PERSPECTIVE

Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology

Jason D Morgenstern,[1] Laura C Rosella,[2,3] Andrew P Costa,[1] Russell J de Souza,[1,4] and Laura N Anderson[1]

[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]Vector Institute, Toronto, Ontario, Canada; and [4]Population Health Research Institute, Hamilton Health Sciences, Hamilton, Ontario, Canada

ABSTRACT

The field of nutritional epidemiology faces challenges posed by measurement error, diet as a complex exposure, and residual confounding. The objective of this perspective article is to highlight how developments in big data and machine learning can help address these challenges. New methods of collecting 24-h dietary recalls and recording diet could enable larger samples and more repeated measures to increase statistical power and measurement precision. In addition, use of machine learning to automatically classify pictures of food could become a useful complimentary method to help improve precision and validity of dietary measurements. Diet is complex due to thousands of different foods that are consumed in varying proportions, fluctuating quantities over time, and differing combinations. Current dietary pattern methods may not integrate sufficient dietary variation, and most traditional modeling approaches have limited incorporation of interactions and nonlinearity. Machine learning could help better model diet as a complex exposure with nonadditive and nonlinear associations. Last, novel big data sources could help avoid unmeasured confounding by offering more covariates, including both omics and features derived from unstructured data with machine learning methods. These opportunities notwithstanding, application of big data and machine learning must be approached cautiously to ensure quality of dietary measurements, avoid overfitting, and confirm accurate interpretations. Greater use of machine learning and big data would also require substantial investments in training, collaborations, and computing infrastructure. Overall, we propose that judicious application of big data and machine learning in nutrition science could offer new means of dietary measurement, more tools to model the complexity of diet and its relations with diseases, and additional potential ways of addressing confounding. *Adv Nutr* 2021;12:621–631.

Keywords: machine learning, big data, nutritional epidemiology, artificial intelligence, nutritional sciences, diet, nutrition, precision nutrition

Introduction

Suboptimal diet recently surpassed smoking as the leading risk factor for noncommunicable disease morbidity and mortality in the Global Burden of Disease Study (1). Therefore, ongoing efforts to improve knowledge of diet's effects on health must be a top priority in efforts to improve public health. Much progress in understanding diet has been made in the past half-century, with scientific findings in nutritional epidemiology leading to policy changes such as trans-fat bans in many countries (2). Despite much progress, challenges remain, including substantial barriers to accurate and precise measurement of diet (3, 4), appropriately modeling the complexity of diet (5, 6), and multicollinearity and residual confounding (7). The objective of this article is to review how the application of big data sets and machine learning may help address challenges in nutritional epidemiology, with a focus on measurement error, dietary complexity, confounding, disease prediction, and inferential studies. First, we define big data and machine learning with respect to nutritional epidemiology, and then we review five specific topics: measurement error, dietary complexity, confounding, disease prediction, and inferential studies (**Table 1**).

Supported by Canadian Institutes of Health Research.
Author disclosures: The authors report no conflicts of interest.
Perspective articles allow authors to take a position on a topic of current major importance or controversy in the field of nutrition. As such, these articles could include statements based on author opinions or point of view. Opinions expressed in Perspective articles are those of the author and are not attributable to the funder(s) or the sponsor(s) or the publisher, Editor, or Editorial Board of *Advances in Nutrition*. Individuals with different positions on the topic of a Perspective are invited to submit their comments in the form of a Perspectives article or in a Letter to the Editor.
Address correspondence to JDM (e-mail: morgensj@mcmaster.ca).
Abbreviations used: AHEI, (alternative) Healthy Eating Index; ASA24, Automated Self-Administered 24-hour Dietary Assessment Tool; DASH, Dietary Approaches to Stop Hypertension; MDS, Mediterranean Diet Score; PCA, principal component analysis; TMLE, targeted maximum likelihood estimation.

Big Data and Machine Learning

"Big data" refers to data sets that usually include both many observations and many variables, making the use of

TABLE 1 Summary of major potential applications of big data and machine learning to nutritional epidemiology

	Measurement error	Dietary complexity	Confounding	Disease prediction	Inferential studies
Potential big data and machine learning applications	New measurement methods Frequent repeated measures Increased statistical power Increased precision Decreased regression dilution bias	Including more complex and comprehensive dietary exposures Improved modeling of interactions and nonlinearity	New data sources could reduce unmeasured confounding Greater opportunities for use of negative controls and instrumental variables Machine learning methods applied in inferential frameworks	Improved disease predictions with greater incorporation of complex dietary exposures, nonlinearity, and interactions in predictive models	Less biased estimation of causal effects Hypothesis generation with methods for interpreting machine learning models
Limitations	Evaluation of the validity and precision of new measurement methods is in early stages Many proposed methods still rely on self-report Selection bias Investments in big data infrastructure and expertise would be needed Privacy concerns must be addressed	Limited interpretability of unsupervised and supervised machine learning methods High sample sizes needed to reliably model nonadditive and nonlinear relations	Potential for limited interpretability of machine learning–derived covariates Potential for worsening model bias and variance if there is data-driven inclusion of covariates in models	Potential for overfitting Limited interpretability of models Careful validation required to ensure reliable predictions May not enhance performance relative to traditional models (e.g., if interactions and nonlinearity are not very important)	Potential for inaccurate data-driven conclusions Interpretability of machine learning models remains limited

traditional statistical methods difficult (8). As a result, there is often a need for more flexible modeling than provided for in classical statistical analysis. The specific size of data sets 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 (9). Big data sets are also often less structured than traditionally collected data, and they may be a byproduct of something, rather than an intentionally collected sample (10). Big data has risen alongside exponential improvement and expansion of computing devices and data storage capacity. Health researchers have begun to leverage new sources of big data, from both primary and secondary sources, such as electronic health records and social media. In addition, researchers now work with big data arising from the investigation of complex biological systems such as the genome and microbiome (11).

