THE ROLE OF ADVANCED COMPUTATION, PREDICTIVE TECHNOLOGIES, AND BIG DATA ANALYTICS RELATED TO FOOD AND NUTRITION RESEARCH: A WORKSHOP

Key Papers

Tab 1	Cote, M., and B. Lamarche. 2021. Artificial intelligence in nutrition research: Perspectives on current and future applications. *Appl Physiol Nutr Metab*:1-8.
Tab 2	Das, S. K., A. J. Miki, C. M. Blanchard, E. Sazonov, C. H. Gilhooly, S. Dey, C. B. Wolk, C. S. H. Khoo, J. O. Hill, and R. P. Shook. 2022. Perspective: Opportunities and challenges of technology tools in dietary and activity assessment: Bridging stakeholder viewpoints. *Adv Nutr* 13(1):1-15.
Tab 3	Ghanbari, R., Y. Li, W. Pathmasiri, S. McRitchie, A. Etemadi, J. D. Pollock, H. Poustchi, A. Rahimi-Movaghar, M. Amin-Esmaeili, G. Roshandel, A. Shayanrad, B. Abaei, R. Malekzadeh, and S. C. J. Sumner. 2021. Metabolomics reveals biomarkers of opioid use disorder. *Transl Psychiatry* 11(1):103.
Tab 4	Gichoya, J. W., I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L. C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S. C. Huang, P. C. Kuo, M. P. Lungren, L. J. Palmer, B. J. Price, S. Purkayastha, A. T. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R. Wang, Z. Zaiman, and H. Zhang. 2022. AI recognition of patient race in medical imaging: A modelling study. *Lancet Digit Health* 4(6):e406-e414.
Tab 5	Karagas, M. R., S. McRitchie, A. G. Hoen, C. Takigawa, B. Jackson, E. R. Baker, J. Madan, S. J. Sumner, and W. Pathmasiri. 2022. Alterations in microbial-associated fecal metabolites in relation to arsenic exposure among infants. *Expo Health* 14(4):941-949.
Tab 6	Kirk, D., E. Kok, M. Tufano, B. Tekinerdogan, E. J. M. Feskens, and G. Camps. 2022. Machine learning in nutrition research. *Adv Nutr* 13(6):2573-2589.
Tab 7	Lindquist, J., D. M. Thomas, D. Turner, J. Blankenship, and T. K. Kyle. 2021. Food for thought: A natural language processing analysis of the 2020 Dietary Guidelines public comments. *Am J Clin Nutr* 114(2):713-720.
Tab 8	Robinson, W. R., A. Renson, and A. I. Naimi. 2020. Teaching yourself about structural racism will improve your machine learning. *Biostatistics* 21(2):339-344.
Tab 9	Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206-215.
Tab 10	Thomas, D. M., S. Kleinberg, A. W. Brown, M. Crow, N. D. Bastian, N. Reisweber, R. Lasater, T. Kendall, P. Shafto, R. Blaine, S. Smith, D. Ruiz, C. Morrell, and N. Clark. 2022. Machine learning modeling practices to support the principles of AI and ethics in nutrition research. *Nutr Diabetes* 12(1):48.

Canadian Science Publishing

REVIEW

Artificial intelligence in nutrition research: perspectives on current and future applications

Mélina Côté and Benoît Lamarche

Abstract: Artificial intelligence (AI) is a rapidly evolving area that offers unparalleled opportunities of progress and applications in many healthcare fields. In this review, we provide an overview of the main and latest applications of AI in nutrition research and identify gaps to address to potentialize this emerging field. AI algorithms may help better understand and predict the complex and non-linear interactions between nutrition-related data and health outcomes, particularly when large amounts of data need to be structured and integrated, such as in metabolomics. AI-based approaches, including image recognition, may also improve dietary assessment by maximizing efficiency and addressing systematic and random errors associated with self-reported measurements of dietary intakes. Finally, AI applications can extract, structure and analyze large amounts of data from social media platforms to better understand dietary behaviours and perceptions among the population. In summary, AI-based approaches will likely improve and advance nutrition research as well as help explore new applications. However, further research is needed to identify areas where AI does deliver added value compared with traditional approaches, and other areas where AI is simply not likely to advance the field.

Novelty:

- Artificial intelligence offers unparalleled opportunities of progress and applications in nutrition.
- There remain gaps to address to potentialize this emerging field.

Key words: artificial intelligence, machine learning, algorithms, nutrition, prediction, dietary assessment, metabolomics, social media.

Résumé : L'intelligence artificielle (« IA ») est un domaine qui évolue rapidement et qui offre des possibilités de progrès inégalées dans de nombreux domaines de la santé. Cette courte revue propose un aperçu des principales applications de l'IA pour la recherche en nutrition et identifie les lacunes dans ce domaine émergent. L'IA permet d'une part de mieux comprendre les interactions complexes et non linéaires entre les données nutritionnelles et le risque de maladies, en particulier dans un contexte de données massives, comme dans le domaine de la métabolomique. L'IA, incluant les approches basées sur la reconnaissance d'images, promet également d'améliorer l'évaluation des apports alimentaires en maximisant l'efficacité et en atténuant les erreurs systématiques et aléatoires associées aux questionnaires alimentaires auto-administrés. Enfin, l'IA permet de recueillir, structurer et analyser de grandes quantités de données tirées des plateformes de médias sociaux afin de mieux comprendre les comportements et les perceptions alimentaires au sein de diverses populations. En résumé, les approches basées sur l'IA sont susceptibles d'améliorer et de faire progresser la recherche en nutrition. Cependant, davantage d'études sont nécessaires afin d'identifier les domaines où l'IA apporte une valeur ajoutée réelle par rapport aux approches traditionnelles et d'autres domaines où ce n'est pas le cas.

Les nouveautés :

- L'intelligence artificielle présente des occasions de progrès inégalés en nutrition.
- Il reste des lacunes à combler dans ce domaine émergent.

Mots-clés : intelligence artificielle, apprentissage automatique, algorithmes, nutrition, prédiction, évaluation alimentaire, métabolomique, médias sociaux.

Introduction

Artificial intelligence (AI) is a rapidly evolving field that has offered unparalleled opportunities of progress in many healthcare fields, like precision medicine, radiology, genetics and molecular medicine (Davenport and Kalakota 2019; Hamet and Tremblay 2017; Mintz and Brodie 2019). AI has the advantage of harnessing voluminous datasets, such as Electronic Health Records, omics data and longitudinal data from large cohorts (Goldstein et al. 2016; Mehta and Devarakonda 2018; van der Ploeg et al. 2014; Wiens and Shenoy 2018). Nutrition research is among the healthcare fields that are increasingly benefiting from these new computational techniques, notably because of the important amount and complexity of data generated in this field. Among others, AI-based methods have enhanced health outcome prediction in the context of various dietary exposures and have contributed to the improvement

Received 17 June 2021. Accepted 9 September 2021.

M. Côté and B. Lamarche. Centre de recherche Nutrition, santé et société (NUTRISS), INAF, Université Laval, Québec, QC, Canada; School of Nutrition, Université Laval, Québec, QC, Canada.

Corresponding author: Benoît Lamarche (email: benoit.lamarche@fsaa.ulaval.ca).

and development of dietary assessment tools (Morgenstern et al. 2021). AI-based methods are used in several "omics"-related nutrition research as well as in processing and analyzing social media information. This article provides a brief overview of such current and future applications of AI in nutrition research and addresses research gaps. The main AI techniques discussed in this review are machine learning (ML), deep learning (DL) and natural language processing (NLP).

Machine leaning

The field of ML, which is considered an important area of AI, shares several similarities with traditional statistical methods. However, while traditional statistical methods are generally more focused on inference due to its probabilistic nature, ML is more oriented towards prediction and classification by using learning algorithms trained in large datasets (Bzdok et al. 2018). Another difference between both approaches is that traditional statistical models are developed based on a priori knowledge of associations among a set of variables while ML algorithms assume that the data are not generated by any particular mechanism (Lavigne et al. 2019). ML can also deal in a much more efficient way with very large and complex datasets than traditional statistics. Popular ML techniques include supervised learning and unsupervised learning.

Supervised learning algorithms are provided with the labeled output results so they can *learn* from the data to find relationships among variables and improve predictions (Alloghani 2020; Lavigne et al. 2019). Predictive performances such as accuracy of various algorithms can then be compared when new data points are presented. Supervised learning algorithms require a lot of data to adequately train and test the algorithms. Main applications of supervised learning are for prediction purposes, such as regression or classification. Popular supervised learning algorithms discussed in this review include decision trees (DT), random forests (RF), support vector machines (SVM) and k-nearest neighbour (KNN). DT make predictions by learning a set of decision rules based on the structure of the data (Song and Lu 2015). Following a flow-like (or tree) structure, each node of a decision tree represents an input variable, each branch represents a decision rule, and each leaf represents a prediction or classification result (Song and Lu 2015). RF is a more sophisticated classification algorithm that replicates (or bootstraps) samples of the data to generate a multitude of decision trees and determines a predicted class by averaging the estimated prediction of each tree (Zhang and Ma 2012). SVM sort data into 2 or more classes of the outcome of interest using multi-dimensional hyperplanes derived from linear, polynomial, radial basis or sigmoid functions (Howley and Madden 2005). KNN is an algorithm that assumes that close (or neighbor) data points in a complex dataset share similar features (Cunningham and Delany 2022). Therefore, a new data point is classified according to the similarities of a predetermined number of closest points in the dataset.

In unsupervised learning, no labels are provided to the algorithms. These algorithms rather search for natural or hidden patterns and relations within the data (Alloghani 2020; Lavigne et al. 2019). Algorithms cannot formally be compared, as model performance cannot be measured. Therefore, unsupervised learning is mostly useful in exploratory analyses. Main applications of unsupervised learning include clustering, feature extraction and visualization. An example of unsupervised learning includes the *k*-means clustering, which finds a fixed number (*k*) of clusters that group similar data points together to identify underlying patterns (Likas et al. 2003).

Deep learning

DL is a subsection of ML that processes several data types and structures multiple times before achieving the output of interest. The main DL algorithms used are artificial neural networks with multiple hidden layers (convolutional neural networks, recurrent neural networks, etc.), which aim to imitate the human brain (Dongare et al. 2012). In other words, the extra depth, compared with traditional ML algorithms, allows carrying out more complex tasks. Broadly, an input dataset is fed to a first layer of multiple algorithms. The outputs (prediction, classification) of the first layer are then fed to a second layer of algorithms. This can be repeated multiple times to reach a final output, such as for example a predicted health status. Main applications of DL algorithms include NLP, information retrieval and image recognition (Deng 2014).

Natural language processing

NLP aims to achieve human-like language processing (Liddy 2001). It is applied among others for speech recognition, text analysis and translation. More specifically, NLP can analyze written or spoken texts to paraphrase, translate, answer questions or identify context or meaning (Liddy 2001). It is particularly useful to extract and structure information from social media, medical files or articles from blogs, among others (Lavigne et al. 2019).

Diet-related health outcome prediction

ML is particularly powerful to analyze multi-dimensional data and to uncover nonlinear relationships and interactions among variables of interest, with the potential to achieve a more in-depth, accurate and sophisticated prediction of diet-related health outcomes (Morgenstern et al. 2021). For example, processing of dietary patterns or nutrient and food intake data using RF, SVM and KNN algorithms have been shown to predict malnutrition among children under 5 as well as early risk factors of developing overweight/obesity later in life among prematurely born children (Talukder and Ahammed 2020; Fu et al. 2020). In some instances, ML algorithms have outperformed traditional predictive and classification statistical models. For example, KNN and RF algorithms outperformed traditional statistical models to predict long-term cardiometabolic risk based on dietary patterns (Panaretos et al. 2018). Similarly, "feeding" nutrition-related data to RF algorithms improved cardiovascular mortality risk prediction compared with standard Cox models (Rigdon and Basu 2019). Another study compared logistic regression with ML algorithms, including a neural network, DT, SVM and KNN, to predict the risk of depression using United States and South Korean population-based data (Oh et al. 2019). The neural network identified participants with depression with the highest accuracy among the ML algorithms and performed significantly better than logistic regression in the South Korean dataset. A large proportion of ML applications in nutrition is oriented towards prediction purposes, because ML alone cannot be used for inferential purposes (Hernán et al. 2019). However, ML can assist in better processing data by analyzing a larger number of variables with more complex relationships and do not require restrictive assumptions (Schuler and Rose 2017). Studies have demonstrated that nonparametric methods, such as ML algorithms, can be used along with doubly robust estimation techniques such as targeted maximum likelihood estimation (TMLE) for estimating causal effects (Schuler and Rose 2017; Naimi and Balzer 2018). For example, TMLE paired with the Super Learner ML algorithm has revealed protective associations between diets high in fruits and vegetables and the risk of adverse pregnancy outcomes that logistic regression models did not reveal (Bodnar et al. 2020). The Super Learner is an ML algorithm that uses cross-validation to first estimate the performance of multiple preselected classification or regression models and then uses the optimal weighted combination of these models to make a prediction (Naimi and Balzer 2018). It must be noted that these studies concern more traditional and less complex statistical models, i.e., simpler, linear models. Other, more complex, statistical models, like cubic splines, may take in

account non-linear associations and perform equally or better than ML algorithms and thus should also be compared with ML algorithms.

Other AI techniques, such as unsupervised ML, can also be used to investigate the intricate associations between diet and health. For example, *k*-means clustering algorithms were used to identify risk factors for low muscle mass based on nutritional and health-related factors among men and women (Kwon et al. 2020). The algorithm generated 5 clusters for men and women separately based on age, total energy, carbohydrate ratio, protein ratio, fat ratio, smoking habits, alcohol consumption, physical activity and a number of chronic diseases, yielding similar characteristics among each cluster. Logistic regression was then used to analyze the associations between each of the 9 variables and low muscle mass index, hence identifying risk factors within each cluster. For example, key characteristics of 2 clusters associated with an increased risk of low muscle mass among men were old age, low energy intake, high carbohydrate intake, low protein and fat intake, low alcohol consumption, less physical activity and a high number of chronic diseases. Clustering algorithms may therefore help identify phenotypes or risk factors related to distinct exposure-disease relationships and overcome some of the limitations of conventionally used models, such as better representing complex data structures (Kwon et al. 2020).

Finally, AI-based approaches are particularly powerful when applied to the field of precision nutrition, as summarized in a recent systematic literature review (Kirk et al. 2021). For example, a study demonstrated that a boosted DT-based algorithm integrating data from blood samples, dietary intake, anthropometry, physical activity and gut microbiota composition was superior to traditional glucose monitoring approaches in predicting postprandial glycemic responses to real-life meals (Zeevi et al. 2015). Data also showed that individualized dietary interventions based on these predictions significantly improved postprandial glucose management through consistent alterations to the gut microbiota (Zeevi et al. 2015). Berry et al. (2020) demonstrated promising applications of RF regression models to predict the postprandial triglyceride, glucose and C-peptide response to food intake based on meal composition, habitual diet, meal context, anthropometry, genetics, microbiome and clinical and biochemical parameters. Hall et al. (2018) used an unsupervised clustering algorithm to identify different types of patterns of elevations in postprandial glucose response, "glucotypes". The temporal profiles of blood glucose responses of 57 participants with different diagnoses of diabetes were clustered into 3 "glucotypes" representing low, moderate and severe variability in glycemic responses. Further research may allow a more accurate identification of patients among whom high glucose variability influences the risk of developing diabetes or cardiovascular diseases. Using information from 96 Single Nucleotide Polymorphisms (SNPs) related to type 2 diabetes as well as sex, body mass index and age among a sample of 677 healthy or diabetic participants, Lopez et al. (2018) have shown that a RF algorithm performed better than logistic regression for classifying the diabetes status. Other studies in precision nutrition focused on predicting body weight or risk of obesity. Ramyaa et al. (2019) have used supervised regression algorithms, including SVM, neural networks and KNN, as well as unsupervised algorithms, including *k*-means clustering, to predict weight or body mass index using self-reported dietary intake and physical activity data. For instance, a KNN algorithm could predict body weight with a mean approximate error of 6.98 kg. When performing *k*-means clustering and then using KNN to predict body weight within the cluster, the algorithm performed better with a mean approximate error of 1.1 kg. Similarly, Babajide et al. (2020) tested different ML algorithms, including SVM, RF and ANN, to predict body weight at the end of 10-week dietary intervention program based on anthropometry, body composition, metabolic rate and dietary intake data. The RF model performed best with the lowest error rate and highest R-square value of 96%. Curbelo Montañez et al. (2017) have used a RF-based feature selection algorithm to identify the most relevant SNPs and a SVM classifier to identify genetic variants predicting susceptibility to obesity. The SVM classifier performed better than logistic regression when using the obesity-associated SNPs identified by the RF-based algorithm.

In summary, it is being increasingly recognized that traditional predictive models may fall short in deciphering the complex interactions and nonlinear associations between diet-related data and health outcomes. AI-based approaches, particularly ML algorithms, show great potential in this field because of their capacity to better capture such complexity in predictive medicine.

Dietary assessment

Assessing dietary intake in research using self-reported methods such as 24-hour recalls, food diaries and food frequency questionnaires is highly challenging because such methods are time-consuming, subjective and prone to nontrivial systematic and random errors (Zhao et al. 2021). Morgenstern et al. (2021) have provided a thorough overview on how ML can help address challenges in nutritional epidemiology, with particular focus on measurement errors, statistical power, increasing precision and validity and reducing bias.

Non-image–based tools

The measurement of dietary intakes and behaviours using non-image–based tools may also be enhanced through the use of novel ML-based approaches. For example, a voice-based mobile nutrition monitoring system that uses NLP techniques has been developed to monitor dietary intake (Hezarjaribi et al. 2018). Once NLP techniques convert spoken words to text, an algorithm links food names to a nutrition database to estimate calorie intake. This monitoring system estimated calorie intake with an accuracy of 92.2%. Kalantarian and Sarrafzadeh (2015) used a smartwatch microphone combined with a RF algorithm to identify chews and swallows. Authors report good classification performance between food items, talking and ambient noise, with an F-measure of 94.5%, a common metric used in ML. Heydarian et al. (2019) have summarized the current evidence on the use of upper limb-mounted motion sensors for passive and objective assessment of certain eating behaviours. DT, RF, SVM and DL algorithms can be used to differentiate between eating activities (e.g., chewing, swallowing) and non-eating activities (e.g., walking, writing, talking on the phone, brushing teeth). In another study, Mertes et al. (2020) used a plate with weight sensors and a bite detection algorithm based on a RF classifier to assess eating. Out of 836 bites, the algorithm detected 602 bites, yielding an accuracy of 74%. Other studies have focused on developing algorithms that improve the accuracy of food intake measurements using data derived from existing self-reported instruments. For example, applying DT-based approaches to the UK National Diet and Nutrition Survey, Rosso and Giabbanelli (2018) have shown that public health dietary surveys can be simplified while improving accuracy in predicting adherence to 5 key dietary recommendations. Authors concluded that the use of simplified surveys to monitor public health nutrition is cost-effective while possibly attenuating biases as participation burden is reduced. Chin et al. (2019) have demonstrated the feasibility of using ML algorithms to substantially decrease time required to estimate nutrient intakes that are not automatically outputted in a web-based 24-hour recall, by combining 24-hour recall data with information from food and nutrient databases. Finally, applying ML algorithms to dietary intake data from the Global Burden of Disease study, Schmidhuber et al. (2018) have developed predictive models that estimate the consumption of each nutrient based on their national availability, with an accuracy greater than 80%. Such data are invaluable to address the nutritional needs of specific populations in the context of particular food systems.

Image-based tools

Development of new dietary assessment methods based on automated food image recognition and analysis is a fast-evolving field of research. Indeed, an increasing number of food image databases, algorithms for food recognition and classification models are being developed and validated for use on multiple devices for food intake monitoring (Vieira Resende Silva and Cui 2017). For example, the Snap-n-eat is a food recognition application based on a SVM classifier that estimates energy, food and nutrient intake using images taken directly with the user's smartphone, without additional intervention by the research team (Zhang et al. 2015). Similarly, the Keenoa food recognition application, which is linked to the Canadian Nutrient File 2015 (Deeks et al. 2017) generates a nutrient analysis that is sent directly to a dietician when a patient or participant in research takes a picture of their meal, thus facilitating the monitoring of dietary habits by a dietician or a research team (Ji et al. 2020). The use of artificial neural networks (ANN), particularly convolutional neural networks (CNN), is also prominent in food recognition algorithms (Vieira Resende Silva and Cui 2017). Merchant and Pande (2019) have developed a smartphone application based on a CNN for food recognition to provide nutritional assessment and suggest food recipes to diabetic patients. The algorithm recognized food with an accuracy of 70%. Fang et al. (2019) have developed a CNN to estimate total energy intake based on a single image of a meal, yielding predictions with an average error of 209 kilocalories. Since such algorithms are training-based, capturing more images with distinct food types and sizes will help generate more accurate predictions. Another study has shown that a trained ANN can reconstruct a 3D point cloud of a single food image taken in depth on a smartphone and predict its volume with an accuracy of up to 93% (Lo et al. 2018). Other studies have developed CNN algorithms to recognize foods and drinks among voluminous datasets of images. For example, CNN algorithms were able to recognize 11 food groups among thousands of images with an accuracy of up to 82% (Gözde Özsert and Özyildirim 2018). Quite similarly, Jia et al. (2019) have developed a CNN algorithm that differentiates food items from non-food items with an accuracy of 86.4% using images passively taken by a wearable camera among free-living individuals and Liu et al. (2016) developed CNN algorithms trained using food images taken by smartphones that recognized different foods with an accuracy of up to 77%.

The performance of ML algorithms to recognize and identify food on images is improving rapidly due to the voluminous and increasing number of food image datasets available (Tahir and Chu 2021). Indeed, the training-based nature of ML algorithms implies that the use of more images of distinct food types and sizes will inevitably improve their performance. Popular food datasets used to develop such algorithms include the Food-101 dataset (11 categories, 101 000 images (Bossard et al. 2014)), the Food-5k (2 categories, 5000 images (Singla et al. 2016)), the Food-11 (11 categories, 16 643 images (Singla et al. 2016)), the UEC-FOOD-100 (100 categories, 14 361 images (Matsuda et al. 2012)) and the UEC-FOOD-256 (256 categories, 25 088 images (Kawano and Yanai 2015)). Other studies are developing databases by extracting images from publicly available data, like Mezgec and Korousic Seljak (2017), who generated a dataset of 225 953 images of food and drinks from the Internet, or Rich et al. (2016) who generated a dataset with 808 864 food-related images from Instagram.

These examples are just a glimpse at the numerous AI applications that exist for collecting nutritional data and assessing food intake (Kirk et al. 2021; Lo et al. 2020, Tahir and Chu 2021). Multiple ML, NLP and DL algorithms are being developed to assess food intake more accurately, by relying more on objectively recorded food intake and behaviours, thus potentially attenuating biases known to be associated with self-reported dietary intake data as well as decreasing participation burden (Zhao et al. 2021; Morgenstern et al. 2021). However, conventional dietary assessment tools and AI-based tools are to date complementary methods because most new AI-based tools cannot yet fully replace conventional tools (Zhao et al. 2021; Morgenstern et al. 2021).

Nutritional metabolomics and biomarkers

Principal component analysis, multivariate data analysis and partial least-squares discriminant analysis are among the most common models used to analyze metabolomics data. However, AI and related algorithms, such as ANN, RF and SVM, are particularly well-tailored to decipher the nonlinear nature of complex associations found in large nutritional metabolomics datasets. More specifically, AI techniques combined with metabolomics are used to identify biomarkers of foods or nutrients. For example, Shinn et al. (2021) have developed a RF algorithm to identify fecal bacteria biomarkers of 6 specific foods. The overall classification accuracy of the 6 foods was 70% using 22 fecal bacteria biomarkers. KNN, SVM and RF algorithms have been developed to predict total antioxidant properties of food matrices based on datasets of flavonoid food content, which may ultimately reveal new roles of these bioactive molecules on health and disease (Guardado Yordi et al. 2019). In the sports nutrition realm, ML has been employed to investigate the impact of nutrient intake on biomarkers of hydration in cyclists (Munoz et al. 2020). Since multicollinearity between variables can increase errors in traditional statistical modelling, thus augmenting the chance of observing null associations, researchers used *k*-means clustering to first identify clusters representing 1 or more nutrients that may serve as mediators of body water status in cyclists.

AI-based methods applied to large metabolomics datasets may help identify biomarkers of health outcomes. For instance, SVM, a supervised ML algorithm, and Random Walks, an unsupervised ML algorithm, were developed to predict anti-cancer molecules within different foods (Veselkov et al. 2019). The researchers were able to show that specific plant-based foods contained molecules with anti-cancer properties that were similar to those seen with existing cancer drugs on molecular networks. The relationship between food addiction and the brain-gut-microbiome axis has been explored using metabolomics and brain imaging among female participants with obesity (Dong et al. 2020). Researchers were able to accurately classify participants according to criteria related to food addiction using a RF algorithm that processed fecal metabolites and brain imaging data. Algorithms like RF, SVM and ANN have also been used to process mass spectrometry-based metabolomics, through selection and processing of peaks, normalizing, imputing and interpreting data, for biomarker detection, classification or regression (Liebal et al. 2020).

Finally, AI-related approaches are widely used to facilitate analysis and interpretation of multiple omics datasets (Khorraminezhad et al. 2020; Liebal et al. 2020; Mendez et al. 2020). For example, Perakakis et al. (2019) applied SVM, KNN and RF algorithms to metabolomics, lipidomics and glycomics data to predict non-alcoholic steatohepatitis (NASH) and non-alcoholic fatty liver (NAFL). The SVM algorithm classified participants as healthy, with NASH or with NAFL with an accuracy of 90%. Hence, there is no doubt that advanced AI-based methods are revolutionizing our capacity to take advantage of the multidimensional nature of datasets available in the nutritional "omics" fields.

Social media content analysis

In social media lies exceptional amounts of data for health research, including a breadth of nutrition-related information to which large segments of the population are exposed to and influenced by. AI-based applications, like NLP and ML, can gather, structure and analyze information from a variety of social media platforms to monitor and better understand nutritional behaviours and perceptions among populations of interest. Models have been developed to analyze the textual and image content of

social media, like Twitter and Reddit, to assess content and perceptions regarding weight loss, emotional eating, diet, diabetes, obesity or exercise (Karami et al. 2018; Hwang et al. 2020, Liu and Yin 2020; Shaw and Karami 2017). For example, NLP topic modelling and clustering algorithms have been used to process and analyze posts on *r/loseit*, an online community built around weight management and weight loss (Liu and Yin 2020). More classical linear regression models were then used to investigate associations between the content of the posts and the user's weight loss. Another study trained ML classification algorithms, including DT, SVM, and KNN, to detect emotional eating posts in the *r/loseit* community (Hwang et al. 2020). NLP was then used to categorize the emotional eating posts in 4 main topics: addressing feelings, sharing physical changes, sharing or asking for dietary information and sharing dietary strategies (Hwang et al. 2020). Karami et al. (2018) analyzed 4.5 million Twitter posts mentioning obesity (51.7%), diet (23.7%), exercise (16.6%) or diabetes (8%) and identified related topics and correlations. For instance, sub-topics for obesity, apart from diet, exercise and diabetes, included Alzheimer, cancer and children. Sub-topics for diet, apart from obesity, exercise and diabetes, included vegetarian, pregnancy, celebrities, weight loss, religious and mental health. Strong correlations existed between topics related to exercise and obesity, while notable correlations were found between topics related to diet and obesity as well as diabetes and obesity.

AI-based methods are also being developed to better understand the spatial pattern of nutrition-related information found in social media. For example, NLP and geocoding of social media posts have provided information on the real-time distribution of nutritional behaviours, habits and health outcomes, which may help in turn identify populations at risk (Ghosh and Guha 2013; Nguyen et al. 2016; Shah et al. 2019; De Choudhury et al. 2016; Cesare et al. 2019; Widener and Li 2014). Such information may be proven to be of great value from a public health perspective. For instance, NLP and ML algorithms have been developed to collect and analyze data from Twitter to assess Canadian's health and nutritional habits (Shah et al. 2019). Using NLP to analyze social media content has the advantage of being less time consuming than traditional dietary assessment methods and allow quick access to information and data collected in real time (Shah et al. 2019). The developed model classified food and non-food posts with an accuracy of 93% and provided information such as approximate caloric ratio (caloric intake vs energy expenditure) of Twitter posts per province as well as foods and activities most tweeted about per province in Canada. Another study estimated the quality of available foods in different geographical locations using data from 3 million food-related posts shared on Instagram to better understand healthy food availability (De Choudhury et al. 2016). Notably, Instagram posts made by people located in food deserts were higher in fat, cholesterol and sugar intake and lower in protein and fiber. Nguyen et al. (2016) used geotagged Twitter posts to create a neighbourhood database with indicators of well-being and health behaviours, to better understand the effects of neighbourhood on health. For example, the study found that social and economic disadvantage, high urbanicity and a higher density of fast-food restaurants was associated with lower happiness and fewer healthy behaviours. A study also used an NLP framework to geolocate social media posts to map out availability of healthy and unhealthy foods (Widener and Li 2014). They demonstrated among others that there was a higher number of unhealthy food-related Twitter posts in disadvantaged areas with low access to healthy food stores.

In sum, social media has the potential to contribute to time-dependent and geospatial surveillance of physical activity, dietary habits and a multitude of other topics. Such information, along with more conventional data on dietary habits, may eventually become an integral part of the evidence considered to craft public health policy and initiatives.

Research gaps

Studies have shown that AI-based approaches do not always outperform conventionally used classification and prediction models. For example, many studies have reported no advantage of using AI in disease and health outcome prediction compared with using traditional statistical models (Christodoulou et al. 2019; Gravesteijn et al. 2020; Lynam et al. 2020; Nusinovici et al. 2020; Kuhle et al. 2018). For outcome variables that can be measured with less signal to noise ratio, like survival rate or readmission rate in hospitals, ML algorithms have been shown to perform better than traditional statistical models (Feng et al. 2019; Mortazavi et al. 2016). However, for outcome variables that have more signal to noise ratio, such as predicting risk of major chronic diseases, depression or prognosticating traumatic brain injury, ML algorithms have not always outperformed conventionally used models (Nusinovici et al. 2020; Gravesteijn et al. 2020; Oh et al. 2019). ML algorithms may also have little to no impact on improving performance when applied to smaller sample sizes and/or used with a small number of variables, because the advantage of added model complexity is not necessary in these cases. There are other examples where ML algorithms did not outperform traditional statistical models. In a study that compared classification performance of 8 different linear and ML approaches on 10 clinical metabolomics datasets, linear classification models performed similarly to ML algorithms, like SVM and ANN, in most datasets, because the outcome was linearly separable (Mendez et al. 2019*b*). Therefore, classification performance depends not only highly on the structure and the amount of data, but also on the need for more complex algorithms (Mendez et al. 2019*b*). The contradictions regarding the best models for disease and health outcome prediction suggest that AI-based modelling should always be compared with traditional statistical models to identify for which types of variables and for which outcomes traditional statistical models or AI algorithms are better suited (Goldstein et al. 2016; Liebal et al. 2020). This is particularly the case in nutrition where there is only a paucity of studies having compared AI-based and traditional approaches.

Another limitation is the fact that datasets, which are necessary to improve performance of AI algorithms, are not always available. As indicated above, voluminous datasets are available for image and social media analysis, as well as in metabolomics analysis. However, the same cannot be said for precision nutrition and other nutrition-related disease and health prediction, which may rely on smaller sample sizes with fewer variables. Generating voluminous datasets for precision nutrition and nutrition-related predictions is inevitable to improve performance of algorithms.

Furthermore, the challenge of interpretability of the output from AI algorithms remains a significant limitation in many instances, including when trying to identify sets of metabolites that predict a given outcome in nutritional metabolomics studies (Mendez et al. 2019*a*, 2020; Sen et al. 2021). Interpretability relates to the understanding of the model's design, i.e., how the model works, and explainability relates to the understanding of what the model is saying, i.e., being able to understand and explain its prediction output (Marcinkevics and Vogt 2020). Like interpretability, explainability is also a limitation of AI, especially in the realm of health where legal, medical and ethical issues must be considered (Amann et al. 2020). Inherent interpretability and explainability must be a priority beyond performance and error rates, especially for algorithms labeled as "black-box models". Therefore, one must adopt a critical stance towards AI when developing and validating algorithms for specific data structures and outcome variables. Such posture applies fully to the nutrition research field.

While AI-based dietary assessment tools may address certain challenges and limitations associated with the use of conventional

dietary assessment instruments such as 24-hour recalls, food diaries and food frequency questionnaires, they are not exempt of important limitations. While AI-based dietary assessment tools may be less subjective to biases associated to self-reported data, this is not always the case as reactive biases remain when participants knowingly wear a sensor or take pictures of their meals (Zhao et al. 2021). For example, Ji et al. (2020) demonstrated that participants using an AI-based dietary assessment smartphone application alone under-reported food intakes compared with when a dietician could adjust the participants intake on the application. Food image analysis is also limited in its capacity to determine the nutrient content of certain foods, such as the fat percentage of dairy products. Identifying hidden or mixed foods and precisely estimating nutrients and calories from food images have also been mentioned as important challenges when using AI-based tools (Vieira Resende Silva and Cui 2017). Hence, AI-based dietary assessment tools cannot currently address all the challenges and limitations inherent to traditional tools. Using AI for dietary assessment is promising, but more research is needed to improve and validate algorithms (Zhao et al. 2021). Therefore, it may be more appropriate and even recommended for the time being at least to combine AI-based and traditional tools to improve dietary assessment and the quality of dietary intake data (Zhao et al. 2021).

Finally, many limitations and ethical concerns exist regarding the use of AI for social media analysis. Social media information comes in large amounts and concerns a wide variety of topics. However, this information can be incomplete, incorrect or biased and algorithms have inherent limitations in this regard (Lanfranchi 2017; Matheny et al. 2020). For example, people may post about what they eat in a different location from where they live, creating a geographical bias (Shah et al. 2019). Also, algorithms may not always be developed to identify food words or posts that were used in a metaphorical way (Shah et al. 2019; Widener and Li 2014). The type and number of posts can also be affected by other factors, like seasons, special events or new social media trends (Shah et al. 2019; Widener and Li 2014). More importantly, social media data are inevitably biased because of missing information concerning people who do not use the social media of interest, with notable underrepresentation of minority and low socioeconomic groups (Hwang et al. 2020; Lanfranchi 2017; Matheny et al. 2020; Safdar et al. 2020; Widener and Li 2014). This is an important concern from a public health nutrition perspective, where we wish to identify populations at risk. There have also been ethical and regulation concerns related to the use of AI to extract and use social media data. Many users may be unaware that their information is used for research because of a misunderstanding of what "publicly available data" means (Kern et al. 2016). Therefore, meticulousness is necessary to avoid training algorithms with strongly biased and inequitable data and results should always be interpreted with nuance. It may be of more interest to use social media as a tool for *taking the pulse*, for more time-dependent monitoring and for comparison purposes with other public health datasets rather than using it as the only data source (Widener and Li 2014).

Conclusion

Has AI revolutionized nutrition research? Not quite yet. Application of AI-based methods may contribute to improving predictive models of diet and disease outcomes, to better collecting, processing and understanding complex nutrition-related data, and to better monitoring of a population's nutritional status. However, several limitations and concerns regarding the use of AI in nutrition research emphasize the importance of further research to develop and identify the algorithms best suited to nutrition data. The remaining gaps between the promises and the true enhancements of research through AI reinforces the importance of deploying AI-based approaches with caution and realism, with emphasis on legal and regulatory issues as well as on prioritization of equity, inclusion, and a human rights lens for this work while addressing implicit and explicit biases (Matheny et al. 2020).

Conflict of interest statement

The authors declare there are no competing interests.

Acknowledgements

M.C. received a scholarship from the Fonds de recherche du Québec-Santé. The funding organizations were not involved in the writing of this article. We would like to thank Emma Dagenais for her contribution to this work.

References

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., and Aljaaf, A.J. 2020. A systematic review on supervised and unsupervised machine learning algorithms for data science. *In* Supervised and Unsupervised Learning for Data Science. *Edited by* M.W. Berry, A.H. Mohamed, and Y.B. Wah. Springer, Cham. pp. 3–21.

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., and Precise, Q.C. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. **20**: 310. doi:10.1186/s12911-020-01332-6. PMID:33256715.

Babajide, O., Hissam, T., Anna, P., Anatoliy, G., Astrup, A., Alfredo, M.J., et al. 2020. A machine learning approach to short-term body weight prediction in a dietary intervention program. *In* Proceedings of the 20th International Conference, Computational Science – ICCS 2020. Amsterdam, the Netherlands, 3–5 June, 2020. *Edited by* V.V. Krzhizhanovskaya, G. Závodszky, M.H. Lees, J.J. Dongarra, P.M.A. Sloot, S. Brissos, and J. Teixeira. Springer, Cham. pp. 441–455.

Berry, S.E., Valdes, A.M., Drew, D.A., Asnicar, F., Mazidi, M., Wolf, J., et al. 2020. Human postprandial responses to food and potential for precision nutrition. Nat. Med. **26**: 964–973. doi:10.1038/s41591-020-0934-0. PMID:32528151.

Bodnar, L.M., Cartus, A.R., Kirkpatrick, S.I., Himes, K.P., Kennedy, E.H., Simhan, H.N., et al. 2020. Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. Am. J. Clin. Nutr. **111**: 1235–1243. doi:10.1093/ajcn/nqaa027. PMID:32108865.

Bossard, L., Guillaumin, M., and Van Gool, L. 2014. Food-101 – mining discriminative components with random forests. *In* Proceedings of the 13th Conference, Computer Vision – ECCV 2014, Zurich, Switzerland, 6–12 September, 2014. *Edited by* D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Springer, Cham. pp. 446–461.

Bzdok, D., Altman, N., and Krzywinski, M. 2018. Statistics versus machine learning. Nat. Methods, **15**: 233–234. doi:10.1038/nmeth.4642. PMID:30100822.

Cesare, N., Dwivedi, P., Nguyen, Q., and Nsoesie, E.O. 2019. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. Palgrave Commun. **5**: 106. doi:10.1057/s41599-019-0314-x. PMID:32661492.

Chin, E.L., Simmons, G., Bouzid, Y.Y., Kan, A., Burnett, D.J., Tagkopoulos, I., and Lemay, D.G. 2019. Nutrient estimation from 24-hour food recalls using machine learning and database mapping: a case study with lactose. Nutrients, **11**: 3045. doi:10.3390/nu11123045. PMID:31847188.

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., and VAN Calster, B. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. Clin. Epidemiol. **110**: 12–22. doi:10.1016/j.jclinepi.2019.02.004. PMID:30763612.

Cunningham, P., and Delany, S.J. 2022. k-nearest neighbour classifiers – a tutorial. ACM Computing Surveys, **54**(6): 128. doi:10.1145/3459665.

Curbelo Montañez, C.A., Fergus, P., Hussain, A., Al-Jumeily, D., Dorak, M.T., and Abdullah, R. 2017. Evaluation of phenotype classification methods for obesity using direct to consumer genetic data. *In* Proceedings of the 13th International Conference, ICIC: International Conference on Intelligent Computing, Liverpool, UK, 7–10 August, 2017. *Edited by* D.-S. Huang, K.-H. Jo, J.C. Figueroa-García. Springer, Cham. pp. 350–362.

Davenport, T., and Kalakota, R. 2019. The potential for artificial intelligence in healthcare. Future Healthc. J. **6**: 94–98. doi:10.7861/futurehosp.6-2-94. PMID:31363513.

De Choudhury, M., Sharma, S., and Kiciman, E. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. *In* CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, San Francisco, Calif., 27 February–2 March, 2016. Association for Computing Machinery, New York, NY. pp. 1157–1170. doi:10.1145/2818048.2819956.

Deeks, J., Verreault, M-F., and Cheung, W. 2017. Canadian Nutrient File (CNF): Update on Canadian food composition activities. J. Food Compost. Anal. **64**(Part 1): 43–47. doi:10.1016/j.jfca.2017.04.009.

Deng, L. 2014. Deep learning: methods and applications. Foundations and Trends in Signal Processing, **7**: 197–387. doi:10.1561/2000000039.

Dong, T.S., Mayer, E.A., Osadchiy, V., Chang, C., Katzka, W., Lagishetty, V., et al. 2020. A distinct brain-gut-microbiome profile exists for females with obesity and food addiction. Obesity, **28**: 1477–1486. doi:10.1002/oby.22870. PMID:32935533.

Dongare, A., Kharde, R.R., and Kachare, A.D. 2012. Introduction to artificial neural network. International Journal of Engineering and Innovative Technology, **2**(1): 189–194.

Fang, S., Shao, Z., Kerr, D.A., Boushey, C.J., and Zhu, F. 2019. An end-to-end image-based automatic food energy estimation technique based on learned energy distribution images: protocol and methodology. Nutrients, **11**: 877. doi:10.3390/nu11040877. PMID:31003547.

Feng, J.Z., Wang, Y., Peng, J., Sun, M.W., Zeng, J., and Jiang, H. 2019. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. J. Crit. Care, **54**: 110–116. doi:10.1016/j.jcrc.2019.08.010. PMID:31408805.

Fu, Y., Gou, W., Hu, W., Mao, Y., Tian, Y., Liang, X., et al. 2020. Integration of an interpretable machine learning algorithm to identify early life risk factors of childhood obesity among preterm infants: a prospective birth cohort. BMC Med. **18**: 184. doi:10.1186/s12916-020-01642-6. PMID:32646442.

Ghosh, D.D., and Guha, R. 2013. What are we 'tweeting' about obesity? Mapping tweets with Topic Modeling and Geographic Information System. Cartogr. Geogr. Inf. Sci. **40**: 90–102. doi:10.1080/15230406.2013.776210. PMID:25126022.

Goldstein, B.A., Navar, A.M., and Carter, R.E. 2016. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur. Heart J. **38**: ehw302. doi:10.1093/eurheartj/ehw302. PMID:27436868.

Gözde Özsert, Y.I., and Özyildirim, B.M. 2018. Comparison of convolutional neural network models for food image classification. J. Inform. Telecommun. **2**: 347–357. doi:10.1080/24751839.2018.1446236.

Gravesteijn, B.Y., Nieboer, D., Ercole, A., Lingsma, H.F., Nelson, D., van Calster, B., et al. 2020. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J. Clin. Epidemiol. **122**: 95–107. doi:10.1016/j.jclinepi.2020.03.005. PMID:32201256.

Guardado Yordi, E., Koelig, R., Matos, M.J., Perez Martinez, A., Caballero, Y., Santana, L., et al. 2019. Artificial intelligence applied to flavonoid data in food matrices. Foods, **8**: 573. doi:10.3390/foods8110573. PMID:31739559.

Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., Mclaughlin, T., and Snyder, M. 2018. Glucotypes reveal new patterns of glucose dysregulation. PLoS Biol. **16**: e2005143. doi:10.1371/journal.pbio.2005143. PMID:30040822.

Hamet, P., and Tremblay, J. 2017. Artificial intelligence in medicine. Metabolism, **69**: S36–S40. doi:10.1016/j.metabol.2017.01.011. PMID:28126242.

Hernán, M.A., Hsu, J., and Healy, B. 2019. A second chance to get causal inference right: a classification of data science tasks. Chance, **32**: 42–49. doi:10.1080/09332480.2019.1579578.

Heydarian, H., Adam, M., Burrows, T., Collins, C., and Rollo, M.E. 2019. Assessing eating behaviour using upper limb mounted motion sensors: a systematic review. Nutrients, **11**: 1168. doi:10.3390/nu11051168. PMID:31137677.

Hezarjaribi, N., Mazrouee, S., and Ghasemzadeh, H. 2018. Speech2Health: a mobile framework for monitoring dietary composition from spoken data. IEEE J. Biomed. Health Inform. **22**: 252–264. doi:10.1109/JBHI.2017.2709333. PMID:29300701.

Howley, T., and Madden, M.G. 2005. The genetic kernel support vector machine: description and evaluation. Artif. Intell. Rev. **24**: 379–395. doi:10.1007/s10462-005-9009-3.

Hwang, Y., Kim, H.J., Choi, H.J., and Lee, J. 2020. Exploring abnormal behavior patterns of online users with emotional eating behavior: topic modeling study. J. Med. Internet Res. **22**: e15700. doi:10.2196/15700. PMID:32229461.

Ji, Y., Plourde, H., Bouzo, V., Kilgour, R.D., and Cohen, T.R. 2020. Validity and usability of a smartphone image-based dietary assessment app compared to 3-day food diaries in assessing dietary intake among Canadian adults: randomized controlled trial. JMIR Mhealth Uhealth, **8**: e16953. doi:10.2196/16953. PMID:32902389.

Jia, W., Li, Y., Qu, R., Baranowski, T., Burke, L.E., Zhang, H., et al. 2019. Automatic food detection in egocentric images using artificial intelligence technology. Public Health Nutr. **22**: 1168–1179. doi:10.1017/S1368980018000538. PMID:29576027.

Kalantarian, H., and Sarrafzadeh, M. 2015. Audio-based detection and evaluation of eating behavior using the smartwatch platform. Comput. Biol. Med. **65**: 1–9. doi:10.1016/j.compbiomed.2015.07.013. PMID:26241487.

Karami, A., Dahl, A.A., Turner-Mcgrievy, G., Kharrazi, H., and Shaw, G. 2018. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. Int. J. Inf. Manage. **38**: 1–6. doi:10.1016/j.ijinfomgt.2017.08.002.

Kawano, Y., and Yanai, K. 2015. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. *In* Proceedings, Part III, Computer Vision – ECCV 2014 Workshops, Zurich, Switzerland, 6–7 and 12 September, 2014. *Edited by* L. Agapito, M.M. Bronstein, and C. Rother. Springer, Cham. pp. 3–17.

Kern, M.L., Park, G., Eichstaedt, J.C., Schwartz, H.A., Sap, M., Smith, L.K., and Ungar, L.H. 2016. Gaining insights from social media language: Methodologies and challenges. Psychol. Methods, **21**: 507–525. doi:10.1037/met0000091. PMID:27505683.

Khorraminezhad, L., Leclercq, M., Droit, A., Bilodeau, J.F., and Rudkowska, I. 2020. Statistical and machine-learning analyses in nutritional genomics studies. Nutrients, **12**: 3140. doi:10.3390/nu1210314. PMID:33066636.

Kirk, D., Catal, C., and Tekinerdogan, B. 2021. Precision nutrition: a systematic literature review. Comput. Biol. Med. **133**: 104365. doi:10.1016/j.compbiomed.2021.104365. PMID:33866251.

Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A.C., Joseph, K.S., and Allen, V.M. 2018. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC Pregnancy Childbirth, **18**: 333. doi:10.1186/s12884-018-1971-2. PMID:30111303.

Kwon, Y.J., Kim, H.S., Jung, D.H., and Kim, J.K. 2020. Cluster analysis of nutritional factors associated with low muscle mass index in middle-aged and older adults. Clin. Nutr. **39**: 3369–3376. doi:10.1016/j.clnu.2020.02.024. PMID:32192777.

Lanfranchi, V. 2017. Machine learning and social media in crisis management: agility vs ethics. *In* WiPe/CoRe Paper – T4 – Ethics Legal and Social Issues, Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management, Albi, France, 21–24 May, 2017. *Edited by* T. Comes, F. Bénaben, C. Hanachi, and M. Lauras. pp. 256–265.

Lavigne, M., Mussa, F., Creatore, M.I., Hoffman, S.J., and Buckeridge, D.L. 2019. A population health perspective on artificial intelligence. Healthc. Manage. Forum, **32**: 173–177. doi:10.1177/0840470419848428. PMID:31106580.

Liddy, E.D. 2001. Natural language processing. *In* Encyclopedia of Library and Information Science. 2nd ed. Marcel Decker, Inc., New York.

Liebal, U.W., Phan, A.N.T., Sudhakar, M., Raman, K., and Blank, L.M. 2020. Machine learning applications for mass spectrometry-based metabolomics. Metabolites, **10**: 243. doi:10.3390/metabo10060243. PMID:32545768.

Likas, A., Vlassis, N., and Verbeek, J.J. 2003. The global k-means clustering algorithm. Pattern Recognit. **36**: 451–461. doi:10.1016/S0031-3203(02)00060-2.

Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., and Ma, Y. 2016. DeepFood: deep learning-based food image recognition for computer-aided fietary asssessment. *In* Inclusive Smart Cities and Digital Health, Proceedings of the 14th International Conference on Smart Homes and Health Telematics, ICOST 2016, Wuhan, China, 25–27 May, 2016. *Edited by* C.K. Chang, L. Chiari, Y. Cao, H. Jin, M. Mokhtari, and H. Aloulou. Springer, Cham. pp. 37–48.

Liu, Y., and Yin, Z. 2020. Understanding weight loss via online discussions: content analysis of Reddit posts using topic modeling and word clustering techniques. J. Med. Internet Res. **22**: e13745. doi:10.2196/13745. PMID:32510460.

Lo, F.P., Sun, Y., Qiu, J., and Lo, B. 2018. Food volume estimation based on deep learning view synthesis from a single depth map. Nutrients, **10**: 2005. doi:10.3390/nu10122005. PMID:30567362.

Lo, F.P.W., Sun, Y., Qiu, J., and Lo, B. 2020. Image-based food classification and volume estimation for dietary assessment: a review. IEEE J. Biomed. Health Inform. **24**: 1926–1939. doi:10.1109/JBHI.2020.2987943. PMID:32365038.

Lopez, B., Torrent-Fontbona, F., Vinas, R., and Fernandez-Real, J.M. 2018. Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. Artif. Intell. Med. **85**: 43–49. doi:10.1016/j.artmed.2017.09.005. PMID:28943335.

Lynam, A.L., Dennis, J.M., Owen, K.R., Oram, R.A., Jones, A.G., Shields, B.M., and Ferrat, L.A. 2020. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. Diagn. Progn. Res. **4**: 6. doi:10.1186/s41512-020-00075-2. PMID:32607451.

Marcinkevics, R., and Vogt, J.E. 2020. Interpretability and explainability: a machine learning zoo mini-tour. arXiv:2012.01805 [cs.LG].

Matheny, M.E., Whicher, D., and Thadaney Israni, S. 2020. Artificial intelligence in health care: a report from the National Academy of Medicine. JAMA, **323**: 509–510. doi:10.1001/jama.2019.21579. PMID:31845963.

Matsuda, Y., Hoashi, H., and Yanai, K. 2012. Recognition of multiple-food images by detecting candidate regions. *In* Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, Melbourne, Victoria, Australia, 9–13 July, 2012. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ.

Mehta, N., and Devarakonda, M.V. 2018. Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey? J. Allergy Clin. Immunol. **141**: 2019–2021.e1. doi:10.1016/j.jaci.2018.02.025. PMID:29518424.

Mendez, K.M., Broadhurst, D.I., and Reinke, S.N. 2019*a*. The application of artificial neural networks in metabolomics: a historical perspective. Metabolomics, **15**: 142. doi:10.1007/s11306-019-1608-0. PMID:31628551.

Mendez, K.M., Reinke, S.N., and Broadhurst, D.I. 2019*b*. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. Metabolomics, **15**: 150. doi:10.1007/s11306-019-1612-4. PMID:31728648.

Mendez, K.M., Broadhurst, D.I., and Reinke, S.N. 2020. Migrating from partial least squares discriminant analysis to artificial neural networks: a comparison of functionally equivalent visualisation and feature contribution tools using jupyter notebooks. Metabolomics, **16**: 17. doi:10.1007/s11306-020-1640-0. PMID:31965332.

Merchant, K., and Pande, Y. 2019. ConvFood: a CNN-based food recognition mobile application for obese and diabetic patients. *In* Emerging Research in Computing, Information, Communication and Applications. *Edited by* N. Shetty, L. Patnaik, H. Nagaraj, P. Hamsavath, and N. Nalini. Springer, Singapore. pp. 493–502.

Mertes, G., Ding, L., Chen, W., Hallez, H., Jia, J., and Vanrumste, B. 2020. Measuring and localizing individual bites using a sensor augmented

plate during unrestricted eating for the aging population. IEEE J. Biomed. Health Inform. **24**: 1509–1518. doi:10.1109/JBHI.2019.2932011. PMID:31380774.

Mezgec, S., and Korousic Seljak, B. 2017. NutriNet: a deep learning food and drink image recognition system for dietary assessment. Nutrients, **9**: 657. doi:10.3390/nu9070657. PMID:28653995.

Mintz, Y., and Brodie, R. 2019. Introduction to artificial intelligence in medicine. Minim Invasive Ther. Allied Technol. **28**: 73–81. doi:10.1080/13645706.2019.1575882. PMID:30810430.

Morgenstern, J.D., Rosella, L.C., Costa, A.P., De Souza, R.J., and Anderson, L.N. 2021. Perspective: big data and machine learning could help advance nutritional epidemiology. Adv. Nutr. **12**: 621–631. doi:10.1093/advances/nmaa183. PMID:33606879.

Mortazavi, B.J., Downing, N.S., Bucholz, E.M., Dharmarajan, K., Manhapra, A., Li, S.-X., et al. 2016. Analysis of machine learning techniques for heart failure readmissions. Circ. Cardiovasc. Qual. Outcomes, **9**: 629–640. doi:10.1161/CIRCOUTCOMES.116.003039. PMID:28263938.

Munoz, C.X., Johnson, E.C., Kunces, L.J., Mckenzie, A.L., Wininger, M., Butts, C.L., et al. 2020. Impact of nutrient intake on hydration biomarkers following exercise and rehydration using a clustering-based approach. Nutrients, **12**: 1276. doi:10.3390/nu12051276. PMID:32365848.

Naimi, A.I., and Balzer, L.B. 2018. Stacked generalization: an introduction to super learning. Eur. J. Epidemiol. **33**: 459–464. doi:10.1007/s10654-018-0390-z. PMID:29637384.

Nguyen, Q.C., Li, D., Meng, H.-W., Kath, S., Nsoesie, E., Li, F., and Wen, M. 2016. Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. JMIR Public Health Surveil. **2**: e158. doi:10.2196/publichealth.5869. PMID:27751984.

Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., et al. 2020. Logistic regression was as good as machine learning for predicting major chronic diseases. J. Clin. Epidemiol. **122**: 56–69. doi:10.1016/j.jclinepi.2020.03.002. PMID:32169597.

Oh, J., Yun, K., Maoz, U., Kim, T.S., and Chae, J.H. 2019. Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm. J. Affect Disord. **257**: 623–631. doi:10.1016/j.jad.2019.06.034. PMID:31357159.

Panaretos, D., Koloverou, E., Dimopoulos, A.C., Kouli, G.M., Vamvakari, M., Tzavelas, G., et al. 2018. 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. **120**: 326–334. doi:10.1017/S0007114518001150. PMID:29789037.

Perakakis, N., Polyzos, S.A., Yazdani, A., Sala-Vila, A., Kountouras, J., Anastasilakis, A.D., and Mantzoros, C.S. 2019. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: a proof of concept study. Metabolism, **101**: 154005. doi:10.1016/j.metabol.2019.154005. PMID:31711876.

Ramyaa, R., Hosseini, O., Krishnan, G.P., and Krishnan, S. 2019. Phenotyping women based on dietary macronutrients, physical activity, and body weight using machine learning tools. Nutrients, **11**: 1681. doi:10.3390/nu11071681. PMID:31336626.

Rich, J., Haddadi, H., and Hospedales, T.M. 2016. Towards bottom-up analysis of social food. *In* DH '16: Proceedings of the 6th International Conference on Digital Health Conference, Montreal, Que., 11–13 April, 2016. Association for Computing Machinery, New York, NY. pp. 111–120. doi:10.1145/2896338.2897734.

Rigdon, J., and Basu, S. 2019. Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the USA using nationally randomly sampled data. BMJ Open, **9**: e032703. doi:10.1136/bmjopen-2019-032703. PMID:31784446.

Rosso, N., and Giabbanelli, P. 2018. Accurately inferring compliance to five major food guidelines through simplified surveys: applying data mining to the UK national diet and nutrition survey. JMIR Public Health Surveil. **4**: e56. doi:10.2196/publichealth.9536. PMID:29848474.

Safdar, N.M., Banja, J.D., and Meltzer, C.C. 2020. Ethical considerations in artificial intelligence. Eur. J. Radiol. **122**: 108768. doi:10.1016/j.ejrad.2019.108768. PMID:31786504.

Schmidhuber, J., Sur, P., Fay, K., Huntley, B., Salama, J., Lee, A., et al. 2018. The Global Nutrient Database: availability of macronutrients and micronutrients in 195 countries from 1980 to 2013. Lancet Planet. Health, **2**: e353–e368. doi:10.1016/S2542-5196(18)30170-0. PMID:30082050.

Schuler, M.S., and Rose, S. 2017. Targeted maximum likelihood estimation for causal inference in observational studies. Am. J. Epidemiol. **185**: 65–73. doi:10.1093/aje/kww165. PMID:27941068.

Sen, P., Lamichhane, S., Mathema, V.B., McGlinchey, A., Dickens, A.M., Khoomrung, S., and Orešič, M. 2021. Deep learning meets metabolomics: a methodological perspective. Brief Bioinform. **22**: 1531–1542. doi:10.1093/bib/bbaa204. PMID:32940335.

Shah, N., Srivastava, G., Savage, D.W., and Mago, V. 2019. Assessing Canadians health activity and nutritional habits through social media. Front. Public Health, **7**: 400. doi:10.3389/fpubh.2019.00400. PMID:31993412.

Shaw, G., and Karami, A. 2017. Computational content analysis of negative tweets for obesity, diet, diabetes, and exercise. Proc. Assoc. Inf. Sci. Technol. **54**: 357–365. doi:10.1002/pra2.2017.14505401039.

Shinn, L.M., Li, Y., Mansharamani, A., Auvil, L.S., Welge, M.E., Bushell, C., et al. 2021. Fecal bacteria as biomarkers for predicting food intake in healthy adults. J. Nutr. **151**: 423–433. doi:10.1093/jn/nxaa285. PMID:33021315.

Singla, A., Yuan, L., and Ebrahimi, T. 2016. Food/non-food image classification and food categorization using pre-trained GoogLeNet model. *In* MADiMa '16: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Amsterdam, the Netherlands, 16 October, 2016. Association for Computing Machinery, New York, NY. pp. 3–11. doi:10.1145/2986035.2986039.

Song, Y.Y., and Lu, Y. 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry, **27**: 130–135. doi:10.11919/j.issn.1002-0829.215044. PMID:26120265.

Tahir, G., and Chu, K.L. 2021. A review of the vision-based approaches for dietary assessment. arXiv:2106.11776 [cs.CV].

Talukder, A., and Ahammed, B. 2020. Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. Nutrition, **78**: 110861. doi:10.1016/j.nut.2020.110861. PMID:32592978.

van der Ploeg, T., Austin, P.C., and Steyerberg, E.W. 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med. Res. Methodol. **14**: 137. doi:10.1186/1471-2288-14-137. PMID:25532820.

Veselkov, K., Gonzalez, G., Aljifri, S., Galea, D., Mirnezami, R., Youssef, J., et al. 2019. HyperFoods: machine intelligent mapping of cancer-beating molecules in foods. Sci. Rep. **9**: 9237. doi:10.1038/s41598-019-45349-y. PMID:31270435.

Vieira Resende Silva, B., and Cui, J. 2017. A survey on automated food monitoring and dietary management systems. J. Health Med. Inform. **8**: 272. doi:10.4172/2157-7420.1000272. PMID:30101038.

Widener, M.J., and Li, W. 2014. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. Appl. Geogr. **54**: 189–197. doi:10.1016/j.apgeog.2014.07.017.

Wiens, J., and Shenoy, E.S. 2018. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin. Infect Dis. **66**: 149–153. doi:10.1093/cid/cix731. PMID:29020316.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. 2015. Personalized nutrition by prediction of glycemic responses. Cell, **163**: 1079–1094. doi:10.1016/j.cell.2015.11.001. PMID:26590418.

Zhang, C., and Ma, Y. 2012. Ensemble machine learning: methods and applications. Springer, New York.

Zhang, W., Yu, Q., Siddiquie, B., Divakaran, A., and Sawhney, H. 2015. Snap-n-Eat": food recognition and nutrition estimation on a smartphone. J. Diabetes Sci. Technol. **9**: 525–533. doi:10.1177/1932296815582222. PMID:25901024.

Zhao, X., Xu, X., Li, X., He, X., Yang, Y., and Zhu, S. 2021. Emerging trends of technology-based dietary assessment: a perspective study. Eur. J. Clin. Nutr. **75**: 582–587. doi:10.1038/s41430-020-00779-0. PMID:33082535.

PERSPECTIVE

Perspective: Opportunities and Challenges of Technology Tools in Dietary and Activity Assessment: Bridging Stakeholder Viewpoints

Sai Krupa Das,[1,2] Akari J Miki,[1] Caroline M Blanchard,[1] Edward Sazonov,[3] Cheryl H Gilhooly,[1,2] Sujit Dey,[4] Colton B Wolk,[1] Chor San H Khoo,[5] James O Hill,[6,7] and Robin P Shook[8,9]

[1] Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University, Boston, MA, USA; [2] Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA, USA; [3] Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, AL, USA; [4] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA; [5] Institute for the Advancement of Food and Nutrition Sciences, Washington, DC, USA; [6] Department of Nutrition Sciences, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, USA; [7] Nutrition Obesity Research Center, University of Alabama at Birmingham, Birmingham, AL, USA; [8] Center for Children's Healthy Lifestyles & Nutrition, Children's Mercy Kansas City, Kansas City, MO, USA; and [9] School of Medicine, University of Missouri-Kansas City, Kansas City, MO, USA

ABSTRACT

The science and tools of measuring energy intake and output in humans have rapidly advanced in the last decade. Engineered devices such as wearables and sensors, software applications, and Web-based tools are now ubiquitous in both research and consumer environments. The assessment of energy expenditure in particular has progressed from reliance on self-report instruments to advanced technologies requiring collaboration across multiple disciplines, from optics to accelerometry. In contrast, assessing energy intake still heavily relies on self-report mechanisms. Although these tools have improved, moving from paper-based to online reporting, considerable room for refinement remains in existing tools, and great opportunities exist for novel, transformational tools, including those using spectroscopy and chemo-sensing. This report reviews the state of the science, and the opportunities and challenges in existing and emerging technologies, from the perspectives of 3 key stakeholders: researchers, users, and developers. Each stakeholder approaches these tools with unique requirements: researchers are concerned with validity, accuracy, data detail and abundance, and ethical use; users with ease of use and privacy; and developers with high adherence and utilization, intellectual property, licensing rights, and monetization. Cross-cutting concerns include frequent updating and integration of the food and nutrient databases on which assessments rely, improving accessibility and reducing disparities in use, and maintaining reliable technical assistance. These contextual challenges are discussed in terms of opportunities and further steps in the direction of personalized health. *Adv Nutr* 2022;13:1–15.

Statement of Significance: This article is the first to discuss the status and challenges of current and emerging technology tools designed to measure individual food intake, eating behavior, and physical activity through the perspectives of 3 stakeholders: researchers, users, and developers. The objective of this work is to bring together experts to address interdisciplinary and cross-cutting issues with the shared mission of improving the measurement of energy intake and expenditure.

Keywords: dietary assessment, food apps, wearable device, physical activity, mobile health, image recognition, image-based dietary records

Introduction

The collective and cross-disciplinary contributions of scientists, engineers, software developers, and experts from multiple technical domains are beginning to arrive at what even a few decades ago was just a dream: personalized health. The fields of personalized nutrition and physical activity have broadly kept pace with other health disciplines in this regard, contributing to deeper understanding of complex, multitiered relations between food, eating behaviors, metabolic regulation, and energy balance. The future of personalized health and the next generation of nutrition and physical activity guidance rely heavily on what we can learn about individual behavior, which requires accurate assessment of these behaviors.

This article discusses the status of and ongoing challenges for current and emerging technology tools designed to measure individual food intake, eating behavior, and physical activity through the perspectives of 3 stakeholders:

researchers, users, and developers. These tools have multiple applications including monitoring outcomes in interventions that strive to alter dietary intake (1) or physical activity (2), and have the potential to transform energy metabolism research and improve health outcomes. With growing interest in determinants that influence individual variability in health outcomes, such as genetic, behavioral, and psychological differences, these tools can enable self-monitoring, allow for detailed research analysis, and provide an avenue for personalized professional recommendations. Previous review articles have summarized the current state of tools for assessing dietary intake (3–6), eating behavior (7, 8), and physical activity (6, 9, 10).

Broadly, current tools tend to be either active (requiring user input) or passive (not requiring user input). Examples include engineered devices such as wearables and sensors, mobile phone applications (apps), and Web-based tools. One promising area of emerging tools is sensor technology that aims to enable more accurate and objective measurement of dietary intake and eating behavior than self-report. These sensor-based tools generally fall into 3 categories: wearable sensors, camera-based devices, and weight scale–based devices. Wearable sensors include devices with sensors on the head or neck to detect chewing or swallowing (11–16), wrist-based inertial sensors to detect hand-to-mouth gestures as a proxy for bites (12, 17, 18), and others (19–21). Camera-based methods (21–25) use food images to recognize consumed food and estimate energy intake. Weight-scale devices are used in dining locations to continuously weigh consumed food (26–28), although eating behaviors can only be captured at the location of the instrument (29).

Multimodal sensing technology has advanced steadily, with the development of devices that have improved estimates of physical activity, energy expenditure, and sleep, and provide important contextual information. For example, for tracking activity, a multimodal sensing device may include traditional actigraphy and $\geq$1 of the following: multiple accelerometers (30), gyroscopes (31), magnetometers (31), inclinometers (32, 33), Global Positioning System (GPS) (34, 35), photovoltaic sensors (36–38), heart-rate sensors (39), wireless proximity sensors (40), galvanic skin sensors (41), and user-friendly screen displays (42). However, few if any devices on the market contain all these features, due in part to manufacturing costs, battery demands, and size limitations. Going forward, advancements will likely involve improving existing features and combining them into a single device (43), much like a commercially available smartwatch (44). Each of these tools, and others discussed in this article, present challenges and opportunities for stakeholders (**Table 1**). Underpinning most new or emerging tools are questions of user burden, validity, and privacy.

Before proceeding to specific challenges, we submit the following underlying premise: that stakeholders share the goal of accurately measuring intake and expenditure by *1*) maximizing the capture of objective data, and/or *2*) minimizing error in the capture of subjective data. For emerging tools, this generally means moving toward technology that can capture data as freely as possible from user input. Further, any new tools should reduce or minimize the burden on users and researchers (45). For researchers, tools should maximize the amount and completeness of data collected, include a reliable system of data storage and retrieval (46), and, when possible, have automated, standardized, and harmonized data coding that uses shared terminology and definitions (45). For users, tools should be simple and intuitive, provide privacy controls (47, 48), and require minimal instruction (49, 50) and time to complete assessments (46). For developers, particularly where monetization opportunities exist, satisfying the demands of researchers and users should ensure use by both groups remains high and continuous. Finally, sustained user adherence is a desirable goal for all stakeholders.

Current Status of Knowledge

State of technology tools: assessing energy intake compared with expenditure

Current physical activity tools are considerably more advanced than dietary intake tools. Although both intake and expenditure methodologies previously relied heavily on subjective data, technology for measuring expenditure has successfully integrated expertise across wide-ranging fields (e.g., optics, electromechanical engineering, inferential statistics) and has advanced in nearly all necessary technical and nontechnical domains, from complex algorithms that can differentiate between psychological or physical stressors (51), to the aesthetic elegance of wearable devices. Meanwhile, intake methodologies still overwhelmingly rely on digital adaptations of paper-based instruments of self-reported intake, including diaries, records, and image-based approaches.

Assessing intake may be more complex than assessing expenditure because intake is a question of measuring not just behavior, but also endless heterogeneous origins, preparations, and combinations of foods. Further, even if people were able to perfectly describe the foods they ate, they would not be able to report their nutritional qualities. There are significant opportunities for advancing the development of intake technologies that, similarly to expenditure-measuring technologies, make use of a wide range of scientific fields to better capture both food intake and eating behavior.

Editorial work for the manuscript was supported by the Institute for the Advancement of Food and Nutrition Sciences (IAFNS; Washington, DC). IAFNS is a nonprofit science organization that pools funding from industry and advances science through in-kind and financial contributions from the public and private sectors. The opinions expressed herein are those of the authors. Author disclosures: CHK is a senior science fellow consultant at the IAFNS. All other 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 SKD (e-mail: sai.das@tufts.edu).

Abbreviations used: CORE, Connected and Open Research Ethics; DLW, doubly labeled water; EMA, Ecological Momentary Assessment; GPS, Global Positioning System; IRB, institutional review board; NIR, near-infrared.

TABLE 1 Summary of opportunities and challenges of emerging technologies in dietary assessment and energy expenditure

	Researcher challenges	User challenges	Developer challenges
Dietary intake and eating behavior	• Upgrade self-reported dietary intake assessments • Enhance portion size estimation • Simplify and maximize food lists • Validate dietary assessment tools • Consider reactivity in self-monitoring	• Streamline user interface of dietary assessments • Increase convenience and decrease burden of dietary assessment • Improve wearability, comfort, and acceptability	• Capitalize on opportunities to improve existing tools • Improve image-based methods of assessment (active image capture, passive image capture and multimodal sensing, segmentation, food recognition, portion size estimation, food-image databases, automatic image analysis) • Preserve privacy in public • Create new tools
Energy expenditure and physical activity	• Improve energy expenditure assessment • Standardize validation studies of activity trackers • Incorporate novel analysis techniques for activity tracker data	• Increase convenience and decrease burden of dietary assessment • Improve wearability, comfort, and acceptability	• Capitalize on opportunities to improve existing tools • Preserve privacy in public
Cross-stakeholder challenges	• Simplify dashboards and enhance communication with users • Enhance user and researcher output of dietary intake • Improve accessibility and reduce disparity • Build motivation to encourage long-term use • Technical assistance • Build collaborations • Preserve user and bystander privacy while maximizing data collection • Maintain data integrity		

Researcher perspectives

Many challenges and opportunities exist for researchers to address both input-focused and output-focused needs. Here, we define input-focused needs as those related to the quality of incoming data (e.g., accuracy of user food intake reports, completeness of food and nutrient databases), whereas output-focused needs include generating research-ready data that are use-compatible across multiple platforms and use commonly agreed-upon terminologies, researcher and/or user dashboards, and other output, such as automated health messaging.

Input-focused needs.

Upgrade self-reported dietary intake assessments. Conventional methods of dietary assessment are interviewer-administered 24-h dietary recalls, FFQs, and dietary records, all of which are self-report (52) and subject to error via limitations of human memory, social desirability bias (52, 53), and reactivity to self-monitoring [i.e., altered energy intake on reporting days (54)]. Development of modern dietary assessment tools has focused on digital adaptations of these conventional methods [e.g., online 24-h dietary records (55), online FFQs (56)] and food-logging apps (57), which are already in widespread use. Although these tools will continue to experience self-report limitations, there is room for other improvements, especially with respect to accuracy of portion size estimation and amount of user burden.

Enhance portion size estimation. A participant's ability to estimate and remember portion sizes of consumed foods has been a large source of error in dietary assessment (58, 59) and thus is a target for improvement. Some new, common self-report methods (e.g., food-logging apps) use reference images of portion sizes to assist users with estimation (3, 50, 60). Flexibility in entering portion size is another consideration. Some software allows users to choose portion sizes from a predefined list, or enter them manually, and choose between different measurement units, such as standardized portions or household measures (3, 50, 61). Additional software improvements would allow for inclusion of dimensions and packaged food amounts (3, 50, 61) and automatic conversion of variably reported portion sizes into standard metric units for research purposes, meeting both user and researcher needs. Ideally, data will be harmonized for use across different platforms, necessarily preceded by the development of common data terminologies, to also allow for accurate comparisons of data points, such as nutrient calculations.

Promising emerging approaches use images of consumed foods, such as image-assisted dietary recall [in which images are used to assist a research participant (62, 63)], image-based dietary record [in which images document eating occasions (63, 64)], and automated image analysis (65). Advantages of such methods are reduced reliance on participant memory and direct visual documentation of eating occasions (66).

In particular, image-assisted 24-h dietary records have been shown to reduce underreporting (62).

Food images can be analyzed with manual, semiautomatic, and automatic approaches (5). Manual image analysis has the most potential for immediate application in research; however, approaches with higher levels of automation require further development. As with image-assisted methods, accurate analysis requires high-quality images (67).

Simplify and maximize food lists. A great deal of user burden in self-report software and apps derives from lists of food items. Presented lists depend on the quality of the underlying food databases and affect the accuracy of user entry and output data (68, 69). Determining the optimal length of the food list has been a challenge (61, 68). Although extensive and highly detailed lists may benefit researchers (70–72), for users, scrolling through long lists can be burdensome (68). Even so, concise food lists also may be problematic (61), even if they produce only small differences in total nutrient intake compared with extensive lists (69, 73), because users may feel frustrated when precise food items cannot be found. There is limited research on how users find the "best match" when an exact match is missing (68). Although barcode entry eases user burden, manufacturer data on which researchers subsequently rely may be incomplete. Hence, researchers must compare the benefits and limitations of different databases, as well as their effects on user-entry behavior, and the specificity of resulting data.

Validate dietary assessment tools. Several reviews have examined the validity [i.e., acceptable levels of accuracy, precision, and reliability (52)] of technology tools for measuring dietary intake (3, 45, 60, 61, 66, 74–76). A recent review (3) of technology-based tools for research, surveillance, or consumer use identified interviewer-administered 24-h dietary records, weighed portions, biomarker data, and direct observation of eating occasions as common reference/validation measures. Although most of the reviewed comparison studies showed acceptable levels of agreement between the technology tool and the traditional self-report method (within ~60 kcal), it was observed that use of validation biomarkers was lacking (3). Such comparisons can provide valuable information, but researchers should be cautious of possible correlated errors and seek validation studies that use objective measures such as doubly labeled water (DLW) or direct observation.

Improve energy expenditure assessment. As with dietary assessment, self-report via diaries or questionnaires was the most common method for measuring physical activity in research (77–79). Although such methods are inexpensive and convenient, they have poor reliability and validity compared with DLW (10). Like intake data, self-reported physical activity is affected by question misinterpretation, recall bias, and social desirability (10, 78, 79). Floor effects have been observed with unstructured or spontaneous activities (e.g., housework, gardening), resulting in failure to capture low-intensity activities (77, 79). Overestimation is another common issue (77, 78).

As noted, the use of physical activity devices has become increasingly common by consumers (80, 81), and in epidemiological (82) and intervention (79) studies. Common activity trackers include pedometers, accelerometers, and heart-rate monitors. Despite their widespread use, these devices are still somewhat limited in capturing physical activities that vary in intensity and displacement (i.e., stationary compared with mobile). Pedometers can measure only walking activity in step counts (83). Accelerometers have limited sensitivity with detecting light-intensity activities and nonambulatory activities such as cycling and weightlifting (9). The perceived relation between heart rate and energy expenditure has been the premise of using heart-rate monitors, but they have poor correlation at low and high intensities (10, 83). Beyond physical activity, energy expenditure can already be measured by direct calorimetry using existing, innovative, portable tools, such as the Personal Calorie Monitor (84). However, the ongoing challenge is to develop devices or analytic methods that can assess all types of physical activities and energy expenditure, as well as associated physiological phenomena (e.g., body temperature, perspiration, heart rate) (85). The field is already moving toward integration. Recent studies show it is possible to distinguish, using a wristband device, between simultaneous psychological and physical stressors (51). Another recent device undergoing validation is a commercial wristband containing a photoplethysmogram, accelerometer, thermometer, capacitive touch sensor, and gyroscope (86, 87).

Additional significant upgrades to existing devices and tools would also address user- and/or population-based differences in activity, which can vary by sex, age, ability, health status, and other characteristics (10). New tools should include feedback and data output that reflect user characteristics such as age, sex, body composition, fitness, and perceived exertion. In addition, given that most validation studies have been done in laboratory settings, key environmental characteristics that influence perceived exertion such as elevation, temperature, and humidity (88–90) would ideally be captured by newer devices and software, and integrated into expenditure estimation algorithms.

Standardize validation studies of activity trackers. Thus far, validation studies of activity trackers have exhibited heterogeneity in study design and activity calculations, posing challenges to comparisons. Variable aspects of study design include definition of "valid" days that are suitable for analysis [e.g., 10 h of wear time (91)], device placement [e.g., hip compared with wrist (9, 92)], and context [laboratory compared with free-living (9)]. As noted, many validation studies are conducted in the laboratory. However, pattern recognition models based on laboratory data have limited validity in free-living settings (9, 93).

In addition to study design, the devices themselves exhibit heterogeneity in sensitivity, sampling frequency, noise-separating filters, and other aspects of data capture (91). Algorithms for obtaining desired output such as steps, energy expenditure, and distance use different underlying calculations, which are further obscured by their proprietary nature and restricted sharing (9, 92). In data analysis, there is little consensus on best practices for data processing, algorithms (94), and data interpretation [e.g., the "cut-point conundrum" (95)]. Given these variables of study design and calculations, standardizing data output and validation methods is logistically difficult, and will likely require significant and ongoing collaboration between researchers and developers.

Consider reactivity in self-monitoring. Reactivity in self-monitoring—the conscious or unconscious changes in behavior as a reaction to the act of self-monitoring (54)—is a recognized phenomenon in both intake and expenditure research. For example, wearing an activity monitor may cause a participant to exercise more than usual (96), or using a food app may shift participant eating behavior away from complex dishes to mitigate the burden of logging foods (53, 68, 97). To date, few studies have examined how technology tools induce this reactivity (66). From a researcher perspective, it is beneficial to have control over the feedback or health messages a user receives from a program. The frequent desire of researchers to minimize reactivity to self-monitoring is often in direct contrast to user preferences to access and use their own health data.

Output-focused needs.

Simplify dashboards and enhance communication with users. Online 24-h dietary records, online FFQs, and food-logging apps should have a customizable dashboard for research participant management tasks such as registering new participants, updating contact information, viewing lists of usernames, and exporting files (49, 70). Such improvements need not be limited to dietary data. Integrating both real-time intake and expenditure data in a live dashboard is aspirational, and would provide researchers (and users, if appropriate) with opportunities to detect and address missing data due to technical issues or participant noncompliance (98).

Immediate communication with participants would be beneficial as well. In particular, Ecological Momentary Assessment (EMA) prompts have been shown to be successful methods of user engagement (98). EMA involves real-time measurements of behaviors and experiences of research participants in their natural settings (99). Advantages of EMA-based communication with participants include the ability to provide feedback on image or input quality and address and edit implausible or incomplete entries (45).

Enhance user and researcher output of dietary intake. Researchers also must specify the output desired from technology tools, including transformations of raw intake data. Some tools, such as the Automated Self-Administered 24-hour (ASA24) Dietary Assessment Tool, already perform automated calculations of food and nutrient intake, including food group and supplement data (61). Researchers have an important role in determining the accuracy of calculations, decisions that should not rest with developers alone (68). Updated tools should improve the accuracy of nutrient intake calculations derived from recipe functions that prompt users to enter ingredients and preparation methods (100), and include foods, food groups, food patterns, and supplement data. Further, these should be equipped to export data in multiple file formats for both users (if desired) and researchers (46). Cross-platform compatibility—or the ability to readily harmonize data across different platforms—to accurately compare the accuracy and validity of multiple inputs, and to integrate outputs, would be an ideal outcome in current and future software/platform iterations. As mentioned, such harmonization requires the development of common data terminology as well as essential metrics that can be easily translated for a variety of end-users (e.g., researchers, clinicians, users).

Incorporate novel analysis techniques for activity tracker data. As noted, many activity trackers have built-in proprietary algorithms for measuring activity counts and translating them to minutes of activity or energy expenditure. Researchers have more recently focused on machine learning to analyze activity counts, as well as raw acceleration data (94, 101). Machine-learning algorithms create a predictive model by associating patterns of raw data based on known reference activities (102), thereby addressing concerns of physical activity as a nonlinear action and heterogeneity of developer-defined activity counts. Identifying the most relevant method of machine learning for a given application is a key consideration, and may include random forest (103), artificial neural network (104), and support vector machine (105, 106) approaches, among others. Distinctions between free-living and laboratory-based activities (94, 107) and consideration of on-body location (12) will be able to further refine estimates of expenditure.

User perspectives

Streamline the user interface of dietary assessments.

Potential users of digital dietary assessments include consumers, research participants, and patients. Accordingly, developing new tools should be an iterative process that involves usability testing and improvements based on user feedback (108), which has often emphasized the importance of aesthetics, simplicity, intuitiveness, and practicality (70, 108–110). Notably, users have expressed preferences for a clean layout with no pop-ups (70) and a flat interface with a single screen for multiple recall activities such as selecting food items, recording times of meals, and specifying portion sizes (108, 111). Some users prefer a predefined list of meals or template that gives structure to the recall (108). As users make entries on the main screen, a side navigation panel with a dynamic list of entered items and options to edit them has

been shown to be helpful (108). Graphics and images, such as examples of portion sizes, also could improve aesthetics, ease of use, and data validity (109).

Increase the convenience and decrease the burden of dietary assessment.

Any new technology tool should be convenient and minimally disruptive to the user's lifestyle (112, 113). In research, investigator preferences for detailed, accurate data often conflict with user needs for convenient reporting methods (70). Users have noted difficulties with logging food intake in various situations such as commuting, at the workplace, and in social gatherings (113), and perceive the recording process as time-consuming and burdensome (113, 114). Hence, tools could have the option to customize the level of detail for dietary assessment (115), or different tools could accommodate specific needs of users (and, ostensibly, researchers).

As mentioned, tools could provide multiple options for data entry, such as image capture, text, selection from databases, and barcode scanning (63), and should be adaptable to different devices including smartphones and others (116), thereby catering to user preferences. Moreover, tools should allow users to either make entries during eating occasions or make all entries in 1 sitting, similar to a recall, although this flexibility may be problematic in research settings (109). Regardless of the data entry method, users should be able to edit entries at any time and review them before final submission (108, 109). Ultimately, features that make tools flexible and convenient help users adhere to long-term reporting of dietary intake.

Improve wearability, comfort, and acceptability.

Comfort and acceptability are important considerations for wearable devices. The ideal wearable intake or expenditure device is portable, lightweight, unobtrusive, and aesthetically pleasing (117). Examples of current intake wearables include cameras worn around the neck (62, 118), a microcamera attached to the ear (119), a badge-like miniature camera (65), and a head-mounted camera (120). Users reported discomfort with using an ear-worn microcamera (119) or neck-worn camera (62, 96) and a preference for small, inconspicuous designs (96).

Another important consideration is creating a device that can be easily worn in the correct orientation such that users' body shapes and postures do not affect data capture and quality (62, 121). Device placement is critical for activity trackers as well; the hip is the most widely used target owing to its proximity to the center of mass and ability to capture most movements. However, many people remove devices before sleeping or showering, resulting in poor compliance, and belts can move and twist throughout the day (92). Innovative "smart clothing" (122, 123)—although eminently wearable—suffers from similar limitations. Device placement on the nondominant wrist has garnered great interest because of its potential to increase compliance and total wear time (92), but wrist-worn trackers may fail to accurately capture energy expenditure of nonambulatory arm movements (44) and may not function properly in populations that use assistive devices (124). Hence, developing a device that is accurate, functional, and acceptable for daily continuous wear by diverse users is an ongoing challenge.

User comfort with devices in public and social settings is another important consideration, especially for image-capture tools. Notably, users have expressed feeling embarrassed or self-conscious taking images or videos of their meals in front of other people (113, 114), and wearable cameras often attract unwanted attention (125). Hence, when designing studies, investigators should weigh the benefits and limitations of attention-drawing tools (e.g., wearable cameras) compared with more discreet ones (e.g., apps).

Improve accessibility and reduce disparity.

Smartphone ownership is growing rapidly worldwide, but growth has been largely restricted to younger and better-educated populations, especially in emerging economies (126). Similarly, users of health apps and health-related wearables tend to be younger, more highly educated, and more affluent than nonusers, indicating possible disparities in access to these tools (127, 128). Disparities render these tools inaccessible to older adults, individuals with lower socioeconomic status, and other populations that may have low digital or eHealth literacy [defined as "the ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to addressing or solving a health problem" (129)]. It falls to developers and entities such as public institutions, nonprofit organizations, and research bodies to facilitate universal access to these tools (128, 130). A promising approach is to develop affordable tools appropriate for a wide range of reading and eHealth literacy levels (128, 131). In dietary assessment, this may entail using images of food items and portion sizes, developing educational material intended to expand nutrition knowledge, and providing assistance with interpreting results. Tools should be available in multiple languages and connected to food databases that are suited to ethnic dietary patterns. These efforts would promote equitable access and potentially support public health efforts.

A major area for improvement is accessibility for older adults. Aging is associated with changes in vision, hearing, motor function, and cognition, and many older adults have limited digital literacy (116, 132). Given these challenges, adoption rates of health apps is low among smartphone owners age 65 y and older, and downloaded health apps are shortly abandoned (133). To encourage wider adoption and long-term use, tool features should include adjustable text size and color contrast between the background, text, and images (111, 133). Buttons should be large enough for easy operation (133), and text and symbols that accompany each icon should be unambiguous and/or explicitly indicate their function, eliminating user guesswork (111). Tools should avoid using symbols that may be unfamiliar to older users with limited technology experience (111). Further, navigation structure should be consistent and simple (133),

and each recording task should minimize the number of steps toward completion (111). Feedback should be available in different modes (e.g., audio, vibrotactile, visual) (133), and the tool should generate messages and warnings to prevent errors due to unintended actions (133). Overall, where possible, tool development should follow the principles of universal design (134).

Build motivation to encourage long-term use.
User burnout, especially for recording dietary intake, is a challenge commonly observed in research settings (68, 110, 135). Tools should minimize user recording fatigue and make the experience enjoyable, incentivizing users to regularly maintain their records. An important motivation is the opportunity to set personal goals and monitor progress (115). Whereas researchers may seek to prevent reactivity to self-monitoring and restrict display metrics for specific hypotheses, users often prefer to see quantification of their health data and behaviors, and to identify opportunities for improvement (127). The process of self-quantifying behavior can boost an individual's confidence and self-efficacy (115, 127), which can be powerful motivation to continue using the tool long-term. Thus, adaptions of full quantification approaches to meet researcher needs may instead include reporting to users abbreviated measures such as adherence to a chosen dietary pattern (e.g., ketogenic or paleo diets), intake of certain nutrients (e.g., calcium, folate), or the balance of recorded dietary intake (e.g., healthy, neutral, unhealthy) (115). Tools may display health behaviors as visually appealing graphs or organized metrics, or in comparison with previous behaviors, personal goals, or peers (115).

Tools should also be interactive and engage users as much as possible. For example, when an app detects a lapse in dietary recording using EMA or similar approaches, it should remind and encourage users to make regular entries (76, 115, 135). Gamification could augment the entertainment value of tools (110, 131), and rewards such as coupons and discounts could be effective incentives (115). Further, a social network where users can share their results, discuss their concerns, and exchange advice could promote camaraderie (110, 127, 131) and motivate users to continue recording dietary intake for sustained periods.

Developer perspectives

Capitalize on opportunities to improve existing tools.
Developers should explore technology-enhanced features that further streamline the process of recording dietary intake. Multiple modes of user entry such as text entry, database browsing, voice recording, speech-to-text, and image capture (63) can decrease user burden. Allowing the user to save favorite foods, view lists of recent items, and copy entries also saves time (46).

Innovative technology including data-driven approaches, augmented reality, and portable systems can further enhance tool features. An online 24-h dietary record or food-logging app with a data-driven algorithm might make suggestions based on user intake history (46) and prompt forgotten items (136). Applications of augmented reality, such as a ruler function embedded in a smartphone camera, would be helpful for estimating portion size (137).

Integrate food, nutrient, and food-image databases.
Developers should focus efforts on maintaining continuous access of apps/software to high-quality, regularly updated food composition databases (138), including public data sets [e.g., the USDA's FoodData Central (139) and the European Food Safety Authority (EFSA) Comprehensive European Food Consumption Database (140)], licensed databases produced by research- or consumer-oriented companies, and nutrition fact labels provided by manufacturers (68). Any comprehensive database would include the most recent data on supplements, branded products, restaurant dishes, nonlabeled food items, culture-specific foods, food groups, food patterns, and product reformulations (68, 141). With new products on the market every year, updating databases remains a challenge (68, 71). The USDA Global Branded Food Products Database, a component of FoodData Central, is one such database that currently incorporates industry-provided nutrient data on labeled food items (139). Any efforts are necessarily ongoing, and should consolidate multiple sources of data, maintain a complete and comprehensive database, and standardize data coding of food intake.

Image-based methods of assessment, discussed below, require large and diverse food-image databases (142, 143). Currently, most image data sets are tailored for specific studies or types of food (142), and no publicly available, general food-image database yet exists. Some initiatives have compiled food images online (144, 145), but photos often vary in lighting, angle, and other characteristics, and may not include food volume or nutritional information (142). Going forward, an organized food-image database expanding on existing food and nutrient databases will be crucial if image-based intake assessment methodologies are to advance beyond their current nascent state.

Improve image-based methods of assessment.
Active and passive image capture. As an emerging set of methods, dietary assessment using images requires further technical refinement.

Both active and passive image capture approaches have challenges with obtaining analysis-ready, high-quality images (118, 146). The ideal methods require minimal user instruction and have high tolerance for user error. However, with active capture, a primary challenge is user burden. Users must follow specific and often demanding steps for high-quality image capture (142), e.g., place food on a brightly colored dish (147) or a container with a specific shape (24), separate food items (148), take pictures at a 45–60° angle (63), and place in the frame fiducial markers (63) of known color and dimension (5, 63, 143). In the realm of cutting-edge technology, virtual reality could eliminate the need for some of these steps, including using fiducial markers (137).

Passive image capture also presents technical and privacy challenges. This approach involves a wearable device that is in continuous operation and takes images at an adjustable rate, such as the badge-like eButton (65) or neck-worn SenseCam (62, 118). Passive capture devices can result in images of suboptimal quality especially under poorly lit conditions (62, 118), tend to require considerable amounts of power (65, 117), and have limited memory capacity (117, 121). Improved devices should thus facilitate passive capture of images under a variety of environmental conditions and more efficiently use battery power and memory. Fortunately, single-unit devices with multimodal gating mechanisms (e.g., including inertial and acoustic sensors to detect chewing sounds) hold promise for preserving battery life, maintaining privacy (117), and avoiding unnecessary data collection (119, 149).

Automatic image analysis. Once captured, images can be analyzed using manual, semiautomatic, or automatic approaches (5). In a manual approach, nutritionists calculate nutritional content from an image using the user descriptions of ingredients and portion sizes, food analysis software, and food databases (150–152). However, manual approaches require extensive user and staff training, time, and resources (63). Automatic approaches use software and classification models to segment, recognize, and calculate volumes of food, thereby reducing user input (153, 154). This strategy currently faces issues with generality, because food databases in automatic approaches are often limited in terms of the number and types of food items (63). Further, the segmentation and recognition phases rely on high-quality images where all food items are clearly visible (63). As an alternative to fully automatic approaches, semiautomatic approaches use classification software that relies on cues provided by users or researchers, such as manually identifying foods or segmenting items (98, 155). However, as with fully manual approaches, the required human input in even semiautomatic approaches may be too burdensome for practical or long-term use.

Segmentation, food recognition, and portion size estimation. After retrieval of necessary images, image analysis consists of 3 main phases: segmentation of food regions and items, extraction and recognition of food properties, and estimation of portion size (5, 117, 143). Segmentation generally uses algorithms that rely on graph-based, color, or spatial representations of the images; algorithmic techniques such as region-growing and edge (100) or circle detection (156) are often used (142, 143). Accuracy decreases as the number of unique foods increases (142), and further decreases if foods are similar in color, contour, or other characteristics (143). The selection of an appropriate segmentation algorithm depends on the types of foods, characteristics of the images, automation level, and amount and type of user input.

Compared with segmentation, food recognition is more complex. The main strategies for recognition are traditional classifiers and deep learning techniques (143). Traditional classifiers extract specific visual features from the images, such as shapes, texture, and pixel color. This approach requires the researcher to manually identify the important features of the image during development (143). This information is then organized and fed into models such as support vector machines (24, 157), Bag of Features (157, 158), and K-Nearest Neighbors (24). However, these machine-learning techniques are poor at recognizing mixed foods or foods with similar appearances (143), which may have different nutritional content (5). As alternatives to traditional classifiers, deep learning techniques could eliminate the need for user or researcher input after training/development (143, 159, 160), and have performed significantly better than traditional techniques (143, 160). However, this is an emerging approach and requires further refinement.

The final step in the analysis process is portion size estimation. In fully automated image analyses, deriving a 3-dimensional quantity from a single 2-dimensional image is a challenge (142, 143). Attempts to measure volume include generating 3-dimensional shape models based on the food type (161, 162), and using multiple pictures or short videos to reconstruct the food item (25, 162). Although these techniques appear promising, they require large amounts of processing power and time (143, 147).

Create new tools.

As noted, there is a dearth of cutting-edge technology tools to assess intake, especially relative to expenditure technologies, with most being digital adaptations of paper-based methodologies. Branches of optics, thermo-sensing, and other technologies are not exclusive to expenditure assessment tools and are currently underutilized for assessing dietary intake. There have been inconsistent advancements in these tools, but several promising ones include portable, handheld near-infrared (NIR) analysis sensors (163) and smart utensils with light spectrophotometers (164) that analyze the nutrient composition of foods. NIR is a long-standing technology in food testing and an established method for quantifying macronutrients in many types of food and agricultural products, notably for food adulteration (165, 166). NIR could move toward wide consumer use, but first it must be miniaturized and a database must be compiled of nutrient profiles for foods against which calibration training must occur (160). It is easy to imagine a future in which handheld food analysis tools integrating chemo-sensing, spectroscopy, optics, etc., become as common as wrist-worn activity trackers.

Ensure technical assistance is available.

Although intake and expenditure assessment tools should be as intuitive as possible and require minimal user training, technical assistance for users will likely always be necessary. Developers should consider tutorials and help guides to accompany apps and devices, tailored to the computer literacy of the target audience (110, 116). Effective assistance is crucial for increasing user comfort with technology and

willingness to continue user engagement as consumers or research participants (116, 167).

Ethical and Legal Considerations

Scientific interest in recording free-living individual behavior has led to rapid growth of digital health research (130) and federally funded studies on pervasive technologies (168). The ability to collect unprecedented amounts of continuous, real-time personal data has contributed to growing ethical and legal concerns (169), recently culminating in 2018 policies such as the European Union's General Data Protection Regulation (170) and the California Consumer Privacy Act (171). User privacy is a concern in research with pervasive technologies (48, 169) and, hence, technologies should comply with ethical guidelines. Researchers have found the current regulatory infrastructure and ethical guidelines to be insufficient (169), and updating them to reflect ongoing technological progress will be challenging (48). Further, standards of data security and privacy largely differ among various stakeholders such as technology companies, engineers, and scientists working with human subjects (48). Variable familiarity with novel technology or privacy risk management could also lead to variability in institutional review board (IRB) reviews of research protocols (48) and under- or overprotection of participants (47). Guidance on app development (e.g., on compliance with the Health Insurance Portability and Accountability Act) from entities such as the US Department of Health and Human Services may be an important resource (172) for responsible development of and research with new digital health tools.

Research ethics of pervasive technologies

Pertinent aspects of research ethics surrounding pervasive technologies include informed consent, participant privacy, bystander rights, and data management (48, 169). Researchers have speculated on the existence of the "privacy paradox," where users express privacy concerns while consenting to broad terms of service and wide sharing of personal information on hundreds of apps and websites (173). This purported discrepancy between stated concerns and actual behavior may suggest users' insufficient understanding of how their data are collected and their inability to protect their own interests (173), suggesting that obtaining meaningful informed consent may be difficult. Hence, the informed consent process should convey information, especially the potential risks of data breach and loss of privacy, in a way that is appropriate for the participant's technological literacy and knowledge about data usage (48).

As for participant privacy, sensitive data such as GPS coordinates and images should be unlinked from personally identifiable information and protected health information (174). Other strategies include providing the user with more control over data collection, such as the option to remove the recording device, a privacy or on-and-off switch (65, 121), and the opportunity to privately review and delete sensitive images (174, 175). Whereas the privacy of the research subject is prioritized, the status of bystander rights under regulations is ambiguous, especially regarding privacy in specific circumstances (e.g., home, workplace, public park) and the participant's responsibility to disclose use of a recording device (48). To prevent possible violations of privacy, past study protocols have instructed participants to confer with family and cohabitants before the start of a study and provided them with a procedure for responding to individuals who did not want to be recorded (174). As technology enhances the granularity of recorded data on free-living behavior, violation of participant and bystander privacy is a growing concern.

Finally, data management has its own set of challenges, and poor practices could increase the risk of data breach (48). Researchers should submit detailed protocols for maximizing data security, and IRBs should consult experts for best practices on technology, data security, and law (48, 174).

Emerging initiatives for ethical practices

There are recommended practices for obtaining informed consent, protecting participant privacy, respecting bystander rights, and maximizing data security. However, there are risks of harm to participants that remain unknown (176). Some initiatives have aimed to help researchers and IRB members navigate this uncertainty. One approach is directly asking research participants about their experiences with pervasive technologies, the extent to which the informed consent process reflected actual experiences, and their perceptions of data confidentiality (125). Another noteworthy initiative is Connected and Open Research Ethics (CORE), an interdisciplinary online community that connects researchers, ethicists, IRB affiliates, and other stakeholders of digital health research (47). CORE features a library and forum for posting questions and sharing resources such as examples of IRB protocols and informed consent forms (47). Such interdisciplinary resource-sharing efforts will promote awareness of the risks of digital health research and, ultimately, responsible and ethical practices.

Conclusions and Directions

Consumer preferences continue to drive developer enhancements to technologies designed to capture health-related data. Opportunities and challenges for researchers and developers abound. Many emerging tools rely on underlying research into technologies unrelated to consumer health behavior, such as artificial intelligence and machine learning, GPS, optics, accelerometry, or image recognition. Adapting these innovations for assessing dietary intake and energy expenditure requires ongoing collaboration between researchers and developers in the context of user acceptability.

Ever closer to personalized health

Knowledge of accurate dietary intake and energy expenditure is expected to provide insight into the etiology of illness and inform tailored preventive and treatment interventions (177, 178). The accelerated adoption of telehealth approaches due to the COVID-19 pandemic (179, 180) will make ongoing

adoption of emerging behavior technologies even more likely in clinical practice. Such tools are already beginning to be implemented by practitioners to support personalized health recommendations (181–183).

New dietary and activity assessment tools provide opportunities for real-time monitoring and guidance. For example, providers may select nutrients or food groups of clinical interest; identify and recommend the optimal amount of exercise based on a patient's age, fitness, and health status (9, 184); and choose to display specific metrics to their patients (181). Messaging features allow health care providers to give immediate feedback or answer patient questions, as well as serve as a vehicle for brief counseling sessions, which, for example, have been shown to increase physical activity in patients (177, 185). In addition, cloud-based systems allow multiple providers to access data and coordinate care (49, 68, 76, 186). However, for practitioners to meaningfully use complex and voluminous nutritional and activity data in clinical practice, they will need efficient, targeted, and clinically effective algorithms. It is not reasonable to expect that small or even large clinical practices or hospital systems will develop their own such algorithms for use with their patient populations; these will need to be generated by researchers in conjunction with developers, with the clinical guidance of expert providers.

Collaboration among stakeholders

Developing a technology tool requires interdisciplinary collaboration and effective communication between developers and other stakeholders, be they researchers or end-users. Given their different training backgrounds, and involvement at different stages of a tool's development and application, collaborators must work toward achieving at least a baseline understanding of their respective needs, limitations, and operations. For example, most developers are trained in engineering, mathematics, and/or computational sciences, and thus researchers must gain a basic understanding of a developer's vocabulary to ensure an effective cross-discipline collaboration. Conversely, because researchers work with human subjects, developers must have some understanding of research ethics involving human subjects (187). If a commercial product is used in a study, its terms of service and privacy policy may conflict with human research protections (48). Researchers are also required to support tool development with scientific evidence, such as theories of behavior change (115) or accurate calculation of nutrient intake (68).

Researchers are similarly encouraged to understand the workflow of typical device or software development processes and challenges. Researchers, whether involved in product development or validity studies, should be prepared to navigate complex legal areas, especially intellectual property and proprietary issues (187). In particular, studies on the validity of consumer activity trackers have encountered difficulties comparing algorithms and evaluating ongoing updates to software and hardware (177, 188). This becomes particularly important in long-term research studies, which should carefully plan for technology updates, product discontinuations, etc. Going forward, with the common goal of developing valid tools, researchers and developers may have to find the delicate balance between protecting ownership rights and establishing a framework for sharing open-source code.

Finally, among the many opportunities may be some obvious ones. For example, given the many devices that can now readily detect various activity types and related physiological phenomena, it would be a natural next step to assess whether these devices may be informative with respect to assessing intake and eating behavior. That is, can these ostensible "activity trackers" also be used to assess hunger by heart rate variability, or macronutrient content of a meal given postprandial body temperature? Collaboration opportunities not just between stakeholders, but between the intake and expenditure sides of the energy balance equation, are evident.

Summary

All emerging technologies require improvements in accessibility, acceptability, and availability. In addition, as technologies become ever-more pervasive, increasing attention must be paid to ethics and responsible use. Current tools in expenditure assessment have successfully integrated diverse scientific domains to accurately capture activity and other physiological phenomena with minimal to no user input. Opportunities for improvement remain, especially with regard to capturing dietary intake, despite improvements rendered from digital adaptations of older methodologies. Although considerable advancements are occurring in image-based assessment approaches, there remains a pressing need for transformational technologies—perhaps still to be discovered—that move the field definitively beyond self-report (189) and integrate advances across the domains of chemo-sensing, spectroscopy, and many others. Such innovations will likely require "out of the box" creativity and engineering from researchers and developers; this is the present and future challenge.

Acknowledgments

We thank Adela Hruby for editorial work. The authors' responsibilities were as follows—SKD: developed and drafted the manuscript with content contributions from AJM, ES, CBW, SD, and RPS; and all authors: participated in reviewing and editing and read and approved the final manuscript.

References

1. Villinger K, Wahl DR, Boeing H, Schupp HT, Renner B. The effectiveness of app-based mobile interventions on nutrition behaviours and nutrition-related health outcomes: a systematic review and meta-analysis. Obes Rev 2019;20(10):1465–84.
2. Kirk MA, Amiri M, Pirbaglou M, Ritvo P. Wearable technology and physical activity behavior change in adults with chronic cardiometabolic disease: a systematic review and meta-analysis. Am J Health Promot 2019;33(5):778–91.
3. Eldridge AL, Piernas C, Illner A-K, Gibney MJ, Gurinović MA, De Vries JHM, Cade JE. Evaluation of new technology-based tools

for dietary intake assessment—an ILSI Europe Dietary Intake and Exposure Task Force evaluation. Nutrients 2019;11:55.
4. Vu T, Lin F, Alshurafa N, Xu W. Wearable food intake monitoring technologies: a comprehensive review. Computers 2017;6(1):4.
5. Doulah A, Mccrory MA, Higgins JA, Sazonov E. A systematic review of technology-driven methodologies for estimation of energy intake. IEEE Access 2019;7:49653–68.
6. McClung HL, Ptomey LT, Shook RP, Aggarwal A, Gorczyca AM, Sazonov ES, Becofsky K, Weiss R, Das SK. Dietary intake and physical activity assessment: current tools, techniques, and technologies for use in adult populations. Am J Prev Med 2018;55(4):e93–e104.
7. Kankanhalli A, Shin J, Oh H. Mobile-based interventions for dietary behavior change and health outcomes: scoping review. JMIR Mhealth Uhealth 2019;7(1):e11312.
8. Bell BM, Alam R, Alshurafa N, Thomaz E, Mondol AS, de la Haye K, Stankovic JA, Lach J, Spruijt-Metz D. Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review. NPJ Digit Med 2020;3(1):38.
9. Ainsworth B, Cahalin L, Buman M, Ross R. The current state of physical activity assessment tools. Prog Cardiovasc Dis 2015;57(4):387–95.
10. Sylvia LG, Bernstein EE, Hubbard JL, Keating L, Anderson EJ. Practical guide to measuring physical activity. J Acad Nutr Diet 2014;114(2):199–208.
11. Sazonov E, Schuckers S, Lopez-Meyer P, Makeyev O, Sazonova N, Melanson EL, Neuman M. Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. Physiol Meas 2008;29(5):525–41.
12. Amft O, Troster G. On-body sensing solutions for automatic dietary monitoring. IEEE Pervasive Comput 2009;8(2):62–70.
13. Sazonov ES, Makeyev O, Schuckers S, Lopez-Meyer P, Melanson EL, Neuman MR. Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior. IEEE Trans Biomed Eng 2010;57(3):626–33.
14. Päßler S, Fischer W-J. Food intake activity detection using a wearable microphone system. In: 7th International Conference on Intelligent Environments (IE); 25–28 July, 2011; Nottingham, United Kingdom. Amsterdam (Netherlands): IOS Press; 2011. p. 298–301.
15. Päßler S, Fischer W-J. Food intake monitoring: automated chew event detection in chewing sounds. IEEE J Biomed Health Inform 2014;18(1):278–89.
16. Fontana JM, Farooq M, Sazonov E. Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior. IEEE Trans Biomed Eng 2014;61(6):1772–9.
17. Dong Y, Scisco J, Wilson M, Muth E, Hoover A. Detecting periods of eating during free-living by tracking wrist motion. IEEE J Biomed Health Inform 2014;18(4):1253–60.
18. Salley JN, Hoover AW, Wilson ML, Muth ER. Comparison between human and bite-based methods of estimating caloric intake. J Acad Nutr Diet 2016;116(10):1568–77.
19. Rahman T, Adams AT, Zhang M, Cherry E, Zhou B, Peng H, Choudhury T. BodyBeat: a mobile system for sensing non-speech body sounds. In: MobiSys '14: Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services; 16–19 June, 2014; Bretton Woods, NH. New York: Association for Computing Machinery; 2014. p. 2–13.
20. Mirtchouk M, Merck C, Kleinberg S. Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 12–16 September, 2016; Heidelberg, Germany. New York: Association for Computing Machinery; 2016. p. 451–62.
21. Doulah A, Ghosh T, Hossain D, Imtiaz MH, Sazonov E. "Automatic ingestion monitor version 2" – a novel wearable device for automatic food intake detection and passive capture of food images. IEEE J Biomed Health Inform 2021;25(2):568–76.
22. Martin CK, Kaya S, Gunturk BK. Quantification of food intake using food image analysis. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2–6 September, 2009; Minneapolis, MN. Manhattan (NY): IEEE; 2009. p. 6869–72.
23. Pouladzadeh P, Shirmohammadi S, Al-Maghrabi R. Measuring calorie and nutrition from food image. IEEE Trans Instrum Meas 2014;63(8):1947–56.
24. Zhu F, Bosch M, Khanna N, Boushey CJ, Delp EJ. Multiple hypotheses image segmentation and classification with application to dietary assessment. IEEE J Biomed Health Inform 2015;19(1):377–88.
25. Kong F, Tan J. DietCam: automatic dietary assessment with mobile camera phones. Pervasive Mob Comput 2012;8(1):147–63.
26. Kissileff HR, Klingsberg G, Van Itallie TB. Universal eating monitor for continuous recording of solid or liquid consumption in man. Am J Physiol 1980;238(1):R14–22.
27. Chang K-h, Liu S-y, Chu H-h, Hsu JY-j, Chen C, Lin T-y, Chen C-y, Huang P. The diet-aware dining table: observing dietary behaviors over a tabletop surface. In: Fishkin KP, Schiele B, Nixon P, Quigley A, editors. Pervasive computing. Berlin and Heidelberg (Germany): Springer Berlin Heidelberg; 2006. p. 366–82.
28. Papapanagiotou V, Diou C, Langlet B, Ioakimidis I, Delopoulos A. Automated extraction of food intake indicators from continuous meal weight measurements. In: Ortuño F, Rojas I, editors. Bioinformatics and biomedical engineering. Cham (Switzerland): Springer International Publishing; 2015. p. 35–46.
29. Papapanagiotou V, Diou C, Ioakimidis I, Sodersten P, Delopoulos A. Automatic analysis of food intake and meal microstructure based on continuous weight measurements. IEEE J Biomed Health Inform 2019;23(2):893–902.
30. Kelly LA, McMillan DG, Anderson A, Fippinger M, Fillerup G, Rider J. Validity of actigraphs uniaxial and triaxial accelerometers for assessment of physical activity in adults in laboratory conditions. BMC Med Phys 2013;13(1):5.
31. Hibbing PR, Lamunion SR, Kaplan AS, Crouter SE. Estimating energy expenditure with Actigraph GT9X inertial measurement unit. Med Sci Sports Exerc 2018;50(5):1093–102.
32. Bassett DR, John D, Conger SA, Rider BC, Passmore RM, Clark JM. Detection of lying down, sitting, standing, and stepping using two activPAL monitors. Med Sci Sports Exerc 2014;46(10): 2025–9.
33. An H-S, Kim Y, Lee J-M. Accuracy of inclinometer functions of the activPAL and Actigraph GT3X+: a focus on physical activity. Gait Posture 2017;51:174–80.
34. Wieters KM, Kim J-H, Lee C. Assessment of wearable global positioning system units for physical activity research. J Phys Act Health 2012;9(7):913–23.
35. Carlson JA, Schipperijn J, Kerr J, Saelens BE, Natarajan L, Frank LD, Glanz K, Conway TL, Chapman JE, Cain KL, et al. Locations of physical activity as assessed by GPS in young adolescents. Pediatrics 2016;137(1):e20152430.
36. Albrecht BM, Stalling I, Recke C, Bammann K. Accelerometer-assessed outdoor physical activity is associated with meteorological conditions among older adults: cross-sectional results from the OUTDOOR ACTIVE study. PLoS One 2020;15(1):e0228053.
37. Flynn JI, Coe DP, Larsen CA, Rider BC, Conger SA, Bassett DR, Jr. Detecting indoor and outdoor environments using the Actigraph GT3X+ light sensor in children. Med Sci Sports Exerc 2014;46(1):201–6.
38. Joyce DS, Zele AJ, Feigl B, Adhikari P. The accuracy of artificial and natural light measurements by actigraphs. J Sleep Res 2020;29(5):e12963.
39. Brage S, Westgate K, Franks PW, Stegle O, Wright A, Ekelund U, Wareham NJ. Estimation of free-living energy expenditure by heart rate and movement sensing: a doubly-labelled water study. PLoS One 2015;10(9):e0137206.
40. Kuzik N, Carson V. Accelerometer Bluetooth proximity validation in parents and early years children. Meas Phys Educ Exerc Sci 2018;22(4):287–93.
41. Correa JB, Apolzan JW, Shepard DN, Heil DP, Rood JC, Martin CK. Evaluation of the ability of three physical activity monitors to predict

weight change and estimate energy expenditure. Appl Physiol Nutr Metab 2016;41(7):758–66.
42. Hurter L, Fairclough SJ, Knowles ZR, Porcellato LA, Cooper-Ryan AM, Boddy LM. Establishing raw acceleration thresholds to classify sedentary and stationary behaviour in children. Children (Basel) 2018;5(12):172.
43. Bagot KS, Matthews SA, Mason M, Squeglia LM, Fowler J, Gray K, Herting M, May A, Colrain I, Godino J, et al. Current, future and potential use of mobile and wearable technologies and social media data in the ABCD study to increase understanding of contributors to child health. Dev Cogn Neurosci 2018;32:121–9.
44. Wright SP, Hall Brown, TS, Collier SR, Sandberg K. How consumer physical activity monitors could transform human physiology research. Am J Physiol Regul Integr Comp Physiol 2017;312(3):R358–67.
45. Sharp DB, Allman-Farinelli M. Feasibility and validity of mobile phones to assess dietary intake. Nutrition 2014;30(11–12):1257–66.
46. Rusin M, Arsand E, Hartvigsen G. Functionalities and input methods for recording food intake: a systematic review. Int J Med Inf 2013;82(8):653–64.
47. Torous J, Nebeker C. Navigating ethics in the digital age: introducing Connected and Open Research Ethics (CORE), a tool for researchers and institutional review boards. J Med Internet Res 2017;19(2):e38.
48. Nebeker C, Harlow J, Espinoza Giacinto R, Orozco-Linares R, Bloss CS, Weibel N. Ethical and regulatory challenges of research using pervasive sensing and other emerging technologies: IRB perspectives. AJOB Empir Bioeth 2017;8(4):266–76.
49. Arens-Volland AG, Spassova L, Bohn T. Promising approaches of computer-supported dietary assessment and management—current research status and available applications. Int J Med Inf 2015;84(12):997–1008.
50. Cade JE. Measuring diet in the 21st century: use of new technologies. Proc Nutr Soc 2017;76(3):276–82.
51. Sevil M, Rashid M, Hajizadeh I, Askari MR, Hobbs N, Brandt R, Park M, Quinn L, Cinar A. Discrimination of simultaneous psychological and physical stressors using wristband biosignals. Comput Methods Programs Biomed 2021;199:105898.
52. Kirkpatrick SI, Baranowski T, Subar AF, Tooze JA, Frongillo EA. Best practices for conducting and interpreting studies to validate self-report dietary assessment methods. J Acad Nutr Diet 2019;119(11):1801–16.
53. Subar AF, Freedman LS, Tooze JA, Kirkpatrick SI, Boushey C, Neuhouser ML, Thompson FE, Potischman N, Guenther PM, Tarasuk V, et al. Addressing current criticism regarding the value of self-report dietary data. J Nutr 2015;145(12):2639–45.
54. Nelson RO, Hayes SC. Theoretical explanations for reactivity in self-monitoring. Behav Modif 1981;5(1):3–14.
55. Subar AF, Kirkpatrick SI, Mittl B, Zimmerman TP, Thompson FE, Bingley C, Willis G, Islam NG, Baranowski T, McNutt S, et al. The Automated Self-Administered 24-hour dietary recall (ASA24): a resource for researchers, clinicians, and educators from the National Cancer Institute. J Acad Nutr Diet 2012;112(8):1134–7.
56. Fallaize R, Forster H, Macready AL, Walsh MC, Mathers JC, Brennan L, Gibney ER, Gibney MJ, Lovegrove JA. Online dietary intake estimation: reproducibility and validity of the Food4Me food frequency questionnaire against a 4-day weighed food record. J Med Internet Res 2014;16(8):e190.
57. Lemacks JL, Adams K, Lovetere A. Dietary intake reporting accuracy of the Bridge2U mobile application food log compared to control meal and dietary recall methods. Nutrients 2019;11(1):199.
58. Cade JE, Warthon-Medina M, Albar S, Alwan NA, Ness A, Roe M, Wark PA, Greathead K, Burley VJ, Finglas P, et al. DIET@NET: best practice guidelines for dietary assessment in health research. BMC Med 2017;15(1):202.
59. Timon CM, Cooper SE, Barker ME, Astell AJ, Adlam T, Hwang F, Williams EA. A comparison of food portion size estimation by older adults, young adults and nutritionists. J Nutr Health Aging 2018;22:230–6.
60. 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(2):268–80.
61. Conrad J, Koch SAJ, Nöthlings U. New approaches in assessing food intake in epidemiology. Curr Opin Clin Nutr Metab Care 2018;21(5):343–51.
62. Gemming L, Rush E, Maddison R, Doherty A, Gant N, Utter J, Ni Mhurchu C. Wearable cameras can reduce dietary under-reporting: doubly labelled water validation of a camera-assisted 24 h recall. Br J Nutr 2015;113(2):284–91.
63. 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(3):283–94.
64. Boushey CJ, Spoden M, Delp EJ, Zhu F, Bosch M, Ahmad Z, Shvetsov YB, DeLany JP, Kerr DA. Reported energy intake accuracy compared to doubly labeled water and usability of the mobile food record among community dwelling adults. Nutrients 2017;9(3):312.
65. Sun M, Burke LE, Mao Z-H, Chen Y, Chen H-C, Bai Y, Li Y, Li C, Jia W. eButton: a wearable computer for health monitoring and personal assistance. In: Proceedings of the 51st Annual Design Automation Conference; 1–5 June, 2014; San Francisco, CA. New York: Association for Computing Machinery; 2014. p. 1–6.
66. Archundia Herrera MC, Chan CB. Narrative review of new methods for assessing food and energy intake. Nutrients 2018;10(8):1064.
67. Howes E, Boushey CJ, Kerr DA, Tomayko EJ, Cluskey M. Image-based dietary assessment ability of dietetics students and interns. Nutrients 2017;9(2):114.
68. Gilhooly CH. Are calorie counting apps ready to replace traditional dietary assessment methods? Nutr Today 2017;52(1):10–8.
69. Blanchard CM, Chin MK, Gilhooly CH, Barger K, Matuszek G, Miki AJ, Côté RG, Eldridge AL, Green H, Mainardi F, et al. Evaluation of PIQNIQ, a novel mobile application for capturing dietary intake. J Nutr 2021;151(5):1347–56.
70. Carter MC, Albar SA, Morris MA, Mulla UZ, Hancock N, Evans CE, Alwan NA, Greenwood DC, Hardie LJ, Frost GS, et al. Development of a UK online 24-h dietary assessment tool: myfood24. Nutrients 2015;7(6):4016–32.
71. Carter MC, Hancock N, Albar SA, Brown H, Greenwood DC, Hardie LJ, Frost GS, Wark PA, Cade JE. Development of a new branded UK food composition database for an online dietary assessment tool. Nutrients 2016;8(8):480.
72. Thompson FE, Subar AF. Dietary assessment methodology. In: Coulston AM, Boushey CJ, Ferruzzi MG, Delahanty LM, editors. Nutrition in the prevention and treatment of disease. 4th ed. Cambridge (MA): Academic Press; 2017. p. 5–48.
73. Evans K, Hennessy Á, Walton J, Timon C, Gibney E, Flynn A. Development and evaluation of a concise food list for use in a web-based 24-h dietary recall tool. J Nutr Sci 2017;6:e46.
74. Illner AK, Freisling H, Boeing H, Huybrechts I, Crispim SP, Slimani N. Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. Int J Epidemiol 2012;41(4):1187–203.
75. Conrad J, Nöthlings U. Innovative approaches to estimate individual usual dietary intake in large-scale epidemiological studies. Proc Nutr Soc 2017;76(3):213–9.
76. Allman-Farinelli M, Gemming L. Technology interventions to manage food intake: where are we now? Curr Diab Rep 2017;17(11):103.
77. Shephard RJ, Aoyagi Y. Measurement of human energy expenditure, with particular reference to field studies: an historical perspective. Eur J Appl Physiol 2012;112(8):2785–815.
78. Steene-Johannessen J, Anderssen SA, van der Ploeg HP, Hendriksen IJM, Donnelly AE, Brage S, Ekelund U. Are self-report measures able to define individuals as physically active or inactive? Med Sci Sports Exerc 2016;48(2):235–44.
79. Silfee VJ, Haughton CF, Jake-Schoffman DE, Lopez-Cepero A, May CN, Sreedhara M, Rosal MC, Lemon SC. Objective measurement of physical activity outcomes in lifestyle interventions among adults: a systematic review. Prev Med Rep 2018;11:74–80.

80. Gartner. Gartner forecasts global spending on wearable devices to total $81.5 billion in 2021. Stamford (CT): Gartner; 2021.
81. Grand View Research. Wearable technology market size, share & trends analysis report by product (wrist-wear, eye-wear & head-wear, foot-wear, neck-wear, body-wear), by application, by region, and segment forecasts, 2020–2027. San Francisco (CA): Grand View Research; 2020.
82. Rowlands AV, Mirkes EM, Yates T, Clemes S, Davies M, Khunti K, Edwardson CL. Accelerometer-assessed physical activity in epidemiology: are monitors equivalent? Med Sci Sports Exerc 2018;50(2):257–65.
83. Ndahimana D, Kim E-K. Measurement methods for physical activity and energy expenditure: a review. Clin Nutr Res 2017;6(2):68–80.
84. Lyden K, Swibas T, Catenacci V, Guo R, Szuminsky N, Melanson EL. Estimating energy expenditure using heat flux measured at a single body site. Med Sci Sports Exerc 2014;46(11):2159–67.
85. O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, Finlayson G, Stubbs J. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. Br J Sports Med 2020;54(6):332–40.
86. Miller DJ, Capodilupo JV, Lastella M, Sargent C, Roach GD, Lee VH, Capodilupo ER. Analyzing changes in respiratory rate to predict the risk of COVID-19 infection. PLoS One 2020;15(12):e0243693.
87. Berryhill S, Morton CJ, Dean A, Berryhill A, Provencio-Dean N, Patel SI, Estep L, Combs D, Mashaqi S, Gerald LB, et al. Effect of wearables on sleep in healthy individuals: a randomized crossover trial and validation study. J Clin Sleep Med 2020;16(5):775–83.
88. Sánchez R, Villena M. Comparative evaluation of wearable devices for measuring elevation gain in mountain physical activities. Proc Inst Mech Eng P J Sport Eng Technol 2020;234(4):312–9.
89. Levine A, Buono MJ. Rating of perceived exertion increases synergistically during prolonged exercise in a combined heat and hypoxic environment. J Therm Biol 2019;84:99–102.
90. Maughan RJ, Otani H, Watson P. Influence of relative humidity on prolonged exercise capacity in a warm environment. Eur J Appl Physiol 2012;112(6):2313–21.
91. Plasqui G, Bonomi AG, Westerterp KR. Daily physical activity assessment with accelerometers: new insights and validation studies. Obes Rev 2013;14(6):451–62.
92. Schrack JA, Cooper R, Koster A, Shiroma EJ, Murabito JM, Rejeski WJ, Ferrucci L, Harris TB. Assessing daily physical activity in older adults: unraveling the complexity of monitors, measures, and methods. J Gerontol A Biol Sci Med Sci 2016;71(8):1039–48.
93. Sardinha LB, Júdice PB. Usefulness of motion sensors to estimate energy expenditure in children and adults: a narrative review of studies using DLW. Eur J Clin Nutr 2017;71(3):331–9.
94. Farrahi V, Niemelä M, Kangas M, Korpelainen R, Jämsä T. Calibration and validation of accelerometer-based activity monitors: a systematic review of machine-learning approaches. Gait Posture 2019;68:285–99.
95. Trost SG. State of the art reviews: measurement of physical activity in children and adolescents. Am J Lifestyle Med 2007;1(4):299–314.
96. Arab L, Estrin D, Kim DH, Burke J, Goldman J. Feasibility testing of an automated image-capture method to aid dietary recall. Eur J Clin Nutr 2011;65(10):1156–62.
97. Thompson FE, Subar AF, Loria CM, Reedy JL, Baranowski T. Need for technological innovation in dietary assessment. J Am Diet Assoc 2010;110(1):48–51.
98. Martin CK, Correa JB, Han H, Allen HR, Rood JC, Champagne CM, Gunturk BK, Bray GA. Validity of the Remote Food Photography Method (RFPM) for estimating energy and nutrient intake in near real-time. Obesity 2012;20(4):891–9.
99. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1–32.
100. Zhang MM. Identifying the cuisine of a plate of food. Technical Report CSE 190. La Jolla (CA): University of California at San Diego; 2011.
101. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med 2014;48(13):1019–23.
102. Narayanan A, Desai F, Stewart T, Duncan S, Mackay L. Application of raw accelerometer data and machine-learning techniques to characterize human movement behavior: a systematic scoping review. J Phys Act Health 2020;17(3):360–83.
103. Pavey TG, Gilson ND, Gomersall SR, Clark B, Trost SG. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. J Sci Med Sport 2017;20(1):75–80.
104. Trost SG, Wong W-K, Pfeiffer KA, Zheng Y. Artificial neural networks to predict activity type and energy expenditure in youth. Med Sci Sports Exerc 2012;44(9):1801–9.
105. Wang Z, Wu D, Chen J, Ghoneim A, Hossain MA. A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection. IEEE Sens J 2016;16(9):3198–207.
106. Mimouna A, Khalifa AB, Ben Amara NE. Human action recognition using triaxial accelerometer data: selective approach. In: 15th International Multi-Conference on Systems, Signals & Devices (SSD); 19–22 March, 2018; Hammamet, Tunisia. Manhattan (NY): IEEE; 2018. p. 491–6.
107. Keadle SK, Lyden KA, Strath SJ, Staudenmayer JW, Freedson PS. A framework to evaluate devices that assess physical behavior. Exerc Sport Sci Rev 2019;47(4):206–14.
108. Simpson E, Bradley J, Poliakov I, Jackson D, Olivier P, Adamson AJ, Foster E. Iterative development of an online dietary recall tool: INTAKE24. Nutrients 2017;9(2):118.
109. Bucher Della Torre S, Carrard I, Farina E, Danuser B, Kruseman M. Development and evaluation of e-CA, an electronic mobile-based food record. Nutrients 2017;9(1):76.
110. Chen YS, Wong JE, Ayob AF, Othman NE, Poh BK. Can Malaysian young adults report dietary intake using a food diary mobile application? A pilot study on acceptability and compliance. Nutrients 2017;9(1):62.
111. Watkins I, Kules B, Yuan X, Xie B. Heuristic evaluation of healthy eating apps for older adults. J Consum Health Internet 2014;18(2):105–27.
112. Kelly P, Marshall SJ, Badland H, Kerr J, Oliver M, Doherty AR, Foster C. An ethical framework for automated, wearable cameras in health behavior research. Am J Prev Med 2013;44(3):314–9.
113. Kerr DA, Dhaliwal SS, Pollard CM, Norman R, Wright JL, Harray AJ, Shoneye CL, Solah VA, Hunt WJ, Zhu F, et al. BMI is associated with the willingness to record diet with a mobile food record among adults participating in dietary interventions. Nutrients 2017;9(3):244.
114. Zang J, Song J, Wang Z, Yao C, Ma J, Huang C, Zhu Z, Smith LP, Du S, Hua J, et al. Acceptability and feasibility of smartphone-assisted 24 h recalls in the Chinese population. Public Health Nutr 2015;18(18):3272–7.
115. Zahry NR, Cheng Y, Peng W. Content analysis of diet-related mobile apps: a self-regulation perspective. Health Commun 2016;31(10):1301–10.
116. Kirkpatrick SI, Gilsing AM, Hobin E, Solbak NM, Wallace A, Haines J, Mayhew AJ, Orr SK, Raina P, Robson PJ, et al. Lessons from studies to evaluate an online 24-hour recall for use with children and adults in Canada. Nutrients 2017;9(2):100.
117. Prioleau T, Moore E, Ghovanloo M. Unobtrusive and wearable systems for automatic dietary monitoring. IEEE Trans Biomed Eng 2017;64(9):2075–89.
118. O'Loughlin G, Cullen SJ, McGoldrick A, O'Connor S, Blain R, O'Malley S, Warrington GD. Using a wearable camera to increase the accuracy of dietary analysis. Am J Prev Med 2013;44(3):297–301.
119. Pettitt C, Liu J, Kwasnicki RM, Yang G-Z, Preston T, Frost G. A pilot study to determine whether using a lightweight, wearable micro-camera improves dietary assessment accuracy and offers information on macronutrients and eating rate. Br J Nutr 2016;115(1):160–7.
120. Schiboni G, Wasner F, Amft O. A privacy-preserving wearable camera setup for dietary event spotting in free-living. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); 19–23 March, 2018; Athens, Greece. Manhattan (NY): IEEE; 2018. p. 872–7.

121. Gemming L, Utter J, Ni Mhurchu C. Image-assisted dietary assessment: a systematic review of the evidence. J Acad Nutr Diet 2015;115(1):64–77.
122. Muhammad Sayem AS, Hon Teay S, Shahariar H, Fink PL, Albarbar A. Review on smart electro-clothing systems (SeCSs). Sensors 2020;20(3):587.
123. E&T Editorial Staff. Sensors woven into clothing could monitor vital signs. [Internet]. London: The Institution of Engineering and Technology; 2020. [Cited 2021 Jul 8]. Available from: https://eandt.theiet.org/content/articles/2020/04/sensors-woven-into-clothing-could-monitor-vital-signs/.
124. Walker RK, Hickey AM, Freedson PS. Advantages and limitations of wearable activity trackers: considerations for patients and clinicians. Clin J Oncol Nurs 2016;20(6):606–10.
125. Nebeker C, Lagare T, Takemoto M, Lewars B, Crist K, Bloss CS, Kerr J. Engaging research participants to inform the ethical conduct of mobile imaging, pervasive sensing, and location tracking research. Transl Behav Med 2016;6(4):577–86.
126. Taylor K, Silver L. Smartphone ownership is growing rapidly around the world, but not always equally. Washington (DC): Pew Research Center; 2019.
127. Régnier F, Chauvel L. Digital inequalities in the use of self-tracking diet and fitness apps: interview study on the influence of social, economic, and cultural factors. JMIR Mhealth Uhealth 2018;6(4):e101.
128. Bol N, Helberger N, Weert JCM. Differences in mobile health app use: a source of new digital inequalities? Inform Soc 2018;34(3):183–93.
129. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9.
130. Müller AM, Maher CA, Vandelanotte C, Hingle M, Middelweerd A, Lopez ML, DeSmet A, Short CE, Nathan N, Hutchesson MJ, et al. Physical activity, sedentary behavior, and diet-related eHealth and mHealth research: bibliometric analysis. J Med Internet Res 2018;20(4):e122.
131. Lee HE, Cho J. What motivates users to continue using diet and fitness apps? Application of the uses and gratifications approach. Health Commun 2017;32(12):1445–53.
132. Takemoto M, Manini TM, Rosenberg DE, Lazar A, Zlatar ZZ, Das SK, Kerr J. Diet and activity assessments and interventions using technology in older adults. Am J Prev Med 2018;55(4): e105–15.
133. Harrington CN, Ruzic L, Sanford JA. Universally accessible mHealth apps for older adults: towards increasing adoption and sustained engagement. In: Antona M, Stephanidis C, editors. Universal access in human–computer interaction. Human and technological environments: 11th International Conference, UAHCI 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part III. Cham (Switzerland): Springer International Publishing; 2017. p. 3–12.
134. Steinfeld E, Maisel J. Universal design: creating inclusive environments. Hoboken (NJ): Wiley; 2012.
135. Gilmore LA, Duhé AF, Frost EA, Redman LM. The technology boom: a new era in obesity management. J Diabetes Sci Technol 2014;8(3):596–608.
136. Osadchiy T, Poliakov I, Olivier P, Rowland M, Foster E. Validation of a recommender system for prompting omitted foods in online dietary assessment surveys. In: PervasiveHealth '19: proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare; 20–23 May 2019; Trento, Italy. New York: Association for Computing Machinery; 2019. p. 208–15.
137. Yang Y, Jia W, Bucher T, Zhang H, Sun M. Image-based food portion size estimation using a smartphone without a fiducial marker. Public Health Nutr 2019;22(7):1180–92.
138. Charrondiere UR, Rittenschober D, Nowak V, Stadlmayr B, Wijesinha-Bettoni R, Haytowitz D. Improving food composition data quality: three new FAO/INFOODS guidelines on conversions, data evaluation and food matching. Food Chem 2016;193: 75–81.
139. USDA, Agricultural Research Service (ARS). FoodData Central. [Internet]. Beltsville (MD): USDA ARS; 2019. [Cited 2021 Feb 20]. Available from: fdc.nal.usda.gov.
140. European Food Safety Authority (EFSA). The EFSA Comprehensive European Food Consumption Database. [Internet]. Parma (Italy): EFSA; 2020. [Cited 2021 Apr 1]. Available from: https://www.efsa.europa.eu/en/food-consumption/comprehensive-database.
141. Steele R. An overview of the state of the art of automated capture of dietary intake information. Crit Rev Food Sci Nutr 2015;55(13):1929–38.
142. Hassannejad H, Matrella G, Ciampolini P, De Munari I, Mordonini M, Cagnoni S. Automatic diet monitoring: a review of computer vision and wearable sensor-based methods. Int J Food Sci Nutr 2017;68(6):656–70.
143. Subhi MA, Ali SH, Mohammed MA. Vision-based approaches for automatic food recognition and dietary assessment: a survey. IEEE Access 2019;7:35370–81.
144. Bossard L, Guillaumin M, Van Gool L. Food-101 – mining discriminative components with random forests. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI. Cham (Switzerland): Springer International Publishing; 2014. p. 446–61.
145. Pouladzadeh P, Yassine A, Shirmohammadi S. FooDD: food detection dataset for calorie measurement using food images. In: Murino V, Puppo E, Sona D, Cristani M, Sansone C, editors. New trends in image analysis and processing – ICIAP 2015 workshops. Cham (Switzerland): Springer International Publishing; 2015. p. 441–8.
146. Rollo ME, Ash S, Lyons-Wall P, Russell A. Trial of a mobile phone method for recording dietary intake in adults with type 2 diabetes: evaluation and implications for future applications. J Telemed Telecare 2011;17(6):318–23.
147. Zhu F, Bosch M, Woo I, Kim S, Boushey CJ, Ebert DS, Delp EJ. The use of mobile devices in aiding dietary assessment and evaluation. IEEE J Sel Top Signal Process 2010;4(4):756–66.
148. Dahl Lassen A, Poulsen S, Ernst L, Kaae Andersen K, Biltoft-Jensen A, Tetens I. Evaluation of a digital method to assess evening meal intake in a free-living adult population. Food Nutr Res 2010;54:5311.
149. Liu J, Johns E, Atallah L, Pettitt C, Lo B, Frost G, Yang G-Z. An intelligent food-intake monitoring system using wearable sensors. In: Proceedings of the 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN); 9–12 May, 2012; London, United Kingdom. Washington (DC): IEEE Computer Society; 2012. p. 154–60.
150. Ptomey LT, Willis EA, Honas JJ, Mayo MS, Washburn RA, Herrmann SD, Sullivan DK, Donnelly JE. Validity of energy intake estimated by digital photography plus recall in overweight and obese young adults. J Acad Nutr Diet 2015;115(9):1392–9.
151. McClung HL, Champagne CM, Allen HR, McGraw SM, Young AJ, Montain SJ, Crombie AP. Digital food photography technology improves efficiency and feasibility of dietary intake assessments in large populations eating ad libitum in collective dining facilities. Appetite 2017;116:389–94.
152. Prinz N, Bohn B, Kern A, Püngel D, Pollatos O, Holl RW. Feasibility and relative validity of a digital photo-based dietary assessment: results from the Nutris-Phone study. Public Health Nutr 2019;22(7):1160–7.
153. Akpro Hippocrate EA, Suwa H, Arakawa Y, Yasumoto K. Food weight estimation using smartphone and cutlery. In: Proceedings of the First Workshop on IoT-enabled Healthcare and Wellness Technologies and Systems; 30 June, 2016; Singapore. New York: Association for Computing Machinery; 2016. p. 9–14.
154. Jia W, Li Y, Qu R, Baranowski T, Burke LE, Zhang H, Bai Y, Mancino JM, Xu G, Mao Z-H, et al. Automatic food detection in egocentric images using artificial intelligence technology. Public Health Nutr 2019;22(7):1168–79.
155. Fang S, Liu C, Tahboub K, Zhu F, Delp EJ, Boushey CJ. cTADA: the design of a crowdsourcing tool for online food image identification and segmentation. In: 2018 IEEE Southwest Symposium on Image

Analysis and Interpretation (SSIAI); 8–10 April, 2018; Las Vegas, NV. Manhattan (NY): IEEE; 2018. p. 25–8.
156. Matsuda Y, Yanai K. Multiple-food recognition considering co-occurrence employing manifold ranking. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012); 11–15 November, 2012; Tsukuba, Japan. Manhattan (NY): IEEE; 2012. p. 2017–20.
157. Hoashi H, Joutou T, Yanai K. Image recognition of 85 food categories by feature fusion. In: Proceedings of the 2010 IEEE International Symposium on Multimedia; 13–15 December, 2010; Taichung, Taiwan. Washington (DC): IEEE Computer Society; 2010. p. 296–301.
158. Kawano Y, Yanai K. Real-time mobile food recognition system. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops; 23–28 June, 2013; Portland, OR. Washington (DC): IEEE Computer Society; 2013. p. 1–7.
159. Christodoulidis S, Anthimopoulos M, Mougiakakou S. Food recognition for dietary assessment using deep convolutional neural networks. In: Murino V, Puppo E, Sona D, Cristani M, Sansone C, editors. New trends in image analysis and processing—ICIAP 2015 Workshops. Cham (Switzerland): Springer International Publishing; 2015. p. 458–65.
160. Pouladzadeh P, Shirmohammadi S, Yassine A. You are what you eat: so measure what you eat! IEEE Instrum Meas Mag 2016;19(1):9–15.
161. Xu C, He Y, Khannan N, Parra A, Boushey C, Delp E. Image-based food volume estimation. In: Proceedings of the 5th International Workshop on Multimedia for Cooking & Eating Activities; 21 October, 2013; Barcelona, Spain. New York: Association for Computing Machinery; 2013. p. 75–80.
162. Hassannejad H, Matrella G, Ciampolini P, Munari I, Mordonini M, Cagnoni S. A new approach to image-based estimation of food volume. Algorithms 2017;10(2):66.
163. Thong YJ, Nguyen T, Zhang Q, Karunanithi M, Yu L. Predicting food nutrition facts using pocket-size near-infrared sensor. In: 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 11–15 July, 2017; Jeju, Republic of Korea. Manhattan (NY): IEEE; 2017. p. 742–5.
164. Huang Q, Yang Z, Zhang Q. Smart-U: smart utensils know what you eat. In: IEEE INFOCOM: IEEE Conference on Computer Communications; 15–19 April, 2018; Honolulu, HI. Manhattan (NY): IEEE; 2018. p. 1439–47.
165. Calibre Control. Why is NIR an effective food testing technique? [Internet]. Warrington (United Kingdom): Calibre Control International; 2019. [Cited 2021 Apr 1]. Available from: https://www.calibrecontrol.com/news-blog/2019/10/28/why-is-nir-an-effective-food-testing-technique.
166. Trimble S. Food-processing relies on near-infrared spectroscopy. [Internet]. Camas (WA): Felix Instruments; 2020. [Cited 2021 Apr 1]. Available from: https://felixinstruments.com/blog/food-processing-relies-on-near-infrared-spectroscopy/.
167. Ward HA, McLellan H, Udeh-Momoh C, Giannakopoulou P, Robb C, Wark PA, Middleton L. Use of online dietary recalls among older UK adults: a feasibility study of an online dietary assessment tool. Nutrients 2019;11(7):1451.
168. Dunseath S, Weibel N, Bloss CS, Nebeker C. NIH support of mobile, imaging, pervasive sensing, social media and location tracking (MISST) research: laying the foundation to examine research ethics in the digital age. NPJ Digit Med 2018;1(1):20171.
169. Klurfeld DM, Hekler EB, Nebeker C, Patrick K, Khoo CSH. Technology innovations in dietary intake and physical activity assessment: challenges and recommendations for future directions. Am J Prev Med 2018;55(4):e117–22.
170. Haug CJ. Turning the tables — the new European General Data Protection Regulation. N Engl J Med 2018;379(3):207–9.
171. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021;4(1):2.
172. Office for Civil Rights, US Department of Health and Human Services. Resources for mobile health apps developers. [Internet]. Washington (DC): Office for Civil Rights; 2020. [Cited 2021 Jun 24]. Available from: https://www.hhs.gov/hipaa/for-professionals/special-topics/health-apps/index.html.
173. Bashir M, Hayes C, Lambert AD, Kesan JP. Online privacy and informed consent: the dilemma of information asymmetry. Proc Assoc Inf Sci Technol 2015;52(1):1–10.
174. Nebeker C, Linares-Orozco R, Crist K. A multi-case study of research using mobile imaging, sensing and tracking technologies to objectively measure behavior: ethical issues and insights to guide responsible research practice. J Res Admin 2015;46(1):118–37.
175. Thomaz E, Parnami A, Essa I, Abowd GD. Feasibility of identifying eating moments from first-person images leveraging human computation. Proceedings of the 4th International SenseCam & Pervasive Imaging Conference; 18–19 November, 2013; San Diego, CA. New York: Association for Computing Machinery; 2013. p. 26–33.
176. Nebeker C, Bartlett Ellis RJ, Torous J. Development of a decision-making checklist tool to support technology selection in digital health research. Transl Behav Med 2020;10(4):1004–15.
177. Lobelo F, Rohm Young D, Sallis R, Garber MD, Billinger SA, Duperly J, Hutber A, Pate RR, Thomas RJ, Widlansky ME, et al. Routine assessment and promotion of physical activity in healthcare settings: a scientific statement from the American Heart Association. Circulation 2018;137(18):e495–522.
178. Chen J, Gemming L, Hanning R, Allman-Farinelli M. Smartphone apps and the nutrition care process: current perspectives and future considerations. Patient Educ Couns 2018;101(4):750–7.
179. Koonin LM, Hoots B, Tsang CA, Leroy Z, Farris K, Jolly T, Antall P, McCabe B, Zelis CBR, Tong I, et al. Trends in the use of telehealth during the emergence of the COVID-19 pandemic — United States, January–March 2020. MMWR Morb Mortal Wkly Rep 2020;69(43):1595–9.
180. Wosik J, Fudim M, Cameron B, Gellad ZF, Cho A, Phinney D, Curtis S, Roman M, Poon EG, Ferranti J, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020;27(6):957–62.
181. Forster H, Walsh MC, Gibney MJ, Brennan L, Gibney ER. Personalised nutrition: the role of new dietary assessment methods. Proc Nutr Soc 2016;75(1):96–105.
182. Michel M, Burbidge A. Nutrition in the digital age—how digital tools can help to solve the personalized nutrition conundrum. Trends Food Sci Technol 2019;90:194–200.
183. Sauceda A, Frederico C, Pellechia K, Starin D. Results of the Academy of Nutrition and Dietetics' consumer health informatics work group's 2015 member app technology survey. J Acad Nutr Diet 2016;116(8):1336–8.
184. Zubin Maslov P, Schulman A, Lavie CJ, Narula J. Personalized exercise dose prescription. Eur Heart J 2018;39(25):2346–55.
185. Crump C, Sundquist K, Sundquist J, Winkleby MA. Exercise is medicine: primary care counseling on aerobic fitness and muscle strengthening. J Am Board Fam Med 2019;32(1):103–7.
186. Spanakis EG, Santana S, Tsiknakis M, Marias K, Sakkalis V, Teixeira A, Janssen JH, de Jong H, Tziraki C. Technology-based innovations to foster personalized healthy lifestyles and well-being: a targeted review. J Med Internet Res 2016;18(6):e128.
187. Buday R, Tapia R, Maze GR. Technology-driven dietary assessment: a software developer's perspective. J Hum Nutr Diet 2014;27:10–7.
188. Lobelo F, Kelli HM, Tejedor SC, Pratt M, McConnell MV, Martin SS, Welk GJ. The Wild Wild West: a framework to integrate mHealth software applications and wearables to support physical activity assessment, counseling and interventions for cardiovascular disease risk reduction. Prog Cardiovasc Dis 2016;58(6):584–94.
189. Ershow AG, Ortega A, Timothy Baldwin J, Hill JO. Engineering approaches to energy balance and obesity: opportunities for novel collaborations and research: report of a joint National Science Foundation and National Institutes of Health workshop. J Diabetes Sci Technol 2007;1(1):95–105.

Ghanbari et al. *Translational Psychiatry* (2021)11:103
https://doi.org/10.1038/s41398-021-01228-7

Translational Psychiatry

ARTICLE Open Access

Metabolomics reveals biomarkers of opioid use disorder

Reza Ghanbari[1,2], Yuanyuan Li[1], Wimal Pathmasiri[1], Susan McRitchie[1], Arash Etemadi[3], Jonathan D. Pollock[4], Hossein Poustchi[2], Afarin Rahimi-Movaghar[5], Masoumeh Amin-Esmaeili[5,6], Gholamreza Roshandel[7], Amaneh Shayanrad[2], Behrouz Abaei[2], Reza Malekzadeh[2] and Susan C. J. Sumner[1]

Abstract

Opioid use disorder (OUD) is diagnosed using the qualitative criteria defined by the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Diagnostic biomarkers for OUD do not currently exist. Our study focused on developing objective biological markers to differentiate chronic opiate users with OUD from chronic opiate users without OUD. Using biospecimens from the Golestan Cohort Study, we compared the metabolomics profiles of high opium users who were diagnosed as OUD positive with high opium users who were diagnosed as OUD negative. High opium use was defined as maximum weekly opium usage greater than or equal to the median usage (2.4 g per week), and OUD was defined as having 2 or more DSM-5 criteria in any 12-month period. Among the 218 high opium users in this study, 80 were diagnosed as OUD negative, while 138 were diagnosed as OUD positive. Seven hundred and twelve peaks differentiated high opium users diagnosed as OUD positive from high opium users diagnosed as OUD negative. Stepwise logistic regression modeling of subject characteristics data together with the 712 differentiating peaks revealed a signature that is 95% predictive of an OUD positive diagnosis, a significant ($p < 0.0001$) improvement over a 63% accurate prediction based on subject characteristic data for these samples. These results suggest that a metabolic profile can be used to predict an OUD positive diagnosis.

Introduction

More than fifty years have passed since Dole and Nyswander described opioid addiction as a metabolic disease, suggesting that opioids disrupt homeostasis to produce drug-seeking behavior in the face of adverse consequences[1]. An important issue in the addiction is that people exposed to opioids may develop dependence, but not Opioid Use Disorder (OUD)[2]. OUD is a chronic recurrent disorder that increasingly causes undesirable emotional states by involving the brain's reward system and could include impaired social functioning[3,4].

Despite significant advances in the genetics and neurobiology of addiction as a brain disease, and preliminary studies to discover biomarkers of OUD, validated systemic biomarkers for OUD do not exist[5]. Differential diagnosis of OUD is obtained through interview or questionnaire to determine if the patients meet the DSM-5 diagnostic criteria. These criteria include impaired control, social impairment, risky use, tolerance, withdrawal, craving, and continued use despite problems. Having at least two of the 11 criteria meets the diagnoses of OUD with the number of criteria met as an indicator of the severity of the OUD[6].

Iran is a country with a high rate of opiate use. Opium is the main opiate used[7]. Our study focused on a random sample of opium users in the Golestan Cohort Study (GCS) in Iran, where more than 8400 individuals (about 17% of the participants) reported chronic opiate use with a median duration of use of 19 years[8]. 75% of the opium

Correspondence: Reza Malekzadeh (dr.reza.malekzadeh@gmail.com) or Susan C. J. Sumner (Susan_sumner@unc.edu)
[1]Department of Nutrition, Nutrition Research Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]Digestive Oncology Research Center, Digestive Diseases Research Institute, Tehran University of Medical Science, Tehran, Iran
Full list of author information is available at the end of the article
These authors contributed equally: Reza Ghanbari, Yuanyuan Li

users in the full cohort used a combination of teriak and shireh with only 4 people reporting heroin use[8].

We investigated urinary metabolomic profiles to reveal biomarkers that could differentiate high opium users who were diagnosed as OUD positive from high opium users who were diagnosed as OUD negative. Our investigation is important because optimized treatment relies on accurate diagnosis of OUD. The DSM-5, the predominant diagnostic instrument in psychiatry, has known limitations for diagnosis of substance abuse disorders[9]. Objective biological markers can improve the diagnosis that is currently based on subjective DSM-5 questionnaire. In addition metabolites that are increased or decreased in opium users diagnosed with OUD (compared with opium users not diagnosed with OUD) can be used to determine pathway perturbations, and lead to the identification of druggable or nutritional targets.

Materials and methods

Study population

The details of the GCS (a cohort of over 50,000 adults aged 40–75 living in Golestan Province, Northeast Iran) have been previously published[10]. The GCS was approved by appropriate ethics committees at Tehran University of Medical Sciences, the US National Cancer Institute (NCI, IRB# 07-C-N120), and the International Agency for Research on Cancer (IARC).

In 2018, a random sample of 451 GCS participants who reported long-term opium use and 92 never-users were recalled. They underwent a detailed interview using modified Persian and Turkman versions of the Section L of WHO Composite International Diagnostic Interview (CIDI, version 2.1) to diagnose lifetime OUD[7], based on the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5). The presence of 2 or more of the 11 criteria during any 12-month period of life was defined as lifetime OUD.

Variables considered for adjustment of the logistic regression models (see below) included age at enrollment, gender, tobacco use (current/former/never), BMI, and route of opium use (ingestion/inhalation). OUD diagnosis was the outcome for the models. Alcohol use was not included in the analysis, because it was rare in this population, and only 3.5% of participants reported ever using alcohol.

No subjects participated in drug-related treatment for addiction as part of the GCS. Cohort participants gave non-fasted spot urine samples which were stored at −20 °C until 2015 when they were transferred on dry ice to the NCI Biorepository and stored at −80 °C. Aliquots were then shipped to UNC Chapel Hill.

Sample selection

In this sample selection, we excluded individuals who had discordant reports of their opiate use compared with baseline (baseline users who reported no lifetime opium use at the recall visit and vice versa, $n = 24$), and those without a urine sample available ($n = 8$). We also restricted the current analysis to high opium users reporting equal to or more than the median intake (2.4 g per week), to reduce the chance of misclassification. The final sample used in the current study included 138 urine samples from high opium users who were diagnosed as OUD positive, and 80 urine samples from high opium users diagnosed as OUD negative. Urine samples were selected from an additional 80 subjects who reported that they had never used opium.

Untargeted metabolomics via ultra-performance liquid chromatography (UPLC) high-resolution mass spectrometry

Details of the sample preparation, data acquisition, data preprocessing and metabolite identification and annotation are provided in the Supplementary Material Section. In brief, urine samples were prepared according to published methods[11], and untargeted metabolomics data were acquired on a Vanquish UHPLC systems coupled with a Q Exactive™ HF-X Hybrid Quadrupole-Orbitrap™ Mass Spectrometer (UPLC-HR-MS; Thermo Fisher Scientific). Data were processed using Progenesis QI (Waters Corporation). Peaks detected by UPLC-HR-MS were identified or annotated. Signals detected on our untargeted platform are matched to an in-house physical standards library that was developed by acquiring data for over 2000 chemical standards run under the same conditions to the study samples. The evidence basis for metabolite identifications and annotations are based on matching to our in-house library physical standards library (Ontology Level, OL), as well as to Public Databases (PD), and are detailed in the supplementary material.

Hypothesis testing

Statistical tests for the normalized peaks in the metabolomics profiles were conducted using a two-tailed *t*-test with the Satterthwaite correction for unequal variances or the chi-square test. Statistical analyses were conducted using SAS 9.4 (SAS Institute Inc., Cary, NC). In this exploratory metabolomics study, *p*-values were not adjusted for multiple testing[12,13]. The nominal *p*-values are reported for the following comparisons 80 high opium users diagnosed as OUD negative versus 138 high opium users diagnosed as OUD positive.

Logistic regression modeling

Logistic regression was used to model which peaks/metabolites were predictive of a positive OUD diagnosis. Several modeling approaches were used that included all normalized metabolomics peaks, or included subsets of peaks. Stepwise logistic regression procedures (criteria:

model entry $p < 0.1$ and model removal $p > 0.05$) with standardization of continuous variables, was used for model selection. The Hosmer-Lemeshow goodness-of-fit test was used to assess the final model for adequacy. Receiver operating characteristics (ROC) curve and the area under the curve (AUC) were used to evaluate metabolites as predictors of OUD. Stepwise models were conducted, with and without subject characteristics as potential covariates, using:

(a) 712 peaks which differentiated high opium users who were diagnosed as OUD positive from high opium users who were diagnosed as OUD negative.
(b) 40 identified/annotated metabolites that differentiated the high opium users who were diagnosed as OUD positive from high opium users who were diagnosed as OUD negative (26 of these 40 metabolites also differentiated opium users from non-opium users).
(c) 14 identified/annotated metabolites that were unique to differentiation of high opium users who were diagnosed as OUD positive from high opium users who were diagnosed as OUD negative (but did not also differentiate opium users from non-opium users).

Pathway enrichment: high opium users who were diagnosed as OUD positive from the high opium users who were diagnosed OUD as negative

Pathway enrichment was conducted using the Mummichog software in Metaboanalyst 4.0[14]. All features (m/z) remaining after filtering data were entered together with the p-value that was calculated for the comparison of high opium users who were diagnosed as OUD positive and high opium users who were diagnosed OUD as negative. A p-value cut-off of 0.01 was used to determine the size of the permutation group that the algorithm used for selecting significant features to match for all possible metabolites. A mass accuracy of 3 ppm was used as the threshold for annotations used in identifying candidate pathways. All possible metabolites which were matched by m/z were searched in the human reference metabolic network (hsa_mfn), and the null distribution of module activities was estimated by using 100 permutations of random lists drawn from the experimental reference feature list. The candidate pathways were based on the similarity of m/z.

Results

Sample characteristics

The subject characteristics for the 218 high Opium users who were diagnosed as OUD positive (138 subjects) or OUD negative (80 subjects) are provided in Table 1. For these study samples, the OUD diagnosis was associated at $p < 0.1$ with age at the time of enrollment ($p = 0.029$, OUD positive were 2 years younger than OUD negative), and route of opium exposure ($p = 0.054$, higher by inhalation than by ingestion), but was not associated with BMI, gender, or tobacco use.

Table 1 Subject characteristics of high opium users diagnosed as OUD positive and high opium users diagnosed as OUD negative.

Characteristic	OUD positive ($n = 138$)	OUD negative ($n = 80$)	p-value[2]
Age at enrollment (yrs), mean (SD) [range]	49.0 (6.1) [39.7, 67.5]	51.0 (6.6) [40.5, 68.6]	**0.029**
Male (count, %)	110 (79.7%)	62 (77.5%)	0.700
Tobacco smoking status			0.141
Current smoker (count, %)	76 (55.1%)	37 (46.2%)	
Former smoker (count, %)	11 (8.0%)	13 (16.3%)	
Never smoker (count, %)	51 (36.9%)	30 (37.5%)	
Opium use, maximum nokhods/week, mean (SD) [range][1]	31.0 (16.8) [12.0, 105.0]	29.8 (23.8) [12.0, 168.0]	0.698
Ever used alcohol			0.790
Yes	35 (25.4%)	19 (23.8%)	
No	103 (74.6%)	61 (76.3%)	
Body mass index, mean (SD) [range]	23.5 (4.2) [15.5, 37.5]	24.2 (4.7) [15.6, 37.3]	0.230
Route of opium administration			**0.054**
Inhalation	73 (52.9%)	53 (66.3%)	
Ingestion	65 (47.1%)	27 (33.7%)	
Severity of opioid use disorder, DSM-5			
Absent	0	80 (100%)	
Mild	65 (47.1%)	0	
Moderate	43 (31.2%)	0	
Severe	30 (21.7%)	0	

[1]Nokhod is the local measurement for the amount of opium used, and is equivalent to approximately 0.2 grams (*42*). The sample of 218 opium users was selected from 430 opium users with the following distribution of maximum nokhods per week: 119 subjects had low opium use (0.3–3.0), 93 subjects had moderate opium use (3.5–10.5), and 218 subjects had high opium use (≥ 12.0).
[2]Bold values indicate statistical significance $P < 0.1$.

Metabolic profiles of high opium user diagnosed as OUD positive versus high opium user diagnosed as OUD negative

Over 7714 UPLC-HRMS signals were obtained after data preprocessing. Hypothesis testing and fold change

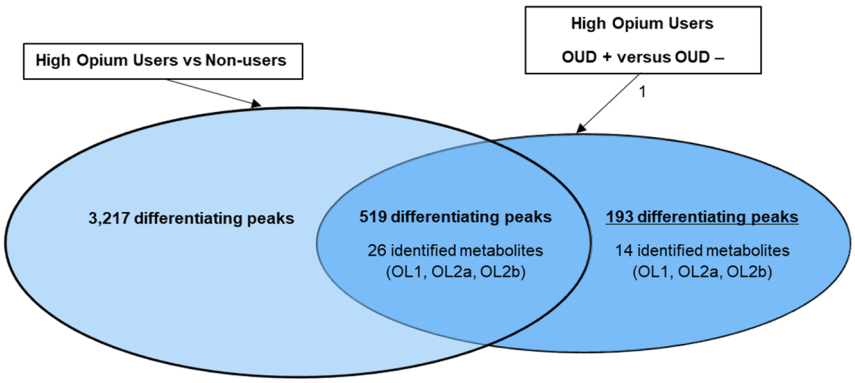

Fig. 1 Over 3700 peaks differentiated high opium users from non-opium users. 712 peaks differentiated high opium users diagnosed as OUD positive from high opium users diagnosed as OUD negative. 193 peaks were unique to the differentiation OUD positive versus OUD negative high opium users. Metabolites were identified or annotated using an in-house physical standards library, and peaks were annotated using big data analytics.

were determined for the normalized peaks in the metabolomics data set for the comparison of (a) high opium users diagnosed as OUD positive vs high opium users diagnosed as OUD negative, and (b) opium users vs non-opium users. Over 700 peaks (712) tested different by *t*-test ($p < 0.10$) between high opium users diagnosed as OUD positive versus high opium users diagnosed as OUD negative (Fig. 1). Forty of the 712 peaks were identified or annotated through matching to the in-house physical standards library (Table 2), while additional peaks were annotated using big data analytics (Table S1).

Pathway enrichment

Pathway enrichment was conducted in Metaboanalyst[14] using all 7714 peaks. A cut-off for pathway significance ($p < 0.01$) was used to determine the size of the permutation group that the algorithm used to determine the enrichment between high opium users diagnosed as OUD positive versus high opium users diagnosed as OUD negative. The candidate pathways based on the match of exact mass (<3 ppm) of key metabolites that are included in the known pathway map are provided in Table S2. The distribution plot of the Enrichment Factor versus −log10 (*P*) is shown in Fig. 2. High opium users diagnosed as OUD positive versus those diagnosed as OUD negative had an enrichment for pathways involving biotin (vitamin B7), folate (vitamin B9), cytochrome P450 metabolism, purine metabolism, keratan sulfate degradation, N-glycan degradation, and R group synthesis. *Vitamin* absorption, bioavailability, and utilization are known to be impacted by drug addiction[15]. *Cytochrome P450s* are involved in the metabolism of opium, and the slow versus fast metabolism has been associated with addiction[16]. Opioid use has been shown to alter *purine* metabolism[17]. *Keratan sulfate* is a glycosaminoglycan that is at significant levels in central and peripheral nervous systems[18]. *N-glycan* is required to express the correctly folded form of the delta-opioid receptor[19]. R group synthesis is associated with the FAD/FADH2 conversion of fatty acids.

Modeling approach 1

Stepwise logistic regression was used to determine which of the 712 peaks were predictive of an OUD positive diagnosis. First, the area under the ROC curve (AUC) was calculated using the subject characteristics (Table 1) of age at the time of enrollment and route of opium use. This base model resulted in an AUC of 0.625 (Figure S1a). Second, all 712 peaks that differentiated ($p < 0.10$) the high opium users diagnosed as OUD positive from high opium users diagnosed as OUD negative was modeled without including subject characteristics. This resulted in an AUC of 0.720 which was significantly different ($p = 0.042$) from the base model. Third, all 712 peaks that differentiated ($p < 0.10$) the high opium users diagnosed as OUD positive from high opium users diagnosed as OUD negative was modeled with age at the time of enrollment and opium use as covariates. This resulted in an AUC of 0.946, which was significantly increased ($p < 0.0001$) over the base model. Using this modeling approach, only 16 peaks were selected that were predictive of an OUD positive diagnosis (Table 3). Two of the 16 peaks matched to pterin (OL1) and tryptophan (OL2b) using the in-house physical standards library. Annotations using public databases are provided for 6 additional peaks, while 8 of the peaks remained unknown unknowns.

Modeling approach 2

Forty of the 712 peaks that differentiated ($p < 0.10$) high opium users diagnosed as OUD positive from high opium users diagnosed as OUD negative could be matched to the in-house physical standards library (Table 2, Fig. 1). Major differentiators included metabolites derived from opium

Table 2 Signals that differentiated ($p < 0.10$) high OUD positive opium users from high OUD negative opium users (Fig. 1) matched to 40 metabolites in the in-house physical standards library.

Ontology	Metabolite	*p*-Value	Fold change	Derivation[a]
OL1[b]	**Pterine**	0.001	1.3 (+)[c]	Vitamin B9 metabolism
OL2A	Deoxyadenosine	0.002	1.6 (+)	Purine metabolism
OL1	Morphine-6-beta-D-glucuronide	0.002	1.6 (+)	Opium use
OL1	Morphine-3-beta-D-glucuronide	0.004	1.6 (+)	Opium use
OL2B	Naloxone-3-beta-D-glucuronide	0.005	2.4 (+)	Opium use
OL1	Morphine	0.005	1.5 (+)	Opium use
OL2B	Octopamine	0.006	1.5 (+)	Tryptophan metabolism
OL1	Cotinine	0.006	1.6 (+)	Tobacco use
OL1	Codeine	0.007	1.4 (+)	Opium use
OL1	Codeine-6-beta-D-glucuronide	0.013	1.4 (+)	Opium use
OL2B	Morphine-3-beta-D-glucuronide	0.015	1.4 (+)	Opium use
OL2B	Serine	0.017	1.7 (−)	Amino acid metabolism
OL2B	Morphine	0.019	1.5 (+)	Opium use
OL2A	**6-Carboxyhexonate**	0.021	1.3 (−)	Fatty acid metabolism
OL2A	**N-Acetyl-S-(3,4-dihydroxybutyl)-L-cysteine**	0.023	3.5 (−)	Butadiene and acrylamide
OL1	**Sarcosine**	0.026	1.2 (+)	Amino acid methylation
OL2A	Codeine isomer or derivative	0.030	1.4 (+)	Opium use
OL1	Hydroxycotinine	0.034	1.4 (+)	Tobacco use
OL1	N-Acetyl-DL-tryptophan	0.035	1.2 (−)	Amino acid acetylation
OL2B	**Tryptophan**	0.038	1.3 (−)	Tryptophan metabolism
OL1	Codeine	0.040	1.3 (+)	Opium use
OL2B	Dihydromorphine	0.042	1.5 (+)	Opium use
OL1	N-Acetylcystine	0.048	1.1 (−)	Amino acid acetylation
OL1	**Mono-isobutyl phthalate**	0.049	2.1 (+)	Environmental exposure
OL1	**N-Methyl-D-aspartic acid**	0.049	1.2 (+)	Amino acid methylation
OL2B	**Lauroylcarnitine**	0.057	1.9 (−)	Carnitine metabolism
OL2A	Kynurenine	0.060	1.4 (−)	Tryptophan metabolism
OL1	Nicotine	0.068	1.5 (+)	Tobacco use
OL2B	N-Acetylproline	0.069	1.3 (−)	Amino acid acetylation
OL1	**2,4-Dihydroxypteridine**	0.069	1.2 (+)	Vitamin B9 metabolism
OL1	**Azelate**	0.071	1.4 (−)	Fatty acid oxidation
OL2B	N-Acetylcysteine	0.072	1.2 (−)	Amino acid acetylation
OL2A	Creatinine	0.079	1.1 (−)	Amino acid metabolism
OL2A	N-Acetylproline	0.081	1.1 (−)	Amino acid acetylation
OL2B	**Glycocholate**	0.082	1.4 (+)	Bile acid metabolism
OL1	N-Acetyl-S-(2-carbamoylethyl)-L-cysteine	0.084	1.2 (−)	Butadiene and acrylamide
OL2B	**Mono ethyl hexyl phthalate**	0.085	1.4 (+)	Environmental exposure

Table 2 continued

Ontology	Metabolite	*p*-Value	Fold change	Derivation[a]
OL2B	**3-Methylhistamine**	0.087	1.4 (−)	Amino acid methylation
OL2A	N-Acetylphenylalanine	0.094	1.2 (−)	Amino acid acetylation
OL1	**Phosphorylcholine**	0.096	1.9 (+)	Choline metabolism

Fourteen of these metabolites (bold) were unique to the differentiation ($p < 0.10$) of high opium users who were diagnosed as OUD positive versus OUD negative.
[a]Derivation: Metabolites were derived from endogenous metabolism, opium use, tobacco use, or environmentally related exposure.
[b]Ontology: OL1, highly confident identification based on matching with in-house physical standard library (IPSL) via retention time (RT, with RT error ≤|0.5| min), exact mass (MS, with mass error <5 ppm), and tandem mass similarity based on experimental fragmentation spectra (experimental MS/MS, with similarity ≥30); OL2a, confident identification based on matching with IPSL via MS and RT; OL2b, annotation for the isomer or derivatives of the compound listed, based on matching with IPSL via MS and MS/MS.
[c]Direction of change. +/−, increased/decreased in OUD positive.

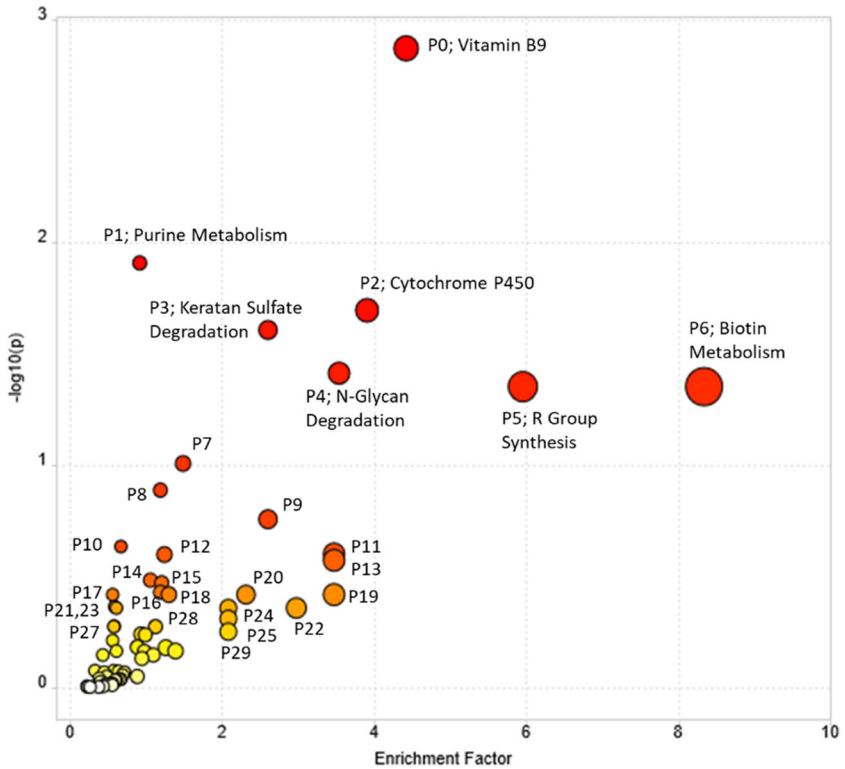

Fig. 2 The distribution plot of the Enrichment Factor versus −log10(P) of the pathway enrichment analysis. Mummichog was used to evaluate the pathway enrichment of all features (*m/z*) with $p < 0.01$ based on the *t*-test for the comparison of high opium users who were diagnosed as OUD positive versus high opium users who were diagnosed OUD as negative was used as the threshold to determine the size of the permutation groups used by the algorithm.

use, tobacco use, involved in biopterins and vitamin B9, tryptophan metabolism, acetylation of amino acids, bile acids, fatty acids, and carnitine metabolism. In addition, *N*-Acetyl-S-(3,4-dihydroxybutyl)-L-cysteine ($p = 0.023$) and *N*-Acetyl-S-(2-carbamoylethyl)-L-cysteine ($p = 0.083$) were lower in urine of high opium users diagnosed as OUD positive vs high opium users diagnosed as OUD negative. N-Acetyl-S-(3,4-dihydroxybutyl)-L-cysteine and N-Acetyl-S-(2-carbamoylethyl)-L-cysteine ($p = 0.083$) are metabolic products of butadiene[20] (BD) and acrylamide[21] (AM), respectively. These metabolites have previously been detected in biospecimens from tobacco users at significantly higher levels than non-tobacco users. They are attributed to the metabolism of the parent compounds (AM and BD) that form during the curation process, or on combustion of tobacco[22]. They could also be formed in the curing and combustion of opium or other plant material.

Stepwise logistic regression using these 40 metabolites resulted in an AUC of 0.76, which was significantly increased ($p = 0.0049$) over the base model (Figure S1b).

Table 3 Metabolites predictive of a positive OUD diagnosis in high opium users.

Ontology[a]	Model 1[b]	Model 2[b]	Model 3
	16 peaks (Table S3)	*8 metabolites* (Table S4)	*5 metabolites* (Table S5)
OL1	Pterine (+)[c]	Pterine (+)	Pterine (+)
OL1	–	Sarcosine (+)	Sarcosine (+)
OL2b	Tryptophan (−)	Tryptophan (−)	Tryptophan (−)
OL1	–	Azelate (−)	Azelate (−)
OL2b	–	N-acetylproline (−)	–
OL2b	–	Octopamine (+)	–
OL2b	–	Serine (−)	–
OL1	–	Nicotine (+)	–
OL2a	–		N-Acetyl-dihydroxybutyl-cysteine (−)
PDd[d]	2-Polyprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinone	–	–
PDa	5α-androst-16-en-3α-ol	–	–
PDa	L-Tyrosinamide	–	–
PDb	D-1-[(3-Carboxypropyl)amino]-1-deoxyfructose	–	–
PDc	Alanyl-Proline	–	–
PDc	Phlorisobutyrophenone 2-glucoside	–	–

Age at time of enrollment and route of opium use were covariates in the model. Model 1 used 712 peaks that differentiated high opium users diagnosed as OUD positive from high opium users diagnosed as OUD negative. Models 2 and 3 included peaks matched using the in-house physical standards library.
[a]Ontology: OL1, highly confident identification based on matching with in-house physical standard library (IPSL) via retention time (RT, with RT error ≤|0.5| min), exact mass (MS, with mass error <5 ppm), and tandem mass similarity based on experimental fragmentation spectra (experimental MS/MS, with similarity ≥30); OL2a, confident identification based on matching with IPSL via MS and RT; OL2b, annotation for the isomer or derivatives of the compound listed, based on matching with IPSL via MS and experimental MS/MS.
[b]Eight of the 16 peaks that predicted OUD under Model 1 were identified or annotated, while 8 peaks (not listed) remained unknown.
[c]Direction of change. +/−, increased/decreased in OUD positive.
[d]PD, Public Data Base. PDa, annotation based on matching with PD via MS and experimental MS/MS (could be the listed compound, or the isomer or derivatives of the listed compound); PDb, annotation based on matching with public database via MS and predict MS/MS; PDc, annotation for the listed compound based on matching with public database via MS and isotopic similarity or adducts; PDd, annotation for the listed compound based on matching with public database via MS.

Including subject characteristics of age at the time of enrollment and route of opium exposure resulted in an AUC of 0.80, also significantly ($p < 0.0001$) increased from the base model. Metabolites that were predictive of an OUD positive diagnosis (Table 3) included tryptophan, pterine, sarcosine, N-acetylproline, azelate, octopamine, serine, and nicotine.

Metabolic profiles unique to the OUD positive versus OUD negative diagnosis

Five hundred and nineteen of the 712 peaks that tested different between OUD positive high opium users versus high OUD negative high opium users, also differentiated the opium users from non-opium users (Fig. 1). To provide a focus on only metabolites that are important to the diagnosis of OUD, the 519 signals that were also important to differentiation of opium users from non-opium users were excluded for this analysis. This resulted in 193 peaks unique to the differentiation ($p < 0.10$) of subjects diagnoses as OUD positive versus those diagnoses as OUD negative. Of these 193 peaks, only 14 peaks that defined the OUD diagnosis matched to the in-house physical standards library. These 14 metabolites are listed in Table 2, while the additional annotated peaks through public databases are provided in Table S1.

Eleven of the 14 peaks that matched to the in-house library that were most important to defining OUD included the following endogenous metabolites: pterine ($p = 0.0011$), 2,4-dihydroxypterine ($p = 0.0695$), sarcosine ($p = 0.0263$), phosphorylcholine ($p = 0.0962$), 6-carboxyhexonate ($p = 0.021$), lauroylcarnitine (0.0574), glycocholate ($p = 0.0816$), 3-methylhistamine (0.087), azelate ($p = 0.0713$), n-methyl-D-aspartic acid ($p = 0.0488$), and tryptophan (0.0378).

The biological significance of these 11 endogenous metabolites is summarized:

(a) *Pterin* is part of biopterin and folate. Biopterins are cofactors for aromatic amino acid hydroxylases, which are involved in the synthesis of dopamine, norepinephrine, epinephrine, and serotonin, and trace amines[23]. The active form of folate (vitamin B9) is tetrahydrofolate which accepts and donates

one carbon unit (methyl group). *Dihydroxypteridine* is involved in folate and riboflavin pathways[24].

(b) *Phosphorylcholine* is derived from phosphorylation of choline[25], and *sarcosine* is an intermediate in the metabolism of choline to glycine[26].

(c) *N-Methyl-d-aspartic acid* (NMDA) is an agonist at the NMDA receptor and mimics the action of glutamate[27], and *tryptophan* is in the neurotransmitter pathway[28]. *Azelaic acid* (AZA) is a competitive inhibitor of tyrosinase in vitro[29].

(d) *Lauroylcarnitine* is associated with fatty oxidation disorders involving acyl CoA dehydrogenase deficiency, and carnitine palmitoyltransferase I and II deficiency[30]. *6-Carboxyhexanoic acid* is a medium-chain fatty acid derived from heptanedioic acid and is involved in the gut microbial biosynthesis of biotin[31].

(e) *Glycocholate* is a secondary bile acid, produced in the microbial flora of the colonic environment by bacteria[32], and is absorbed and recirculated. Bile acids are important for absorption of hydrophobic nutrients, dietary fats and vitamins, and the regulation enzymes involved in cholesterol homeostasis.

(f) *3-Methylhistamine* is a prominent metabolite of histamine, which has a role in allergy, inflammation, gastric acid secretion, and neurotransmission[33].

Three metabolites (Table 2) derived from exogenous exposures were also important to the differentiation of the high opium users who were diagnosed as OUD positive from high opium users who were diagnosed as OUD negative. These included mono-isobutyl phthalate ($p = 0.0485$, +) and mono ethyl hexyl phthalate ($p = 0.0852$, +), which could arise as metabolic products following ingestion of phthalates that leach from plastics used in inhalation of opium. N-Acetyl-S-(3,4-dihydroxybutyl)-L-cysteine was also a differentiator ($p = 0.0228$, −), and is presumably derived as a metabolic product of BD intake associated with the curing or combustion of plant matter.

Modeling approach 3

Stepwise logistic regression using the 14 metabolites unique to the differentiation of OUD positive versus OUD negative, together with covariates of age at the time of enrollment and route of opium use resulted in an AUC of 0.751, which was significantly increased ($p < 0.0005$) over the base model (Figure S1c). Stepwise logistic regression using only the 14 metabolites (with no subject characteristics) resulted in an ACU of 0.706, which was not significantly increased ($p = 0.127$) over the base model. Results from Model 3 (with or without the covariates of age at time of enrollment and route of opium use) show 5 of the 14 metabolites (pterine, sarcosine, tryptophan, azelate, and N-Acetyl-S-(3,4-dihydroxybutyl)-L-cysteine) as predictive of a positive OUD (Table 3).

Discussion

Our study revealed metabolomics signatures of OUD in a predominantly Turkmen population of chronic high opium users. We provide metabolite identifications and annotations for 712 features detected using untargeted mass spectrometry that are important to the differentiation of high opium users diagnosed as OUD positive from high opium users diagnosed as OUD negative. None of these identifications are known metabolites derived from other drugs of abuse. Pathway enrichment analysis points to a general disruption in vitamin B9 (folate), vitamin B7 (biotin), cytochrome P450, purine, and glycan metabolism, and FAD/FADH2 conversion of fatty acids.

Stepwise logistic regression analysis of these 712 peaks, together with subject characteristics, resulted in 16 candidate peaks that predict 95% of the high opium users who were diagnosed as OUD positive.

Forty of the 712 features which differentiate high opium users who were diagnosed as OUD positive from high opium users who were diagnosed as OUD negative matched to an in-house physical standards library. Models constructed with only these 40 metabolites predicted 80% of the subjects diagnosed as OUD positive, selecting 8 metabolites as predictors. Predictors of an OUD diagnosis in these high opium users included an increase in three endogenous compounds (pterine, sarcosine, and octopamine), a decrease in four endogenous compounds (tryptophan, azelate, N-acetylproline, and serine), and an increase in nicotine. Fourteen metabolites were determined to be unique to OUD diagnosis, after subtracting analytes known to overlap with opium use. Models using these 14 metabolites predicted 75% of the subjects testing OUD positive and replicated pterine, sarcosine, tryptophan, and azelate as metabolite predictors.

Many identified or annotated metabolites that differentiated high opium users who were OUD positive from high opium users who were OUD negative play a significant role in neurotransmitter synthesis and signal transduction[34]. Tryptophan is the major amino acid precursor of serotonin (5HT). 5HT deficits have been implicated in physical symptoms and emotional dysphoria following withdrawal from opioids[35]. Alterations in sarcosine, serine, kyneurate, NMDA found in this study are consistent with the observations that glutamatergic signaling is disrupted by opioids[36]. Sarcosine (methyl-glycine) acts as an NMDA receptor agonist and a glycine receptor agonist[37]. Serine is converted to D-serine by serine racemace. D-serine acts as co-agonist with glutamate to activate NMDA receptors[38]. Kyneurate a metabolite of tryptophan metabolized to quinolinic acid acts as a NMDA receptor agonist[39]. Octopamine is a trace amine that is an agonist of TAAR1 receptors implicated in mediating the actions of drugs of abuse[40]. Methylhistamine is a histamine receptor (H3) agonist that inhibits the

firing of cholinergic neurons in the ventral striatum and decreases dopamine release[41].

Limitations of the current study include the (a) use of self-report for the amount of opium consumed, (b) assumption that symptoms over any 12-month period-of-time are accurately recalled, (c) estimated amount/grams of opium may vary within or among regions, and be underestimated[42], (d) analyses were not stratified by the route of administration or type of opiate used due to the sample size, and because this paper focuses on a marker of OUD independent of route, (e) that the current study was not powered for multiple testing, and (f) that the metabolomic profiles are not quantitated, or replicated in this cohort or across cohorts.

The year that the individuals met a DSM-5 OUD diagnosis during their history of opium use is unknown because the DSM-5 interview was not conducted at the time of baseline urine collection. These urinary baseline metabolomic profiles presented herein could result from chronic opium use, and/or from inherent individual metabolic differences present prior to the acquisition of OUD.

Chronic use of opiates and opioids without meeting the criteria for DSM-5 OUD is not unique to the Turkman population for opiate use. Some chronic pain patients treated with prescription opioids and chronic users of illicit opioids do not meet the criteria for DSM-5 OUD. This suggests that the approaches used in this study are likely to be generalizable to other cohorts. This is also consistent with other substance use disorders where heavy use does not necessarily imply a substance used disorder[43].

Research on biomarkers for OUD and other substance use disorders has focused on neuroimaging (MRI, fMRI, and PET) and EEG studies[44]. While these biomarkers may eventually be clinically validated in other populations, they will be costly to implement. In contrast, validation of biomarkers in other populations in accessible biological fluids (e.g., urine, blood, saliva) will be less costly, and easier to implement in general medical practice. In addition, these non-invasive biomarkers will be important complements to results from neuroimaging studies.

In conclusion, if the current results are replicated, the identification of peripheral biomarkers for OUD would represent a significant advancement in defining and managing the disease. It would further validate the Dole and Nyswander hypothesis that OUD is a brain disease in which metabolism is disrupted, and would provide biomarkers for OUD that could be used to optimize treatment. In addition, validation of the discovered metabolic perturbations related to vitamins and fatty acids could lead to the development of a nutrient cocktail to test in clinical settings for efficacy to mitigate symptoms that lead to the diagnosis of OUD.

Acknowledgements
Dr. Blake Rushing and Dr. Krissy Kay contributed to quality control review of figures and tables. Mr. Justin Chandler assisted with literature review, and editorial review. The metabolomics platform was developed with funding in part by the NIH Common Fund Phase 1 Program (1U24DK097193, Sumner PI), and the NIEHS CHEAR program (U2CES026544, Fennell PI). This metabolomics investigation was funded in part through the NIDA Invest Fellowship (Dr. Reza Ghanbari), and the Sumner-Lab. The Golestan Cohort Study was supported in part by Tehran University of Medical Sciences (grant no. 81/15); Cancer Research UK (grant no. C20/A5860); the Intramural Research Program of the NCI, NIH; and various collaborative research agreements with IARC.

Author details
[1]Department of Nutrition, Nutrition Research Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Digestive Oncology Research Center, Digestive Diseases Research Institute, Tehran University of Medical Science, Tehran, Iran. [3]Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), Bethesda, MD, USA. [4]Genetics, Epigenetics, and Developmental Neuroscience Branch, National Institute on Drug Abuse (NIDA), Bethesda, MD, USA. [5]Iranian National Center for Addiction Studies (INCAS), Tehran University of Medical Sciences (TUMS), Tehran, Iran. [6]Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [7]Golestan Research Center of Gastroenterology and Hepatology, Golestan University of Medical Sciences, Gorgan, Iran

Conflict of interest
The authors declare that they have no conflict of interest.

Disclosure
The views and opinions expressed in this manuscript are those of the authors only and do not necessarily represent the views, official policy or position of the U.S. Department of Health and Human Services or any of its affiliated institutions or agencies.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information
The online version contains supplementary material available at https://doi.org/10.1038/s41398-021-01228-7.

Received: 12 August 2020 Revised: 8 January 2021 Accepted: 18 January 2021
Published online: 04 February 2021

References

1. Dole, V. P. & Nyswander, M. E. Heroin addiction—a metabolic disease. *Arch. Intern. Med.* **120**, 19–24 (1967).
2. Volkow, N. D. & Blanco, C. Medications for opioid use disorders: clinical and pharmacological considerations. *J. Clin. Invest.* **130**, 10–13 (2020).
3. Strang, J. et al. Opioid use disorder. *Nat. Rev. Dis. Prim.* **6**, 1–28 (2020).
4. Ruffle, J. K. Molecular neurobiology of addiction: what's all the (Δ) FosB about? *Am. J. Drug Alcohol Abus.* **40**, 428–437 (2014).
5. Bough, K. J. & Pollock, J. D. Defining substance use disorders: the need for peripheral biomarkers. *Trends Mol. Med.* **24**, 109–120 (2018).
6. Hasin, D. S. et al. DSM-5 criteria for substance use disorders: recommendations and rationale. *Am. J. Psychiatry* **170**, 834–851 (2013).
7. Amin-Esmaeili, M. et al. Epidemiology of illicit drug use disorders in Iran: prevalence, correlates, comorbidity and service utilization results from the Iranian Mental Health Survey. *Addiction* **111**, 1836–1847 (2016).
8. Khademi, H. et al. Opium use and mortality in Golestan Cohort Study: prospective cohort study of 50,000 adults in Iran. *BMJ* **344**, e2502 (2012).
9. Gorfinkel, L., Voon, P., Wood, E. & Klimas, J. Diagnosing opioid addiction in people with chronic pain. *BMJ* **362**, k3949. https://doi.org/10.1136/bmj.k3949 (2018).

10. Pourshams, A. et al. Cohort Profile: The Golestan Cohort Study-a prospective study of oesophageal cancer in northern Iran. *Int. J. Epidemiol.* **39**, 52–59 (2010).
11. Want, E. J. et al. Global metabolic profiling procedures for urine using UPLC–MS. *Nat. Protoc.* **5**, 1005–1018 (2010).
12. Xi, B., Gu, H. & Baniasadi, H. & Raftery, D. in *Mass Spectrometry in Metabolomics* 333–353 (Springer, 2014).
13. Bender, R. & Lange, S. Adjusting for multiple testing—when and how? *J. Clin. Epidemiol.* **54**, 343–349 (2001).
14. Chong, J. et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **46**, W486–W494 (2018).
15. Grotzkyj-Giorgi, M. Nutrition and addiction-can dietary changes assist with recovery? *Drugs Alcohol Today* **9**, 24 (2009).
16. Kramlinger, V. M., Rojas, M. A., Kanamori, T. & Guengerich, F. P. Cytochrome P450 3A enzymes catalyze the O6-demethylation of thebaine, a key step in endogenous mammalian morphine biosynthesis. *J. Biol. Chem.* **290**, 20200–20210 (2015).
17. Mannelli, P. et al. Opioid use affects antioxidant activity and purine metabolism: preliminary results. *Hum. Psychopharmacol.* **24**, 666–675 (2009).
18. Caterson, B. & Melrose, J. Keratan sulfate, a complex glycosaminoglycan with unique functional capability. *Glycobiology* **28**, 182–206 (2018).
19. Markkanen, P. M. & Petäjä-Repo, U. E. N-glycan-mediated quality control in the endoplasmic reticulum is required for the expression of correctly folded δ-opioid receptors at the cell surface. *J. Biol. Chem.* **283**, 29086–29098 (2008).
20. Carmella, S. G. et al. Effects of smoking cessation on eight urinary tobacco carcinogen and toxicant biomarkers. *Chem. Res. Toxicol.* **22**, 734–741 (2009).
21. Bjellaas, T., Janak, K., Lundanes, E., Kronberg, L. & Becher, G. Determination and quantification of urinary metabolites after dietary exposure to acrylamide. *Xenobiotica* **35**, 1003–1018 (2005).
22. Chang, C. M. et al. Biomarkers of tobacco exposure: summary of an FDA-sponsored public workshop. *Cancer Epidemiol. Biomark. Prev.* **26**, 291–302 (2017).
23. Nagatsu, T. & Ichinose, H. Regulation of pteridine-requiring enzymes by the cofactor tetrahydrobiopterin. *Mol. Neurobiol.* **19**, 79–96 (1999).
24. Brown, E. G. *Ring Nitrogen and Key Biomolecules: The Biochemistry of N-heterocycles* (Springer Science & Business Media, 2012).
25. Jimenez, B., del Peso, L., Montaner, S., Esteve, P. & Lacal, J. C. Generation of phosphorylcholine as an essential event in the activation of Raf-1 and MAP-kinases in growth factors-induced mitogenic stimulation. *J. Cell. Biochem.* **57**, 141–149 (1995).
26. Ducker, G. S. & Rabinowitz, J. D. One-carbon metabolism in health and disease. *Cell Metab.* **25**, 27–42 (2017).
27. Paoletti, P. & Neyton, J. NMDA receptor subunits: function and pharmacology. *Curr. Opin. Pharm.* **7**, 39–47 (2007).
28. Gheorghe, C. E. et al. Focus on the essentials: tryptophan metabolism and the microbiome-gut-brain axis. *Curr. Opin. Pharm.* **48**, 137–145 (2019).
29. Passi, S., Picardo, M., Mingrone, G., Breathnach, A. S. & Nazzaro-Porro, M. Azelaic acid-biochemistry and metabolism. *Acta Derm.-venereologica. Supplementum* **143**, 8–13 (1989).
30. Yamada, K. & Taketani, T. Management and diagnosis of mitochondrial fatty acid oxidation disorders: focus on very-long-chain acyl-CoA dehydrogenase deficiency. *J. Hum. Genet.* **64**, 73–85 (2019).
31. Estrada, P. et al. The pimeloyl-CoA synthetase BioW defines a new fold for adenylate-forming enzymes. *Nat. Chem. Biol.* **13**, 668–674 (2017).
32. Tanaka, H., Doesburg, K., Iwasaki, T. & Mierau, I. Screening of lactic acid bacteria for bile salt hydrolase activity. *J. Dairy Sci.* **82**, 2530–2535 (1999).
33. Lieberman, P. The basics of histamine biology. *Ann. Allergy, Asthma Immunol.* **106**, S2–5 (2011).
34. Dinis-Oliveira, R. J. Metabolism and metabolomics of opiates: a long way of forensic implications to unravel. *J. Forensic Leg. Med.* **61**, 128–140 (2019).
35. Welsch, L., Bailly, J., Darcq, E. & Kieffer, B. L. The negative affect of protracted opioid abstinence: progress and perspectives from rodent models. *Biol. Psychiatry* **87**, 54–63 (2020).
36. Hearing, M., Graziane, N., Dong, Y. & Thomas, M. J. Opioid and psychostimulant plasticity: targeting overlap in nucleus accumbens glutamate signaling. *Trends Pharm. Sci.* **39**, 276–294 (2018).
37. Zhang, H. X., Lyons-Warren, A. & Thio, L. L. The glycine transport inhibitor sarcosine is an inhibitory glycine receptor agonist. *Neuropharmacology* **57**, 551–555 (2009).
38. Wolosker, H. The neurobiology of d-serine signaling. *Adv. Pharmacol.* **82**, 325–348 (2018).
39. Schwarcz, R., Bruno, J. P., Muchowski, P. J. & Wu, H. Q. Kynurenines in the mammalian brain: when physiology meets pathology. *Nat. Rev. Neurosci.* **13**, 465–477 (2012).
40. Liu, J. F. & Li, J. X. TAAR1 in addiction: looking beyond the tip of the iceberg. *Front Pharm.* **9**, 279 (2018).
41. Varaschin, R. K. et al. Histamine H3 receptors decrease dopamine release in the ventral striatum by reducing the activity of striatal cholinergic interneurons. *Neuroscience* **376**, 188–203 (2018).
42. Mohebbi, E. et al. An exploratory study of units of reporting opium usein Iran: implications for epidemiologic studies. *Arch. Iran. Med.* **22**, 541–545 (2019).
43. Degenhardt, L. et al. Agreement between definitions of pharmaceutical opioid use disorders and dependence in people taking opioids for chronic non-cancer pain (POINT): a cohort study. *Lancet Psychiatry* **2**, 314–322 (2015).
44. Moningka, H. et al. Can neuroimaging help combat the opioid epidemic? A systematic review of clinical and pharmacological challenge fMRI studies with recommendations for future research. *Neuropsychopharmacology* **44**, 259–273 (2019).

HHS Public Access

Author manuscript

Lancet Digit Health. Author manuscript; available in PMC 2022 November 11.

Published in final edited form as:

Lancet Digit Health. 2022 June ; 4(6): e406–e414. doi:10.1016/S2589-7500(22)00063-2.

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya,

Imon Banerjee,

Ananth Reddy Bhimireddy,

John L Burns,

Leo Anthony Celi,

Li-Ching Chen,

Ramon Correa,

Natalie Dullerud,

Marzyeh Ghassemi,

Shih-Cheng Huang,

Po-Chih Kuo,

Matthew P Lungren,

Correspondence to: Dr Judy Wawira Gichoya, Department of Radiology, Emory University, Atlanta, GA 30322, USA, judywawira@emory.edu.

Contributors

IB was responsible for the conceptualisation of the study, data curation from Emory, supervision of trainees, as well as writing, reviewing, and editing the manuscript. ARB was responsible for training the race prediction model for the Digital Hand Atlas, which was supervised by SP, as well as reviewing the manuscript and preparing the code repository accompanying the manuscript under the supervision of JWG. JLB participated in writing and reviewing the manuscript. LAC was responsible for the overall study design, critical review of the manuscript, as well as synthesis of the results, literature review, and writing and reviewing the manuscript. LC conducted the experiments on the MIMIC-CXR dataset under supervision of PK, reported results, and reviewed the manuscript. RC prepared the Emory chest x-ray dataset under supervision of JWG and IB, as well as contributing to the literature review and the review of the manuscript. ND conducted the experiments on anatomic and phenotypic confounders, specifically on predictions based on age and sex. MG was responsible for the overall study design, supervision of experiments, results analysis, manuscript writing, and literature review. JWG was responsible for the overall study design, Emory datasets extraction and curation, design and supervision of experiments, results analysis, literature review, and manuscript writing. JWG also provided qualitative review of the saliency maps to evaluate any localising information. SH created the Stanford RSPECT dataset under supervision of MPL and conducted external validation of the CT chest prediction model. PK was responsible for designing and conducting experiments on the MIMIC-CXR dataset for race prediction, exploration of anatomic and phenotypic confounders (including body-mass index [BMI]), segmenting the dataset into lung and non-lung segments, noisy and blurred images, ablation experiments, as well as writing and reviewing the manuscript. MPL was responsible for supervising the creation of the RSPECT Stanford dataset, extracting race labels for the CheXpert dataset, conducting qualitative review of saliency maps, as well as participating in the literature review and writing of the manuscript. BJP, supervised by LO-R and JWG, trained race prediction models on Emory cervical spine radiographs, Emory chest x-ray, MIMIC-CXR, and CheXpert datasets. BJP also conducted experiments on anatomic and phenotype confounders, specifically on the effect of resolution change on prediction and BMI. ATP assisted manuscript preparation and review. SP participated in the overall study and experiment design, supervised ARB, JLB, and BJP, and helped edit the manuscript. LO-R trained the CT chest race prediction model on the NLST dataset and supervised BJP on experiments. LO-R also designed the experiments and summarised results from multiple experiments, as well as manuscript writing and review. LO-R also conducted qualitative review of saliency maps. CO prepared the Emory chest x-ray dataset, under the supervision of JWG and IB, and conducted race prediction experiments on this dataset. LS-K assisted with the design of the overall study and experiments, critical review of results, literature review, and manuscript writing. HT prepared the Emory cervical spine radiographs and mammogram datasets with IB and JWG, and also contributed to writing and editing the manuscript. RW conducted the experiments on the MIMIC-CXR dataset, under the supervision of PK, reported results, and also reviewed the manuscript. ZZ, under the supervision of IB and JWG, prepared the Emory CT dataset, conducted the external validation of the CT experiments on the Emory dataset, trained race prediction models on the Emory mammogram dataset, summarised results, and reviewed the manuscript. HZ conducted the experiments on high and low filter image manipulations and reviewed the manuscript. LJP assisted with overall study design, critical review of results, and manuscript writing. All authors had access to the datasets used in this study. JWG, PK, IB, and HT verified the data.

Lyle J Palmer,
Brandon J Price,
Saptarshi Purkayastha,
Ayis T Pyrros,
Lauren Oakden-Rayner,
Chima Okechukwu,
Laleh Seyyed-Kalantari,
Hari Trivedi,
Ryan Wang,
Zachary Zaiman,
Haoran Zhang

(J W Gichoya MD, A R Bhimireddy MS, H Trivedi MD) **and Department of Computer Science** (Z Zaiman), **Emory University, Atlanta, GA, USA; School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA** (I Banerjee PhD, R Correa BS); **School of Informatics and Computing, Indiana University–Purdue University, Indianapolis, IN, USA** (J L Burns MS, S Purkayastha PhD); **Institute for Medical Engineering and Science** (L A Celi MD, M Ghassemi PhD) **and Department of Electrical Engineering and Computer Science** (M Ghassemi), **Massachusetts Institute of Technology, Cambridge, MA, USA; Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA** (L A Celi); **Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan** (L-C Chen BS, P-C Kuo PhD, R Wang BS); **Department of Computer Science, University of Toronto, Toronto, ON, Canada** (N Dullerud MS, L Seyyed-Kalantari PhD, H Zhang MS); **Stanford University School of Medicine, Palo Alto, CA, USA** (S-C Huang, M P Lungren MD); **Australian Institute for Machine Learning** (L Oakden-Rayner MD, L J Palmer PhD) **and School of Public Health** (L J Palmer), **University of Adelaide, Adelaide, SA, Australia; Florida State University College of Medicine, Tallahassee, FL, USA** (B J Price MD); **Dupage Medical Group, Hinsdale, IL, USA** (A T Pyrros MD); **Department of Computer Science, Georgia Institute of Technology, Atlanta, GA, USA** (C Okechukwu MS); **Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, ON, Canada** (L Seyyed-Kalantari); **Vector Institute for Artificial Intelligence, Toronto, ON, Canada** (L Seyyed-Kalantari)

Summary

Background—Previous studies in medical imaging have shown disparate abilities of artificial intelligence (AI) to detect a person's race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images. We aimed to conduct a comprehensive evaluation of the ability of AI to recognise a patient's racial identity from medical images.

Methods—Using private (Emory CXR, Emory Chest CT, Emory Cervical Spine, and Emory Mammogram) and public (MIMIC-CXR, CheXpert, National Lung Cancer Screening Trial, RSNA Pulmonary Embolism CT, and Digital Hand Atlas) datasets, we evaluated, first, performance quantification of deep learning models in detecting race from medical images, including the ability of these models to generalise to external environments and across multiple

imaging modalities. Second, we assessed possible confounding of anatomic and phenotypic population features by assessing the ability of these hypothesised confounders to detect race in isolation using regression models, and by re-evaluating the deep learning models by testing them on datasets stratified by these hypothesised confounding variables. Last, by exploring the effect of image corruptions on model performance, we investigated the underlying mechanism by which AI models can recognise race.

Findings—In our study, we show that standard AI deep learning models can be trained to predict race from medical images with high performance across multiple imaging modalities, which was sustained under external validation conditions (x-ray imaging [area under the receiver operating characteristics curve (AUC) range 0·91–0·99], CT chest imaging [0·87–0·96], and mammography [0·81]). We also showed that this detection is not due to proxies or imaging-related surrogate covariates for race (eg, performance of possible confounders: body-mass index [AUC 0·55], disease distribution [0·61], and breast density [0·61]). Finally, we provide evidence to show that the ability of AI deep learning models persisted over all anatomical regions and frequency spectrums of the images, suggesting the efforts to control this behaviour when it is undesirable will be challenging and demand further study.

Interpretation—The results from our study emphasise that the ability of AI deep learning models to predict self-reported race is itself not the issue of importance. However, our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging.

Funding—National Institute of Biomedical Imaging and Bioengineering, MIDRC grant of National Institutes of Health, US National Science Foundation, National Library of Medicine of the National Institutes of Health, and Taiwan Ministry of Science and Technology.

Introduction

Bias and discrimination in artificial intelligence (AI) systems has been studied in multiple domains,[1–4] including in many health-care applications, such as detection of melanoma,[5,6] mortality prediction,[7] and algorithms that aid the prediction of health-care use,[8] in which the performance of AI is stratified by self-reported race on a variety of clinical tasks.[9] Several studies have shown disparities in the performance of medical AI systems across race. For example, Seyyed-Kalantari and colleagues showed that AI models produce significant differences in the accuracy of automated chest x-ray diagnosis across racial and other demographic groups, even when the models only had access to the chest x-ray itself.[9] Importantly, if used, such models would lead to more patients who are Black and female being incorrectly identified as healthy compared with patients who are White and male. Moreover, racial disparities are not simply due to under-representation of these patient groups in the training data, and there exists no statistically significant correlation between group membership and racial disparities.[10]

In related work, several groups reported that AI algorithms can identify various demographic patient factors. One study[11] found that an AI model could predict sex and distinguish between adult and paediatric patients from chest x-rays, while other studies[12] reported

reasonable accuracy at predicting the chronological age of patients from various imaging studies. In ophthalmology, retinal images have been used to predict sex, age, and cardiac markers (eg, hypertension and smoking status).[13–15] These findings, which show that demographic factors that are strongly associated with disease outcomes (eg, age, sex, and racial identity), are also strongly associated with features of medical images and might induce bias in model results, mirroring what is known from over a century of clinical and epidemiological research on the importance of covariates and potential confounding.[16,17] Many published AI models have conceptually amounted to simple bivariate analyses (ie, image features and their ability to predict clinical outcomes). Although more recent AI models have begun to consider other risk factors that conceptually approach multivariate modelling, which is the mainstay of clinical and epidemiological research, key demographic covariates (eg, age, sex, and racial identity) have been largely ignored by most deep learning research in medicine.

Findings regarding the possibility of confounding of racial identity in deep learning models suggest a possible mechanism for racial disparities resulting from AI models: that AI models can directly recognise the race of a patient from medical images. However, this hypothesis is largely unexplored[18] and, in contrast to other demographic factors (eg, age and sex), there is a widely held, but tacit, belief among radiologists that the identification of a patient's race from medical images is almost impossible, and that most medical imaging tasks are essentially race agnostic (ie, the task is not affected by the patient's race). Given the possibility for discriminatory harm in a key component of the medical system that is assumed to be race agnostic, understanding how race has a role in medical imaging models is of high importance[19] as many AI systems that use medical images as the primary inputs are being cleared by the US Food and Drug Administration and other regulatory agencies.[20–22]

In this study, we aimed to investigate how AI systems are able to detect a patient's race to differing degrees of accuracy across self-reported racial groups in medical imaging. To do so, we aimed to investigate large publicly and privately available medical imaging datasets to examine whether AI models are able to predict an individual's race across multiple imaging modalities, various datasets, and diverse clinical tasks.

Methods

Definitions of race and racial identity

Race and racial identity can be difficult attributes to quantify and study in health-care research[23] and are often incorrectly conflated with biological concepts (eg, genetic ancestry).[24] In this modelling study, we defined race as a social, political, and legal construct that relates to the interaction between external perceptions (ie, "how do others see me?") and self-identification, and specifically make use of self-reported race of patients in all of our experiments. We variously use the terms race and racial identity to refer to this construct throughout this study.

Datasets

We obtained public and private datasets (table 1, appendix p 2) that covered several imaging modalities and clinical scenarios. No one single race was consistently dominant across the datasets (eg, the proportion of Black patients was between 6% and 72% across the datasets). For all datasets, ethical approval was obtained from the relevant institutional ethical boards.

Investigation of possible mechanisms of race detection

We conduced three main groups of experiments to investigate the cause of previously established AI performance disparities by patient race. These experiments were: (1) to assess the ability of deep learning AI models to recognise race from medical images, including the ability of these models to generalise to new environments and across multiple imaging modalities; (2) to examine possible confounding anatomic and phenotype population features as explanations for these performance scores, and (3) to investigate the underlying mechanisms by which AI models can recognise race. The full list of experiments are summarised in table 2 and the appendix (pp 22–23).

We did not present measures of performance variance or null hypothesis tests because these data are uninformative given the large dataset sizes and the large effect sizes reported (ie, even in experiments in which a hypothesis could be defined, all p values were <0·001).

Race detection in radiology imaging

To investigate the ability of deep learning systems to detect race from radiology images, first, we developed models for the detection of racial identity on three large chest x-ray datasets—MIMIC-CXR (MXR),[25] CheXpert (CXP),[26] and Emory-chest x-ray (EMX) with both internal validation (ie, testing the model on an unseen subset of the dataset used to train the model) and external validation (ie, testing the model on a completely different dataset than the one used to train the model) to establish baseline performance. Second, we trained racial identity detection models for non-chest x-ray images from multiple body locations, including digital radiography, mammograms, lateral cervical spine radiographs, and chest CTs, to evaluate whether the model's performance was limited to chest x-rays.

After establishing that deep learning models could detect a patient's race in medical imaging data, we generated a series of competing hypotheses to explain how this process might occur. First, we assessed differences in physical characteristics between patients of different racial groups (eg, body habitus[27] or breast density[28]). Second, we assessed whether there was a difference in disease distribution among patients of different racial groups (eg, previous studies provide evidence that Black patients have a higher incidence of particular diseases, such as cardiac disease, than White patients).[29,30] Third, we assessed whether there were location-specific or tissue-specific differences (eg, there is evidence that Black patients have a higher adjusted bone mineral density and a slower age-adjusted annual rate of decline in bone mineral density than White patients).[31,32] Fourth, we assessed whether there were effects of societal bias and environmental stress on race outcomes from medical imaging data, as shown by differences in race detection by age and sex (reflecting cumulative and occupational differences in exposures). Last, we assessed whether there was an effect on the

ability of AI deep learning systems to detect race when multiple demographic and patient factors were combined, including age, sex, disease, and body habitus.

We also investigated potential explanations of race detection that could target the known shortcut mechanisms that deep models might be using as proxies for race[33] by evaluating, first, frequency domain differences in the high frequency image features (ie, textural) and low frequency image features (ie, structural) that could be predictive of race; second, how differences in image quality might influence the recognition of race in medical images (given the possibility that image acquisition practices might differ for patients with different racial identities); and, last, whether specific image regions contribute to the recognition of racial identity (eg, specific patches or regional variations in the images, such as radiographic markers in the top right corner).

Role of the funding source

Grant support was used to pay for data collection, data analysis, data interpretation, and writing of the manuscript. The funders did not influence the decision to publish or the target journal for publication.

Results

The deep learning models assessed in this study showed a high ability to detect patient race using chest x-ray scans, with sustained performance on other modalities and strong external validations across datasets (table 3).

The ability of deep learning models that were trained on the CXP dataset to predict patient race from the body-mass index (BMI) alone was much lower than the image-based chest x-ray models (area under the receiver operating characteristics curve [AUC] 0·55), indicating that race detection is not due to obvious anatomic and phenotypic confounder variables. Similar results were observed across stratified BMI groups (0·92–0·99; appendix p 24).

The ability of logistic regression models to classify race on the basis of tissue density (AUC 0·54) and on the combination of age and tissue density (0·61) was far lower than the ability of the image models on the breast mammograms in the EM-Mammo dataset (0·81; appendix p 25). These findings suggest that breast density and age did not account for most image model performance when detecting race.

Moreover, the ability of models to predict race from the diagnostic labels alone was much lower than the chest x-ray image-based models, with AUC values between 0·54 and 0·61 for MXR, and between 0·52 and 0·57 for CXP (appendix p 30). AUC values for race detection in the no finding class of 0·914 (95% CI 0·901–0·926) were obtained for Asian patients, 0·949 (0·945–0·953) for Black patients, and 0·941 (0·937–0·945) for White patients, versus 0·944 (0·938–0·950 [Asian patients]), 0·940 (0·937–0·942 [Black patients]), and 0·933 (0·930–0·936 [White patients]) for the entire dataset containing all disease classes, including the no finding class. These results suggest that high AUC values for racial identity recognition were not caused by disease labels.

We found that deep learning models effectively predicted patient race even when the bone density information was removed for both MXR (AUC value for Black patients: 0·960 [CI 0·958–0·963]) and CXP (AUC value for Black patients: 0·945 [CI 0·94–0·949]) datasets. The average pixel thresholds for different tissues did not produce any usable signal to detect race (AUC 0·5). These findings suggest that race information was not localised within the brightest pixels within the image (eg, in the bone).

For patients in different age groups, there was no appreciable difference in racial identity recognition performance (appendix p 15). Similarly, there was also no appreciable difference in racial identity recognition performance between male and female patients (appendix p 17).

The performance of a logistic regression model (AUC 0·65), a random forest classifier (0·64), and an XGBoost model (0·64) to classify race on the basis of age, sex, gender, disease, and body habitus performed much worse than the race classifiers trained on imaging data (AUC >0·95; appendix p 20). This finding suggests that the combination of these confounders did not significantly affect the imaging model's ability to classify race.

We also examined whether race information persisted in all spectral ranges and in the presence of highly degraded images. As shown in figure 1, we tested the effect on model performance of adding a low-pass filter and a high-pass filter for various diameters in the MXR dataset, and show samples of the transformed images in figure 2. The addition of a low-pass filter resulted in significantly degraded performance at around diameter ten, which corresponded to high levels of visual degradation. A high performance (up to diameter 100) in the absence of discernible anatomical features was maintained with the addition of a high-pass filter (ie, model performance was maintained despite extreme degradation of the image visually). Further experiments that used band-pass and notch filtering are reported in the appendix (pp 25–26), with the transformed images visualised also given in the appendix (pp 7–8).

The AUC of various image resolutions, from 1 pixel resolution to 320 × 320 images in the MXR dataset, are shown in the appendix (p 12). For images at 160 × 160 resolution or higher, AUC values were >0·95. There was a reduction in performance for images below this resolution, which demonstrates that race information persisted more than random chance even for resolutions as small as 4 × 4 (appendix p 28). Similar results were observed for the perturbed images, with AUC values of 0·74 to 0·80 for the noisy images and 0·64 to 0·72 for the blurred images (appendix p 29).

Concerning whether race information was localised to a specific anatomical region or body segment, using data from multiple experiments from several datasets, there was no evidence of a clear contribution of any anatomical regions or body segments on race identity. Models tested on non-lung segmentations of images were better able to identify race compared with models tested on lung segmentations, but segmented predictions were lower than the original image predictions (appendix p 29). Therefore, the race information utilised by artificial intelligence was likely to be determined from a combination of information from all image segments, including both lung and non-lung segments. Similar findings were observed in

slice-wise analysis of CT scans. Occluding the image regions identified by saliency maps (appendix p 9) caused a decrease in AUC values in race identification but still led to AUC values ≥0·67 (appendix p 29).

Race prediction was robust to the removal of any particular patch from images in the MXR dataset, indicating that race information was not localised within a specific part of the 3 × 3 grid (appendix p 30). We observed that there are parts of the image with little race information (appendix p 30). However, in most cases, using only one ninth of the image was sufficient to obtain prediction performance that was almost identical to using the entire image (appendix p 30).

Race prediction performance was also robust across models trained on single equipment and single hospital location on the chest x-ray and mammogram datasets (appendix pp 30–31). We observed a decrease in performance (although the outputs were better than random) on the digitised chest x-ray in the CheXphoto dataset compared with the digital CXP dataset, implying that some signal still persisted with different image acquisitions (appendix p 31).

Discussion

In this modelling study, which used both private and public datasets, we found that deep learning models can accurately predict the self-reported race of patients from medical images alone. This finding is striking as this task is generally not understood to be possible for human experts. We also showed that the ability of deep models to predict race was generalised across different clinical environments, medical imaging modalities, and patient populations, suggesting that these models do not rely on local idiosyncratic differences in how imaging studies are conducted for patients with different racial identities. Beyond these findings, in two of the datasets (MXR and CXP) analysed, all patients were imaged in the same locations and with the same processes, presumably independently of race.

We also provide evidence that disease distribution and body habitus of patients in the CXP, MXR, and EMX datasets were not strongly predictive of racial group, implying that the deep learning models were not relying on these features alone. Although an aggregation of these and other features could be partially responsible for the ability of AI models to detect racial identity in medical images, we could not identify any specific image-based covariates that could explain the high recognition performance presented here.

Our findings conflict with data from Jabbour and colleagues' study,[34] which measured the extent to which models learned potentially sensitive attributes (eg, age, race, and BMI) from an institutional dataset (the AHRF dataset) of 1296 patient chest x-rays. Their findings led to an AUC value of 0·66 (0·54–0·79). Possible explanations for this discrepant performance compared with our experiment could be due to the use of transfer learning in Jabbour and colleagues' study, in which the MXR and CXP datasets were used for initial training, and the final layers were fine-tuned on the AHRF dataset. This possible contamination in the dataset might have degraded performance due to label misalignment. We do not have access to the AHRF dataset for further external validation and Jabbour and colleagues did not extend their experiments to MXR and CXP datasets.

The results of the low-pass filter and high-pass filter experiments done in our study suggest that features relevant to the recognition of racial identity were present throughout the image frequency spectrum. Models trained on low-pass filtered images maintained high performance even for highly degraded images. More strikingly, models that were trained on high-pass filtered images maintained performance well beyond the point that the degraded images contained no recognisable structures; to the human coauthors and radiologists it was not clear that the image was an x-ray at all. Furthermore, experiments that were involved in patch-based training, slice-based error analysis, and saliency mapping were non-contributory: no specific regions of the images consistently informed race recognition decisions. Overall, we were unable to isolate specific image features that were responsible for the recognition of racial identity in medical images, either by spatial location, in the frequency domain, or that were caused by common anatomic and phenotype confounders associated with racial identity.

Although the ability to accurately detect self-reported race from highly degraded x-ray images is not meaningful on its own, this ability is important in the larger sociotechnical context that AI models operate in for medical imaging. One commonly proposed method to mitigate the known disparity in AI model performance is through the selective removal of features that encode sensitive attributes to make AI models “colorblind”.[35] Although this approach has already been criticised as being ineffective, or even harmful in some circumstances,[36] our work suggests that such an approach could be impossible in medical imaging because racial identity information appears to be incredibly difficult to isolate. The ability to detect race was not mitigated by any reasonable reduction in resolution or by the addition of noise, nor by frequency spectrum filtering or patch-based masking. Even ignoring the question of whether these approaches were beneficial, it seems plausible that technical solutions along these lines are unlikely to succeed and that strategies designed to detect racial bias,[37] paired with the intentional design of models to equalise racial outcomes,[38] should be considered to be the default approach to optimise the safety and fairness of AI in this context. The regulatory environment in particular, while evolving, has not yet produced strong processes to guard against unexpected racial recognition by AI models; either to identify these capabilities in models or to mitigate the harms that might be caused.

There were several limitations to this work. Most importantly, we relied on self-reported race as the ground truth for our predictions. There has been extensive research into the association between self-reported race and genetic ancestry, which has shown that there is more genetic variation within races than between races, and that race is more a social construct than a biological construct.[24] We note that in the context of racial discrimination and bias, the vector of harm is not genetic ancestry but the social and cultural construct that of racial identity, which we have defined as the combination of external perceptions and self-identification of race. Indeed, biased decisions are not informed by genetic ancestry information, which is not directly available to medical decision makers in almost any plausible scenario. As such, self-reported race should be considered a strong proxy for racial identity.

Our study was also limited by the availability of racial identity labels and the small cohorts of patients from many racial identity categories. As such, we focused on Asian, Black, and White patients, and excluded patient populations that were too small to adequately analyse (eg, Native American patients). Additionally, Hispanic patient populations were also excluded because of variations in how this population was recorded across datasets. Moreover, our experiments to exclude bone density involved brightness clipping at 60% and evaluating average body tissue pixels, with no methods to evaluate if there was residual bone tissue that remained on the images. Future work could look at isolating different signals before image reconstruction.

We finally note that this work did not establish new disparities in AI model performance by race. Our study was instead informed by previously published literature that has shown disparities in some of the tasks we investigated.[10,39] The combination of reported disparities and the findings of this study suggest that the strong capacity of models to recognise race in medical images could lead to patient harm. In other words, AI models can not only predict the patients' race from their medical images, but appear to make use of this capability to produce different health outcomes for members of different racial groups.

To conclude, our study showed that medical AI systems can easily learn to recognise self-reported racial identity from medical images, and that this capability is extremely difficult to isolate. We found that patient racial identity was readily learnable from medical imaging data alone, and could be generalised to external environments and across multiple imaging modalities. We strongly recommend that all developers, regulators, and users who are involved in medical image analysis consider the use of deep learning models with extreme caution as such information could be misused to perpetuate or even worsen the well documented racial disparities that exist in medical practice. Our findings indicate that future AI medical imaging work should emphasise explicit model performance audits on the basis of racial identity, sex, and age, and that medical imaging datasets should include the self-reported race of patients when possible to allow for further investigation and research into the human-hidden but model-decipherable information related to racial identity that these images appear to contain.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

JWG and ATP are funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) MIDRC grant of the National Institutes of Health (75N92020C00008 and 75N92020C00021). JWG and SP are funded by US National Science Foundation (grant number 1928481) from the Division of Electrical, Communication & Cyber Systems. MPL was funded by the National Library of Medicine of the National Institutes of Health (R01LM012966). LAC is funded by the National Institute of Health through a NIBIB grant (R01 EB017205). PK is funded by the Ministry of Science and Technology (Taiwan; MOST109-2222-E-007-004-MY3).

Declaration of interests

MG has received speaker fees for a Harvard Medical School executive education class. HT has received consulting fees from Sirona medical, Arterys, and Biodata consortium. HT also owns lightbox AI, which provides expert annotation of medical images for radiology AI. MPL has received consulting fees from Bayer, Microsoft, Phillips,

and Nines. MPL also owns stocks in Nines, SegMed, and Centaur. LAC has received support to attend meetings from MISTI Global Seed Funds. ATP has received payment for expert testimony from NCMIC insurance company. ATP also has a pending institutional patent for comorbidity prediction from radiology images. All other authors declare no competing interests.

Data sharing

The MIMIC-CXR dataset, CheXpert dataset, National lung cancer screening trial, RSNA Pulmonary Embolism CT, and the Digital Hand Atlas are all publicly available. The Emory University datasets (Emory CXR, Emory Chest CT, Emory Cervical Spine, and Emory Mammogram) are available on request after signing a data use agreement. All code is available at https://github.com/Emory-HITI/AI-Vengers.

References

1. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? FAccT '21; March 3–10, 2021. 10.1145/3442188.3445922.
2. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed april 25, 2022).
3. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. Proc Natl Acad Sci USA 2020; 117: 7684–89. [PubMed: 32205437]
4. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. PMLR 2018; 81: 77–91.
5. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018; 154: 1247–48. [PubMed: 30073260]
6. Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? J Invest Dermatol 2018; 138: 2277–79. [PubMed: 29864435]
7. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. Lancet Digit Health 2021; 3: e241–49. [PubMed: 33766288]
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366: 447–53. [PubMed: 31649194]
9. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021; 27: 2176–82. [PubMed: 34893776]
10. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. arXiv 2020; published online Oct 16. 10.48550/arXiv.2003.00827 (preprint).
11. Yi PH, Wei J, Kim TK, et al. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. Emerg Radiol 2021; 28: 949–54. [PubMed: 34089126]
12. Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. Radiology 2021; 301: 692–99. [PubMed: 34581608]
13. Rim TH, Lee G, Kim Y, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. Lancet Digit Health 2020; 2: e526–36. [PubMed: 33328047]
14. Munk MR, Kurmann T, Márquez-Neila P, Zinkernagel MS, Wolf S, Sznitman R. Assessment of patient specific information in the wild on fundus photography and optical coherence tomography. Sci Rep 2021; 11: 8621. [PubMed: 33883573]
15. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018; 2: 158–64. [PubMed: 31015713]

16. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986; 15: 413–19. [PubMed: 3771081]
17. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. SSO Schweiz Monatsschr Zahnheilkd 1999; 14: 29–46.
18. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform 2021; 28: e100289.
19. Tariq A, Purkayastha S, Padmanaban GP, et al. Current clinical applications of artificial intelligence in radiology and their best supporting evidence. J Am Coll Radiol 2020; 17: 1371–81. [PubMed: 33153541]
20. FDA cleared AI algorithms. https://report.acr.org/t/PUBLIC/views/CascadeReport/Commercial (accessed May 3, 2022).
21. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020; 3: 118. [PubMed: 32984550]
22. Tadavarthi Y, Vey B, Krupinski E, et al. The State of radiology AI: considerations for purchase decisions and current market offerings. Radiol Artif Intell 2020; 2: e200004. [PubMed: 33937846]
23. Krieger N Shades of difference: theoretical underpinnings of the medical controversy on black/white differences in the United States, 1830–1870. Int J Health Serv 1987; 17: 259–78. [PubMed: 3294621]
24. Cooper R, David R. The biological concept of race and its application to public health and epidemiology. J Health Polit Policy Law 1986; 11: 97–116. [PubMed: 3722786]
25. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019; 6: 317. [PubMed: 31831740]
26. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. AAAI 2019; 33: 590–97.
27. Wagner DR, Heyward VH. Measures of body composition in blacks and whites: a comparative review. Am J Clin Nutr 2000; 71: 1392–402. [PubMed: 10837277]
28. del Carmen MG, Halpern EF, Kopans DB, et al. Mammographic breast density and race. AJR Am J Roentgenol 2007; 188: 1147–50. [PubMed: 17377060]
29. Office of Minority Health. Heart disease and African Americans. Jun 27, 2021. https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlid=19 (accessed April 25, 2022).
30. Graham G Disparities in cardiovascular disease risk in the United States. Curr Cardiol Rev 2015; 11: 238–45. [PubMed: 25418513]
31. Ettinger B, Sidney S, Cummings SR, et al. Racial differences in bone density between young adult black and white subjects persist after adjustment for anthropometric, lifestyle, and biochemical differences. J Clin Endocrinol Metab 1997; 82: 429–34. [PubMed: 9024231]
32. Hochberg MC. Racial differences in bone strength. Trans Am Clin Climatol Assoc 2007; 118: 305–15. [PubMed: 18528512]
33. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 2021; 3: 610–619.
34. Jabbour S, Fouhey D, Kazerooni E, Sjoding MW, Wiens J. Deep learning applied to chest X-rays: exploiting and preventing shortcuts. PMLR 2020; 126: 750–82.
35. Ioannidis JPA, Powe NR, Yancy C. Recalibrating the use of race in medical research. JAMA 2021; 325: 623–24. [PubMed: 33492329]
36. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020; 383: 874–82. [PubMed: 32853499]
37. Brown S, Davidovic J, Hasan A. The algorithm audit: scoring the algorithms that score us. Big Data Soc 2021; published online Jan 28. 10.1177/2053951720983865.
38. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 2021; 27: 136–40. [PubMed: 33442014]

39. Seyyed-Kalantari L, Liu G, McDermott M, Chen I, Ghassemi M. Medical imaging algorithms exacerbate biases in underdiagnosis. Research Square 2021; published online Jan 2021. DOI:10.21203/rs.3.rs-151985/v1.

Research in context

Evidence before this study

We used three different search engines to do our review. For PubMed, we used the following search terms: "(((disparity OR bias OR fairness) AND (classification)) AND (x-ray OR mammography)) AND (machine learning [MeSH Terms])." For IEEE Xplore, we used the following search terms: "((disparity OR bias OR fairness) AND (mammography OR x-ray) AND (machine learning))". For ACM, we used the following search terms: "[Abstract: mammography x-ray] AND [Abstract: classification prediction] AND [All: disparity fairness]". All queries were limited to dates between Jan 1, 2010, and Dec 31, 2020. We included any studies that were published in English, focused on medical images, and that were original research. We also reviewed commentaries and opinion articles. We excluded articles that were not written in English or that were outside of the medical imaging domain. To our knowledge, there is no published meta-analysis or systematic review on this topic. Most published papers focused on measuring disparities in tabular health data without much emphasis on imaging-based approaches.

Although previous work has shown the existence of racial disparities, the mechanism for these differences in medical imaging is, to the best of our knowledge, unexplored. Pierson and colleagues noted that an artificial intelligence (AI) model that was designed to predict severity of osteoarthritis using knee x-rays could not identify the race of the patients. Yi and colleagues conducted a forensics evaluation on chest x-rays and found that AI algorithms could predict sex, distinguish between adult and paediatric patients, and differentiate between US and Chinese patients. In ophthalmology, retinal scan images have been used to predict sex, age, and cardiac markers (eg, hypertension and smoking status). We found few published studies that explicitly targeted the recognition of racial identity from medical images, possibly because radiologists do not routinely have access to, nor rely on, demographic information (eg, race) for diagnostic tasks in clinical practice.

Added value of this study

In this study, we investigated a large number of publicly and privately available large-scale medical imaging datasets and found that self-reported race is accurately predictable by AI models trained with medical image pixel data alone as model inputs. First, we showed that AI models are able to predict race across multiple imaging modalities, various datasets, and diverse clinical tasks. This high level of performance persisted during external validation of these models across a range of academic centres and patient populations in the USA, as well as when the models were optimised to do clinically motivated tasks. Second, we conducted ablations that showed that this detection was not due to trivial proxies, such as body habitus, age, tissue density, or other potential imaging confounders for race (eg, underlying disease distribution in the population). Finally, we showed that the features learned appear to involve all regions of the image and frequency spectrum, suggesting the efforts to control this behaviour when it is undesirable will be challenging and demand further study.

Implications of all the available evidence

In our study, we emphasise that the ability of AI to predict racial identity is itself not the issue of importance, but rather that this capability is readily learned and therefore is likely to be present in many medical image analysis models, providing a direct vector for the reproduction or exacerbation of the racial disparities that already exist in medical practice. This risk is compounded by the fact that human experts cannot similarly identify racial identity from medical images, meaning that human oversight of AI models is of limited use to recognise and mitigate this problem. This issue creates an enormous risk for all model deployments in medical imaging: if an AI model relies on its ability to detect racial identity to make medical decisions, but in doing so produced race-specific errors, clinical radiologists (who do not typically have access to racial demographic information) would not be able to tell, thereby possibly leading to errors in health-care decision processes.

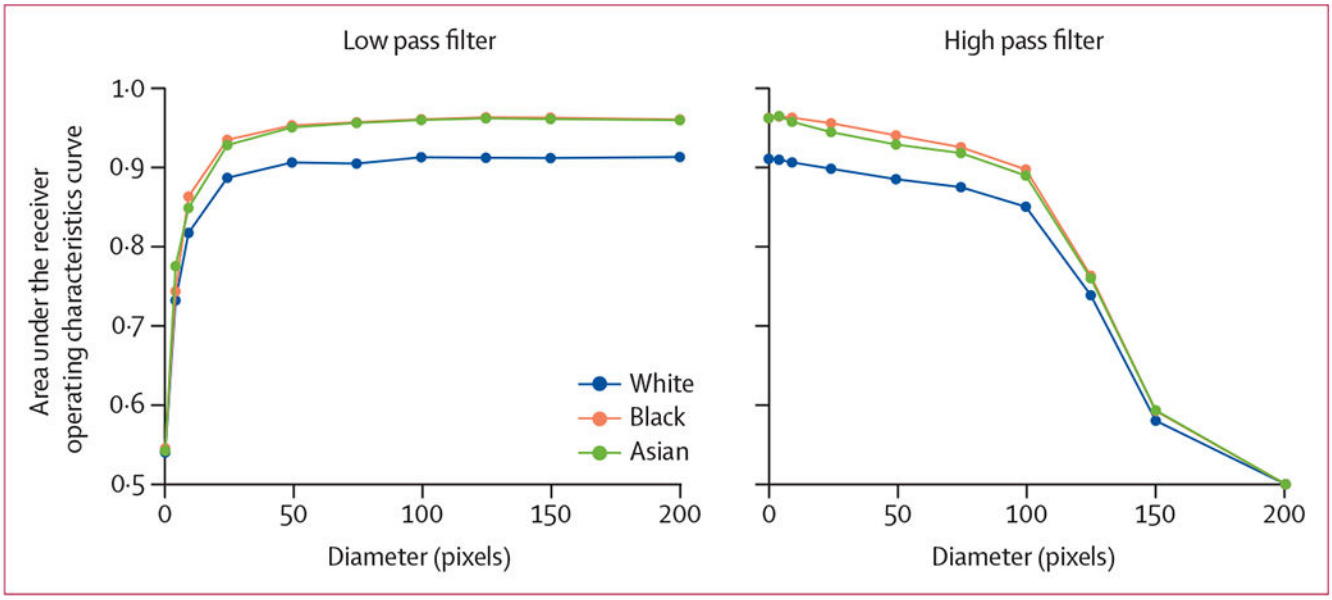

Figure 1: The effect on model performance of adding a low-pass filter and a high-pass filter for various diameters in the MXR dataset

MXR=MIMIC-CXR dataset.

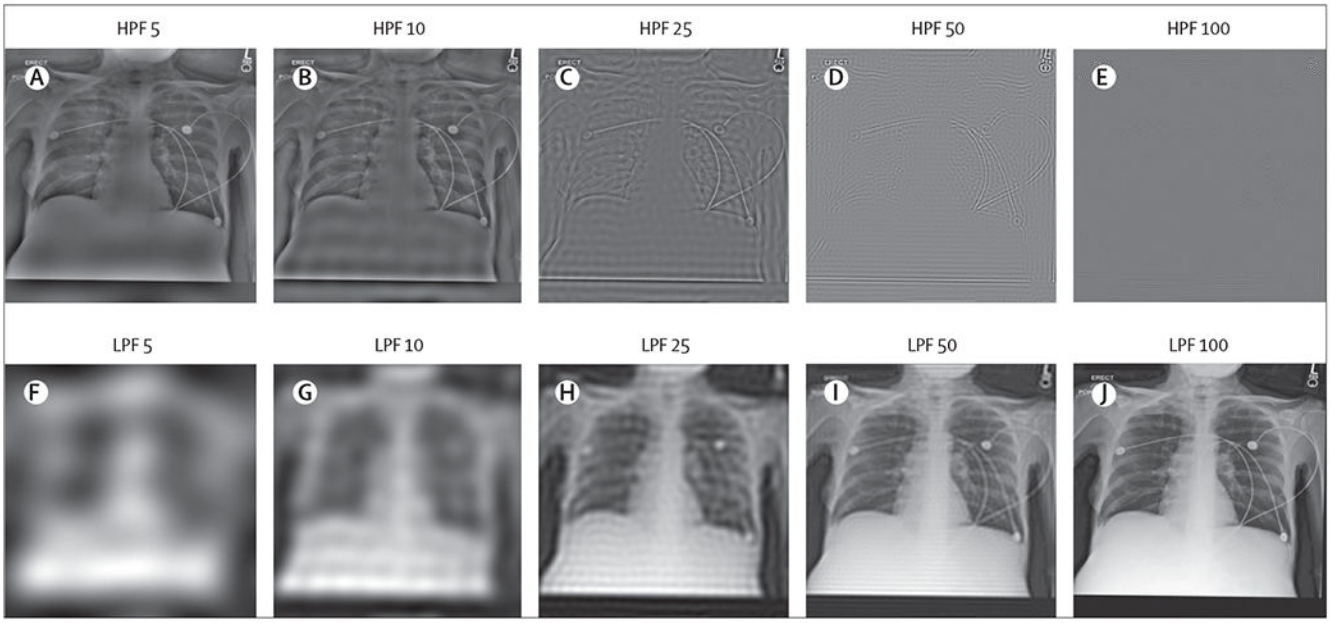

Figure 2: Samples of the images after low-pass filters and high-pass filters in MXR dataset
HPF=high-pass filtering. LPF=low-pass filtering. MXR=MIMIC-CXR dataset.

Table 1:

Summary of datasets used for race prediction experiments

	MXR	CXP	EMX	NLST	RSPECT (Stanford subset)	EM-CT	DHA	EM-Mammo	EM-CS
Data type	Chest x-ray	Chest x-ray	Chest x-ray	Chest CT	Chest CT (PE protocol)	Chest CT	Digital radiography x-ray	Breast mammograms	Lateral c-spine x-ray
Number of patients (number of images)	53073 (228 915)	65400 (223 414)	90518 (227 872)	512 (198 475)	254 (72 329)	560 (187 513)	691 (691)	27160 (86 669)	997 (10 358)
Sex									
Female	27532 (51·9%)	29090 (44·5%)	48477 (53·6%)	184 (36·0%)	135 (53·1%)	286 (51·1%)	400 (49·2%)	27160 (100%)	535 (53·7%)
Male	25541 (48·1%)	36310 (55·5%)	42041 (46·4%)	328 (64·0%)	119 (46·9%)	274 (48·9%)	391 (56·6%)	0	462 (46·3%)
Race									
Black	8957 (16·9%)	3147 (4·8%)	42373 (46·8%)	241 (47·1%)	23 (9·1%)	403 (72·0%)	333 (48·2%)	13696 (50·4%)	247 (24·8%)
Asian	1935 (3·6%)	7096 (10·8%)	3293 (3·6%)	0	0	0	0	0	0
White	34035 (64·1%)	36765 (56·2%)	38071 (42·1%)	271 (53·0%)	231 (90·9%)	157 (28·0%)	358 (51·8%)	13464 (49·6%)	750 (75·2%)
Unknown	8146 (15·3%)	18420 (28·2%)	6781 (7·5%)	0	0	0	0	0	0
Dataset split									
Training, %	60·0%	60·0%	75·0%	78·0%	0	0	70·0%	60·0%	80·0%
Validation, %	10·0%	10·0%	12·5%	10·0%	0	0	10·0%	20·0%	10·0%
Test, %	30·0%	30·0%	12·5%	12·0%	100·0%	100·0%	20·0%	20·0%	10·0%

CXP=CheXpert dataset. DHA=Digital Hand Atlas. EM-CS=Emory Cervical Spine radiograph dataset. EM-CT=Emory Chest CT dataset. EM-Mammo=Emory Mammogram dataset. EMX=Emory chest x-ray dataset. MXR=MIMIC-CXR dataset. NLST=National Lung Cancer Screening Trial dataset. RSPECT=RSNA Pulmonary Embolism CT dataset.

Table 2:

Summary of experiments conducted to investigate mechanisms of race detection in Black patients

	Area under the receiver operating characteristics curve
Race detection in radiology imaging	
Chest x-ray (internal validation) *	
MXR (Resnet34, Densenet121)	0·97, 0·94
CXP (Resnet 34)	0·98
EMX (Resnet34, Densenet121, EfficientNet-B0)	0·98, 0·97, 0·99
Chest x-ray (external validation) *	
MXR to CXP, MXR to EMX	0·97, 0·97
CXP to EMX, CXP to MXR	0·97, 0·96
EMX to MXR, EMX to CXP	0·98, 0·98
Chest x-ray (comparison of models) †	
MXR, CXP, EMX	Multiple results (appendix p 26)
CT chest (internal validation) *	
NLST (slice, study)	0·92, 0·96
CT chest (external validation) *	
NLST to EM-CT (slice, study)	0·80, 0·87
NLST to RSPECT (slice, study)	0·83, 0·90
Limb x-ray (internal validation) *	
DHA	0·91
Mammography *	
EM-Mammo (image, study)	0·78, 0·81
Cervical spine x-ray *	
EM-CS	0·92
Experiments on anatomic and phenotypic confounders	
BMI *	
CXP	0·55, 0·52
Image-based race detection stratified by BMI †	
EMX, MXR	Multiple results (appendix p 24)
Breast density *	
EM-Mammo	0·54
Breast density and age *	
EM-Mammo	0·61
Disease distribution *	
MXR, CXP	0·61, 0·57
Image-based race detection for the no finding class *	
MXR	0·94

	Area under the receiver operating characteristics curve
Model prediction after training on dataset with equal disease distribution †	
MXR	0·75
Removal of bone density features *	
MXR, CXP	0·96, 0·94
Impact of average pixel thresholds †	
MXR	0·50
Impact of age †	
MXR	Multiple results (appendix p 27)
Impact of patient sex †	
MXR	Multiple results (appendix p 28)
Combination of age, sex, disease, and body habitus *	
EMX (logistic regression model, random forest classifier, XGBoost model)	0·65, 0·64, 0.64
Experiments to evaluate the mechanism of race detection	
Frequency domain filtering	
High-pass filtering *	
MXR	Multiple results (appendix p 26)
Low-pass filtering *	
MXR	Multiple results (appendix p 26)
Notch filtering †	
MXR	Multiple results (appendix p 26)
Band-pass filtering †	
MXR	Multiple results (appendix p 25)
Image resolution and quality *	
MXR	Multiple results (appendix p 28)
Anatomical localisation	
Lung segmentation experiments †	
MXR	Multiple results (appendix p 29)
Saliency maps †	
MXR, CXP, EMX, NLST, DHA, EM-Mammo, EM-CS	Multiple results (appendix pp 13–18)
Occlusion experiments †	
MXR	Multiple results (appendix p 30)
Patch-based training *	
MXR	Multiple results (appendix p 30)
Image acquisition differences †	
EMX, EM-Mammo, ChexPhoto	Multiple results (appendix p 31)

BMI=body-mass index. CXP=CheXpert dataset. DHA=Digital Hand Atlas. EM-CS=Emory Cervical Spine radiograph dataset. EM-CT=Emory Chest CT dataset. EM-Mammo=Emory Mammogram dataset. EMX=Emory CXR dataset. MXR=MIMIC-CXR dataset. NLST=National Lung Cancer Screening Trial dataset. RSPECT=RSNA Pulmonary Embolism CT dataset.

[*] Results located in main text.

[†] Results located in the appendix.

Table 3:

Performance of deep learning models to detect race from chest x-rays

	Area under the receiver operating characteristics curve value for race classification		
	Asian (95% CI)	**Black (95% CI)**	**White (95% CI)**
Primary race detection in chest x-ray imaging			
MXR Resnet34	0·986 (0·984–0·988)	0·982 (0·981–0·983)	0·981 (0·979–0·982)
CXP Resnet34	0·981 (0·979–0·983)	0·980 (0·977–0·983)	0·980 (0·978–0·981)
EMX Resnet34	0·969 (0·961–0·976)	0·992 (0·991–0·994)	0·988 (0·986–0·989)
External validation of race detection models in chest x-ray imaging			
MXR Resnet34 to CXP	0·947 (0·944–0·951)	0·962 (0·957–0·966)	0·948 (0·945–0·951)
MXR Resnet34 to EMX	0·914 (0·899–0·928)	0·983 (0·981–0·985)	0·975 (0·973–0·978)
CXP Resnet34 to MXR	0·974 (0·971–0·977)	0·955 (0·952–0·957)	0·956 (0·954–0·958)
CXP Resnet34 to EMX	0·915 (0·901–0·929)	0·968 (0·965–0·971)	0·954 (0·951–0·958)
EMX Resnet34 to MXR	0·966 (0·962–0·969)	0·970 (0·968–0·972)	0·964 (0·962–0·965)
EMX Resnet34 to CXP	0·949 (0·946–0·952)	0·973 (0·970–0·977)	0·947 (0·945–0·950)
Race detection in non-chest x-ray imaging modalities: binary race detection (Black or White)			
NLST	0·92 (slice; 0·910–0·918), 0·96 (study; 0·926–0·982)	··	··
NLST to EM-CT	0·80 (slice; 0·796–0·800), 0·87 (study; 0·829–0·904)	··	··
NLST to RSPECT	0·83 (slice; 0·825–0·834), 0·90 (study; 0·836–0·958)	··	··
EM-Mammo	0·78 (slice; 0·773–0·786), 0·81 (study; 0·794–0·818)	··	··
EM-CS	0·913 (0·892–0·931)	··	··
DHA	0·87 (0·752–0·894)	··	··

Values reflect the area under the receiver operating characteristics curve for each model on the test set per slice and per study (by averaging the predictions across all slices). CXP=CheXpert dataset. DHA=Digital Hand Atlas. EM-CS=Emory Cervical Spine radiograph dataset. EM-CT=Emory Chest CT dataset. EM-Mammo=Emory Mammogram dataset. EMX=Emory CXR dataset. MXR=MIMIC-CXR dataset. NLST=National Lung Cancer Screening Trial dataset. RSPECT=RSNA Pulmonary Embolism CT dataset.

HHS Public Access

Author manuscript

Expo Health. Author manuscript; available in PMC 2023 February 10.

Published in final edited form as:

Expo Health. 2022 December ; 14(4): 941–949. doi:10.1007/s12403-022-00468-2.

Alterations in Microbial-Associated Fecal Metabolites in Relation to Arsenic Exposure Among Infants

Margaret R. Karagas[1], **Susan McRitchie**[2], **Anne G. Hoen**[1], **Cindy Takigawa**[1], **Brian Jackson**[3], **Emily R. Baker**[4], **Juliette Madan**[1,5], **Susan J. Sumner**[2], **Wimal Pathmasiri**[2]

[1]Department of Epidemiology, Geisel School of Medicine at Dartmouth College, Hanover, NH, USA

[2]Nutrition Research Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3]Department of Earth Sciences, Dartmouth College, Hanover, NH, USA

[4]Department of Obstetrics and Gynecology, Geisel School of Medicine at Dartmouth College, Hanover, NH, USA

[5]Department of Pediatrics & Psychiatry, Children's Hospital at Dartmouth, Dartmouth Hitchcock Medical Center, Lebanon, NH, USA

Abstract

In utero and early life exposure to inorganic arsenic (iAs) alters immune response in experimental animals and is associated with an increased risk of infant infections. iAs exposure is related to differences in the gut microbiota diversity, community structure, and the relative abundance of individual microbial taxa both in laboratory and human studies. Metabolomics permits a direct measure of molecular products of microbial and host metabolic processes. We conducted NMR metabolomics analysis on infant stool samples and quantified the relative concentrations of 34 known microbial-related metabolites. We examined these metabolites in relation to both in utero and infant $\log_2$ urinary total arsenic concentrations (utAs, the sum of iAs and iAs metabolites) collected at approximately 6 weeks of age using linear regression models, adjusted for infant sex, age at sample collection, type of delivery (vaginal vs. cesarean section), feeding mode (breast milk vs. any formula), and specific gravity. Increased fecal butyrate ($b = 214.24$), propionate ($b = 518.33$), cholate ($b = 8.79$), tryptophan ($b = 14.23$), asparagine ($b = 28.80$), isoleucine ($b = 65.58$), leucine ($b = 95.91$), malonate ($b = 50.43$), and uracil ($b = 36.13$), concentrations were associated with a doubling of infant utAs concentrations ($p < 0.05$). These associations were largely among infants who were formula fed. No clear associations were observed with maternal utAs and infant fecal metabolites. Metabolomic analyses of infant stool samples lend further evidence that the infant gut microbiota is sensitive to As exposure, and these effects may have functional consequences.

✉Margaret R. Karagas, margaret.karagas@dartmouth.edu, Wimal Pathmasiri, margaret.karagas@dartmouth.edu.

Declarations We declare this manuscript to be original, has not yet been published, and includes information relevant to your journal. This manuscript is not being considered for publication elsewhere. Data can be made available on reasonable request.

Supplementary Information The online version of this article (https://doi.org/10.1007/s12403-022-00468-2) contains supplementary material, which is available to authorized users.

Keywords

Arsenic; Metabolomics; In utero exposure; Postnatal exposure

Introduction

Arsenic (As) is naturally present in the earth's crust and has been used for a broad range of industrial purposes, including as antibacterial and immune modulating medicines in the nineteenth and early twentieth centuries (Kapp 2018). Anthropogenic and natural sources of As contribute to widespread exposure primarily from the food and drinking water systems. The World Health Organization estimates that at least 140 million people worldwide across 50 countries are exposed to drinking water exceeding the provisional guideline of 10 μg/L As. Populations relying on private, unregulated water systems remain vulnerable to As concentrations above regulatory limits. Fish and seafood contain As in forms that are not metabolized by humans and therefore considered non-toxic such as arsenobetaine. However, other dietary sources, especially rice and rice products, fruits and fruit juices, and seaweeds, may contain appreciable concentrations of inorganic As (iAs) and other potentially detrimental forms of As (EFSA Panel on Contaminants in the Food Chain 2009). Diet is the primary source of As exposure for the majority of people and contributes to disproportionately higher intakes of As among young children for their body weight (EFSA Panel on Contaminants in the Food Chain 2009). This poses a health concern, especially for infants consuming formula mixed with potentially contaminated water and when transitioning to solid foods such as infant rice cereal (EFSA Panel on Contaminants in the Food Chain 2009; Carignan et al. 2016; Signes-Pastor et al. 2018).

Evidence both from laboratory experiments and epidemiologic investigations support immunotoxic effects of As on both innate and adaptive immunity (Dangleben et al. 2013). In our earlier work, we observed an increased risk of infant infection associated with higher in utero As concentrations, and specifically lower respiratory infections Farzan et al. 2016. Immunity develops in utero and evolves during the first years of life (Dietert et al. 2010). The gut microbiota plays an essential, bidirectional role in this process—gut microbiota stimulate the maturation of the neonatal immune system, and in turn, an infant's immune response helps to shape the composition of microbes inhabiting the gut (Madan et al. 2012). We previously reported on the relation between infants' urinary As concentrations and gut microbiota composition in a pregnancy cohort of maternal–child dyads from New Hampshire (Hoen et al. 2018; Laue et al. 2020). Notably, we observed decreased relative abundance of keystone taxa in the genera *Bacteroides* and *Bifidobacterium* involved in immune maturation (Hoen et al. 2018; Laue et al. 2020). Analysis of fecal samples using next-generation sequence-based methods allowed us to observe the genetic make-up of microbes in the gut and estimate its diversity, community structure, and composition. However, it only can be used to make inferences about the collective function of the gut microflora. Metabolomics complements genetic sequence-based profiling by providing a critical bridge from microbiota composition to the complex array of small molecules that directly influence biologic activities. As-induced functional changes in the gut microbiota as reflected in the metabolome have been characterized from mouse experiments (Li et al.

2019). By comprehensively profiling the fecal metabolome in the context of a longitudinal pregnancy cohort study, we sought to gain a clearer window into the phenotype of the human infant gut microbiota related to As exposure.

Materials and Methods

Study Population

Our study is based on the ongoing New Hampshire Birth Cohort Study of women recruited during pregnancy and whose offspring are followed to update exposure and health information (Hoen et al. 2018). Reproductive and medical history, health, diet, and lifestyle factors were ascertained from questionnaires and medical record review during pregnancy, and a maternal urine sample was collected at approximately 24 to 28 weeks gestation. Newborn infant characteristics were documented from the delivery medical records, and both infant urine and stool samples were collected at approximately 6 weeks of life.

Assessment of Maternal Pregnancy (In Utero) and Infant Postnatal As Exposure

iAs undergoes a series of reduction and oxidative methyl processes from iAsIV to iAsIII to MMAV to monmethyl-arsonus acid (MMAIII) to DMAV and is excreted in these forms (Abuawad et al. 2021). Therefore, to estimate As exposure, we analyzed urine samples collected during pregnancy at 24–28 weeks gestation and during infancy at approximately 6 weeks of life for As species [arsenite (iAsIII), arsenate (iAsV), monomethylarsonic acid (MMA), dimethylarsinic acid (DMA) and arsenobetaine (AsB)] using high-performance liquid chromatography (HPLC)–inductively coupled plasma mass spectrometry, ICPMS (Hoen et al. 2018). To compute total urinary As we summed the individual fractions of iAs (iAsIII and iAsIV) and the metabolites of iAs, MMA, and DMA (utAs = iAs + MMA + DMA), excluding arsenobetaine found in fish and seafood, which is unmetabolized and considered non-toxic. Although MMAIII is excreted it is rapidly converted to DMA and therefore typically undetectable using standard approaches. Specific gravity of urine was determined by a handheld refractometer with automatic temperature compensation (PAL-10S; ATAGO Co. Ltd.) to adjust for urinary dilution. We further divided MMA by iAs and DMA by MMA to calculate the primary and secondary methylation indices (PMI and SMI), respectively, as indicators of iAs metabolic capacity (Shen et al. 2016). Detection limits for iAs, MMA, and DMA were 0.1, 0.02 and 0.02 μg/L. The spiked recovery rates averaged 82% for iAs, 91% for MMA, 89% for DMA and 94% for AsB.

Stool Collection

Infant diapers containing stool were collected at home by caregivers, stored in a home freezer, and then brought frozen to the 6-week postpartum visit in thermal transport bags on ice packs or transported directly on ice packs. Upon receipt, diapers were stored at − 80 °C until processing. Following an overnight thaw at 4 °C, the stool was aliquoted and frozen at − 80 °C. Stool samples aliquoted into tubes certified as trace element free were used for metabolomic analysis.

Sample Preparation and Data Acquisition

De-identified aliquots of infant stool samples, along with replicates for quality control (QC), were shipped to the NIH Eastern Regional Comprehensive Metabolomics Research Core on dry ice and immediately stored at − 80 °C after being logged in for metabolomics analysis. The metabolomics analysis were adapted from previously described procedures (Brim et al. 2012, 2017; Livanos et al. 2016). Briefly, samples were randomized into batches. In each batch, samples were thawed, and ~150 mg of stool was transferred to MagNA Lyser tubes after recording the weight; samples were then homogenized with 50% acetonitrile in water by using a bead homogenizer (100 mg fecal mass/mL). Homogenized samples were centrifuged at 16,000 rcf, and the supernatant was separated into another tube. An aliquot (1000 μL, 100 mg equivalent of fecal mass) was transferred into an Eppendorf tube and lyophilized overnight. The dried extract was reconstituted in 700 μL of NMR master mix (containing 0.2 M phosphate, 0.5 mM DSS-d6, and 0.2% sodium azide), vortexed on a multitube vortexer at speed 5 for 2 min, and centrifuged at 16,000 rcf for 5 min. A 600 μL aliquot of the supernatant was transferred into a pre-labeled 5 mm NMR tube for data acquisition on a 700 MHz spectrometer. Additionally, pooled QC samples (study pools created from randomly selected study samples and batch pools) were generated from supernatants of study samples, and aliquots of pooled QC samples were dried and reconstituted similar to study samples described above and used for further QC purposes. ^{1}H NMR spectra of feces samples were acquired on a Bruker 700 MHz NMR spectrometer using a 5 mm cryogenically cooled ATMA inverse probe and ambient temperature of 25 °C. A 1D NOESY presaturation pulse sequence (noesygppr1d, [recycle delay, RD]-90°-f_1-90°-f_m-90°-acquire free induction decay (FID)]) was used for data acquisition (Beckonert et al. 2007; Dona et al. 2014). For each sample, 64 transients were collected into 64k data points using a spectral width of 12.02 ppm, 2 s relaxation delay, 10 ms mixing time, and an acquisition time of 3.899 s per FID. The water resonance was suppressed using resonance irradiation during the relaxation delay and mixing time. NMR spectra were processed using TopSpin 3.5 software (Bruker-Biospin, Germany). Spectra were zero-filled, and Fourier transformed after exponential multiplication with a line broadening factor of 0.5. Phase and baseline of the spectra were manually corrected for each spectrum. Spectra were referenced internally to the DSS-d6 signal (d=0 ppm). The quality of each NMR spectrum was assessed for the level of noise and alignment of identified markers. Spectra were assessed for missing data and underwent quality checks. NMR bins (0.5–9.0 ppm) were created excluding water (4.73–4.85 ppm) using intelligent bucket integration of 0.04 ppm bucket width with 50% looseness using ACD Spectrus Processor (ACD Labs, Inc., Toronto, Canada). Integrals of each of the bins were normalized to the total integral of each of the spectra. Chenomx NMR Suite 8.1 Professional (Chenomx, Inc., Edmonton, AB, Canada) (Weljie et al. 2006) was used to determine the relative concentrations of library-matched metabolites previously identified as associating with host and gut microbes co-metabolism (Li et al. 2008; Zheng et al. 2011; Nicholson et al. 2012).

Statistical Analyses

Statistical analyses were conducted using data from infants at approximately 6 weeks who had fecal metabolomics data and complete covariate data. Descriptive statistics were calculated for the participant characteristics, As concentrations, and relative concentration

of metabolites. The masked QC replicates were used to calculate the intra-class correlation (ICC) for each metabolite with a detectable relative concentration. Any metabolite with an ICC < 0.2 was excluded from the analysis.

Normalized binned NMR data were Pareto-scaled and centered prior to multivariate analysis using SIMCA 14 (Sartorius Data Analytics, Umeå, Sweden). The scores plot from the principal component analysis (PCA) was inspected to ensure that the laboratory QC pool samples were clustered in the center of study samples used to create the pools, a method widely applied to metabolomic studies (Chan et al. 2011; Masson et al. 2011; Broadhurst et al. 2018).

Spearman correlations were calculated for each metabolite's relative concentration with urinary total As (utAs). Metabolites with p-value < 0.1 were used as the outcome in multivariable linear regression models to determine the associations between $\log_2$ transform of urinary As concentration (maternal and infant utAs separately). Models with infant urinary As concentrations were adjusted for infant age, sex, feeding method (exclusive breastfeeding or formula/mixed feeding at 6 weeks of age), urine-specific gravity, and delivery mode (vaginal or cesarean section). Models with maternal urinary As concentrations were adjusted for urine-specific gravity, infant age, sex, feeding method, and delivery mode. Continuous variables were centered prior to modeling. We also performed the analysis stratified by below or equal to or above the median infant urinary PMI (0.35) and SMI (8.06).

SAS 9.4 (SAS Institute, Inc., Cary, NC) was used for calculating descriptive statistics, hypothesis tests, correlations, and linear regression, and ICC coefficients of replicate samples. For this exploratory study, p-values < 0.05 were considered to be statistically significant and were not adjusted for multiple testing (Bender and Lange 2001; Xi et al. 2014).

Pathway Enrichment Analysis

GeneGo MetaCore (Clarivate Analytics, PA) was used to assess the enrichment of perturbed metabolic pathways derived from the concentration data. MetaCore uses the hypergeometric test, which represents the enrichment of certain metabolites in a pathway, together with the false discovery rate (FDR). A q-value < 0.05 is considered indicative of significant enrichment in pathways.

Results

Study Population

A total of 83 infants with NMR metabolomics data acquired from infant stool samples also had maternal urinary As species measured at approximately 24 to 28 weeks gestation and complete data for model covariates. Eighty-one infants with NMR metabolomics data also had an analyzed postnatal, approximately 6-week urine sample for As species and complete data for model covariates. The mean age of the 81 infants at stool and urine sample collection was 46 days, 62% were boys, 26% were delivered by cesarean section, and 42% were exclusively breast-fed (Table 1). The utAs concentration during pregnancy

was on average 4.1 mg/L, with a range of 0.2 to 21.0 mg/L (Table 1). Among infants at approximately 6 weeks of age, the average utAs concentration was 0.6 μg/L with a range of 0.1 to 5.2 μg/L (Table 1). The mean (range) of the individual As species were 0.1 μg/L (undetectable to 1.0 μg/L) for iAs, 0.1 μg/L (undetectable to 0.4 μg/L) for MMA, 0.4 μg/L for DMA (undetectable to 4.5 μg/L) and 0.1 μg/L for AsB (undetectable to 1.2 μg/L). Maternal utAs concentrations averaged 4.1 μg/L, with a range of 0.2 to 21.0 μg/L (Table 1). Urinary specific gravity was within a narrow range in both maternal and infant samples (mean = 1.01; range 1.00, 1.03 and mean 1.00; range 1.00, 1.02, respectively).

Quality Control

ICCs for formate and fumarate fell below 0.2 and thus were excluded from further analyses. The average ICC coefficients for the remaining 34 metabolites ranged from 0.2 for isobutyrate to 0.99 for succinate (Supplemental Table 1) and averaged 0.75 for those metabolites used in the analysis. Laboratory QC pools were centered in the PCA plots of the samples from which the pools were created, further indicating that the NMR data were of high quality (data not shown).

Linear Regression

Eighteen of the 34 concentration fitted metabolites with ICC ≥ 0.2 had p-values < 0.1 for Spearman correlations with utAs (Supplemental Table 2). These 18 metabolites measured from infant fecal samples were used as the dependent variables in multivariable linear regression models to examine the associations between maternal and infant urinary utAs (independent variables) after adjusting for covariates. A doubling of infant utAs concentrations was associated with statistically significant increases ($p < 0.05$) in the relative concentrations of infant fecal short-chain fatty acids (SCFAs) butyrate ($b = 214.24$) and propionate ($b = 518.33$); the bile acid cholate ($b = 8.79$); the amino acids asparagine ($b = 28.80$), isoleucine ($b = 65.58$), leucine ($b = 95.91$); and tryptophan ($b = 14.23$), the pyrimidine uracil ($b = 36.13$), and the organic acid malonate ($b = 50.43$, Table 2). Positive associations tended to occur among infants fed formula, with negative associations for certain metabolites among exclusively breast-fed infants (Supplemental Table 3). Additional interactions were observed with phenylalanine and proline (Supplemental Table 3). Maternal urinary As was related only to infant fecal acetate concentration in unadjusted models ($r_s = -0.24$, $p = 0.031$) but was no longer statistically significant after adjustment ($b = -847.51$, $p = 0.115$). Associations with infant urinary As and tryptophan were largely among those with high PMI (p for interaction = 0.0228, Supplemental Table 4), and those for cholate were largely among those with low SMI (p for interaction = 0.0037, Supplemental Table 5). However, most interaction terms did not reach statistical significance.

Pathway Analysis of Concentration Data

MetaCore pathway enrichment analysis using the metabolites associated with utAs in infant urine by Spearman correlation ($p < 0.1$, Supplemental Table 6). Top enriched pathways identified included aminoacyl-t-RNA biosynthesis, amino acid dependent mTORC1 activation (signal transduction), amino acid metabolism (lysine, branched chain amino acids, BCAAs (isoleucine, leucine, valine), and methionine), saturated fatty acid synthesis to hexadecenoic acid, regulation of lipid metabolism (by niacin and isoprenaline), immune

responses [through myeloid-derived suppressor cells (MDSC), M2 macrophages, and Treg cell-mediated modulation of antigen-presenting cell (APC) functions] were among the top enriched pathways (Fig. 1).

Discussion

As exposure remains a major public health concern problem worldwide. The gut microbiota biochemically transforms As compounds (Coryell et al. 2019; McDermott et al. 2020) and at the same time may be affected by As exposure (Chi et al. 2018; McDermott et al. 2020). In our prospective pregnancy cohort study, we observed positive associations between infant urinary As concentrations and the relative concentration of nine infant fecal metabolites (asparagine, butyrate, cholate, isoleucine, leucine, malonate, propionate, tryptophan, and uracil) in multivariable regression analyses. Maternal urinary As concentrations during pregnancy were unrelated to infant fecal metabolites overall. Only a weakly negative association (adjusted $p = 0.115$) was observed between maternal urinary As concentrations and infant fecal acetate concentrations. However, as As concentrations change over the course of pregnancy (Hopenhayn et al. 2003; Tseng 2009; Gardner et al. 2011; Gao et al. 2019a, b; Gao et al. 2019a, b), it is conceivable that our sampling of ~ 24–28 weeks gestation did not capture the relevant time window to influence maternal–fetal transfer of the microbiome during delivery or otherwise influence either microbe or host metabolic products. We found concentrations of SCFAs, bile acids, amino acids, organic acids, and pyrimidines associated with infant urinary As concentrations. These changes may reflect impacts on the microbial composition of the gut or activation or detoxification pathways of either the host or microbes, epigenetic effects or other mechanisms. Thus, alterations in the metabolic pathways associated with these metabolites by As may provide insights on the mechanistic interplay between As and the gut microbiota with functional health consequences.

We previously identified associations between infant urinary As concentration and the gut microbiota at about 6 weeks of age (Hoen et al. 2018). Findings were especially evident among those receiving formula, as in the current study. Based on earlier work of our study and others (Carignan et al. 2016), formula results in higher As exposure due to both the formula powder and the water used to mix the formula. Household tap water can contain high levels of As in our rural cohort which whose households were served by private unregulated water systems such as bedrock wells as an eligibility criterion (Hoen et al. 2018). Eight genera, six within the phylum *Firmicutes*, were enriched with higher As exposure. This is consistent with our current results of a positive relationship with fecal butyrate and propionate concentrations which are produced by anaerobic fermentation of dietary carbohydrates by *Firmicutes* (Louis and Flint 2017; Appert et al. 2020).

SCFAs, succinate, formate, acetate, butyrate, and propionate. are important energy sources for intestinal epithelial cells, have diverse regulatory functions, and impact host physiology and immunity (Louis and Flint 2017; Appert et al. 2020). Of these, acetate has the highest systemic concentrations, as they can be produced by most gut anaerobes, whereas propionate and butyrate are products of only a select group of gut microbiota (Louis and Flint 2017). Butyrate, a SCFA produced by bacterial fermentation of dietary fiber, is considered a

key metabolite during infant gut development in part for its immunoregulatory effects (Roduit et al. 2019). Microbial metabolites, including butyrate, also are hypothesized to play an important role in the gut-brain axis by modulating the functional and signaling activity of brains cells and the blood–brain barrier and influence risk of neurodevelopmental outcomes such as autism (Smith 2015; Liu et al. 2019; Silva et al. 2020). There is evidence of functional redundancy of butyrate producers with co-occurrence of *Clostridiaceae, Ruminococcaceae,* and *Lachnospiraceae,* dominated by the endospore-forming *Clostridiaceae* (Appert et al. 2020). Further, proteolytic microbiota in the gut produce butyrate and propionate from peptide and amino acid fermentation (Louis and Flint 2017).

Propionate is a metabolite produced by genera in both *Firmicutes* and *Bacteroidetes,* including *Bacteroides* and *Clostridium* G2 (Gonzalez-Garcia et al. 2017; Louis and Flint 2017). Like propionate, malonate, which also was increased in relation to higher As exposures, plays a role in tricarboxylic acid cycle and other bacterial metabolic processes (Suvorova et al. 2012). In addition, malonate competitive inhibits succinate dehydrogenase and is involved in the metabolism of propionate (Suvorova et al. 2012). Thus, perturbations in either gut microbes that produce propionate, or in Krebs cycle metabolism that produces malonate, could in part explain the observed differences in propionate and malonate metabolism. However, our results need to be replicated in further studies.

In pathway analyses, we found alterations in metabolites enriched in MDSCs and M2 macrophages in cancer, and immune response, Treg cell-mediated modulation of APC functions) in relation to infant urinary As concentrations. MDSCs inhibit T cell function. T cell alterations occur with leukemia treatment with As (Gao et al. 2017) and have been observed both in highly drinking water-exposed populations (Burchiel et al. 2020), and in our own US cohort of infants exposed prenatally (Nygaard et al. 2017), we observed decreased cord blood naïve T-cells in relation to utAs concentrations during pregnancy.

Enriched pathways identified in our analyses included aminoacyl-tRNA biosynthesis and amino acid metabolism [lysine, BCAAs (isoleucine, leucine, valine), and methionine]. Martin and colleagues likewise found urinary As concentrations associated with differences in aminoacyl-tRNA-biosynthesis in plasma of adult diabetics from the Chihuahua cohort (Martin et al. 2015). In an As-exposed pregnancy cohort from Durango, Mexico, Laine, and colleagues found that maternal urinary iAs% and MMA% related to newborn cord blood aminoacyl-tRNA biosynthesis. Thus, our findings based on fecal metabolites likely reflect host as well as microbial metabolism.

In humans, As undergoes methyl metabolism, and methionine is involved in producing S-adenosylmethionine, the major methyl donor. Interestingly, methionine metabolism was among the top pathways enriched in our study. Both laboratory and population-based studies report changes in DNA methylation as well as H3 lysine 9 dimethylation (Howe and Gamble 2016) in relation to As exposure, which in turn could influence gene expression and cell fate (Tollervey and Lunyak 2012). Seventeen different types of PTMs on more than 30 amino acids have been identified for human H3 alone including the acetylation (ac) and methylation (me) of lysine residues (K) (Howe et al. 2017). Rats deprived of methionine

lived longer and had attenuated age-related T cell changes (Miller et al. 2005). Thus, higher fecal methionine concentrations could affect immune function and downstream gut microbiota composition. We observed few differences in our associations by primary or secondary methylation status; and not for methionine as found in a prior study of prenatal urinary As in relation to cord blood metabolites (Laine et al. 2017). Further, in our study, higher infant urinary As associated with higher fecal uracil. Arsenic trioxide was found to perturb SUMO- and folate-dependent nuclear de novo thymidylate (dTMP) biosynthesis, which can lead to misincorporation of uracil into DNA and genome instability (Kamynina et al. 2017).

We also found increases in isoleucine and leucine concentrations in relation to an infant's urinary As concentrations. In a *Caenorhabditis elegans* model, response to As toxicity was in part driven by genetic variation in *dbt-1* (Zdraljevic et al. 2019), which encodes the E2 subunit of the branched-chain keto acid dehydrogenase (BCKDH) complex involved in BCAA metabolism. BCAA changes following As treatment further suggested BCAA metabolism as a target of As toxicity. This is consistent with our findings that isoleucine and leucine may be altered in infants with higher As exposure, although additional mechanistic and epidemiologic studies are needed.

We found that tryptophan, which is converted by gut bacteria to indole (Lu et al. 2014) was positively associated with As concentrations, especially among those with higher PMIs; this suggests a possible reduction in gut microbial conversion of tryptophan into indole containing metabolites. Cholate is a primary bile acid synthesized in the liver from the oxidation of cholesterol. Bile acids further undergo deconjugation and dihydroxylation by gut microbes (Tian et al. 2020). In our study, the positive association with fecal cholate was stronger among those with lower SMIs. Metabolism of iAs, and accumulation of MMA, in particular, has been associated with a myriad of health outcomes (Abuawad et al. 2021). Whether fecal metabolic differences exist by As metabolic capacity (measured by urinary metabolites or genetic characterization) and whether these differences influence children's later risk of disease merits further investigation. Thus, while preliminary, our findings align with known processes and may inform new avenues of mechanistic exploration.

In summary, our exploratory study indicates that As exposure in infants may have functional perturbations in the infant gut microbiota–host interactions consistent with our previous microbiota analysis. These perturbations could be attributed primarily to bile acid metabolism, SCFA and organic acid metabolism, amino acid metabolism, and pyrimidine metabolism. Our findings further support earlier findings that As exposure in infants affects the developing infant gut microbiota.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This study was supported by National Institute of Environmental Health Sciences (Grant No. P01ES022832), National Library of Medicine (Grant No. R01LM012723) and National Institute of Diabetes and Digestive and Kidney Diseases (Grant No. U24DK097193).

References

Abuawad A, Bozack AK, Saxena R, Gamble MV (2021) Nutrition, one-carbon metabolism and arsenic methylation. Toxicology 457:152803 [PubMed: 33905762]

Appert O, Garcia AR, Frei R, Roduit C, Constancias F, Neuzil-Bunesova V, Ferstl R, Zhang J, Akdis C, Lauener R, Lacroix C, Schwab C (2020) Initial butyrate producers during infant gut microbiota development are endospore formers. Environ Microbiol 22(9):3909–3921 [PubMed: 32686173]

Beckonert O, Keun HC, Ebbels TM, Bundy J, Holmes E, Lindon JC, Nicholson JK (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nat Protoc 2(11):2692–2703 [PubMed: 18007604]

Bender R, Lange S (2001) Adjusting for multiple testing—when and how? J Clin Epidemiol 54(4):343–349 [PubMed: 11297884]

Brim H, Lee EL, Nelson KE, Smoot DT, Sears CL, Hassanzadeh H, Pathmasiri W, Sumner SC, Ashktorab H (2012) Mol671 a comprehensive taxonomic, metagenomic and metabolomic gut flora analysis reveals distinct profiles in healthy and colon adenoma African Americans. Gastroenterology 142(5):S655

Brim H, Yooseph S, Lee E, Sherif ZA, Abbas M, Laiyemo AO, Varma S, Torralba M, Dowd SE, Nelson KE, Pathmasiri W, Sumner S, de Vos W, Liang Q, Yu J, Zoetendal E, Ashktorab H (2017) A microbiomic analysis in African Americans with colonic lesions reveals *Streptococcus* sp. VT162 as a marker of neoplastic transformation. Genes (Basel) 8(11):314 [PubMed: 29120399]

Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, Dunn WB (2018) Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics 14(6):72 [PubMed: 29805336]

Burchiel SW, Lauer FT, Factor-Litvak P, Liu X, Islam T, Eunus M, Abu Horayara M, Islam MT, Rahman M, Ahmed A, Cremers S, Nanda-kumar R, Ahsan H, Olopade C, Graziano J, Parvez F (2020) Arsenic exposure associated T cell proliferation, smoking, and vitamin D in Bangladeshi men and women. PLoS ONE 15(6):e0234965 [PubMed: 32574193]

Carignan CC, Punshon T, Karagas MR, Cottingham KL (2016) Potential exposure to arsenic from infant rice cereal. Ann Glob Health 82(1):221–224 [PubMed: 27325082]

Chan EC, Pasikanti KK, Nicholson JK (2011) Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. Nat Protoc 6(10): 1483–1499 [PubMed: 21959233]

Chi L, Gao B, Tu P, Liu CW, Xue J, Lai Y, Ru H, Lu K (2018) Individual susceptibility to arsenic-induced diseases: the role of host genetics, nutritional status, and the gut microbiome. Mamm Genome 29(l-2):63–79 [PubMed: 29429126]

Coryell M, Roggenbeck BA, Walk ST (2019) The human gut microbiome's influence on arsenic toxicity. Curr Pharmacol Rep 5(6):491–504 [PubMed: 31929964]

Dangleben NL, Skibola CF, Smith MT (2013) Arsenic immunotoxicity: a review. Environ Health 12(1):73 [PubMed: 24004508]

Dietert RR, DeWitt JC, Germolec DR, Zelikoff JT (2010) Breaking patterns of environmentally influenced disease for health risk reduction: immune perspectives. Environ Health Perspect 118(8): 1091–1099 [PubMed: 20483701]

Dona AC, Jimenez B, Schafer H, Humpfer E, Spraul M, Lewis MR, Pearce JT, Holmes E, Lindon JC, Nicholson JK (2014) Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. Anal Chem 86(19):9887–9894 [PubMed: 25180432]

EFSA Panel on Contaminants in the Food Chain (2009) Scientific opinion on arsenic in food. EFSA J 7(10): 1351

Farzan SF, Li Z, Korrick SA, Spiegelman D, Enelow R, Nadeau K, Baker E, Karagas MR (2016) Infant infections and respiratory symptoms in relation to in utero arsenic exposure in a U.S. Cohort: Environ Health Perspect 124(6):840–847 [PubMed: 26359651]

Gao Q, Jiang J, Chu Z, Lin H, Zhou X, Liang X (2017) Arsenic trioxide inhibits tumor-induced myeloid-derived suppressor cells and enhances T-cell activity. Oncol Lett 13(4):2141–2150 [PubMed: 28454374]

Gao S, Lin P-I, Mostofa G, Quamruzzaman Q, Rahman M, Rahman ML, Su L, Hsueh Y-M, Weisskopf M, Coull B, Christiani DC (2019a) Determinants of arsenic methylation efficiency and urinary arsenic level in pregnant women in Bangladesh. Environ Health 18(1):94 [PubMed: 31690343]

Gao S, Mostofa MG, Quamruzzaman Q, Rahman M, Rahman M, Su Hsueh YM, Weisskopf M, Coull B, Christiani DC (2019b) Gene–environment interaction and maternal arsenic methylation efficiency during pregnancy. Environ Int 125:43–50 [PubMed: 30703610]

Gardner RM, Nermell B, Kippler M, Grandér M, Li L, Ekström EC, Rahman A, Lönnerdal B, Hoque AM, Vahter M (2011) Arsenic methylation efficiency increases during the first trimester of pregnancy independent of folate status. Reprod Toxicol 31 (2):210–218 [PubMed: 21078382]

Gonzalez-Garcia RA, McCubbin T, Navone L, Stowers C, Nielsen LK, Marcellin E (2017) Microbial propionic acid production. Fermentation 3(2):21

Hoen AG, Madan JC, Li Z, Coker M, Lundgren SN, Morrison HG, Palys T, Jackson BP, Sogin ML, Cottingham KL, Karagas MR (2018) Sex-specific associations of infants' gut microbiome with arsenic exposure in a US population. Sci Rep 8(1): 12627 [PubMed: 30135504]

Hopenhayn C, Huang B, Christian J, Peralta C, Ferreccio C, Atallah R, Kalman D (2003) Profile of urinary arsenic metabolites during pregnancy. Environ Health Perspect 111(16): 1888–1891 [PubMed: 14644662]

Howe CG, Gamble MV (2016) Influence of arsenic on global levels of histone posttranslational modifications: a review of the literature and challenges in the field. Curr Environ Health Rep 3(3):225–237 [PubMed: 27352015]

Howe CG, Liu X, Hall MN, Ilievski V, Caudill MA, Malysheva O, Lomax-Luu AM, Parvez F, Siddique AB, Shahriar H, Uddin MN, Islam T, Graziano JH, Costa M, Gamble MV (2017) Sex-specific associations between one-carbon metabolism indices and post-translational histone modifications in arsenic-exposed Bangladeshi adults. Cancer Epidemiol Biomark Prev 26(2):261–269

Kamynina E, Lachenauer ER, DiRisio AC, Liebenthal RP, Field MS, Stover PJ (2017) Arsenic trioxide targets MTHFD1 and SUMO-dependent nuclear de novo thymidylate biosynthesis. Proc Natl Acad Sci USA 114(12):E2319–E2326 [PubMed: 28265077]

Kapp RW (2018) Arsenic toxicology ☆. In: Reference module in biomedical sciences. Elsevier, Amsterdam

Laine JE, Bailey KA, Olshan AF, Smeester L, Drobná Z, Stýblo M, Douillet C, García-Vargas G, Rubio-Andrade M, Pathmasiri W, McRitchie S, Sumner SJ, Fry RC (2017) Neonatal metabolomic profiles related to prenatal arsenic exposure. Environ Sci Technol 51(1):625–633 [PubMed: 27997141]

Laue HE, Moroishi Y, Jackson BP, Palys TJ, Madan JC, Karagas MR (2020) Nutrient-toxic element mixtures and the early postnatal gut microbiome in a United States Longitudinal Birth Cohort. Environ Int 138:105613 [PubMed: 32142916]

Li M, Wang B, Zhang M, Rantalainen M, Wang S, Zhou H, Zhang Y, Shen J, Pang X, Zhang M, Wei H, Chen Y, Lu H, Zuo J, Su M, Qiu Y, Jia W, Xiao C, Smith LM, Yang S, Holmes E, Tang H, Zhao G, Nicholson JK, Li L, Zhao L (2008) Symbiotic gut microbes modulate human metabolic phenotypes. Proc Natl Acad Sci USA 105(6):2117–2122 [PubMed: 18252821]

Li X, Brejnrod AD, Ernst M, Rykaer M, Herschend J, Olsen NMC, Dorrestein PC, Rensing C, Sorensen SJ (2019) Heavy metal exposure causes changes in the metabolic health-associated gut microbiome and metabolites. Environ Int 126:454–467 [PubMed: 30844581]

Liu S, Li E, Sun Z, Fu D, Duan G, Jiang M, Yu Y, Mei L, Yang P, Tang Y, Zheng P (2019) Altered gut microbiota and short-chain fatty acids in Chinese children with autism spectrum disorder. Sci Rep 9(1):287 [PubMed: 30670726]

Livanos AE, Greiner TU, Vangay P, Pathmasiri W, Stewart D, McRitchie S, Li H, Chung J, Sohn J, Kim S, Gao Z, Barber C, Kim J, Ng S, Rogers AB, Sumner S, Zhang XS, Cadwell K, Knights D, Alekseyenko A, Backhed F, Blaser MJ (2016) Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. Nat Microbiol 1(11): 16140 [PubMed: 27782139]

Louis P, Flint HJ (2017) Formation of propionate and butyrate by the human colonic microbiota. Environ Microbiol 19(1):29–41 [PubMed: 27928878]

Lu K, Abo RP, Schlieper KA, Graffam ME, Levine S, Wishnok JS, Swenberg JA, Tannenbaum SR, Fox JG (2014) Arsenic exposure perturbs the gut microbiome and its metabolic profile in mice: an integrated metagenomics and metabolomics analysis. Environ Health Perspect 122(3):284–291 [PubMed: 24413286]

Madan JC, Farzan SF, Hibberd PL, Karagas MR (2012) Normal neonatal microbiome variation in relation to environmental factors, infection and allergy. Curr Opin Pediatr 24(6):753–759 [PubMed: 23111681]

Martin E, Gonzalez-Horta C, Rager J, Bailey KA, Sanchez-Ramirez B, Ballinas-Casarrubias L, Ishida MC, Gutierrez-Torres DS, Hernandez Ceron R, Viniegra Morales D, Baeza Terrazas FA, Saunders RJ, Drobna Z, Mendez MA, Buse JB, Loomis D, Jia W, Garcia-Vargas GG, Del Razo LM, Styblo M, Fry R (2015) Metabolomic characteristics of arsenic-associated diabetes in a prospective cohort in Chihuahua, Mexico. Toxicol Sci 144(2):338–346 [PubMed: 25577196]

Masson P, Spagou K, Nicholson JK, Want EJ (2011) Technical and biological variation in UPLC–MS-based untargeted metabolic profiling of liver extracts: application in an experimental toxicity study on galactosamine. Anal Chem 83(3): 1116–1123 [PubMed: 21241057]

McDermott TR, Stolz JF, Oremland RS (2020) Arsenic and the gastrointestinal tract microbiome. Environ Microbiol Rep 12(2): 136–159 [PubMed: 31773890]

Miller RA, Buehner G, Chang Y, Harper JM, Sigler S, Smith-Wheelock M (2005) Methionine-deficient diet extends mouse lifespan, slows immune and lens aging, alters glucose, T4, IGF-I and insulin levels, and increases hepatocyte MIF levels and stress resistance. Aging Cell 4(3): 119–125 [PubMed: 15924568]

Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, Pettersson S (2012) Host-gut microbiota metabolic interactions. Science 336(6086): 1262–1267 [PubMed: 22674330]

Nygaard UC, Li Z, Palys T, Jackson B, Subbiah M, Malipatlolla M, Sampath V, Maecker H, Karagas MR, Nadeau KC (2017) Cord blood T cell subpopulations and associations with maternal cadmium and arsenic exposures. PLoS ONE 12(6):e0179606 [PubMed: 28662050]

Roduit C, Frei R, Ferstl R, Loeliger S, Westermann P, Rhyner C, Schiavi E, Barcik W, Rodriguez-Perez N, Wawrzyniak M, Chassard C, Lacroix C, Schmausser-Hechfellner E, Depner M, von Mutius E, Braun-Fahrlander C, Karvonen AM, Kitjavainen PV, Pekkanen J, Dalphin JC, Riedler J, Akdis C, Lauener R, O'Mahony L, PASTURE/EFRAIM Study Group (2019) High levels of butyrate and propionate in early life are associated with protection against atopy. Allergy 74(4):799–809 [PubMed: 30390309]

Shen H, Niu Q, Xu M, Rui D, Xu S, Feng G, Ding Y, Li S, Jing M (2016) Factors affecting arsenic methylation in arsenic-exposed humans: a systematic review and meta-analysis. Int J Environ Res Public Health 13(2):205 [PubMed: 26861378]

Signes-Pastor AJ, Cottingham KL, Carey M, Sayarath V, Palys T, Meharg AA, Folt CL, Karagas MR (2018) Infants' dietary arsenic exposure during transition to solid food. Sci Rep 8(1):7114 [PubMed: 29739998]

Silva YP, Bernardi A, Frozza RL (2020) The role of short-chain fatty acids from gut microbiota in gut-brain communication. Front Endocrinol 11:25

Smith PA (2015) The tantalizing links between gut microbes and the brain. Nature 526(7573):312–314 [PubMed: 26469024]

Suvorova IA, Ravcheev DA, Gelfand MS (2012) Regulation and evolution of malonate and propionate catabolism in proteobacteria. J Bacteriol 194(12):3234–3240 [PubMed: 22505679]

Tian Y, Gui W, Koo I, Smith PB, Allman EL, Nichols RG, Rimal B, Cai J, Liu Q, Patterson AD (2020) The microbiome modulating activity of bile acids. Gut Microbes 11(4):979–996 [PubMed: 32138583]

Tollervey JR, Lunyak VV (2012) Epigenetics. Epigenetics 7(8): 823–840 [PubMed: 22805743]

Tseng CH (2009) A review on environmental factors regulating arsenic methylation in humans. Toxicol Appl Pharmacol 235(3):338–350 [PubMed: 19168087]

Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM (2006) Targeted profiling: quantitative analysis of ^{1}H NMR metabolomics data. Anal Chem 78(13):4430–4442 [PubMed: 16808451]

Xi B, Gu H, Baniasadi H, Raftery D (2014) Statistical analysis and modeling of mass spectrometry-based metabolomics data. Methods Mol Biol (Clifton NJ) 1198:333–353

Zdraljevic S, Fox BW, Strand C, Panda O, Tenjo FJ, Brady SC, Crombie TA, Doench JG, Schroeder FC, Andersen EC (2019) Natural variation in *C. elegans* arsenic toxicity is explained by differences in branched-chain amino acid metabolism. eLife 8:e40260 [PubMed: 30958264]

Zheng X, Xie G, Zhao A, Zhao L, Yao C, Chiu NH, Zhou Z, Bao Y, Jia W, Nicholson JK, Jia W (2011) The footprints of gut microbial-mammalian co-metabolism. J Proteome Res 10(12):5512–5522 [PubMed: 21970572]

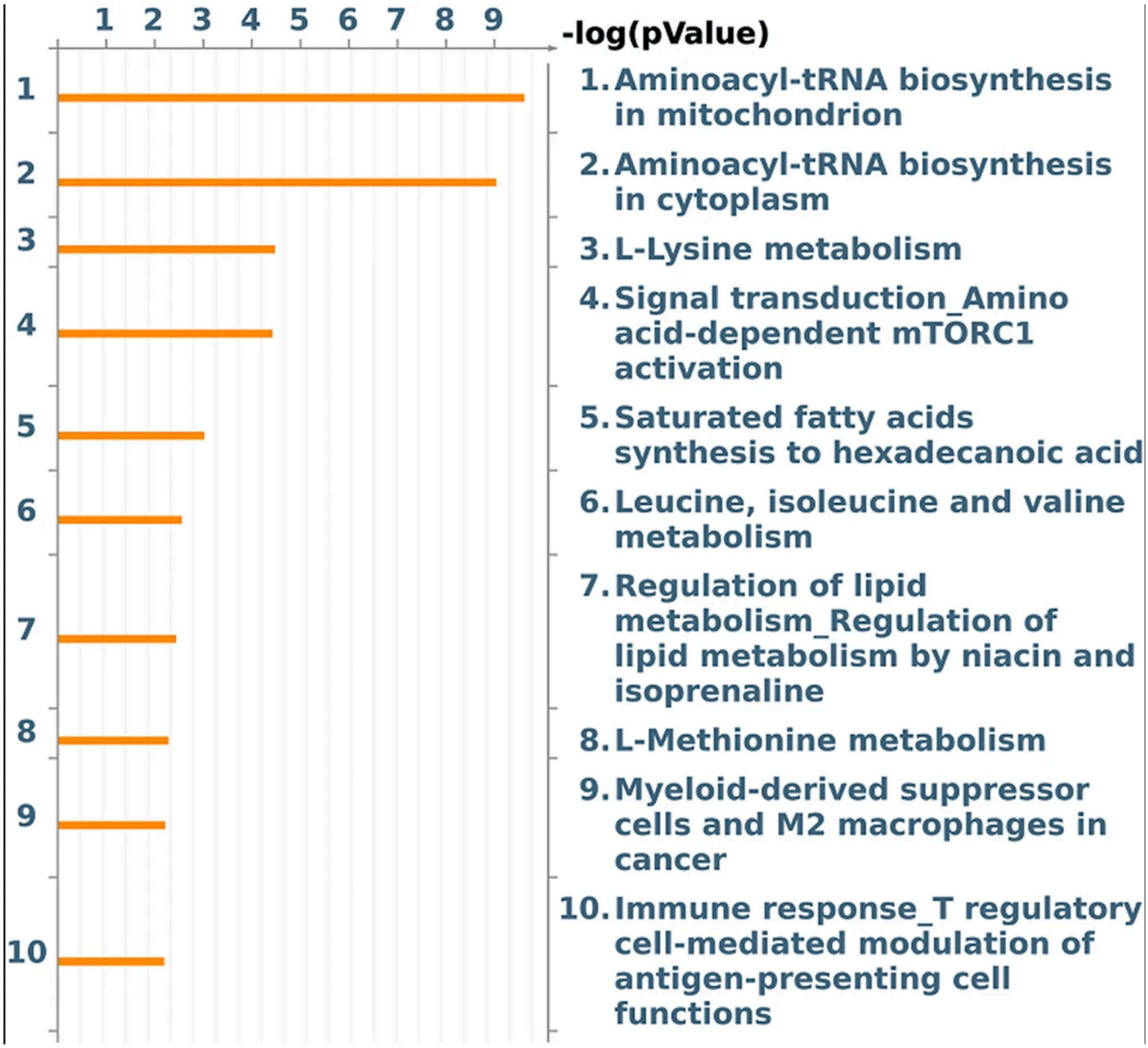

Fig. 1.

Top 10 enriched pathways in the MetaCore pathway enrichment analysis for the metabolites (concentration data) associated with infant utAs. See Supplementary Table 6 or the full list of enriched pathways

Table 1

Selected characteristics of infants from the New Hampshire Birth Cohort Study

	Maternal utAs analysis (*n* = 83)	**Infant utAs analysis (*n* = 81)**
Characteristic	Mean [range] or *N*(%)	Mean [range] or *N*(%)
Infant age at sample collection (days)	45.5 [28.0, 115.0]	45.6 [28.0, 115.0]
Sex		
Male	53 (63.9%)	50(61.7%)
Female	30 (36.1%)	31 (38.3%)
Delivery mode		
C-section	22 (26.5%)	21 (25.9%)
Vaginal	61 (73.5%)	60(74.1%)
Feeding at ~ 6 weeks		
Exclusively breast fed	35 (42.2%)	34 (42.0%)
Formula/mixed fed	48 (57.8%)	47 (58.0%)
Missing	0	0
Maternal age at enrollment (years)	31.5 [21.1,43.7]	N/A*
Maternal urine, total As excluding AsB (mg/L)	4.1 [0.2,21.0]	N/A*
Infant urine, total As excluding AsB (mg/L)	N/A*	0.6 [0.1, 5.2]
iAs	N/A*	0.1 [0, 1.01
MMA	N/A*	0.4 [0, 41
DMA	N/A*	0.1 [0, 4.5]

***NIA* not applicable

Table 2

Changes in microbe related metabolite concentrations with a doubling (log_2) of infant urinary arsenic concentrations ($n = 81$) from multivariable linear regression models

Metabolite	β (95% Q)	*p*-value
Short-chain fatty acids		
Butyrate	214.24 (83.78, 344.70)	0.0016
Propionate	518.33 (94.39,942.28)	0.0172
Isobutyrate	2.61 (−0.86,6.07)	0.1378
Lipids		
Glycerol	−67.89 (−139.16, 3.38)	0.0616
Bile acid		
Cholate	8.79(4.21, 13.36)	0.0003
Amino acids		
Tryptophan	14.23 (3.71,24.74)	0.0087
Lysine	63.29 (−42.79, 169.36)	0.2383
Asparagine	28.80(10.27,47.33)	0.0028
Methionine	16.91 (−5.32, 39.14)	0.1338
Proline	17.09 (−28.28, 62.46)	0.4553
Isoleucine	65.58 (10.56, 120.60)	0.0201
Leucine	95.91 (8.88, 182.95)	0.0312
Glutamate	119.69 (−22.02,261.40)	0.0966
Phenylalanine	18.38 (−12.46,49.23)	0.2388
Sugars		
Fucose	−126.31 (−307.21,54.59)	0.1683
Organic acid		
Malonate	50.43 (3.14,97.72)	0.0369
Pyrimidine		
Uracil	36.13 (2.74,69.52)	0.0343
Other		
Propylene glycol	−62.58 (−128.62, 3.46)	0.0629

Adjusted for infant sex, age, type of delivery (vaginal vs. C-section), feeding mode (breast milk vs. any formula), and specific gravity

REVIEW

Machine Learning in Nutrition Research

Daniel Kirk,[1] Esther Kok,[1] Michele Tufano,[1] Bedir Tekinerdogan,[2] Edith JM Feskens,[1] and Guido Camps[1,3]

[1] Division of Human Nutrition and Health, Wageningen University and Research, Wageningen, The Netherlands; [2] Information Technology Group, Wageningen University and Research, Wageningen, The Netherlands; and [3] OnePlanet Research Center, Wageningen, The Netherlands

ABSTRACT

Data currently generated in the field of nutrition are becoming increasingly complex and high-dimensional, bringing with them new methods of data analysis. The characteristics of machine learning (ML) make it suitable for such analysis and thus lend itself as an alternative tool to deal with data of this nature. ML has already been applied in important problem areas in nutrition, such as obesity, metabolic health, and malnutrition. Despite this, experts in nutrition are often without an understanding of ML, which limits its application and therefore potential to solve currently open questions. The current article aims to bridge this knowledge gap by supplying nutrition researchers with a resource to facilitate the use of ML in their research. ML is first explained and distinguished from existing solutions, with key examples of applications in the nutrition literature provided. Two case studies of domains in which ML is particularly applicable, precision nutrition and metabolomics, are then presented. Finally, a framework is outlined to guide interested researchers in integrating ML into their work. By acting as a resource to which researchers can refer, we hope to support the integration of ML in the field of nutrition to facilitate modern research. *Adv Nutr* 2022;13:2573–2589.

Statement of Significance: Many problems in nutrition are complex, multifactorial, and unlikely to be solved with data analysis methods that have been used traditionally; however, the capabilities of machine learning may be able to. For nutrition researchers to fully capitalize on the types of data that will be generated in the coming years, we provide a guide to machine learning in nutrition for nutrition researchers.

Keywords: machine learning, personalized nutrition, omics, obesity, diabetes, cardiovascular disease, models, random forest, XGBoost

Introduction

There is a high prevalence of nutritionally mediated chronic diseases that have multifaceted origins and require complex and diverse data to be solved. Traditional research applications have approached these questions with focused and mechanistic techniques that may not fully capture the complexity of the interaction between nutrition and disease. Technological and computational advances have recently allowed investigators to utilize high-dimensional data approaches to better understand these diseases and other complex questions. Topical themes, such as obesity (1, 2), omics (3, 4), and the microbiome (5–7), as well as older subjects such as epidemiology, are deriving benefits from these developments (8–10). Due to the increasing complexity of the data generated, new trends in nutrition research, such as precision nutrition (PN) (11) and data-driven disease modeling (12, 13), require an increasing complexity in algorithms to make sense of these data; artificial intelligence (AI) and its subdivision, machine learning (ML), have been important for this.

The terms AI and ML are used interchangeably in some of the literature (13), exposing some of the conceptual confusion that surrounds these topics. The overall goal of AI is to simulate human-like intelligence in a computer system (14). ML is the overarching term used for a subset of algorithms that help achieve this goal. These algorithms are self-learning from the data with which they are presented and can identify complex underlying patterns in data; they are also capable of processing unstructured types of data that traditional statistical techniques are incapable of doing, such as free text, images, video, and audio. Making such unstructured data available for use by ML algorithms increases the amount and potentially the quality of information available, which can lead to better predictive capacity.

By using ML algorithms, the work in the field of AI so far has been able to build systems that are performing well in a specific task. However, outside the scope of that task, most of these systems perform poorly, meaning that true intelligence has yet to be achieved (14). For nutritionists, the application of ML algorithms to their data is not

to approximate human intelligence but rather to process complex data to generate results relevant to health and disease or to process large volumes of complex data—in other words, to apply these algorithms to a very specific scope of the task. Within the field of nutrition, specific tasks that have already benefited from ML algorithms are related to finding causes and potential solutions for many nutrition-related noncommunicable diseases, such as obesity, diabetes, cancer, and CVD, all of which have a complex and multifactorial etiology (3, 4, 14–19). These works have shown promise to the application of ML in solving the biggest challenges in nutrition and to the opportunities before us.

The direction of research in the discipline of nutrition science is going increasingly toward one that would benefit from the use of advanced tools, from data generation through to explanation and prediction. ML has the potential to supplement existing techniques to generate and analyze complex data, but to do so, it must be applied appropriately. Although the use of AI and ML does not require extensive background knowledge in computer science or mathematics, the application of ML without appropriate understanding can lead to biased models and results that do not represent real-world representative. For a nutritionist without prior ML experience, approaching the subject area can be overwhelming, which subsequently hampers adoption of its use by interested researchers.

This article aims to deal with this by providing a resource to which nutrition researchers can refer to guide their efforts. First, ML itself and its distinction from traditional techniques are explained. Next, an overview of ML is provided, and key concepts are elucidated. This covers core ideas in ML, such as ML types, tasks, data types, common algorithms, explainable AI (xAI), and ML performance evaluation. Throughout the aforementioned sections, application examples from the literature are provided. Related terminology can be found in Supplemental Table 1. Following this, a short review of nutrition-orientated literature utilizing ML is elaborated in the form of case studies of areas in nutrition science that are currently employing ML. Finally, practical application is supported by providing a framework for implementing ML in nutrition research. By providing a key reference, we enable the continuation of groundbreaking research by circumventing the problem that ML-naive researchers encounter when dealing with complex problems.

This project was partially funded by the 4TU–Pride and Prejudice program (4TU-UIT-346; 4 Dutch Technical Universities).
Author disclosures: The authors report no conflicts of interest.
Supplemental Tables 1 and 2, Supplemental Figure 1, and Supplemental Material are available from the "Supplementary data" link in the online posting of the article and from the same link in the online table of contents at https://academic.oup.com/advances/.
Address correspondence to DK (e-mail: daniel.kirk@wur.nl).
Abbreviations used: AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; CV, cross-validation; IBS, irritable bowel syndrome; kNN, k-nearest neighbors; LIME, local interpretable model-agnostic explanations; ML, machine learning; NAFLD, nonalcoholic fatty liver disease; NLP, natural language processing; PCA, principal component analysis; PN, precision nutrition; PPGR, postprandial glucose response; RF, random forest; SHAP, Shapley additive explanations; SVM, support vector machine; T2D, type 2 diabetes; xAI, explainable AI.

Machine Learning Capabilities

ML is a subdivision of AI that employs algorithms to complete a task by learning from patterns in the data, rather than being explicitly programmed to do so. This is achieved by defining an objective (e.g., predicting a numerical value), evaluating performance, and then performing experiments recursively to optimize the model. Whereas AI falls short of replicating the complexity of human thinking, it excels in certain aspects of learning, making it faster, able to deal with high-dimensional data, and able to learn abstract patterns (15, 16). These aspects of ML make it more suitable for tasks than traditional statistical techniques and certain domain-specific techniques and so gift ML with practical advantages that make it attractive, as discussed throughout.

Whereas the current section focuses on situations in which ML may be advantageous over traditionally used methods, researchers should be encouraged to consider each option as another tool in the box rather than using one or the other. Understanding the advantages and disadvantages of various methods and learning when and how to apply each one can lead to synergy and more fruitful results by allowing methods to complement one another. Researchers should be encouraged to think deeply about their problem and the research questions that they would like to answer and to select the appropriate techniques that best do this. Whenever possible, researchers should also consider experimenting with multiple options and selecting those most suitable. The idea of selecting the pool of solutions to suit the problem at hand is discussed in detail in the Framework for Applying ML in Nutrition Science section.

Machine Learning and Traditional Statistical Methods

When the goal is inference; interpretability is paramount; and the features are well established, simple, known a priori, and low-dimensional, traditional statistical techniques such as regression methods may suffice (1). However, researchers often choose these approaches due to familiarity, despite that ML techniques can be more suitable and efficacious in certain circumstances. ML is suited for high-dimensional data and when the goal is predictive performance (17). The capability of ML to learn from patterns in the data means that precise and premeditated variable selection is not a necessity; instead, many variables can be trialed, and repeated experimentation can quickly identify those of most relevance. That is, ML can be applied exploratively (17), which may lead to the discovery of novel predictive features, sometimes serendipitously (18–24).

By perceiving data to have been generated by a stochastic model, traditional statistics is limited by the assumptions that it makes: assumptions that are increasingly unjust as data are increasingly complex in domains relevant to nutrition such as health (25). When predicting mortality with epidemiologic data sets, Song et al. (26) noted that the nonlinear capabilities of more sophisticated ML techniques explain their consistently superior performance. Stolfi and Castiglione (27) integrated metabolic, nutritional,

and lifestyle data in an emulator for a handheld device application in the context of precision medicine that predicts processes in the development of type 2 diabetes (T2D). They noted that such a dynamic and high-dimensional system is too computationally demanding for statistical methods traditionally used for emulation. Compared with classifiers or regressors that assume linearity, ensemble predictors (seeSupplemental Figure S1, for example) applied to medical data consistently perform better (1, 21, 25, 28). Whereas ensemble predictors are considered uninterpretable, Breiman (25) made the interesting case that although the mechanism by which outputs are generated is not entirely transparent, this is counterbalanced with the advantage of making a more accurate prediction and in this way better represents the true process of data generation than do white box models. Other authors noted the inadequacy of statistical techniques for dealing with complex data derived from such subject areas as obesity (1), omics (29), and the microbiota (5). Ultimately, as data become more complex, the advantage of presenting a simplified representation of the data comes at a cost, a characteristic of classic statistical techniques from which ML suffers less.

Machine Learning and Traditional Statistical Methods

ML can be used to supplement domain-specific data analysis techniques. Gou et al. (20) linked 2 domains suitable for ML application—diabetes and the microbiome—by using ML to generate risk scores for T2D development based on microbiome composition. After pointing out that analysis of microbiota data is beyond the capabilities of classical statistical tools, they proceeded with ML to predict T2D better than traditional, domain-specific diabetes risk factors while identifying 11 novel microbial taxa predictive for T2D risk. Tap et al. (22) showed that conventional ecologic approaches did not find differences in microbiota signatures between patients with irritable bowel syndrome (IBS) and controls, whereas their ML approach (Lasso) was able to link intestinal microbiota signatures with IBS symptom severity.

Activity tracking utilizes unstructured data of movement generated from wearable devices to predict activity types and calorie expenditure, making it suitable for ML applications. Compared with domain-specific cut points, classification via ML techniques reduces misclassification rate, increases generalizability, allows grading of movement quality, and simplifies experimental design (30–32). Energy expenditure estimation traditionally uses methods that are expensive (e.g., doubly labeled water), impractical (e.g., indirect calorimetry with breathing masks), or non–free-living (e.g., direct calorimetry). Systems that analyze accelerometer data, with or without other physiologic data, can be adequate alternatives for the prediction of energy expenditure in a free-living, practical, and cost-effective manner (33, 34). In CVD research, CVD risk scores may be generated by using various biomarkers and are deployed in clinical practice; here too, ML techniques outperform traditional risk scores, making use of and identifying novel biomarkers in the process (19, 23, 35–38). ML is also a promising alternative for domain-specific techniques that are expensive, invasive, or both, as with nonalcoholic fatty liver disease (NAFLD) (39–41) and cancer (42–47).

Machine Learning: Practical Advantages

There are practical advantages of ML. Since the computer learns itself to complete a task, time and effort need not be invested in instructing the computer on what to do. This not only saves time and effort that would otherwise be spent on programming but also increases adaptability to solve various problems. That is, the same algorithm can be retrained on various data sets and problems. On a similar note, ML accepts various data types as input, including structured (e.g., tabular data) and unstructured (e.g., image based). In some cases, the same ML algorithms can be applied to different problems, perform different tasks, and take as input different data types; neural networks and k-nearest neighbors (kNN) are such examples.

By predicting an outcome based on existing data, ML algorithms can save on the time and cost of having to verify such outcomes experimentally. For example, Sorino et al. (48) concluded that incorporating ML algorithms into the analysis of noninvasive and comparatively cheaper variables could avoid 81.9% of unnecessary ultrasound scans in NAFLD, which are expensive and have long waiting times for results.

The tools required for performing ML experiments are minimal; other than a computer and a virtual environment to work in, all that is needed is a data set. To this end, data are being increasingly generated, and recent pushes for data sharing are meaning that more and more data sets are publicly available, suggesting that researchers across the world are able to run experiments and derive meaningful results in their field without the need for grants, research equipment, or generation of the data themselves. This is an extremely empowering aspect of ML, and if more researchers were better able to mine their own and others' data, more scientific progress could be expected.

Machine Learning Overview

Types

Four types of ML exist, each differing in the way that it learns, the algorithms that it employs, and its uses. A graphical depiction of each learning type is provided in **Figure 1**.

Supervised learning.

In supervised learning, the data come with labels in which the value or class being predicted by the algorithm is known, meaning that performance can be objectively verified. This is commonly observed in predictive models utilizing data sets with health variables and a disease outcome, such as CVD, T2D, and plasma nutrient prediction (49, 50). Since the labels of the data are required, human intervention plays a larger role than in other ML types, which can increase costs and time (51).

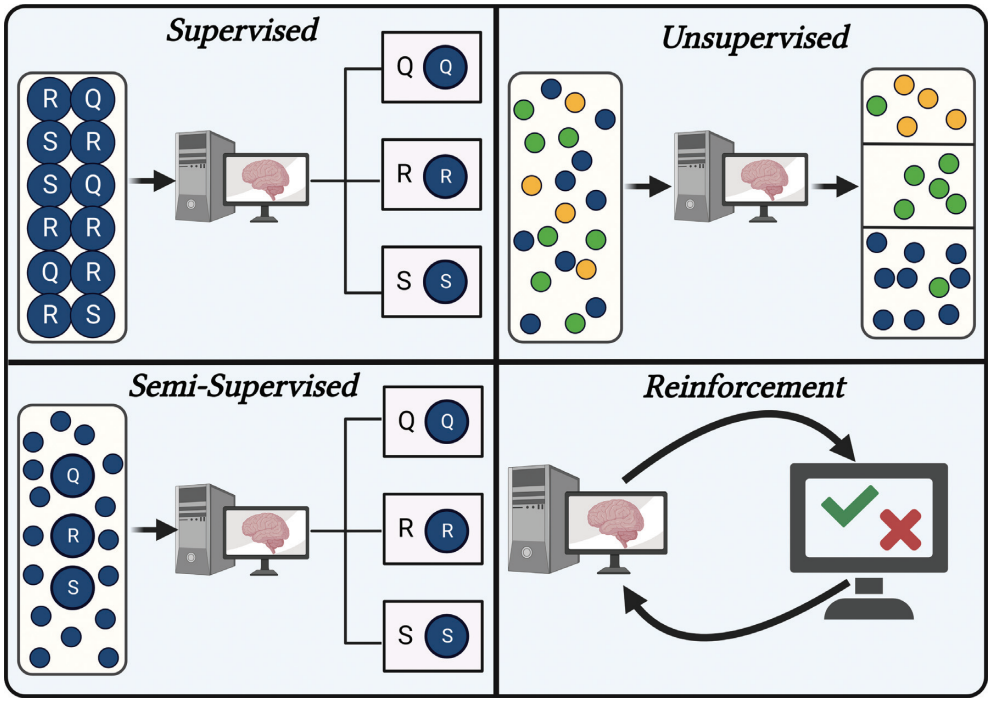

FIGURE 1 Four types of machine learning. In supervised learning, labels are provided in the data for objective evaluation of algorithm performance, whereas in unsupervised learning, the algorithm partitions the data based on similarity. In semi-supervised, only a portion of the data comes with labels, although all data are eventually classified. Reinforcement learning makes use of penalties and rewards in a dynamic environment to train the algorithm.

Unsupervised learning.
Unsupervised learning occurs without labels; instead, the algorithms seek to find patterns in the data and partition them based on similarity. This reduces human intervention, saving time on feature engineering and labeling. The most common use of unsupervised learning is clustering, dimensionality reduction and anomaly detection are also unsupervised. Unsupervised learning has been applied extensively in phenotyping, such as grouping individuals for PN (11). Unsupervised learning can also be used as a processing step before a supervised task to homogenize the data, as evidenced by Ramyaa et al. (52), who predicted BMI in women more accurately after phenotyping than when using the data as a whole. Another attractive use of unsupervised learning is hypothesis generation; because this type of learning works on the detection of patterns, this may lead to the formation of previously unidentified groups in the data.

Semi-supervised learning.
Semi-supervised is somewhat in between the 2 previously defined learning types in that labels are partially present but usually mostly absent. Providing labels on a subset of the data has the advantages of improving accuracy and generalizability while sparing the time and financial costs of labeling an entire data set (53). Consequently, semi-supervised learning has been used to study the influence of genes on disease outcomes when the known genes (i.e., the labeled data) are few (54).

Another example of semi-supervised learning is constrained clustering, which expects that certain criteria are satisfied during cluster formation, such as given data points being necessarily in the same or different clusters (55). This can be a way to circumvent potential issues that can arise when dealing with biological or health data in unsupervised learning, such as the grouping or separation of data points that violate plausibility—for example, the clustering of data of one biological sex with another in a system where this is not possible. However, it should be kept in mind that such findings may also provide interesting information about the data and adding such constraints may mask this.

Reinforcement learning.
In reinforcement learning, the algorithm exists in a dynamic environment and is penalized or rewarded for the decisions that it makes within the environment. The algorithm then updates its behavior to maximize reward, minimize penalization, or both. This allows the algorithm to become proficient in a task without being explicitly programmed to behave in a certain way. A famous example of reinforcement learning is Alpha Go Zero, which was able to achieve superhuman performance playing Go (Weiqi) with only a few hours of training by playing against itself (56). The complex nature of reinforcement learning limits application in simple classification or regression tasks and is instead used where the

integration of complex and varied data is concerned, such as recommender systems (57) or mobile-based fitness apps (58).

Tasks

ML algorithms are employed to complete tasks, which are distinguished into various categories. To complete these tasks, algorithms are used. Various of these are mentioned throughout the current subsection; for a more detailed description of each algorithm, see the Supplemental Material.

Regression.

Regression involves the prediction of a continuous variable based on one or more input variables. In the case of linear regression, a linear relationship between the input variables and the dependent variable is assumed, whereas in nonlinear regression, relationships can be more complex. As well as basic linear regression and its variants (e.g., Ridge, Lasso), more complex algorithms can be used, including some that are more often associated with classification, such as random forest (RF) and support vector machines (SVMs) (59, 60).

Classification.

Classification tasks aim to predict the class labels of data based on their independent variables. Data of the same class will likely have similar characteristics, at least for variables that contribute most to the classification decision; this forms the basis for how an algorithm learns to assign classes to a data point. In binary classification, there are only 2 labels, whereas with multiclass classification, there can be many.

Sample uses of classification include predicting adherence to exercise regimes (61) and image-based food recognition for dietary intake monitoring (11). Classification has significant overlap with regression in that oftentimes a regression problem can be converted to a classification task with only slight modifications and vice versa. This is reflected in the algorithms that can do both, such as SVM, RF, decision trees, and kNN.

Clustering.

Similar to classification, algorithms in clustering split the data based on similar characteristics, but clustering differs in that it is unsupervised, meaning that there is no ground truth or class labels to which the data points should belong. Thus, the goal is to obtain clusters that are more homogeneous than the data as a whole. Because it is typically unsupervised, clustering can be performed with or without expectations, leading to new discoveries and hypothesis generation. A well-known and popular application of clustering in nutrition and health research is that of phenotyping individuals based on shared characteristics, such as microbiome profiles (62) or identifying activity patterns (63). The most common clustering algorithm is *k*-means (for numerical data), with adaptations including *k*-modes (for categorical data) and *k*-prototypes (for mixed data). Other examples include density-based spatial clustering and mean-shift clustering (51).

Recommendation.

Recommender systems use data to generate a recommendation on a decision to be taken and have been used in nutrition to suggest meals to help manage chronic diseases (64–67). Recommender systems can be further classified into subtypes such as collaborative filtering, content based, and popularity based, as well as hybridizations of each. Recommender systems can be complex and may require the integration of multiple components, each of which may involve different ML tasks and algorithms. For example, Baek et al. (68) described a recommender method that clusters individuals based on chronic disease status, suggests suitable foods for each cluster, and considers the preferences of the individual and on the universal level. Because recommendation systems can involve lots of data, deep neural networks are often utilized.

Dimensionality reduction.

When working with data sets with many features, dimensionality can be problematic; it slows computation, may reduce accuracy, and can cause overfitting (69). This is particularly relevant in the modern age where high-dimensional data are being generated (70). Dimensionality reduction techniques aim to reduce dimensionality while maintaining the most important characteristics of the data or the variance. Whereas at times this may cause a small reduction in predictive capability, it may be preferred in exchange for data with drastically fewer (irrelevant) features, which can enhance computational efficiency and interpretability. Conversely, by reducing noise and simplifying learning for the model, dimensionality reduction may sometimes even improve performance.

Modern techniques such as microarrays can generate high-dimensional data with few samples and thus benefit from dimensionality reduction techniques (69, 71). Principal component analysis (PCA) and t-distributed stochastic neighborhood embedding are linear and nonlinear dimensionality reduction techniques, respectively. Because of its capacity to eliminate redundant features, Lasso regression can also be used as a dimensionality reduction technique.

Explainable AI

The concept of xAI is concerned with not only generating an output but also how it was generated. Technological developments have enabled the creation of sophisticated algorithms such as ensemble methods and deep neural networks that, although usually superior, are not interpretable. If predictive ability is most relevant to the problem being solved, then this may not be an issue; however, in some practical applications of ML, it is important to know how the output was generated. This is understandable in medical situations where the predicting output can have serious consequences for, say, patient lifestyle or treatment avenues. The results of xAI can be informative in that they reveal which features contributed most to the algorithm output (72).

In certain situations, interpretable algorithms are preferred over ensemble methods, despite better performance

in the latter, as is the case with nutrition care by Maduri et al. (73) and in the prediction of nutrient content in infant milk by Wong et al. (74). However, methods exist that facilitate interpretability without sacrificing performance. xAI techniques such as Shapley additive explanations (SHAP) and Shapley values (75), partial dependence plots (76), and local interpretable model-agnostic explanations (LIME) (77) exist to make transparent black box models. Choi et al. (28) emphasized the importance of being able to capture complex nonlinear interactions when predicting refeeding hypophosphatemia and thus chose to opt for XGBoost in place of linear models, especially since classification was much better in the former. Instead, they used SHAP values to elucidate the features most influential for classification decision by XGBoost. Zeevi et al. (78) used partial dependence plots to extract the relative contribution of their features for the prediction of postprandial glucose response (PPGR). This enabled the use of a gradient-boosting RF that, although black box, was able to capture the nonlinear relationships inherent to their complex feature set. Davagdorj et al. (79) made use of LIME to explain predictions of artificial neural networks and XGBoost, the best performers, when predicting hypertension in the Korean population. They emphasized the importance not only of prediction quality but also explainability for decision making in public health.

In conclusion, not only is the eventual output of ML relevant but so is the means by which it was produced. After the proof-of-principle stages of model development, researchers can further their fields by incorporating xAI into their work, therein providing transparency and encouraging public understanding.

Evaluating Performance

Metrics for evaluating ML algorithms are broad and can be task, type, or model specific (Supplemental Table 2). For example, evaluation approaches of supervised algorithms are often not suitable for unsupervised techniques since the data might be without labels. In clustering, metrics are instead used that focus on the purity of the data partitioning or similarity of the data after grouping (80). For PCA for dimensionality reduction, cumulative variance with a predetermined, arbitrary cutoff point is used (e.g., 95%).

Additionally, although some evaluation metrics reflect model performance similarly, the specification of evaluation metrics should not be made arbitrarily. For example, where higher accuracy at the cost of specificity might be less problematic in applications categorizing food for dietary intake purposes, the same trade-off can have serious consequences in disease prediction, as in the incorrect assignment of a serious disease (false positives) such as cancer (81). Accuracy is the most common classification metric but too often it is presented or interpreted at face value, whilst other metrics are neglected. The consequences of this can be easily witnessed in data sets with class imbalances in the target variable, as is common in health data. For example, a classifier that predicts negative for all data points on a given NHANES data set where the target variable is undiagnosed T2D would have an accuracy of approximately 97% (82), without actually having real predictive power. For these reasons, multiple metrics or meta-metrics, such as F1 score or area under the receiver operating characteristic curve (AUROC) (83), should be considered.

Evaluation metrics should be specific to the problem at hand. For example, the coefficient of determination R^2 measures how well a continuous target variable is estimated by a set of predictors and is thus generalizable across problems, models, and data sets. However, at times it may be more relevant to know how well the model performs for a specific problem, such as mean error in prediction of plasma cholesterol or cost of meeting a healthy diet; in such cases, a metric such as mean absolute error would be preferred. Likewise, in circumstances where the consequences of the model output are less severe and the treatment response is risk-free, the emphasis would be on accuracy rather than specificity. An example of such a case could be the prediction of risk for overweight, with the treatment response being free admission to an education healthy eating course. These examples demonstrate the influence that the problem has on metric selection.

In sum, many methods exist for evaluating the performance of ML models. Metrics should be chosen thoughtfully, keeping in mind the task, the model being used, and the specific problem trying to be solved.

Validation

ML models will most likely perform better on the training data on which they were trained than on unseen data. Whilst this is to be expected, it can give a deceptive reflection of how well the algorithm has learned a task. If performance drops substantially when going from the training to unseen data, overfitting has occurred, and the model is therefore not generalizable or useful in real applications. Overfitting is a widespread problem and often unaccounted for in the literature. Whereas virtually all supervised techniques suffer from overfitting, the degree to which they do so can be lessened with validation techniques such as data splitting, cross-validation (CV), external validation, and combinations of these, as discussed later and shown in **Figure 2**. It is imperative that robust methods of validation are used to preserve generalizable model conclusions.

Data split.

Also known as the holdout method, the data split method splits the data into a training set and a test set, where the model is first trained on the training data and then applied to the test set to gauge generalizability. In this way, the test data act as the "unseen" data, although, since the data set came from the same source and was processed in the same way, it is not truly unseen. Splitting the data like this can be problematic in small data sets by reducing the instances from which the model has to learn. It can also increase vulnerability to outliers. Finally, unless data are sufficiently large, there can be large variations in results between different splits of the data. In most circumstances, a simple data split is

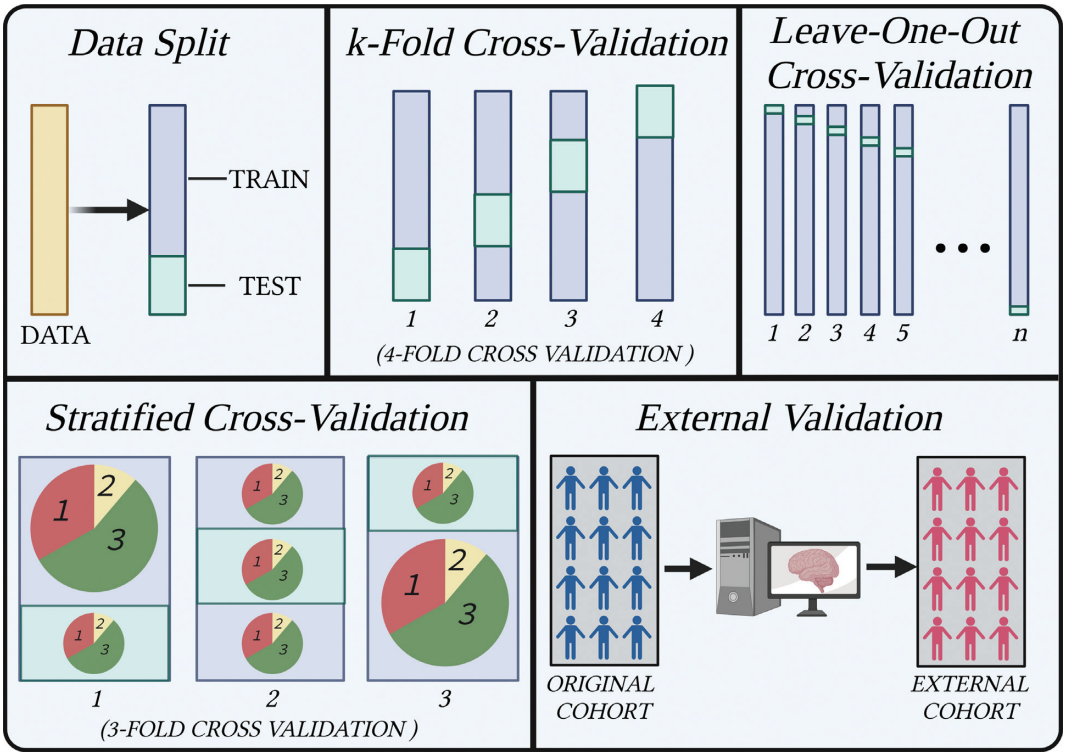

FIGURE 2 Various validation techniques. Data split simply consists of excluding a portion of the data for testing after training. In *k*-fold cross-validation, the data are split into *k* number of folds, and each fold is used once for training and $k - 1$ times for training. Leave-one-out cross-validation uses the same concept except that *k* is equal to the number of data samples, so each individual sample is used once for testing and $n - 1$ times for training. Stratified cross-validation ensures that the proportions of classes remain the same in each split (training and test) and each fold. Finally, external validation consists of using data different from those on which the algorithm was trained.

not a sufficiently powerful tool for assessing the generalizable performance of ML models.

Cross-validation.

CV and its variations run the model multiple times with different splits in the data so that every split is used for training and once for testing. This is most often achieved with *k*-fold CV, where *k* is an arbitrarily selected value by which to split the data, with a minimum value of 2 and a maximum value of $n - 1$. In the case of the latter, this represents another CV variation known as leave-one-out CV, where all instances but 1 are used to train the model, with the remaining data point representing the test data. The aforementioned instances use random sampling to split the data; in another variation, stratified CV, the data are split in a way that maintains the proportions of classes of the original data. This is useful in preventing issues arising when certain classes are particularly low since otherwise the model might be left with too few instances from which to learn. Eventually, results from each fold after performing CV should be averaged to give a balanced CV score, although examining the scores of each fold can also be informative; if they differ wildly, this can be indicative of problems such as outliers or class imbalances. CV provides a more robust way to validate a model and should be selected over a simple data split whenever possible.

Despite these advantages, biased models can still occur when the same CV scheme is used for hyperparameter tuning and model evaluation. Hyperparameter optimization schemes such as those discussed in the Supplemental Material (Hyperparameter Optimization section) often use CV, and although overfitting is reduced, the model is still yet to be tested on a pristine test sample that was not involved in either model training or hyperparameter tuning (84). Nested CV aims to overcome this by splitting the data into an outer loop, which itself is split into training and testing, and an inner loop, which is composed of the training folds of the outer loop. CV is used on the inner folds to select model hyperparameters; then, the outer loop is run by the optimal model identified in the inner loop. This is repeated *k* times, where *k* represents the number of folds of the outer loop. This prevents that all of the data are used for model selection, evaluation, and feature selection and maintains the ability to evaluate cross-validated performance on unseen data. Although this increases computational demands substantially, it provides a much more honest representation of model performance.

External validation.
For an even truer representation of model generalizability, a different data set from that used to train and test the model can be used. For example, a model predicting glycemic response trained on an Israeli cohort was tested in an American cohort, meaning that different dietary elements, societal influences, and genetics are introduced (85). Capturing the effect of intercohort variation in this way informs the degree to which the model can perform its task in different populations. This is a common approach in health studies deploying ML that use cohorts (22, 36, 37, 40, 42, 78, 86, 87).

Case Studies: Applications of Machine Learning in Nutrition Domains

The current section briefly presents two case studies within the discipline of nutrition that are suitable for the application of ML techniques. It is hoped that readers will derive ideas and inspiration from the case studies, which they can then apply to their own research domains.

Precision Nutrition

PN concerns the use of personal information to generate nutritional advice that, in theory, leads to superior health outcomes than generic advice (11, 88). It rests on the basis that differences in a myriad of factors among individuals ultimately necessitate specific nutritional requirements that population-level guidelines cannot capture. The diversity, complexity, and, at times, high dimensionality of the data that represent these factors have created expectations for ML in PN. Such expectation is reflected in the commitment of the National Institutes of Health to supply $170 million in funding algorithm development in PN over the next 5 years (89). A detailed systematic review of ML in PN is provided by Kirk et al. (11); here, a short overview and recent developments are provided.

A model example of the application of ML in PN was in the high-impact study of Zeevi et al. (78), which made use of a gradient-boosting RF model to integrate plasma, microbiome, anthropometric, personal, and dietary data to predict PPGR to the challenge meal, with accuracy comfortably exceeding established methods. A striking finding was the remarkable interindividual variation in PPGR seen in response to the same foods, substantiating the claims of PN for improving health. Berry et al. (90) used a similar design to predict not only postprandial glucose but also triglyceride and C-peptide with an RF regressor. In both studies, the features most relevant to the decision outcomes were estimated, and the contribution of modifiable factors was shown to be large. These studies show not only that the influence of specific foods on metabolic parameters in an individual can be known but also the main factors that can be modified to change this. Such information is of great value to those attempting to manage metabolic health.

Obesity and overweight constitute another important subject area in nutrition, thus attracting attention in PN. Ramyaa et al. (52) found homogeneous phenotypes within a population of women and then proceeded to predict their body weight. The clusters associated with different dietary and physical activity variables suggested that the women responded differently to macronutrients and exercise in their propensity to gain weight and thus that personalized diets and exercise regimes would be effective. Zellerbach and Ruiz (91) aimed to predict instances of overeating based on macronutrient composition of an individual's diet. Although they were unsuccessful, the concept may have merit with the inclusion of other relevant variables (e.g., stress, sleep, and alcohol or drug use) and with higher-quality diet data than the self-recorded publicly available food logs used in their study, where data were collected outside a scientifically controlled setting and liable to bias. That Wang et al. (5) could predict obesity using gut flora data is interesting to PN because of the known relationship between diet and microbiota composition. Because the microbiota is involved in various health conditions (92), its targeting by PN interventions could be fruitful in subsets of individuals.

Malnutrition has been targeted by PN in various ways. