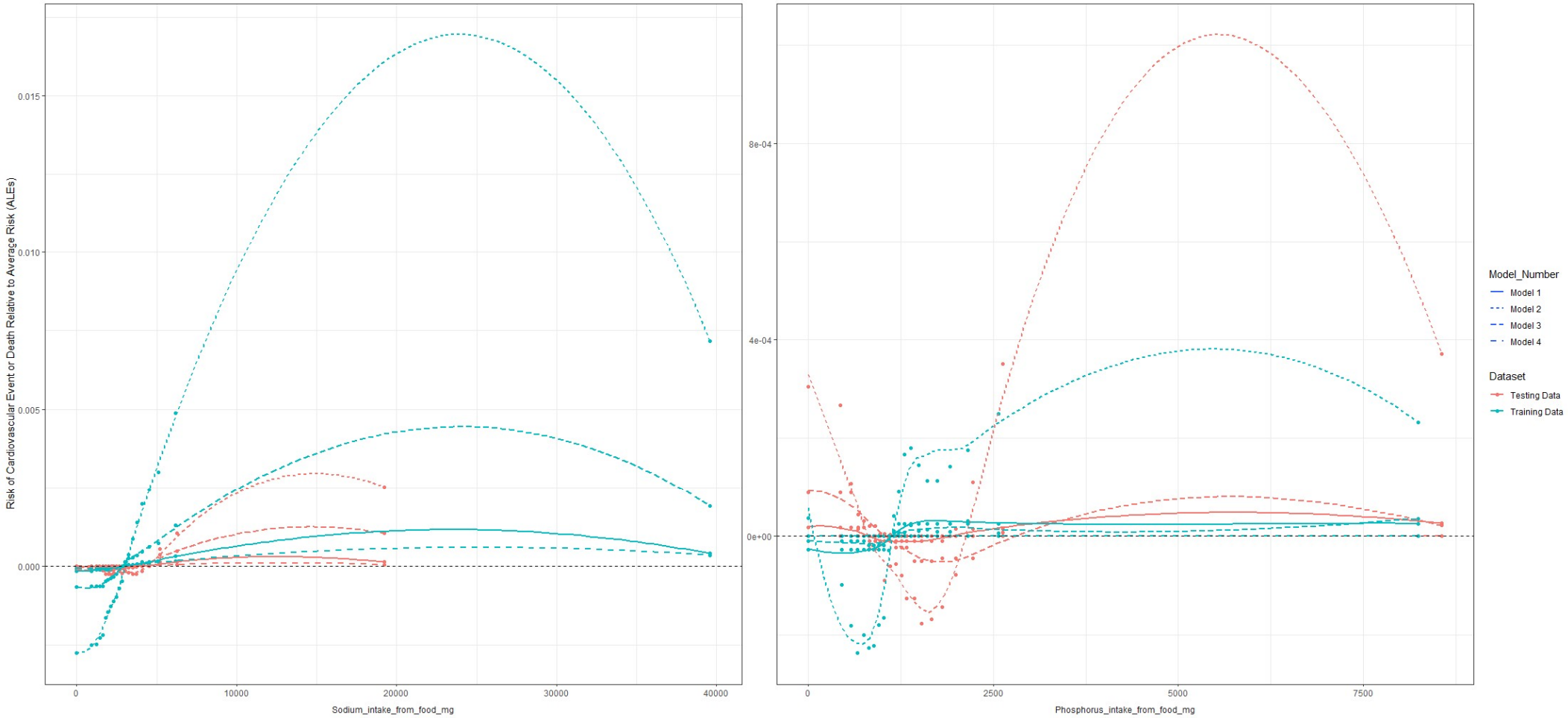
Supplementary figure 7: Accumulated local effects of vitamin D and vitamin B12 supplementation



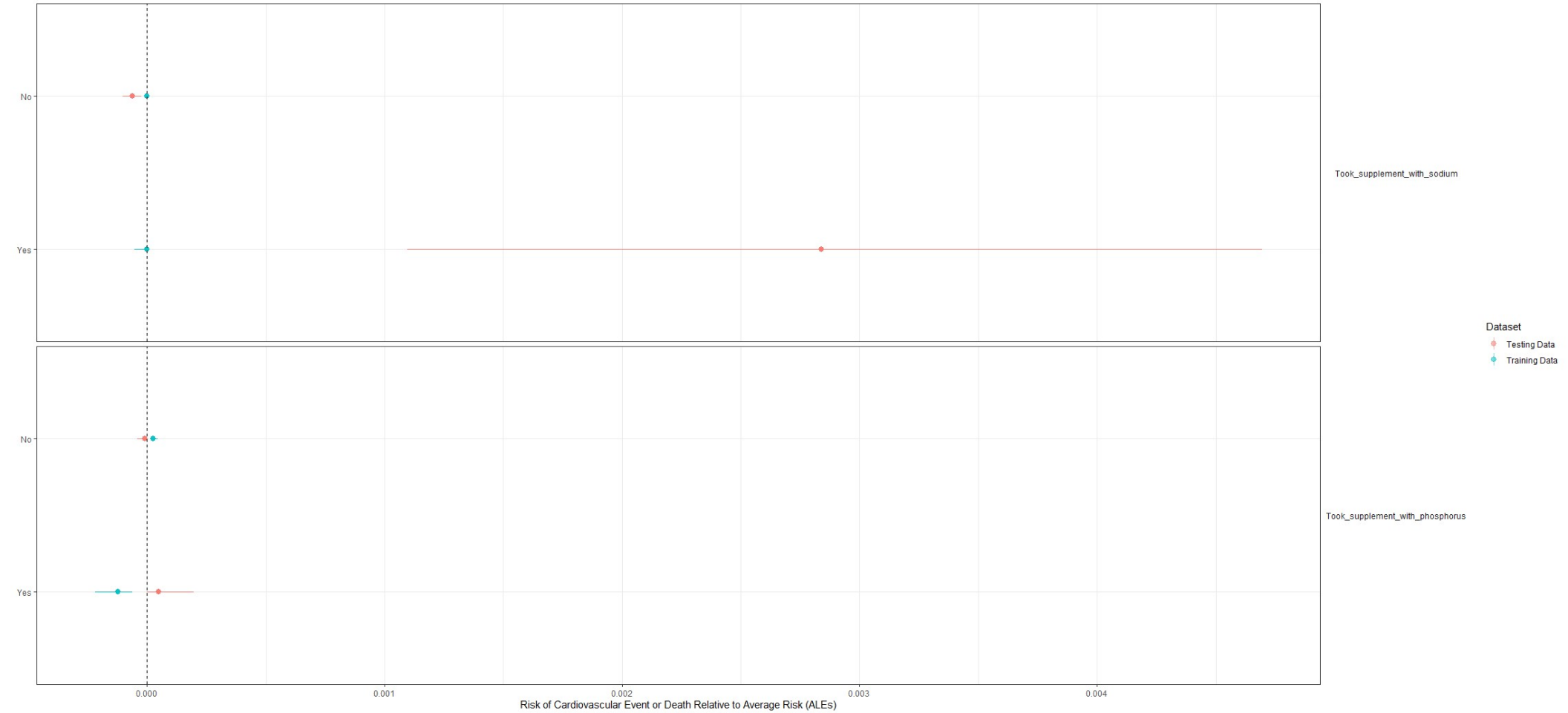
Supplementary figure 8: Accumulated local effects of the food categories with a permutation feature importance greater than zero



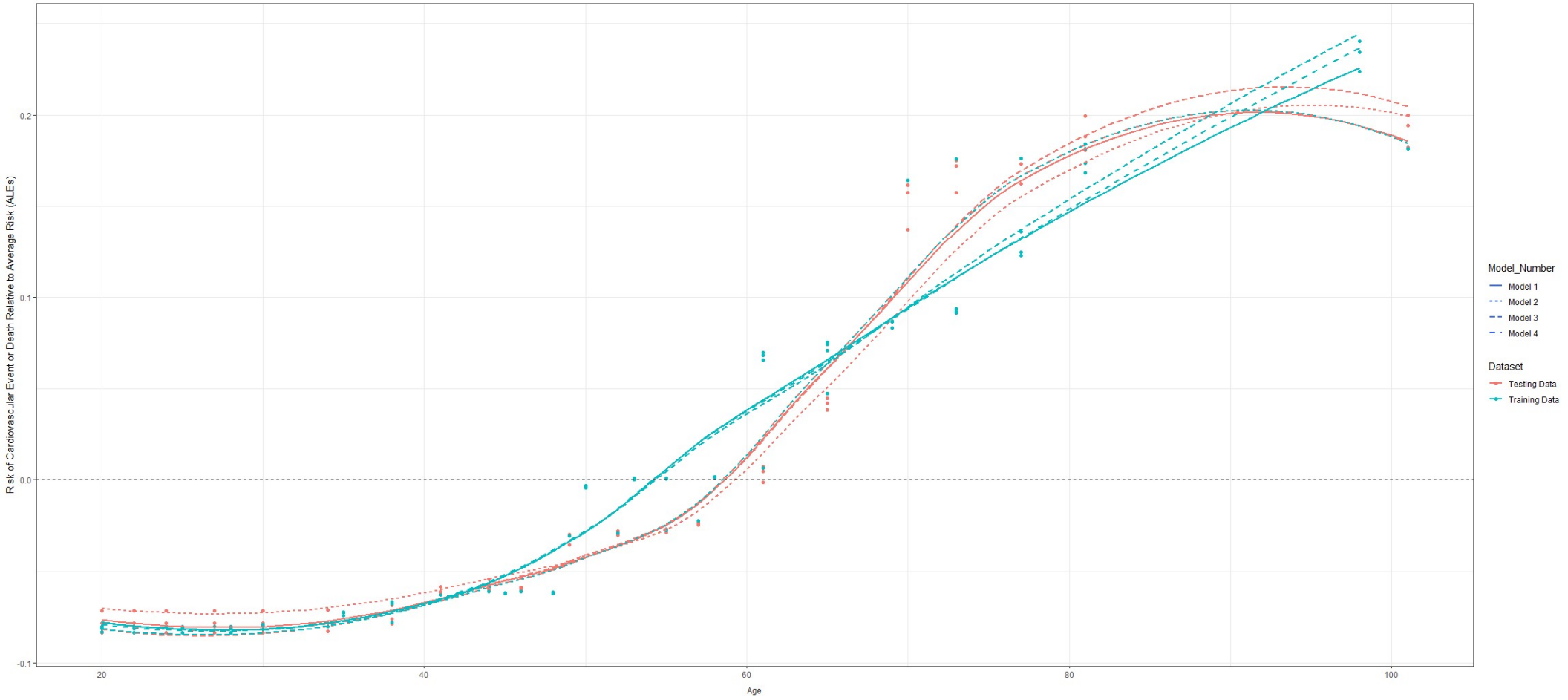
Supplementary figure 9: Accumulated local effects of the minerals from food with a permutation feature importance greater than zero



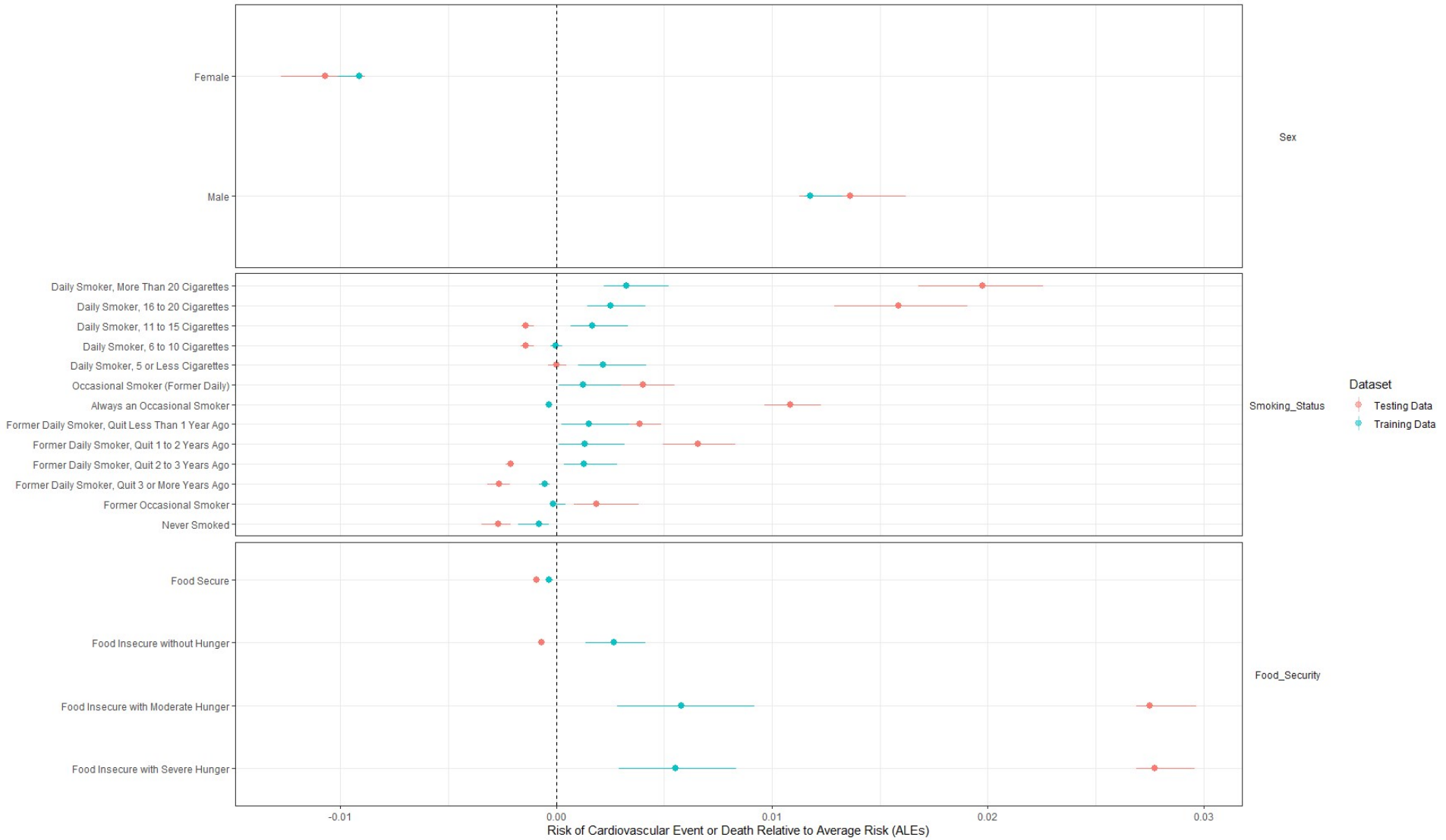
Supplementary figure 10: Accumulated local effects of sodium and phosphorous supplementation



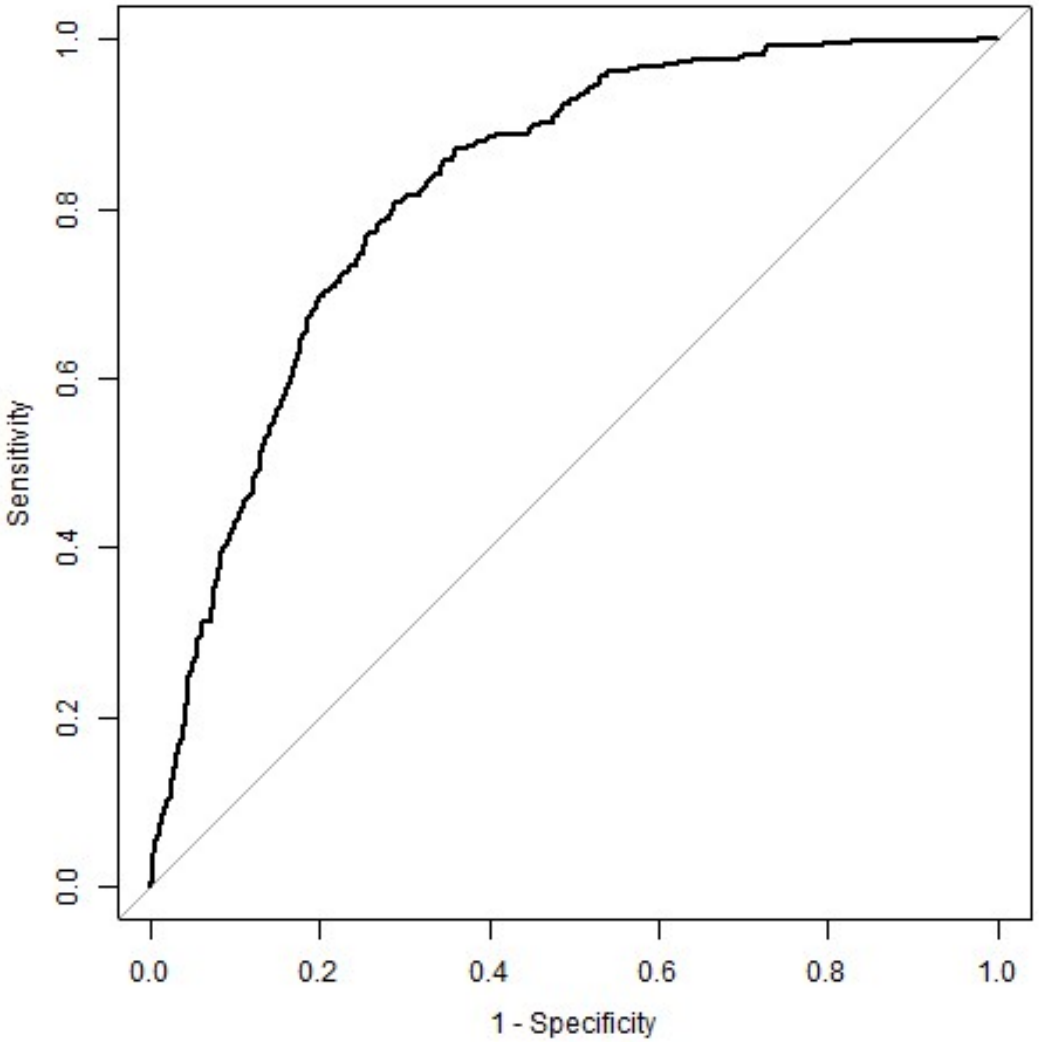
Supplementary figure 11: Accumulated local effects of age



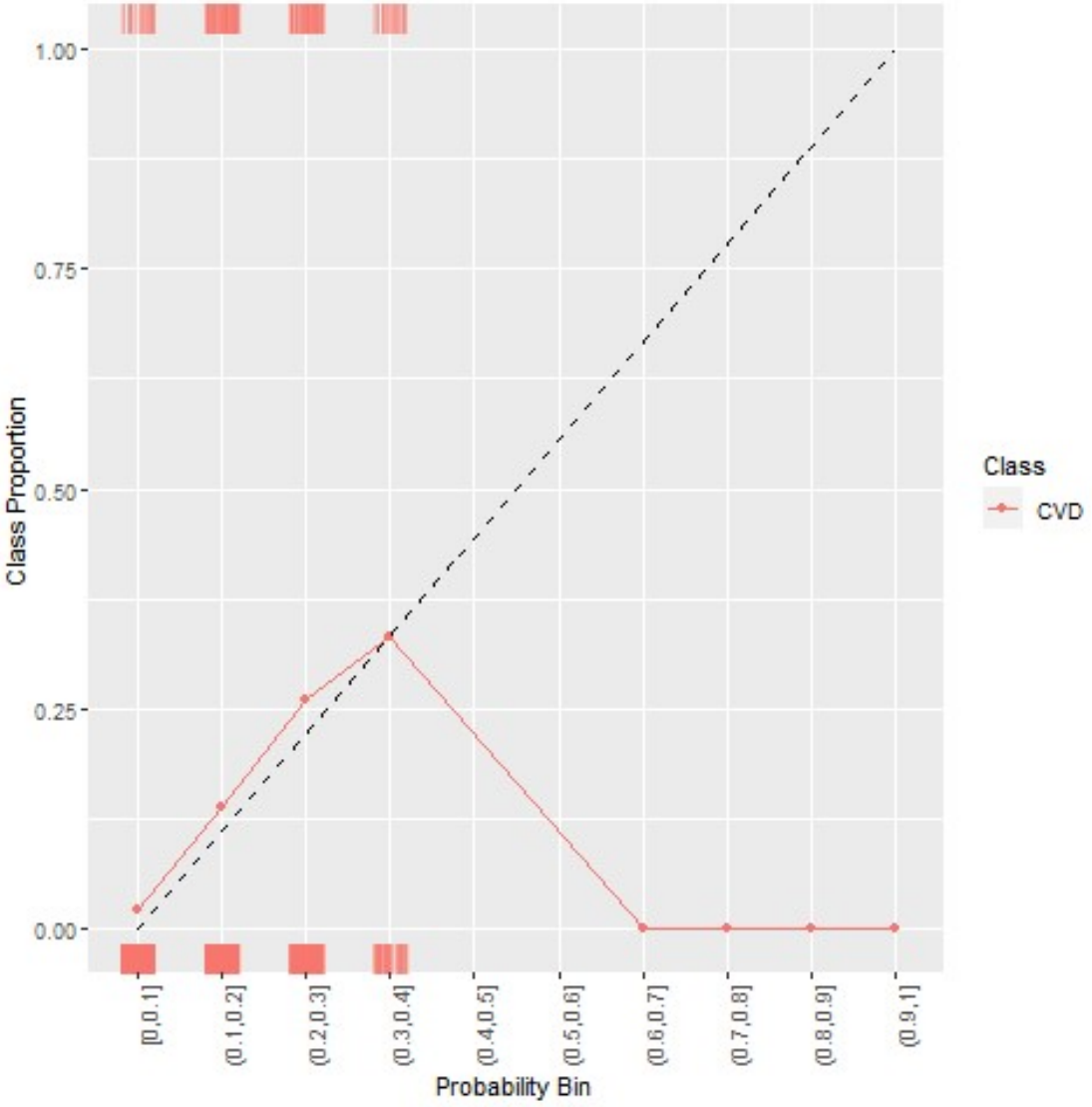
Supplemental figure 12: Accumulated local effects of features not related to nutrition with the highest permutation feature importance after age



Supplemental figure 13: Receiver operator curve plot



Supplemental figure 14: Calibration Plot



Supplementary table 1: All features included in models with descriptions

Feature (unit)	**Description (CCHS 2.2 variable name)**[182,183]
Age (a)	The age of each participant (DHHD_AGE).
Amount of food compared to usual	Reported amount of food consumed compared to usual during the period of the 24-hour dietary recall: much more than usual; usual; much less than usual (R24D_CON). Don't know and not stated considered missing.
Caffeine intake (mg)	Caffeine intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDCAF). Not stated considered missing.
Calcium intake (mg)	Calcium intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDCAL). Not stated considered missing.
Calcium supplement-use	Whether any supplement containing calcium was taken in the past month: yes; no (VSDDFCAL). Not stated considered missing.
Carbohydrate supplement-use	Whether any supplement containing carbohydrate was taken in the past month: yes; no (VSDDFCAR). Not stated considered missing.
Cholesterol intake (mg)	Cholesterol intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDCHO). Not stated considered missing.
Cultural/racial origin	The cultural or racial origin reported by participants: White; Chinese; Aboriginal Peoples of North America; South Asian; Other Racial or Cultural Origin; Multiple Racial/Cultural Origins (SDCDDRAC). The original CCHS 2.2 feature levels "Black," "Korean," "Filipino," "Japanese," Southeast Asian," "Arab," "West Asian," and "Latin American" were combined with "Other Racial or Cultural Origin" due to containing small

	numbers of participants. Not stated was considered missing.
Daily carrot consumption	The number of times per day that carrots are usually consumed (FVCDDCAR). Not stated considered missing.
Daily consumption of other vegetables	The number of times per day that other vegetables (i.e. anything except for carrots, potatoes, and green salad) are usually consumed (FVCDDVEG). Not stated considered missing.
Daily fruit consumption	The number of times per day that fruit is usually consumed (FVCDDFRU). Not stated considered missing.
Daily fruit juice consumption	The number of times per day that fruit juice is usually consumed (FVCDDJUI). Not stated considered missing.
Daily green salad consumption	The number of times per day that green salad is usually consumed (FVCDDSAL). Not stated considered missing.
Daily potato consumption	The number of times per day that potatoes (excluding French fries, fried potatoes, and potato chips) are usually consumed (FVCDDPOT). Not stated considered missing.
Educational level	Highest level of education completed: Grade 8 or lower; Grade 9 – 10; Grade 11 – 13; Secondary school, no post-secondary; Some post-secondary; Trades certificate or diploma; Diploma/certificate – college; University certificate below bachelor's; Bachelor's degree; University degree above bachelor's (EDUDDR10).

Energy intake (kcal)	Energy intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDEKC). Not stated considered missing.
Folic acid intake (µg)	Folic acid intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDFOA). Not stated considered missing.
Folic acid supplement-use	Whether any supplement containing folic acid was taken in the past month: yes; no (VSDDFFOA). Not stated considered missing.
Food security level	Reported household food security status: food secure; food insecure without hunger; food insecure with moderate hunger; food insecure with severe hunger (FSCDDHFS). Not stated considered missing.
Frequency of drinking alcohol	Each participant was asked how often they drank alcoholic beverages over the previous year: never; less than once a month; once a month; 2 to 3 times a month; once a week; 2 to 3 times a week; 4 to 6 times a week; every day (ALCD_1 and ALCD_2). Don't know, refusal, and not stated considered missing.
Household income	Total household income reported from all sources: less than $5000; $5000 to $9000; $10 000 to $14 999; $15 000 to $19 999; $20 000 to $29 999; $30 000 to $39 999; $40 000 to $49 999; $50 000 to $59 999; $60 000 to $79 999; $80 000 or more (INCDDHH). The original CCHS 2.2 feature level "no income" was combined with the feature "less than $5000" due to containing a small number of participants.

Immigration status	Whether each participant immigrated to Canada or was born in Canada: yes; no (SDCD_2). Not stated considered missing.
Iron intake (mg)	Iron intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDIRO). Not stated considered missing.
Iron supplement-use	Whether any supplement containing iron was taken in the past month: yes; no (VSDDFIRO). Not stated considered missing.
Linoleic fatty acid supplement-use	Whether any supplement containing linoleic fatty acids was taken in the past month: yes; no (VSDDFFAL). Not stated considered missing.
Linolenic fatty acid supplement-use	Whether any supplement containing linolenic fatty acids was taken in the past month: yes; no (VSDDFFAN). Not stated considered missing.
Magnesium intake (mg)	Magnesium intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDMAG). Not stated considered missing.
Magnesium supplement-use	Whether any supplement containing magnesium was taken in the past month: yes; no (VSDDFMAG). Not stated considered missing.
Marital status	The marital status of each participant: married; common-law; widowed; separated; divorced; single; never married (DHHD_MS). Don't know and refusal considered missing.
Moisture intake (g)	Water intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDMOI). Not stated considered missing.
Naturally occurring folate intake (µg)	Naturally occurring folate intake derived from all food and drink sources reported in the 24-hour

	dietary recall (FSDDDFON). Not stated considered missing.
Niacin intake (mg)	Niacin intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDNIA). Not stated considered missing.
Niacin supplement-use	Whether any supplement containing niacin was taken in the past month: yes; no (VSDDFNIA). Not stated considered missing.
Percent of life spent in Canada (%)	Percent of the participants life spent in Canada (DHHD_AGE, SDCD_3). Don't know, refusal, and not stated considered missing.
Percent of total energy from alcohol (%)	Percent of total energy intake from alcohol, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDEAL). Not stated considered missing.
Percent of total energy from carbohydrates	Percent of total energy intake from carbohydrate, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDECA). Not stated considered missing.
Percent of total energy from fat (%)	Percent of total energy intake from fat, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDELI). Not stated considered missing.
Percent of total energy from linoleic fatty acids (%)	Percent of total energy intake from linoleic fatty acids, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDEEI). Not stated considered missing.
Percent of total energy from linolenic fatty acids (%)	Percent of total energy intake from linolenic fatty acids, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDENI). Not stated considered missing.

Percent of total energy from monounsaturated fatty acids (%)	Percent of total energy intake from monounsaturated fatty acids, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDEMO). Not stated considered missing.
Percent of total energy from polyunsaturated fatty acids (%)	Percent of total energy intake from polyunsaturated fatty acids, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDEPO). Not stated considered missing.
Percent of total energy from protein (%)	Percent of total energy intake from protein, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDEPR). Not stated considered missing.
Percent of total energy from saturated fatty acids (%)	Percent of total energy intake from saturated fatty acids, derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDESA). Not stated considered missing.
Phosphorous intake (mg)	Phosphorous intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDPHO). Not stated considered missing.
Phosphorous supplement-use	Whether any supplement containing phosphorous was taken in the past month: yes; no (VSDDFPHO). Not stated considered missing.
Physical activity level (kcal/kg/h)	Average daily energy expenditure of the participant in the last 3 months (PACDDEE). Not stated considered missing.
Potassium intake (mg)	Potassium intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDPOT). Not stated considered missing.

Potassium supplement-use	Whether any supplement containing potassium was taken in the past month: yes; no (VSDDFPOT). Not stated considered missing.
Residence location type	Whether participants reported a rural or urban residence, based on reported postal code and 2001 Census geography: urban; rural (GEODDUR2).
Riboflavin intake (mg)	Riboflavin intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDRIB). Not stated considered missing.
Riboflavin supplement-use	Whether any supplement containing riboflavin was taken in the past month: yes; no (VSDDFRIB). Not stated considered missing.
Sense of belonging	Reported sense of belonging to local community: very strong; somewhat strong; somewhat weak; and very weak (GEND_10). Don't know, refusal, and not stated considered missing.
Sex	The reported sex of each participant: male; female (DHHD_SEX).
Smoking status	Amount, frequency, and recency of smoking cigarettes: Daily Smoker, More Than 20 Cigarettes; Occasional Smoker (Former Daily); Always an Occasional Smoker; Former Daily Smoker, Quit 3 or More Years Ago; Former Occasional Smoker; Never Smoked; Former Daily Smoker, Quit Less Than 1 Year Ago; Former Daily Smoker, Quit 1 to 2 Years Ago; Former Daily Smoker, Quit 2 to 3 Years Ago; Daily Smoker, 5 or Less Cigarettes; Daily Smoker, 6 to 10 Cigarettes; Daily Smoker, 11 to 15 Cigarettes; Daily Smoker, 16 to 20 Cigarettes (SMKD_202, SMKD_204,

	SMKD_05D, SMKDDSTY, SMKDDSTP). Don't know, refusal, and not stated considered missing.
Sodium intake (mg)	Sodium intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDSOD). Not stated considered missing.
Sodium supplement-use	Whether any supplement containing sodium was taken in the past month: yes; no (VSDDFSOD). Not stated considered missing.
Stress level	Reported self-perceived stress: not at all stressful; not very stressful; a bit stressful; quite a bit stressful; and extremely stressful (GEND_07). Don't know, refusal, and not stated considered missing.
Thiamin intake (mg)	Thiamin intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDTHI). Not stated considered missing.
Thiamin supplement-use	Whether any supplement containing thiamin was taken in the past month: yes; no (VSDDFTHI). Not stated considered missing.
Total dietary fibre intake (g)	Dietary fibre intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDFI). Not stated considered missing.
Total sugar intake (g)	Total sugar intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDSUG). Not stated considered missing.
Vitamin A intake (µg RAE)	Vitamin A intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDRAE). Not stated considered missing.
Vitamin A supplement-use	Whether any supplement containing vitamin A was taken in the past month: yes; no (VSDDFA). Not stated considered missing.

Vitamin B12 intake (µg)	Vitamin B12 intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDB12). Not stated considered missing.
Vitamin B12 supplement-use	Whether any supplement containing vitamin B12 was taken in the past month: yes; no (VSDDFB12). Not stated considered missing.
Vitamin B6 intake (mg)	Vitamin B6 intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDB6). Not stated considered missing.
Vitamin B6 supplement-use	Whether any supplement containing vitamin B6 was taken in the past month: yes; no (VSDDFB6). Not stated considered missing.
Vitamin C intake (mg)	Vitamin C intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDC). Not stated considered missing.
Vitamin C supplement-use	Whether any supplement containing vitamin C was taken in the past month: yes; no (VSDDFC). Not stated considered missing.
Vitamin D intake (µg)	Vitamin D intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDDMG). Not stated considered missing.
Vitamin D supplement-use	Whether any supplement containing vitamin D was taken in the past month: yes; no (VSDDFDMG). Not stated considered missing.
Vitamin E supplement-use	Whether any supplement containing vitamin E was taken in the past month: yes; no (VSDDFATE). Not stated considered missing.
Vitamin or mineral supplement-use	Whether any vitamin or mineral supplement was taken in the past month: yes; no (VSDD_01). Don't know and refusal considered missing.

Zinc intake (mg)	Zinc intake derived from all food and drink sources reported in the 24-hour dietary recall (FSDDDZIN). Not stated considered missing.
Zinc supplement-use	Whether any supplement containing zinc was taken in the past month: yes; no (VSDDFZIN). Not stated considered missing.

Supplementary table 2: Percent of observations with missing values for each feature included in models

Feature	Percent with Missing Data (%)
Any	9.3
Age	0.0
Sex	0.0
Marital_Status	0.1
Urban_or_Rural	0.0
Stress	0.1
Sense_of_Belonging	0.3
Smoking_Status	0.1
Food_Security	0.4
Immigrant_Status	0.1
Race	0.0
Percent_Life_Canada	0.1
Education	0.8
Household_Income	7.8
Takes_Supplements	0.0
Daily_consumption_fruit_juice	0.1
Daily_consumption_fruit	0.0
Daily_consumption_green_salad	0.1
Daily_consumption_potatoes	0.0
Daily_consumption_carrots	0.1
Daily_consumption_other_vegetables	0.2
Energy_intake_from_food_kcal	0.0
Total_dietary_fibre_from_food_g	0.0
Total_sugars_intake_from_food_g	0.0
Cholesterol_intake_from_food_mg	0.0
Percent_total_energy_from_carbohydrates	0.0
Percent_total_energy_from_fat	0.0

Percent_total_energy_satur_Fatty_acids	0.0
Percent_total_energy_mono_Fatty_acids	0.0
Percent_total_energy_poly_Fatty_acids	0.0
Percent_total_energy_linoleic_fatty_acid	0.0
Percent_total_energy_from_linolenic_acid	0.0
Percent_total_energy_from_proteins	0.0
Percent_total_energy_from_alcohol	0.0
Vit_A_from_food_sources_RAE_mcg	0.0
Vitamin_D_intake_from_food_mcg	0.0
Vitamin_C_intake_from_food_mg	0.0
Thiamin_intake_from_food_mg	0.0
Riboflavin_intake_from_food_mg	0.0
Niacin_intake_from_food_NE_mg	0.0
Vitamin_B6_intake_from_food_mg	0.0
Vitamin_B12_intake_from_food_mcg	0.0
Naturally_occurring_folate_mcg	0.0
Folic_acid_intake_from_food_mcg	0.0
Calcium_intake_from_food_mg	0.0
Phosphorus_intake_from_food_mg	0.0
Magnesium_intake_from_food_mg	0.0
Iron_intake_from_food_sources_mg	0.0
Zinc_intake_from_food_mg	0.0
Sodium_intake_from_food_mg	0.0
Potassium_intake_from_food_mg	0.0
Caffeine_intake_from_food_mg	0.0
Moisture_intake_from_food_g	0.0
Took_supplement_with_carbohydrate	0.0
Took_supplement_with_calcium	0.1

Took_supplement_with_iron	0.0
Took_supplement_with_magnesium	0.0
Took_supplement_with_phosphorus	0.0
Took_supplement_with_potassium	0.0
Took_supplement_with_sodium	0.0
Took_supplement_with_zinc	0.0
Took_suppl_With_vit_D_mcg	0.0
Took_supplement_with_vit_C	0.1
Took_supplement_with_thiamin	0.1
Took_supplement_with_riboflavin	0.1
Took_supplement_with_niacin	0.1
Took_supplement_with_vit_B6	0.1
Took_supplement_with_vit_B12	0.1
Took_suppl_With_folic_acid	0.1
Took_suppl_With_linoleic_acid	0.0
Took_suppl_Linolenic_acid	0.0
Supplement_with_alpha_tocoph	0.0
Took_supplement_with_vit_A	0.0
METS_per_Day	0.0
Frequency_of_Drinking_EtOH	0.1
Amount_food_consumed_compared_to_usual	0.1

Supplementary table 3: Hyperparameter tuning results
Available upon request

Supplemental table 4: All feature importance values for all features in all models
Available upon request

Supplementary table 5: Accumulated local effects of all included features in all models
Available upon request

CHAPTER 4: Conclusion

The major potential benefits of big data and machine learning applied to nutritional epidemiology have been discussed. These include new ways of measuring diet that may help mitigate measurement error; better modelling of non-linearity, non-additivity, and the complexity of diet; new ways of controlling for confounders; and applications to both predictive and causal models. Then, we applied conditional inference forests, a machine learning method, to a linked Canadian population-based survey to predict CVD. Interpretable machine learning methods, permutation feature importance and accumulated local effects, were applied to our models to determine the nutrients that were most predictive of CVD and also to ascertain the nature of those relationships. Some of the predictive nutrients identified were alcohol, sodium, fruits, vegetables, supplement-use, caffeine, B vitamins, protein, moisture, PUFAs, and zinc. Accumulated local effects plots revealed a mixture of threshold-linear, j-shaped, and u-shaped relationships between dietary factors and CVD. Some of these relationships had previously been reported in the literature, while others were counter to findings in the broad literature or had not been described previously. Our model also achieved a competitive level of predictive discrimination and calibration, despite lacking access to many of the biomedical features that prediction models typically include.

How machine learning has been applied in the current work mainly addresses our identified opportunity to better model non-linearity, non-additivity, and the complexity of diet in relation to health outcomes. Our findings of many non-linear relationships among many different nutrition-related variables lends support to the proposition that non-

linearity and addressing more of the richness of diet may be important in nutritional epidemiology. However, without comparison to a traditional parametric model such as logistic regression, it is impossible to know how much this added model complexity improved prediction performance, if at all. An important next step is to perform this comparison, and also to assess the performance of more machine learning algorithms (e.g. boosted trees) that may achieve better predictivity. Additional avenues identified in chapter 2 that we did not explore in our analysis in chapter 3, but may be useful for future nutritional epidemiology studies, include use of big data, the importance of interactions, new ways of controlling for high-dimensional confounders, and causal inference applications. With a larger dataset and repeated measures of diet or food frequency questionnaires, we anticipate that we could better leverage the advantages of machine learning towards detection of smaller, non-linear, and non-additive relationships between diet and disease. Having access to only a single 24-hour dietary recall was a potential limitation of this work, as substantial measurement error may be present. For a subset of the CCHS population a second dietary recall is available, however, for this exploratory machine learning work we did not further explore methods to incorporate the second recall to adjust for intra- and inter-individual variations as have been described previously.[184] Furthermore, when available, we are interested in developing models that also include the foods consumed, rather than just the derived nutrients. This would help to determine if some of the observed effects stemmed from meat, for example, rather than individual nutrients. The application of H-statistics[29] is something we intend to pursue in the future for identifying important interactions incorporated into machine learning

models. Furthermore, the use of unsupervised machine learning methods could help identify important *a posteriori* dietary patterns that could further improve model predictions and interpretability. Finally, in the future, linking this data to higher-dimensional sets of covariates and the use of machine learning for causal inference are other promising opportunities.

A unique aspect of this work is the use of conditional inference forests, which have uncommonly been applied in health research. As we identified in our second manuscript, this algorithm reduces bias in permutation feature importance results and allows higher predictive performance in the context of high mtry parameters, through adjustment of the mincriterion parameter. This may be important in health research when interpretability is important.

Nutritional epidemiology continues to be a challenging area of research with substantial impacts on public health, in which machine learning may enable better use of observational data. Our work is an initial step, showing the plausibility and potential importance of modelling greater dietary complexity and non-linear relationships between diet and disease, that deserves further study. Additionally, our work highlights important developments in interpretable machine learning methods, that will be instrumental in applying these methods to public health more broadly, when understanding how predictions are made is of utmost importance.

4.1 References

1. GBD 2017 Diet Collaborators A, Sur PJ, Fay KA, et al. Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet (London, England)*. 2019;393(10184):1958-1972. doi:10.1016/S0140-6736(19)30041-8
2. Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol*. 2017;70(1):1-25. doi:10.1016/j.jacc.2017.04.052
3. Mozaffarian D. Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity. *Circulation*. 2016;133(2):187-225. doi:10.1161/CIRCULATIONAHA.115.018585
4. Dehghan M, Mente A, Zhang X, et al. Associations of fats and carbohydrate intake with cardiovascular disease and mortality in 18 countries from five continents (PURE): a prospective cohort study. *Lancet*. 2017;390(10107):2050-2062. doi:10.1016/S0140-6736(17)32252-3
5. Seidelmann SB, Claggett B, Cheng S, et al. Dietary carbohydrate intake and mortality: a prospective cohort study and meta-analysis. *Lancet Public Heal*. 2018;3(9):e419-e428. doi:10.1016/S2468-2667(18)30135-X
6. Zhong VW, Van Horn L, Cornelis MC, et al. Associations of Dietary Cholesterol or Egg Consumption with Incident Cardiovascular Disease and Mortality. *JAMA - J Am Med Assoc*. 2019;321(11):1081-1095. doi:10.1001/jama.2019.1572
7. Key TJ, Appleby PN, Bradbury KE, et al. Consumption of Meat, Fish, Dairy Products, and Eggs and Risk of Ischemic Heart Disease. *Circulation*.

2019;139(25):2835-2845. doi:10.1161/CIRCULATIONAHA.118.038813

8. Johnston BC, Zeraatkar D, Han MA, et al. Unprocessed Red Meat and Processed Meat Consumption: Dietary Guideline Recommendations From the Nutritional Recommendations (NutriRECS) Consortium. *Ann Intern Med*. October 2019. doi:10.7326/m19-1621
9. *IARC Monographs Evaluate Consumption of Red Meat and Processed Meat*.; 2015. http://www.iarc.fr/en/media-centre/iarcnews/pdf/Monographs-Q&A_Vol114.pdf. Accessed November 4, 2019.
10. Ioannidis JPA. Unreformed nutritional epidemiology: a lamp post in the dark forest. *Eur J Epidemiol*. 2019;34(4):327-331. doi:10.1007/s10654-019-00487-5
11. Giovannucci E. Nutritional epidemiology: forest, trees and leaves. *Eur J Epidemiol*. 2019;34(4):319-325. doi:10.1007/s10654-019-00488-4
12. Trepanowski JF, Ioannidis JPA. Perspective: Limiting Dependence on Nonrandomized Studies and Improving Randomized Trials in Human Nutrition Research: Why and How. *Adv Nutr*. 2018;9(4):367-377. doi:10.1093/advances/nmy014
13. Ioannidis JPA. The Challenge of Reforming Nutritional Epidemiologic Research. *JAMA*. 2018;320(10):969. doi:10.1001/jama.2018.11025
14. Ioannidis JP. We need more randomized trials in nutrition—preferably large, long-term, and with negative results. *Am J Clin Nutr*. 2016;103(6):1385-1386. doi:10.3945/ajcn.116.136085
15. Ioannidis JPA. The Challenge of Reforming Nutritional Epidemiologic Research.

JAMA. 2018;320(10):969. doi:10.1001/jama.2018.11025

16. Hastie T, Tibshirani R, Witten D, Gareth J. *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer; 2013.
17. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random Survival Forests for High-Dimensional Data. 2011. doi:10.1002/sam.10103
18. Kong YW, Baqar S, Jerums G, Ekinci EI. Sodium and its role in cardiovascular disease - The debate continues. *Front Endocrinol (Lausanne)*. 2016;7(DEC). doi:10.3389/fendo.2016.00164
19. Mente A, Yusuf S. Evolving evidence about diet and health. *Lancet Public Heal*. 2018;3(9):e408-e409. doi:10.1016/S2468-2667(18)30160-9
20. Zeevi D, Korem T, Zmora N, et al. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*. 2015;163(5):1079-1094. doi:10.1016/j.cell.2015.11.001
21. Panaretos D, Koloverou E, Dimopoulos AC, et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): the ATTICA study. *Br J Nutr*. 2018;120(03):1-9. doi:https://dx.doi.org/10.1017/S0007114518001150
22. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 2006;7(1):3. doi:10.1186/1471-2105-7-3
23. Szabo de Edelenyi F, Goumidi L, Bertrais S, et al. Prediction of the metabolic syndrome status based on dietary and genetic parameters, using Random Forest. *Genes Nutr*. 2008;3(3-4):173-176. doi:10.1007/s12263-008-0097-y

24. Nau C, Ellis H, Huang H, et al. Exploring the forest instead of the trees: An innovative method for defining obesogenic and obesoprotective environments. *Health Place*. 2015;35:136-146. doi:10.1016/j.healthplace.2015.08.002

25. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*. 2007;8(1):25. doi:10.1186/1471-2105-8-25

26. Panaretos D, Koloverou E, Dimopoulos AC, et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): the ATTICA study. *Br J Nutr*. 2018:1-9. doi:https://dx.doi.org/10.1017/S0007114518001150

27. Biesbroek S, van der A DL, Brosens MC, et al. Identifying cardiovascular risk factor–related dietary patterns with reduced rank regression and random forest in the EPIC-NL cohort. *Am J Clin Nutr*. 2015;102(1):146-154. doi:10.3945/ajcn.114.092288

28. Pearl J. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*. New York, New York, USA: Association for Computing Machinery (ACM); 2018:3-3. doi:10.1145/3159652.3176182

29. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub; 2019. https://christophm.github.io/interpretable-ml-book/.

30. Statistics Canada. Leading causes of death, total population (age standardization

using 2011 population). https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310080101. Published 2017. Accessed July 8, 2019.

31. Sidney S, Quesenberry CP, Jaffe MG, et al. Recent Trends in Cardiovascular Mortality in the United States and Public Health Goals. *JAMA Cardiol*. 2016;1(5):594. doi:10.1001/jamacardio.2016.1326

32. Trepanowski JF, Ioannidis JPA. Perspective: Limiting Dependence on Nonrandomized Studies and Improving Randomized Trials in Human Nutrition Research: Why and How. *Adv Nutr*. 2018;9(4):367-377. doi:10.1093/advances/nmy014

33. Hébert JR, Frongillo EA, Adams SA, et al. Perspective: Randomized Controlled Trials Are Not a Panacea for Diet-Related Research. *Adv Nutr*. 2016;7(3):423-432. doi:10.3945/an.115.011023

34. Zeilstra D, Younes JA, Brummer RJ, Kleerebezem M. Perspective: Fundamental Limitations of the Randomized Controlled Trial Method in Nutritional Research: The Example of Probiotics. *Adv Nutr*. 2018;9(5):561-571. doi:10.1093/advances/nmy046

35. Mozaffarian D. Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity. *Circulation*. 2016;133(2):187-225. doi:10.1161/CIRCULATIONAHA.115.018585

36. Alpers DH, Bier DM, Carpenter KJ, McCormick DB, Miller AB, Jacques PF. History and Impact of Nutritional Epidemiology. *Adv Nutr*. 2014;5(5):534-536.

doi:10.3945/an.114.006353

37. Harcombe Z, Baker JS, DiNicolantonio JJ, Grace F, Davies B. Evidence from randomised controlled trials does not support current dietary fat guidelines: a systematic review and meta-analysis. *Open Hear*. 2016;3(2):e000409. doi:10.1136/openhrt-2016-000409

38. Mente A, de Koning L, Shannon HS, Anand SS. A Systematic Review of the Evidence Supporting a Causal Link Between Dietary Factors and Coronary Heart Disease. *Arch Intern Med*. 2009;169(7):659. doi:10.1001/archinternmed.2009.38

39. Astrup A, Bertram HC, Bonjour J-P, et al. WHO draft guidelines on dietary saturated and trans fatty acids: time for a new approach? *BMJ*. 2019;366:l4137. doi:10.1136/bmj.l4137

40. Sacks FM, Lichtenstein AH, Wu JHY, et al. Dietary Fats and Cardiovascular Disease: A Presidential Advisory From the American Heart Association. *Circulation*. 2017;136(3). doi:10.1161/CIR.0000000000000510

41. Hamley S. The effect of replacing saturated fat with mostly n-6 polyunsaturated fat on coronary heart disease: a meta-analysis of randomised controlled trials. *Nutr J*. 2017;16(1):30. doi:10.1186/s12937-017-0254-5

42. Hooper L, Martin N, Abdelhamid A, Davey Smith G. Reduction in saturated fat intake for cardiovascular disease. *Cochrane Database Syst Rev*. 2015;(6):CD011737. doi:10.1002/14651858.CD011737

43. Schwab U, Lauritzen L, Tholstrup T, et al. Effect of the amount and type of dietary fat on cardiometabolic risk factors and risk of developing type 2 diabetes,

cardiovascular diseases, and cancer: a systematic review. *Food Nutr Res*. 2014;58(1):25145. doi:10.3402/fnr.v58.25145

44. Chowdhury R, Warnakula S, Kunutsor S, et al. Association of Dietary, Circulating, and Supplement Fatty Acids With Coronary Risk. *Ann Intern Med*. 2014;160(6):398. doi:10.7326/M13-1788

45. Mozaffarian D, Micha R, Wallace S. Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: A systematic review and meta-analysis of randomized controlled trials. *PLoS Med*. 2010;7(3). doi:10.1371/journal.pmed.1000252

46. Mozaffarian D. Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity. *Circulation*. 2016;133(2):187-225. doi:10.1161/CIRCULATIONAHA.115.018585

47. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *Am J Epidemiol*. 2007;166(6):646-655. doi:10.1093/aje/kwm165

48. Loken E, Gelman A. Measurement error and the replication crisis. *Science (80-)*. 2017;355(6325):584-585. doi:10.1126/science.aal3618

49. Grewal R, Cote JA, Baumgartner H. Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Mark Sci*. 2004;23(4):519-529. doi:10.1287/mksc.1040.0070

50. Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: progress in addressing complexity in diet and health. *Annu Rev Nutr*. 2012;32:183-202.

doi:https://dx.doi.org/10.1146/annurev-nutr-072610-145159

51. Qi L. Mendelian randomization in nutritional epidemiology. *Nutr Rev*. 2009;67(8):439-450. doi:10.1111/j.1753-4887.2009.00218.x

52. Aula Médica España Corella G, Aula Médica Madrid G. Biomarkers: background, classification and guidelines for applications in nutritional epidemiology. *Nutr Hosp*. 2015;31:177-188. doi:10.3305/nh.2015.31.sup3.8765

53. Davey Smith G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes Nutr*. 2011;6(1):27-43. doi:10.1007/s12263-010-0181-y

54. Ference BA, Yoo W, Alesh I, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: A Mendelian randomization analysis. *Ration Pharmacother Cardiol*. 2013;9(1):90-98. doi:10.1016/j.jacc.2012.09.017

55. Sacerdote C, Guarrera S, Smith GD, et al. Lactase Persistence and Bitter Taste Response: Instrumental Variables and Mendelian Randomization in Epidemiologic Studies of Dietary Factors and Cancer Risk. *Am J Epidemiol*. 2007;166(5):576-581. doi:10.1093/aje/kwm113

56. "Big Data" : big gaps of knowledge in the field of internet science — Eindhoven University of Technology research portal. https://research.tue.nl/en/publications/big-data-big-gaps-of-knowledge-in-the-field-of-internet-science. Accessed November 4, 2019.

57. Lacey D. 3D data management: Controlling data volume, velocity and variety.

META Gr Res note. 2001;6(70):1.

58. Dedić N, Stanier C. Towards differentiating business intelligence, big data, data analytics and knowledge discovery. In: *Lecture Notes in Business Information Processing*. Vol 285. Springer Verlag; 2017:114-122. doi:10.1007/978-3-319-58801-8_10

59. Xia F, Yang LT, Wang L, Vinel A. Internet of Things. *Int J Commun Syst*. 2012;25(9):1101-1102. doi:10.1002/dac.2417

60. Shukla SK, Murali NS, Brilliant MH. Personalized medicine going precise: From genomics to microbiomics. *Trends Mol Med*. 2015;21(8):461-462. doi:10.1016/j.molmed.2015.06.002

61. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. *IBM J Res Dev*. 1959;3(3):210-229. doi:10.1147/rd.33.0210

62. Optimization for Machine Learning - Google Books. https://books.google.co.in/books?hl=en&lr=&id=JPQx7s2L1A8C&oi=fnd&pg=PA403&dq=#v=onepage&q&f=false. Accessed November 4, 2019.

63. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018;15(4):233-234. doi:10.1038/nmeth.4642

64. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer series in statistics.; 2009.

65. Friedman JH. *Data Mining and Statistics: What's the Connection?*

66. Bermingham ML, Pong-Wong R, Spiliopoulou A, et al. Application of high-

dimensional feature selection: Evaluation for genomic prediction in man. *Sci Rep*. 2015;5. doi:10.1038/srep10312

67. Trunk G V. A Problem of Dimensionality: A Simple Example. *IEEE Trans Pattern Anal Mach Intell*. 1979;PAMI-1(3):306-307. doi:10.1109/TPAMI.1979.4766926

68. S.J. M. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu Rev Public Health*. 2018;39:95-112. doi:http://dx.doi.org/10.1146/annurev-publhealth-040617-014208

69. Health P, Bi MQ, Goodman KE. What Is Machine Learning: a Primer for the Epidemiologist. *Am J Epidemiol*. 2019.

70. Apple Women's Health Study – Harvard T. H. Chan School of Public Health. https://www.hsph.harvard.edu/applewomenshealthstudy/. Accessed January 25, 2020.

71. Maruvada P, Lampe JW, Wishart DS, et al. Perspective: Dietary Biomarkers of Intake and Exposure—Exploration with Omics Approaches. *Adv Nutr*. 2019:1-16. doi:10.1093/advances/nmz075

72. Hoi AS. *Food Image Recognition by Deep Learning*. www.moh.gov.sg/budget2016. Accessed November 4, 2019.

73. Image-Based Calorie Estimation using Deep Learning. https://www.lftechnology.com/blog/ai/image-calorie-estimation-deep-learning/. Accessed November 4, 2019.

74. Sahoo D, Hao W, Ke S, et al. FoodAI: Food Image Recognition via Deep Learning for Smart Food Logging. September 2019. http://arxiv.org/abs/1909.11946.

Accessed November 4, 2019.

75. Dillet R. Foodvisor automatically tracks what you eat using deep learning. TechCrunch. https://techcrunch.com/2019/10/14/foodvisor-automatically-tracks-what-you-eat-using-deep-learning/. Published 2019. Accessed November 4, 2019.

76. Min W, Jiang S, Liu L, Rui Y, Jain R. A Survey on Food Computing. 2018. doi:10.1145/3329168

77. Chin CL, Huang CC, Lin BJ, Wu GR, Weng TC, Chen HF. Smartphone-based food category and nutrition quantity recognition in food image with deep learning algorithm. In: *2016 International Conference on Fuzzy Theory and Its Applications, IFuzzy 2016*. Institute of Electrical and Electronics Engineers Inc.; 2017. doi:10.1109/iFUZZY.2016.8004962

78. Nguyen QC, Li D, Meng H-W, et al. Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. *JMIR public Heal Surveill*. 2016;2(2):e158. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=27751984.

79. Ocké MC. Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis. *Proc Nutr Soc*. 2013;72(2):191-199. doi:10.1017/S0029665113000013

80. McCullough ML, Feskanich D, Stampfer MJ, et al. Diet quality and major chronic disease risk in men and women: Moving toward improved dietary guidance. *Am J Clin Nutr*. 2002;76(6):1261-1271. doi:10.1093/ajcn/76.6.1261

81. Overview & Background of The Healthy Eating Index. https://epi.grants.cancer.gov/hei/. Accessed January 25, 2020.

82. Miller PE, Cross AJ, Subar AF, et al. Comparison of 4 established DASH diet indexes: Examining associations of index scores and colorectal cancer. *Am J Clin Nutr*. 2013;98(3):794-803. doi:10.3945/ajcn.113.063602

83. Walter S, Tiemeier H. Variable selection: Current practice in epidemiological studies. *Eur J Epidemiol*. 2009;24(12):733-736. doi:10.1007/s10654-009-9411-2

84. Kastorini C-MM, Papadakis G, Milionis HJ, et al. Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: A case/case-control study. *Artif Intell Med*. 2013;59(3):175-183. doi:10.1016/j.artmed.2013.08.005

85. Newby P, Muller D, Hallfrisch J, Qiao N, Andres R, Tucker KL. Dietary patterns and changes in body mass index and waist circumference in adults. *Am J Clin Nutr*. 2003;77(6):1417-1425. doi:10.1093/ajcn/77.6.1417

86. Zhang F, Tapera TM, Gou J. Application of a new dietary pattern analysis method in nutritional epidemiology. *BMC Med Res Methodol*. 2018;18(1):119. doi:10.1186/s12874-018-0585-8

87. Koliaki C, Katsilambros N. Dietary sodium, potassium, and alcohol: key players in the pathophysiology, prevention, and treatment of human hypertension. *Nutr Rev*. 2013;71(6):402-411. doi:10.1111/nure.12036

88. Brown IJ, Stamler J, Van Horn L, et al. Sugar-sweetened beverage, sugar intake of individuals, and their blood pressure: international study of macro/micronutrients

and blood pressure. *Hypertens (Dallas, Tex 1979)*. 2011;57(4):695-701. doi:10.1161/HYPERTENSIONAHA.110.165456

89. Kotchen TA, Kotchen JM. Dietary sodium and blood pressure: interactions with other nutrients. *Am J Clin Nutr*. 1997;65(2):708S-711S. doi:10.1093/ajcn/65.2.708S

90. Panaretos D, Koloverou E, Dimopoulos AC, et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): The ATTICA study. *Br J Nutr*. 2018;120(3):326-334. doi:10.1017/S0007114518001150

91. Rigdon J, Basu S. Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the USA using nationally randomly sampled data. *BMJ Open*. 2019;9(11):e032703. doi:10.1136/bmjopen-2019-032703

92. Olstad DL, McIntyre L. Reconceptualising precision public health. *BMJ Open*. 2019;9(9):e030279. doi:10.1136/bmjopen-2019-030279

93. Maharana A, Nsoesie EO. Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity. *JAMA Netw Open*. 2018;1(4):e181535. doi:10.1001/jamanetworkopen.2018.1535

94. Lynch KE, Whitcomb BW, DuVall SL. How Confounder Strength Can Affect Allocation of Resources in Electronic Health Records. *Perspect Heal Inf Manag*. 2018;15(Winter).

95. Phillips SM, Cadmus-Bertram L, Rosenberg D, Buman MP, Lynch BM. Wearable

Technology and Physical Activity in Chronic Disease: Opportunities and Challenges. *Am J Prev Med*. 2018;54(1):144-150. doi:10.1016/j.amepre.2017.08.015

96. Lemstra M, Mackenbach J, Neudorf C, Nannapaneni U. High health care utilization and costs associated with lower socio-economic status: Results from a linked dataset. *Can J Public Heal*. 2009;100(3):180-183. doi:10.1007/bf03405536

97. Hernán M, Robins J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.

98. Lleras-Muney A. The relationship between education and adult mortality in the United States. *Rev Econ Stud*. 2005;72(1):189-221. doi:10.1111/0034-6527.00329

99. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative Controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21(3):383-388. doi:10.1097/EDE.0b013e3181d61eeb

100. Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav*. 1995;Spec No:80-94. doi:10.2307/2626958

101. Arnold BF, Ercumen A, Benjamin-Chung J, Colford JM. Brief report: Negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology*. 2016;27(5):637-641. doi:10.1097/EDE.0000000000000504

102. Low YS, Gallego B, Shah NH. Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *J Comp Eff Res*. 2016;5(2):179-192. doi:10.2217/cer.15.53

103. Schnitzer ME, Lok JJ, Gruber S. Variable Selection for Confounder Control,

Flexible Modeling and Collaborative Targeted Minimum Loss-Based Estimation in Causal Inference. *Int J Biostat*. 2016;12(1):97-115. doi:10.1515/ijb-2015-0017

104. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512-522. doi:10.1097/EDE.0b013e3181a663cc

105. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346. doi:10.1002/sim.3782

106. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9(4):403-425. doi:10.1037/1082-989X.9.4.403

107. Wyss R, Ellis AR, Brookhart MA, et al. The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score. *Am J Epidemiol*. 2014;180(6):645-655. doi:10.1093/aje/kwu181

108. Mccaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*. 2013;32(19):3388-3414. doi:10.1002/sim.5753

109. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010;63(8):826-833. doi:10.1016/j.jclinepi.2009.11.020

110. Toh S, García Rodríguez LA, Hernán MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: An application to electronic medical records. *Pharmacoepidemiol Drug Saf*. 2011;20(8):849-857. doi:10.1002/pds.2152

111. Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur J Clin Pharmacol*. 2013;69(3):549-557. doi:10.1007/s00228-012-1334-2

112. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate Selection in High-Dimensional Propensity Score Analyses of Treatment Effects in Small Samples. *Pract Epidemiol* . 2011;173(12):1404-1413. doi:10.1093/aje/kwr001

113. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353. doi:10.1136/bmj.i2416

114. Manuel DG, Tuna M, Bennett C, et al. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the Cardiovascular Disease Population Risk Tool (CVDPoRT). *CMAJ*. 2018;190(29):E871-E882. doi:10.1503/cmaj.170914

115. Fisher S, Hsu A, Mojaverian N, et al. Dementia Population Risk Tool (DemPoRT): Study protocol for a predictive algorithm assessing dementia risk in the community. *BMJ Open*. 2017;7(10). doi:10.1136/bmjopen-2017-018018

116. Ng R, Sutradhar R, Wodchis WP, Rosella LC. Chronic Disease Population Risk

Tool (CDPoRT): a study protocol for a prediction model that assesses population-based chronic disease incidence. *Diagnostic Progn Res*. 2018;2(1):19. doi:10.1186/s41512-018-0042-5

117. Rosella LC, Manuel DG, Burchill C, Stukel TA. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). doi:10.1136/jech.2009.102244

118. Joseph P, Yusuf S, Lee SF, et al. Prognostic validation of a non-laboratory and a laboratory based cardiovascular disease risk score in multiple regions of the world. *Heart*. 2018;104(7):581-587. doi:10.1136/heartjnl-2017-311609

119. Harre FE, Lee KL, Pollock BG. Regression models in clinical studies: Determining relationships between predictors and response. *J Natl Cancer Inst*. 1988;80(15):1198-1202. doi:10.1093/jnci/80.15.1198

120. Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Appl Stat*. 1994;43(3):429. doi:10.2307/2986270

121. Road Map for Choosing Between Statistical Modeling and Machine Learning | Statistical Thinking. https://www.fharrell.com/post/stat-ml/. Accessed November 4, 2019.

122. Harrell Jr FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer; 2015.

123. Steyerberg EW. *Clinical Prediction Models*. New York: Springer; 2009.

124. Martínez-González MÁ, Hershey MS, Zazpe I, Trichopoulou A. Transferability of

the Mediterranean diet to non-Mediterranean countries. What is and what is not the Mediterranean diet. *Nutrients*. 2017;9(11). doi:10.3390/nu9111226

125. Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *Am J Epidemiol*. 2017;185(1):65-73. doi:https://dx.doi.org/10.1093/aje/kww165

126. Stephan KE, Penny WD, Moran RJ, den Ouden HEM, Daunizeau J, Friston KJ. Ten simple rules for dynamic causal modeling. *Neuroimage*. 2010;49(4):3099-3109. doi:10.1016/j.neuroimage.2009.11.015

127. T. V, C. A, M. A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388(10053):1545-1602. doi:http://dx.doi.org/10.1016/S0140-6736%2816%2931678-6

128. Lopez AD, Adair T. Is the long-term decline in cardiovascular-disease mortality in high-income countries over? Evidence from national vital statistics. *Int J Epidemiol*. 2019;48(6):1815-1823. doi:10.1093/ije/dyz143

129. Dehghan M, Mente A, Zhang X, et al. Associations of fats and carbohydrate intake with cardiovascular disease and mortality in 18 countries from five continents (PURE): a prospective cohort study. *Lancet (London, England)*. 2017;390(10107):2050-2062. doi:10.1016/S0140-6736(17)32252-3

130. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD):

the TRIPOD statement. *J Clin Epidemiol*. 2015;68(2):134-143. doi:https://dx.doi.org/10.1016/j.jclinepi.2014.11.010

131. Canadian Community Health Survey, Cycle 2.2, Nutrition Focus - Food and Nutrition Surveillance - Health Canada - Canada.ca. https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs/canadian-community-health-survey-cycle-2-2-nutrition-focus-food-nutrition-surveillance-health-canada. Accessed December 2, 2018.
132. Discharge Abstract Database metadata (DAD) | CIHI. https://www.cihi.ca/en/discharge-abstract-database-metadata. Accessed February 15, 2020.
133. Sanmartin C, Decady Y, Trudeau R, et al. Linking the Canadian community health survey and the canadian mortality database: An enhanced data source for the study of mortality. *Heal Reports*. 2016;27(12):10-18.
134. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018). https://ethics-gc-ca.libaccess.lib.mcmaster.ca/eng/policy-politique_tcps2-eptc2_2018.html. Accessed February 15, 2020.
135. Tu J V., Ghali WA, Pilote L, Brien S. *Canadian Cardiovascular Atlas*.; 2006.
136. Tu J V., Chu A, Donovan LR, et al. The Cardiovascular Health in Ambulatory Care Research Team (CANHEART): Using Big Data to Measure and Improve Cardiovascular Health and Healthcare Services. *Circ Cardiovasc Qual Outcomes*. 2015;8(2):204-212. doi:10.1161/CIRCOUTCOMES.114.001416

137. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res*. 2016;18(12):e323. doi:10.2196/jmir.5870

138. Josse J, Husson F. missMDA: A package for handling missing values in multivariate data analysis. *J Stat Softw*. 2016;70(1):1-31. doi:10.18637/jss.v070.i01

139. Schiffner J, Bischl B, Lang M, et al. mlr Tutorial. *arXiv*. September 2016. http://arxiv.org/abs/1609.06146. Accessed May 3, 2020.

140. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York: Springer; 2013.

141. Strobl C, Hothorn T, Zeileis A. *Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the Party Package*. http://cran.r-project. Accessed November 24, 2019.

142. Molnar C, Casalicchio G, Bischl B. iml: An R package for Interpretable Machine Learning Software • Review • Repository • Archive. doi:10.21105/joss.00786

143. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77

144. Costanzo S, Di Castelnuovo A, Donati MB, Iacoviello L, de Gaetano G. Alcohol Consumption and Mortality in Patients With Cardiovascular Disease. A Meta-Analysis. *J Am Coll Cardiol*. 2010;55(13):1339-1347. doi:10.1016/j.jacc.2010.01.006

145. Zhu Y, Zhang J, Li Z, et al. Association of sodium intake and major cardiovascular outcomes: A dose-response meta-analysis of prospective cohort studies. *BMC*

Cardiovasc Disord. 2018;18(1). doi:10.1186/s12872-018-0927-9

146. Jayedi A, Ghomashi F, Zargar MS, Shab-Bidar S. Dietary sodium, sodium-to-potassium ratio, and risk of stroke: A systematic review and nonlinear dose-response meta-analysis. *Clin Nutr*. 2019;38(3):1092-1100. doi:10.1016/j.clnu.2018.05.017

147. Cogswell ME, Mugavero K, Bowman BA, Frieden TR. Dietary sodium and cardiovascular disease risk - Measurement matters. *N Engl J Med*. 2016;375(6):580-586. doi:10.1056/NEJMsb1607161

148. Mozaffarian D, Fahimi S, Singh GM, et al. Global sodium consumption and death from cardiovascular causes. *N Engl J Med*. 2014;371(7):624-634. doi:10.1056/NEJMoa1304127

149. O'Donnell M, Mente A, Rangarajan S, et al. Urinary Sodium and Potassium Excretion, Mortality, and Cardiovascular Events. *N Engl J Med*. 2014;371(7):612-623. doi:10.1056/NEJMoa1311889

150. Schwingshackl L, Schwedhelm C, Hoffmann G, et al. Food groups and risk of all-cause mortality: a systematic review and meta-analysis of prospective studies. *Am J Clin Nutr*. 2017;105(6):ajcn153148. doi:10.3945/ajcn.117.153148

151. Wang X, Ouyang Y, Liu J, et al. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: Systematic review and dose-response meta-analysis of prospective cohort studies. *BMJ*. 2014;349. doi:10.1136/bmj.g4490

152. Aune D, Giovannucci E, Boffetta P, 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*. 2017;46(3):1029-1056. doi:10.1093/ije/dyw319

153. Mente A, de Koning L, Shannon HS, Anand SS. A Systematic Review of the Evidence Supporting a Causal Link Between Dietary Factors and Coronary Heart Disease. *Arch Intern Med*. 2009;169(7):659. doi:10.1001/archinternmed.2009.38
154. Myung SK, Ju W, Cho B, et al. Efficacy of vitamin and antioxidant supplements in prevention of cardiovascular disease: Systematic review and meta-analysis of randomised controlled trials. *BMJ*. 2013;346(7893). doi:10.1136/bmj.f10
155. Kim J, Choi J, Kwon SY, et al. Association of Multivitamin and Mineral Supplementation and Risk of Cardiovascular Disease: A Systematic Review and Meta-Analysis. *Circ Cardiovasc Qual Outcomes*. 2018;11(7):e004224. doi:10.1161/CIRCOUTCOMES.117.004224
156. Jeon J, Park K. Dietary vitamin B6 intake associated with a decreased risk of cardiovascular disease: A prospective cohort study. *Nutrients*. 2019;11(7). doi:10.3390/nu11071484
157. Rimm EB, Willett WC, Hu FB, et al. Folate and vitamin B6 from diet and supplements in relation to risk of coronary heart disease among women. *J Am Med Assoc*. 1998;279(5):359-364. doi:10.1001/jama.279.5.359
158. Jayedi A, Zargar MS. Intake of vitamin B6, folate, and vitamin B12 and risk of coronary heart disease: a systematic review and dose-response meta-analysis of prospective cohort studies. *Crit Rev Food Sci Nutr*. 2019;59(16):2697-2707.

doi:10.1080/10408398.2018.1511967

159. Poole R, Kennedy OJ, Roderick P, Fallowfield JA, Hayes PC, Parkes J. Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes. *BMJ*. 2017;359:5024. doi:10.1136/bmj.j5024

160. Zulli A, Smith RM, Kubatka P, et al. Caffeine and cardiovascular diseases: critical review of current research. *Eur J Nutr*. 2016;55(4):1331-1343. doi:10.1007/s00394-016-1179-z

161. Grosso G, Godos J, Galvano F, Giovannucci EL. Coffee, Caffeine, and Health Outcomes: An Umbrella Review. 2017. doi:10.1146/annurev-nutr-071816

162. Zhu Y, Bo Y, Liu Y. Dietary total fat, fatty acids intake, and risk of cardiovascular disease: A dose-response meta-analysis of cohort studies. *Lipids Health Dis*. 2019;18(1):91. doi:10.1186/s12944-019-1035-2

163. Chen GC, Yang J, Eggersdorfer M, Zhang W, Qin LQ. N-3 long-chain polyunsaturated fatty acids and risk of all-cause mortality among general populations: A meta-analysis. *Sci Rep*. 2016;6(1):1-9. doi:10.1038/srep28165

164. Abdelhamid AS, Martin N, Bridges C, et al. Polyunsaturated fatty acids for the primary and secondary prevention of cardiovascular disease. *Cochrane Database Syst Rev*. 2018;2018(7). doi:10.1002/14651858.CD012345.pub2

165. Liao L zhen, Li W dong, Liu Y, Li J ping, Zhuang X dong, Liao X xue. Exploring the causal pathway from omega-6 levels to coronary heart disease: A network Mendelian randomization study. *Nutr Metab Cardiovasc Dis*. 2020;30(2):233-240. doi:10.1016/j.numecd.2019.09.013

166. Qi XX, Shen P. Associations of dietary protein intake with all-cause, cardiovascular disease, and cancer mortality: A systematic review and meta-analysis of cohort studies. *Nutr Metab Cardiovasc Dis*. 2020;30(7):1094-1105. doi:10.1016/j.numecd.2020.03.008

167. Naghshi S, Sadeghi O, Willett WC, Esmaillzadeh A. Dietary intake of total, animal, and plant proteins and risk of all cause, cardiovascular, and cancer mortality: systematic review and dose-response meta-analysis of prospective cohort studies. *BMJ*. 2020;370:m2412. doi:10.1136/bmj.m2412

168. Chu A, Foster M, Samman S. Zinc status and risk of cardiovascular diseases and type 2 diabetes mellitus—A systematic review of prospective cohort studies. *Nutrients*. 2016;8(11). doi:10.3390/nu8110707

169. Milton AH, Vashum KP, McEvoy M, et al. Prospective study of dietary zinc intake and risk of cardiovascular disease in women. *Nutrients*. 2018;10(1). doi:10.3390/nu10010038

170. de Oliveira Otto MC, Alonso A, Lee D-H, et al. Dietary Intakes of Zinc and Heme Iron from Red Meat, but Not from Other Sources, Are Associated with Greater Risk of Metabolic Syndrome and Cardiovascular Disease. *J Nutr*. 2012;142(3):526-533. doi:10.3945/jn.111.149781

171. Kant AK, Graubard BI, Atchison EA. Intakes of plain water, moisture in foods and beverages, and total water in the adult US population-nutritional, meal pattern, and body weight correlates: National Health and Nutrition Examination Surveys 1999-2006. *Am J Clin Nutr*. 2009;90(3):655-663. doi:10.3945/ajcn.2009.27749

172. Willett W. Nutritional epidemiology. *Monogr Epidemiol Biostat*. 1998.

173. Lee DS, Chiu M, Manuel DG, et al. Trends in risk factors for cardiovascular disease in Canada: temporal, socio-demographic and geographic factors. *CMAJ*. 2009;181(3-4):E55-66. doi:10.1503/cmaj.081629

174. Gundersen C, Tarasuk V, Cheng J, de Oliveira C, Kurdyak P. Food insecurity status and mortality among adults in Ontario, Canada. Sichieri R, ed. *PLoS One*. 2018;13(8):e0202642. doi:10.1371/journal.pone.0202642

175. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*. 2010;11(1):110. doi:10.1186/1471-2105-11-110

176. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008;9(1):307. doi:10.1186/1471-2105-9-307

177. Albarqouni L, Doust JA, Magliano D, Barr ELM, Shaw JE, Glasziou PP. External validation and comparison of four cardiovascular risk prediction models with data from the Australian Diabetes, Obesity and Lifestyle study. *Med J Aust*. 2019;210(4):161-167. doi:10.5694/mja2.12061

178. Damen JA, Pajouheshnia R, Heus P, et al. Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: A systematic review and meta-analysis. *BMC Med*. 2019;17(1):1-16. doi:10.1186/s12916-019-1340-7

179. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable

importance for random forests. *BMC Bioinformatics*. 2008;9(1):307. doi:10.1186/1471-2105-9-307

180. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2

181. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Popul Health Metr*. 2012;10(1):20. doi:10.1186/1478-7954-10-20

182. Statistics Canada. *CCHS Cycle 2.2 - Nutrition - General Health and 24-Hour Dietary Recall: Data Dictionary Master File (Rounded)*. Vol Wave 3.; 2008.

183. Statistics Canada. *Cycle 2.2 (2004) Nutrition: General Health File (Including Vitamin and Mineral Supplements) and 24-Hour Dietary Recall MASTER AND SHARE FILES Derived Variables Documentation*.; 2008. doi:10.1016/S1474-4422(08)70132-7

184. Davis KA, Gonzalez A, Loukine L, et al. Early experience analyzing dietary intake data from the canadian community health survey—nutrition using the national cancer institute (NCI) method. *Nutrients*. 2019;11(8). doi:10.3390/nu11081908