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 (12). These approaches rely on algorithms that derive patterns from data with little human input (13). This contrasts with statistical techniques that rely more on human knowledge for verification of model assumptions and variable selection (14). Statistical techniques also emphasize a theoretical approach to hypothesis testing and uncertainty estimation, which is not common in machine learning. 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 (15). For supervised approaches, an example data set 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. In contrast, for unsupervised approaches, there are no human-supplied examples for the observations in a data set, and the algorithm searches for latent patterns or groupings (16). Subsets of unsupervised approaches include dimensionality reduction and clustering (15). An additional subfield of machine learning is feature selection, which aims to remove variables that are less relevant to outcome prediction in supervised problems (17). In health research, machine learning has been applied to the analysis of genome- and microbiome-derived data, where conventional analyses are limited by high dimensionality (17) 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 (18, 19).

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

Measurement error

Description.

Diet is a difficult exposure to measure accurately and precisely. Common methods of dietary assessment widely used in large observational studies are FFQs, 24-h dietary recalls, and biomarkers (2, 20). Each method has its own strengths and limitations and is subject to both random and systematic error to various degrees (21). These measurement methods are often complementary, such as the use of biomarkers to calibrate self-report instruments. Overall, significant progress has been made in the measurement of nutrition, but there is still substantial room for improvement (20). Some sources of error are thought to be nondifferential with respect to outcomes (2); however, this still presents significant problems. Nondifferential measurement error often leads to a diminution of nutrients' associations with health outcomes, but not always (7, 22). Nondifferential measurement error also results in a loss of statistical power (23, 24). Furthermore, nondifferential measurement error can sometimes result in exaggeration of associations between diet and disease when the assumptions of the classical model of measurement error are not met (e.g., lack of error in covariates) (25, 26). This has been highlighted as limiting reproducibility in other fields (26). Last, differential measurement error, when it exists, could compromise the internal validity of studies.

Using big data to improve 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. The ubiquity of internet-connected computers and smartphones opens many new means of active data collection. In addition, increased big data repositories such as those in consumer rewards programs and diet-tracking applications offer opportunities for the use of secondary dietary data. In many cases, these new data sources entail self-report and have many similar limitations to FFQs and 24-h dietary recalls. Their main value could stem from increased scalability and corresponding improvements in statistical power. New electronic measurement modalities may also facilitate more longitudinal, repeated dietary measurements, assuming these tools are less expensive and burdensome than traditional methods. With repeated measurements, dietary variables can be more precise and regression dilution bias can be reduced. For example, 4 repeated 24-h dietary recalls were shown to improve attenuation factors of protein for men and women from 0.32–0.40 with an FFQ to 0.40–0.50 (27). Also, more detailed dietary instruments have been found to reduce bias toward the null relative to FFQs (28–30), and these types of instruments could become more feasible using electronic measurement methods. Therefore, new means of electronic data collection could help improve statistical power by increasing sample sizes, while also potentially improving measurement precision by enabling repeated and more detailed measures.

There has been some validation of automated, electronic dietary measurement modalities. For example, the Web-based Automated Self-Administered 24-Hour Dietary Assessment Tool (ASA24) captures 24-h recalls without the time and expense required by trained interviewers (19); however, there is a significant burden of collection on the respondent, who may not be willing. The ASA24 performed similarly to an interviewer-administered 24-h dietary recall, with 80% of foods classified correctly compared with 83% and no difference in bias (31). Other Web-based, self-administered 24-h dietary recalls have also been shown to have good agreement with interviewer-administered 24-h dietary recalls and other reference measures (i.e., correlation coefficients of 0.4–0.5 between Web-based recalls and biomarkers and 0.3–0.9 compared with interviewer-administered recalls) (32–34). Less user-burdensome electronic dietary measurement methods include the large and detailed grocery purchase habits of populations generated by consumer rewards programs and the eating patterns already recorded in smartphone tracking applications. Grocery purchase data have been useful for ecological studies (35). There has been less validation of purchasing data for individual-level consumption, but 1 study found that the use of household food purchase data had moderate agreement with interviewer-administered 24-h dietary recalls (concordance correlation of 0.57) and showed little bias (36). Smartphone-based dietary records have also been evaluated. For example, MyFitnessPal had correlations with measured food inventories ranging from 0.963 to 0.999 in a small study; however, other similar measurement methods had correlations with reference measurements that varied widely from 0.16 to 0.82 (37–39). Importantly, existing smartphone applications have been found to score well for usability (40). Overall, early assessments of new electronic dietary measurement methods are promising.

Completely new means of dietary measurement enabled by machine learning and modern data infrastructures could improve both scalability and precision. Machine learning models can be used to automatically classify pictures of food (41–46). Such techniques may facilitate less effortful, more regular, and more accurate diet records, improving both precision and validity. This has been an active area of research, showing rapid improvement (47), but advances in both the algorithms used and the size of fully annotated training data sets are needed to improve performance (45, 48, 49). Most published research focuses on categorizing individual foods, achieving accuracies ranging from 50% to 90%. More recently, a larger data set with full recipe information was used to identify multiple foods and ingredients in a single photograph with 65% accuracy, which was similar to human-level performance (50). Another problem is accurately estimating food volumes to obtain accurate absolute energy and nutrient estimates, with current

approaches having reported error rates of 10–50% (49). For example, 1 deep learning–based study estimated energy with an average error of 209 kcal per eating occasion (51). New ways of improving accuracy are being explored, such as using location data to correlate images with nutritional information in online restaurant menus (52) and relating home-prepared foods to online recipe databases (45, 49). If combined with grocery purchase data, accuracy of image-based food records could also be improved through restricting analyses based on known purchases. Despite many limitations with modern deep learning–based food analysis systems, private companies (44) and government health promotion programs (43) have begun to make use of them. These methods may prove most useful in combination with participants' self-report. Deep learning–based image classification could derive much of the dietary information while incorporating other contextual data and then target specific questions for clarification from the respondent. Alternatively, images and natural language relating to diet can be obtained from social media and Web search platforms, which often include integrated food and health-related information (53). Existing studies using dietary data from social media are primarily ecological and have successfully linked derived dietary features to community-level health outcomes and aspects of the built environment (45, 54). With new data collection modalities enabled by machine learning, observational studies could be rapidly scaled either passively or through dissemination of relevant applications, potentially increasing statistical power, measurement precision, and accuracy.

Limitations of new measurement methods.

New dietary measurement tools using big data infrastructures and machine learning techniques must be rigorously evaluated to determine their validity and precision. Larger cohorts facilitated by new data collection methods could improve statistical power, but only if they are sufficiently precise. Furthermore, improved statistical power alone will not alleviate regression dilution bias (7, 22). Overall, many of the same limitations faced with FFQs and 24-h dietary recalls could be expected, with improved scalability and more repeated measures being the major potential benefits. The validity of novel measurement methods is also a concern. For example, passive records from social media and grocery store purchases will likely not be comprehensive nor representative of true dietary consumption. Also, it may be some time before deep learning classification of food achieves practically useful levels of accuracy for nutritional epidemiologic studies. Initial results have been promising, but this requires much more investigation (45, 48, 49). In addition, studies using new means of both passive and active dietary measurement would need to be carefully assessed for impacts on selection bias, which would likely be exacerbated compared with usual research practice (55). Another practical limitation that could hinder application of new dietary measurement modalities, even if valid and precise, is the expertise and investment required for their development and implementation (56). Enrolling hundreds of thousands or millions of participants would require nontrivial database management expertise and infrastructure. In addition, skills in the use of machine learning, as well as access to high-performance computing, would need to be acquired in many cases. These obstacles could be overcome through collaborations and large investments in training and infrastructure, but these may not be practical for much research. Finally, many of these approaches entail major privacy concerns. Careful, collaborative work will be needed to ensure research projects involving these 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 the data.

Modeling the complexity of diet

Description.

Diet is a complex exposure, which makes defining exposure variables and specifying models challenging. Foods are not consumed in isolation but, rather, in varying combinations and proportions. If 1 item is reduced or increased, other parts of the diet must change correspondingly to meet overall energy needs. In addition, nutrients and foods can interact with one another in synergistic and antagonistic ways, making the "whole" very different from the sum of its parts (5, 6). Given this complexity, approaches to modeling diet can focus on individual nutrients, foods, food groups, or dietary patterns. Current dietary patterns are often based on a priori knowledge of important aspects of diet and condensed into 1-dimensional measures, such as the Mediterranean Diet Score (MDS) (57), (alternative) Healthy Eating Index (AHEI) (58, 59), or Dietary Approaches to Stop Hypertension (DASH) score (60). When condensed into unidimensional scores, the multidimensional character of dietary patterns is lost. These dietary patterns can account for some synergy, but only when interactions are known and accounted for during score construction. Such interactions are rarely known (5, 6, 61). Furthermore, in studies of nutrients, foods, and food groups, interactions are often implicitly assumed to be absent in model specification (61). Interactions could be included in parametric models, but only if known a priori. Finally, many nutritional epidemiologic studies assume linear models of associations between diet and disease (61). There is emerging evidence that nonlinear relations may be more common than previously thought. For example, salt (62), carbohydrate (63), and fats (64) may all have U- or J-shaped relations with cardiovascular diseases. In addition, 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 (65–67). Improper specification of models due to erroneous or incomplete exposure characterizations, assumptions regarding interactions, and/or assumptions regarding linearity can lead to masked or spurious associations and biased effect estimates (61, 68–70).

Machine learning methods to model the complexity of diet in relation to disease.

Machine learning could afford inclusion of more complex and more numerous dietary explanatory variables in nutritional epidemiologic models and help identify the most predictive ones empirically (71). Many dimensionality reduction techniques are already often used in nutritional epidemiology, such as principal component analysis (PCA) (72), *k*-means clustering (73), and partial least-squares regression (74). Except for *k*-means clustering, linear dimensionality reduction methods such as these are used in both machine learning and classical statistical analysis. However, there has been less use of nonlinear dimensionality reduction methods in nutritional epidemiology, such as autoencoders, t-distributed stochastic neighbor embedding, and manifold learning (75). Although such approaches may create more representative and comprehensive dietary patterns, they are also likely to suffer from even poorer interpretability than linear dimensionality reduction methods. Corresponding approaches are being developed to improve interpretability of the resulting dimensions (75, 76).

Feature selection methods are another means of addressing dietary complexity. These methods can restrict rich dietary data to a subset more relevant for prediction of the health outcome of interest. Again, there has been some use of these methods in nutritional epidemiology already, such as the use of least absolute shrinkage and selection operator, which was found to better predict cardiometabolic indicators with dietary data compared with traditional methods (77). Other common feature selection algorithms, such as regularized trees, genetic algorithms, and recursive feature elimination, have been used less. In addition, there has been little use of machine learning to analyze multiple levels of food classification simultaneously, such as micro- and macronutrient content, specific food types, and food groups. This could allow the most predictive aspects of diet to be determined empirically for a given problem, which was called for in a recent commentary (3). We are aware of 1 study that applied survival gradient boosted machines and survival random forests to predict cardiovascular mortality with NHANES dietary data (that included multilevel dietary data) (78). These models showed improved predictive calibration and discrimination when including all 103 dietary variables on top of traditional clinical predictors. When the only added dietary variables were a priori dietary scores (MDS, Healthy Eating Index, AHEI, and DASH), there was no performance improvement. Overall, although initial applications appear promising, the use of machine learning in the context of high-dimensional, rich dietary data is 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 nonlinear and nonadditive relations more flexibly. In addition, these relations do not need to be known a priori. Although limited, there are some studies that have applied machine learning to more flexibly model diet–health relations. 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 (79). 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 (a machine learning algorithm) to linear regression (80). This study incorporated rich dietary independent variables and used PCA for dimensionality reduction. Another recent study examined the associations between diet and adverse pregnancy outcomes using Super Learner (an ensemble machine learning algorithm) for targeted maximum likelihood estimation, compared with logistic regression (61). There were predominantly null associations in the logistic regression model. Conversely, the machine learning model demonstrated protective associations between vegetable and fruit intake and preterm births, small-for-gestational-age births, and pre-eclampsia outcomes, in addition to more precise estimates. The authors attributed this difference between the machine learning method and logistic regression to improved modeling of dietary synergy in the machine learning model. Last, the previously discussed study that used machine learning models to predict cardiovascular mortality with NHANES nutrition data showed improved predictive calibration and discrimination compared with Cox proportional hazards models (78). Interestingly, addition of nutrition data to the statistical model did not improve its predictive discrimination or calibration, but when the data were added to the machine learning models, both measures improved. This lends support to the proposition that machine learning models may better leverage the full richness of diet in modeling health outcomes, perhaps both by incorporation of more dietary variables and by accounting for nonlinear and nonadditive relations.

Residual confounding and multicollinearity

Description.

Residual confounding and multicollinearity can limit interpretability of nutritional epidemiologic studies. Nutrients and foods are often strongly correlated with one another and also with other important determinants of health (7). These dense correlations can make it difficult to ascertain the most relevant dietary exposure and to confidently address residual confounding. Residual confounding can be addressed by using a priori knowledge to specify confounders in models (81). However, it is impossible to guarantee the absence of residual confounding in observational studies. In addition, even when confounders are appropriately included in models, residual confounding can remain if measurement

error or unspecified nonadditivity/nonlinearity are present. Multicollinearity can be partially addressed through the use of dietary patterns and a posteriori dimensionality reduction methods, such as factor analysis (82). However, such methods may not encapsulate all important dietary variation, may be difficult to interpret, and may obfuscate attempts to better understand more granular aspects of diet. Overall, both residual confounding and multicollinearity can make it challenging to draw valid inferences. These issues are also strongly related to dietary measurement and model specification because both nondifferential measurement error and inappropriate modeling can exacerbate problems with multicollinearity and residual confounding (25, 83).

New tools to address confounding.

Incorporating data with both higher numbers of observations and more available variables into nutritional epidemiologic studies, alongside machine learning analytical techniques, could possibly reduce residual confounding. It is important to highlight that there has been less application of machine learning in studies providing evidence for causal effects, in which confounding is relevant (55), and that this is distinct from predictive modeling, in which machine learning has been applied more often (84). However, in addition to careful application of domain knowledge, 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 (55); and using new machine learning methods to help control for confounding with high-dimensional data, applied within causal frameworks. Big data sets including variables related to microbiota, genetics, metabolomics, lifestyle, environment, and social determinants of health could enhance analyses by helping avoid missing unmeasured confounders (85). 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 (86). Further big data types that could be considered include medical information from free-text clinical notes (87); physiological data from wearable devices (88); and populations' demographic, socioeconomic, and health records from linked government data sets (89). Another potential advantage of incorporating big data is the greater availability of negative controls. Negative control outcomes are variables expected to be related to the same confounder as the primary outcome but unrelated to the exposure of interest (55). As such, a control model can be developed to assess whether inclusion of the confounder variable removes the association between the exposure and negative control outcome, which would suggest that the confounder variable is adequate to mitigate residual confounding. High-dimensional data sets could also provide more potential instrumental variables, which can allow observational studies to mimic randomized trials under certain assumptions (90–94). Finally, machine learning methods are being developed that may help reduce residual confounding when applied appropriately (95–97). These approaches have often performed comparably to or better than solely expert-based propensity scores when appropriately validated against randomized controlled trials and in simulations (95, 97–105). Such methods can be robust to widely varying covariate sets (97). However, these methods should be used with caution due to their early stage of development and their potential for worsening model bias and variance through data-driven inclusion of mediator, collider, and/or instrumental variables in models (96). Another limitation of machine learning–assisted confounder adjustment is that the type and degree of confounder adjustment achieved may be less interpretable. Examination of the underlying variables included in partially automated methods could yield some interpretability, but this would be challenging with very high-dimensional and ensemble methods (97). 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 data sets.

Improving disease prediction

Relatively few clinical or public health prediction models include dietary data (106). Incorporation of such data into predictive models could improve health outcome predictions. Predictive models are distinct from most research in nutritional epidemiology because they generally include all variables thought to be relevant for prediction of an outcome, are more concerned with the global prediction characteristics of the model than individual exposure variables' associations (e.g., area under the receiving operator curve instead of relative risk), and are less concerned with interpretability (107). Prediction models for cardiovascular disease, one of the major focuses of the science of diet, have been extensively studied for the past 5 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 (106). 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 (108–111). Typically, prediction models have included very few dietary components (106); when these are included, greatly simplified dietary factors are typically used (e.g., only a small number of foods or nutrient ratios) (108, 112). Reasons for excluding dietary variables from current predictive models may include absence of dietary data in many commonly used data sources, difficulty in their collection, and limited or no added predictive performance (e.g., due to high measurement error, use of oversimplified dietary pattern scores, or inappropriate modeling methods). Therefore, inclusion of rich dietary data in predictive models, particularly together with the use of novel data collection and machine learning methods, could be an important and largely untapped avenue for improved performance. As discussed previously, new measurement tools could mitigate

measurement error and allow predictive models to take advantage of relatively small associations. In addition, the use of machine learning models could better leverage complex dietary exposures, nonadditive relations, and nonlinear associations to improve prediction models. A recent cohort study supports this idea because it demonstrated synergistic prediction performance improvements for cardiovascular mortality when combining rich dietary data with machine learning methods (78). A further advantage of applying the machine learning paradigm is that cross-validation makes many algorithms more resistant to the effects of multicollinearity in the context of prediction (82). 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.

Limitations of big data and machine learning in disease prediction.

Notwithstanding the potential positive impacts on predictive modeling, the application of big data and machine learning has several potential pitfalls. First, selection bias and systematic measurement error in novel data sources are a concern (55). If excluded from training data sets, vulnerable populations could be further marginalized by predictive algorithms that are inaccurate for them. In addition, 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 (15). Overfitting is error that can occur with more flexible models when they fit too closely to the limited observed data points, which may lead to worse performance on new data (i.e., fitting to noise rather than signal) (113). Their relative advantage depends on the importance of interactions and nonlinearity for a given problem. Ideally, many machine learning and statistical models should be trialed and evaluated using cross-validation for a given prediction problem. Nonlinear parametric statistical models such as fractional polynomials and restricted cubic splines should also be considered (114, 115). A related issue with most machine learning approaches is that they typically require more observations per variable to make robust predictions (15). Therefore, it may often not be appropriate to apply machine learning techniques in smaller data sets. Alternatively, the numerous feature selection and dimensionality reduction techniques in the machine learning corpus can be used, along with 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 in which only a small sample size is available and both nonlinear and nonadditive relations are not very influential. Finally, it is important to note that modeling health outcomes is distinct from the application domains in which machine learning was originally developed (116). For example, in most computer vision contexts there is a very high signal-to-noise ratio. On the other hand, in medical domains a significant proportion of prediction error likely comes from unmodifiable stochasticity, posing a lower ceiling on possible prediction accuracy. Thus, in health research, uncertainty estimates and probability predictions are more important than they often have been in machine learning. Although 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 probabilities across the full spectrum of risk (81, 107). This contrasts with the more frequent use of discriminative performance metrics such as area under the receiver operator curve in machine learning research.

Informing inferential studies

Although most machine learning and big data research has focused on prediction or classification, it could also help inform inferential studies in nutritional epidemiology. First, if successful in mitigating nondifferential measurement error and increasing sample sizes, new dietary measurement methods could aid detection of smaller effect sizes and reduce the effects of multicollinearity on coefficient stability (7, 25). Furthermore, 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 relations between predictors and outcomes, as well as important interactions (117). In addition, 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 (118). 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. This would not always be advisable, because hypothesis-driven studies would require a much more selective analysis, but could be a useful approach

for exploratory studies. 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 television).

Machine learning could also enhance observational studies that search for evidence of causal relations in nutritional epidemiology, within a potential outcomes framework. New ways of using machine learning to partially automate generation of propensity scores and select confounders from high-dimensional data have already been described (95, 96). In addition, 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 (119). In concert with Super Learner, TMLE has demonstrated less biased estimation of causal effects than traditional approaches. The primary difference is the use of the machine learning ensemble during a secondary targeting phase to better balance the bias–variance tradeoff in estimation of the causal effect (120). As described previously, an initial application of TMLE in nutritional epidemiology found relations between fruit and vegetable intake and pregnancy outcomes that were not uncovered by logistic regression (61). The TMLE effect estimates were also more precise.

Limitations of big data and machine learning for inferential studies.

Although big data and machine learning may be helpful for informing inferential 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 exist that could explain observed relations (120). Therefore, experts' domain knowledge is essential for informing a priori causal models, interpreting results generated by algorithms, and putting findings into the wider evidence context. In particular, although 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 (96). Further issues when using big data and machine learning to inform evidence for causal relations are selection bias and systematic measurement error. Both must be better understood to ensure valid and generalizable results. Last, 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.

Conclusions

Overall, greater use of big data and machine learning could help improve the reliability and validity of nutritional epidemiologic findings. 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 relations between diet and disease. In turn, such improvements could help improve both predictions and inferences regarding the relations between diet and disease. With increased use of big data and machine learning, some challenges facing nutritional epidemiology could potentially be addressed.

Acknowledgments

JDM acknowledges support from the Canadian Institutes of Health Research through the Canada Graduate Scholarships–Masters Award. The authors' responsibilities were as follows—JDM and LNA: conceptualized the study; JDM and LNA: wrote the manuscript; and all authors: reviewed and edited the manuscript and read and approved the final manuscript.

References

1. GBD 2017 Diet Collaborators; Sur PJ, Fay KA, Cornaby L, Ferrara G, Salama JS, Mullany EC, Abate KH, Abbafati C, Abebe Z, et al. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2019;393:1958–72.
2. Satija A, Yu E, Willett WC, Hu FB. Understanding nutritional epidemiology and its role in policy. Adv Nutr 2015;6:5–18.
3. Ioannidis JPA. Unreformed nutritional epidemiology: a lamp post in the dark forest. Eur J Epidemiol 2019;34:327–31.
4. Giovannucci E. Nutritional epidemiology: forest, trees and leaves. Eur J Epidemiol 2019;34:319–25.
5. Krishnan S, Ramyaa R. When two heads are better than one: nutritional epidemiology meets machine learning. Am J Clin Nutr 2020; 111:1124–6.
6. Reedy J, Subar AF, George SM, Krebs-Smith SM. Extending methods in dietary patterns research. Nutrients 2018; 10:571.
7. Trepanowski JF, Ioannidis JPA. Perspective: limiting dependence on nonrandomized studies and improving randomized trials in human nutrition research: why and how. Adv Nutr 2018;9:367–77.
8. Snijders CCP, Matzat U, Reips UD. "Big data" : big gaps of knowledge in the field of internet science. Int J Internet Sci 2012;7:1–5 .
9. Lacey D. 3D data management: controlling data volume, velocity and variety. META Gr Res Note 2001;6:1.
10. Dedić N, Stanier C. Towards differentiating business intelligence, big data, data analytics and knowledge discovery. In: Piazolo F, Geist V, Brehm L, Schmidt R, editors. Innovations in enterprise information systems management and engineering. Cham (Switzerland) : Springer; 2017. pp. 114–22.
11. Shukla SK, Murali NS, Brilliant MH. Personalized medicine going precise: from genomics to microbiomics. Trends Mol Med 2015;21:461–2.
12. Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 1959;3:210–29.
13. Sra, S, Nowozin, S, Wright, SJ. Optimization for machine learning. MIT Press; 2012.
14. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods 2018;15:233–4.

15. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
16. Friedman JH. Data mining and statistics: what is the connection? Computing science and statistics 1998;29:3–9.
17. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, Wright AF, Wilson JF, Agakov F, Navarro P, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci Rep 2015;5 :10312.
18. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Health 2018;39:95–112.
19. Health P, Bi MQ, Goodman KE. What is machine learning: a primer for the epidemiologist. Am J Epidemiol 2019;188:2222–39.
20. Hu FB, Willett WC. Current and future landscape of nutritional epidemiologic research. JAMA 2018;320:2073–4.
21. Naska A, Lagiou A, Lagiou P. Dietary assessment methods in epidemiological research: current state of the art and future prospects. F1000Res 2017;6:926 .
22. Ioannidis JPA. The challenge of reforming nutritional epidemiologic research. JAMA 2018;320:969.
23. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. J Natl Cancer Inst 2011;103:1086–92.
24. Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, Sharbaugh CO, Trabulsi J, Runswick S, Ballard-Barbash R, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. Am J Epidemiol 2003;158:1–13.
25. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol 2007;166:646–55.
26. Loken E, Gelman A. Measurement error and the replication crisis. Science 2017;355:584–5.
27. Schatzkin A, Kipnis V, Carroll RJ, Midthune D, Subar AF, Bingham S, Schoeller DA, Troiano RP, Freedman LS. A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. Int J Epidemiol 2003;32:1054–62.
28. Bingham SA, Luben R, Welch A, Wareham N, Khaw KT, Day N. Are imprecise methods obscuring a relation between fat and breast cancer? Lancet North Am Ed 2003;362:212–4.
29. Dahm CC, Keogh RH, Spencer EA, Greenwood DC, Key TJ, Fentiman IS, Shipley MJ, Brunner EJ, Cade JE, Burley VJ, et al. Dietary fiber and colorectal cancer risk: a nested case–control study using food diaries. J Natl Cancer Inst 2010;102:614–26.
30. Freedman LS, Potischman N, Kipnis V, Midthune D, Schatzkin A, Thompson FE, Troiano RP, Prentice R, Patterson R, Carroll R, et al. A comparison of two dietary instruments for evaluating the fat–breast cancer relationship. Int J Epidemiol 2006;35:1011–21.
31. Kirkpatrick SI, Subar AF, Douglass D, Zimmerman TP, Thompson FE, Kahle LL, George SM, Dodd KW, Potischman N. Performance of the automated self-administered 24-hour recall relative to a measure of true intakes and to an interviewer-administered 24-h recall. Am J Clin Nutr 2014;100:233–40.
32. Timon CM, Van Den Barg R, Blain RJ, Kehoe L, Evans K, Walton J, Flynn A, Gibney ER. A review of the design and validation of web- and computer-based 24-h dietary recall tools. Nutr Res Rev 2016;29:268–80.
33. Wark PA, Hardie LJ, Frost GS, Alwan NA, Carter M, Elliott P, Ford HE, Hancock N, Morris MA, Mulla UZ, et al. Validity of an online 24-h recall tool (myfood24) for dietary assessment in population studies: comparison with biomarkers and standard interviews. BMC Med 2018;16:1–14.
34. Greenwood DC, Hardie LJ, Frost GS, Alwan NA, Bradbury KE, Carter M, Elliott P, Evans CEL, Ford HE, Hancock N, et al. Validation of the Oxford WebQ online 24-hour dietary questionnaire using biomarkers. Pract Epidemiol 2019;188:11858–67.
35. Bandy L, Adhikari V, Jebb S, Rayner M. The use of commercial food purchase data for public health nutrition research: a systematic review. PLoS One 2019;14:e0210192.
36. Appelhans BM, French SA, Tangney CC, Powell LM, Wang Y. To what extent do food purchases reflect shoppers' diet quality and nutrient intake? Int J Behav Nutr Phys Act 2017;14.
37. Pendergast FJ, Ridgers ND, Worsley A, McNaughton SA. Evaluation of a smartphone food diary application using objectively measured energy expenditure. Int J Behav Nutr Phys Act 2017;14: 1–10.
38. Wellard-Cole L, Chen J, Davies A, Wong A, Huynh S, Rangan A, Allman-Farinelli M. Relative validity of the Eat and Track (EaT) smartphone app for collection of dietary intake data in 18-to-30-year olds. Nutrients 2019;11:621.
39. Recio-Rodriguez JI, Rodriguez-Martin C, Gonzalez-Sanchez J, Rodriguez-Sanchez E, Martin-Borras C, Martínez-Vizcaino V, Arietaleanizbeaskoa MS, Magdalena-Gonzalez O, Fernandez-Alonso C, Maderuelo-Fernandez JA, et al. EVIDENT smartphone app, a new method for the dietary record: comparison with a food frequency questionnaire. JMIR Mhealth Uhealth 2019;7:e11463.
40. Ferrara G, Kim J, Lin S, Hua J, Seto E. A focused review of smartphone diet-tracking apps: usability, functionality, coherence with behavior change theory, and comparative validity of nutrient intake and energy estimates. JMIR Mhealth Uhealth 2019;7:e9232.
41. Hoi AS. Food image recognition by deep learning. [Internet]. 2019. [Cited 2021 Jan 28]. Available from: http://images.nvidia.com/content/APAC/events/ai-conference/resource/ai-for-research/FoodAI-Food-Image-Recognition-with-Deep-Learning.pdf.
42. Leapfrog. Image-based calorie estimation using deep learning. [Internet]. [Cited 2019 Nov 4]. Available from: https://www.lftechnology.com/blog/ai/image-calorie-estimation-deep-learning.
43. Sahoo D, Hao W, Ke S, Xiongwei W, Le H, Achananuparp P, Lim E-P, Hoi SCH. FoodAI: food image recognition via deep learning for smart food logging. [Internet]. 2019. [Cited 2019 Nov 4]. Available from: http://arxiv.org/abs/1909.11946.
44. Dillet R. Foodvisor automatically tracks what you eat using deep learning. [Internet]. TechCrunch; 2019. [Cited 2019 Nov 4]. Available from: https://techcrunch.com/2019/10/14/foodvisor-automatically-tracks-what-you-eat-using-deep-learning.
45. Min W, Jiang S, Liu L, Rui Y, Jain R. A survey on food computing. [Internet]. 2018. [Cited 2020 Feb 25]. Available from: http://arxiv.org/abs/1808.07202.
46. 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: 2016 International Conference on Fuzzy Theory and Its Applications, iFuzzy 2016. New York: Institute of Electrical and Electronics Engineers; 2017.
47. Boushey CJ, Spoden M, Zhu FM, Delp EJ, Kerr DA. New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. Proc Nutr Soc 2017;76: 283–94.
48. Alshurafa N, Lin AW, Zhu F, Ghaffari R, Hester J, Delp E, Rogers J, Spring B. Counting bites with bits: expert workshop addressing calorie and macronutrient intake monitoring. J Med Int Res 2019; 21:e14904.
49. Lo FPW, Sun Y, Qiu J, Lo B. Image-based food classification and volume estimation for dietary assessment: a review. IEEE J Biomed Health Inform 2020; 24:1926–39.
50. Marin J, Biswas A, Ofli F, Hynes N, Salvador A, Aytar Y, Weber I, Torralba A. Recipe1M+: a dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE Trans Pattern Anal Mach Intel 2019 [Internet]. [Cited 2020 Aug 23]. Available from: http://pic2recipe.csail.mit.edu.
51. Fang S, Shao Z, Kerr DA, Boushey CJ, Zhu F. An end-to-end image-based automatic food energy estimation technique based on learned energy distribution images: protocol and methodology. Nutrients 2019;11:877.
52. Myers A, Johnston N, Rathod V, Korattikara A, Gorban A, Silberman N, Guadarrama S, Papandreou G, Huang J, Murphy K. Im2Calories:

towards an automated mobile vision food diary. Proc IEEE Int Conf Comput Vis IEEE 2015;2015:1233–41.
53. Nguyen QC, Li D, Meng H-W, Kath S, Nsoesie E, Li F, Wen M. Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. JMIR Public Health Surveill 2016;2:e158.
54. Shah N, Srivastava G, Savage DW, Mago V. Assessing Canadians' health activity and nutritional habits through social media. Front Public Health 2020;7:400.
55. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Health 2018;39:95–112.
56. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KSL, Myles P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368: I627.
57. Ocké MC. Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis. Proc Nutr Soc 2013;72:191–9.
58. McCullough ML, Feskanich D, Stampfer MJ, Giovannucci EL, Rimm EB, Hu FB, Spiegelman D, Hunter DJ, Colditz GA, Willett WC. Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance. Am J Clin Nutr 2002;76:1261–71.
59. National Cancer Institute. Overview & background of the Healthy Eating Index. [Internet]. [Cited 2020 Jan 25]. Available from: https://epi.grants.cancer.gov/hei.
60. Miller PE, Cross AJ, Subar AF, Krebs-Smith SM, Park Y, Powell-Wiley T, Hollenbeck A, Reedy J. Comparison of 4 established DASH diet indexes: examining associations of index scores and colorectal cancer. Am J Clin Nutr 2013;98:794–803.
61. Bodnar LM, Cartus AR, Kirkpatrick SI, Himes KP, Kennedy EH, Simhan HN, Grobman WA, Duffy JY, Silver RM, Parry S, et al. Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. Am J Clin Nutr 2020;111:1235–43.
62. Kong YW, Baqar S, Jerums G, Ekinci EI. Sodium and its role in cardiovascular disease—the debate continues. Front Endocrinol 2016;7: 164.
63. Dehghan M, Mente A, Zhang X, Swaminathan S, Li W, Mohan V, Iqbal R, Kumar R, Wentzel-Viljoen E, Rosengren A, 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 North Am Ed 2017;390:2050–62.
64. Mente A, Yusuf S. Evolving evidence about diet and health. Lancet Public Health 2018;3:e408–9.
65. Koliaki C, Katsilambros N. Dietary sodium, potassium, and alcohol: key players in the pathophysiology, prevention, and treatment of human hypertension. Nutr Rev 2013;71:402–11.
66. Brown IJ, Stamler J, Van Horn L, Robertson CE, Chan Q, Dyer AR, Huang C-C, Rodriguez BL, Zhao L, Daviglus ML, et al. Sugar-sweetened beverage, sugar intake of individuals, and their blood pressure: International Study of Macro/Micronutrients and Blood Pressure. Hypertension 2011;57:695–701.
67. Kotchen TA, Kotchen JM. Dietary sodium and blood pressure: interactions with other nutrients. Am J Clin Nutr 1997;65: 708S–11S.
68. García-Magariños M, López-De-Ullibarri I, Cao R, Salas A. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP–SNP interaction. Ann Hum Genet 2009;73:360–9.
69. Grömping U. Variable importance assessment in regression: linear regression versus random forest. Am Stat 2009;63:308–19.
70. Yang P, Hwa Yang Y, Zhou BB, Zomaya A. A review of ensemble methods in bioinformatics. Curr Bioinform 2010;5:296–308.
71. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol 2009;24:733–6.
72. Kastorini C-MM, Papadakis G, Milionis HJ, Kalantzi K, Puddu P-EE, Nikolaou V, Vemmos KN, Goudevenos JA, Panagiotakos DB. 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;59:175–83.
73. 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;77:1417–25.
74. Melaku YA, Gill TK, Taylor AW, Adams R, Shi Z. A comparison of principal component analysis, partial least-squares and reduced-rank regressions in the identification of dietary patterns associated with bone mass in ageing Australians. Eur J Nutr 2018;57:1969–83.
75. Hosseini B, Hammer B. Interpretable discriminative dimensionality reduction and feature selection on the manifold. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet]. New York (NY): Springer; 2020. pp. 310–26. [Cited 2020 Aug 23]. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0585-8.
76. Tian TS, James GM. Interpretable dimension reduction for classifying functional data. Comput Stat Data Anal 2013;57:282–96.
77. Zhang F, Tapera TM, Gou J. Application of a new dietary pattern analysis method in nutritional epidemiology. BMC Med Res Methodol 2018;18:119.
78. 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:e032703.
79. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, et al. Personalized nutrition by prediction of glycemic responses. Cell 2015;163:1079–94.
80. Panaretos D, Koloverou E, Dimopoulos AC, Kouli GM, Vamvakari M, Tzavelas G, Pitsavos C, Panagiotakos DB. 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;120:326–34.
81. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. New York (NY) Springer; 2015.
82. Garg A, Tai K. Comparison of regression analysis, artificial neural network and genetic programming in handling the multicollinearity problem. In: Proceedings of the 2012 International Conference on Modelling, Identification and Control, ICMIC 2012. New York (NY): IEEE; 2012. pp. 353–8.
83. Grewal R, Cote JA, Baumgartner H. Multicollinearity and measurement error in structural equation models: implications for theory testing. Mark Sci 2004;23:519–29.
84. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. CHANCE 2019;32:42–9.
85. Olstad DL, McIntyre L. Reconceptualising precision public health. BMJ Open 2019;9:e030279.
86. 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:e181535.
87. Lynch KE, Whitcomb BW, DuVall SL. How confounder strength can affect allocation of resources in electronic health records. Perspect Heal Inf Manag 2018;15.
88. 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: 144–50.
89. 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 Health 2009;100:180–3.
90. Hernán M, Robins J. Causal inference: what if. Boca Raton (FL): Chapman & Hall; 2020.
91. Lleras-Muney A. The relationship between education and adult mortality in the United States. Rev Econ Studies 2005;72:189–221.
92. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010;21:383–8.

93. Link BG, Phelan J. Social conditions as fundamental causes of disease. J Health Soc Behav 1995;35:80–94.
94. 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;27:637–41.
95. 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;5:179–92.
96. 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: 97–115.
97. 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;20:512–22.
98. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statist Med 2010;29:337–46.
99. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004;9:403–25.
100. Wyss R, Ellis AR, Brookhart MA, Girman CJ, Funk MJ, Locasale R, Stürmer T. 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;180:645–55.
101. 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. Statist Med 2013;32:3388–414.
102. 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;63:826–33.
103. 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;20:849–57.
104. 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;69:549–57.
105. 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;173:1404–13.
106. Damen J, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016; 353:i2416.
107. Steyerberg EW. Clinical prediction models. New York (NY): Springer; 2009.
108. Manuel DG, Tuna M, Bennett C, Hennessy D, Rosella L, Sanmartin C, Tu J V, Perez R, Fisher S, Taljaard M. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the Cardiovascular Disease Population Risk Tool (CVDPoRT). CMAJ 2018;190:E871–82.
109. Fisher S, Hsu A, Mojaverian N, Taljaard M, Huyer G, Manuel DG, Tanuseputro P. Dementia Population Risk Tool (DemPoRT): study protocol for a predictive algorithm assessing dementia risk in the community. BMJ Open 2017;7: e018018.
110. 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;2:19.
111. 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). J Epidemiol Community Health 2011;65:613–20 .
112. Joseph P, Yusuf S, Lee SF, Ibrahim Q, Teo K, Rangarajan S, Gupta R, Rosengren A, Lear SA, Avezum A, et al. Prognostic validation of a non-laboratory and a laboratory based cardiovascular disease risk score in multiple regions of the world. Heart 2018;104:581–7.
113. Hastie T, Tibshirani R, Witten D, Gareth J. An introduction to statistical learning: with applications in R. New York (NY): Springer; 2013.
114. Harre FE, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 1988;80:1198–202.
115. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat 1994;43:429.
116. Statistical Thinking. Road map for choosing between statistical modeling and machine learning. [Internet]. [Cited 2019 Nov 4]. Available from: https://www.fharrell.com/post/stat-ml.
117. Molnar C. Interpretable machine learning: a guide for making black box models explainable . [Internet]. Leanpub; 2019. [Cited 2020 Aug 23]. Available from: https://christophm.github.io/interpretable-ml-book.
118. 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;9:1226.
119. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. Am J Epidemiol 2017;185:65–73.
120. Stephan KE, Penny WD, Moran RJ, den Ouden HEM, Daunizeau J, Friston KJ. Ten simple rules for dynamic causal modeling. Neuroimage 2010;49:3099–109.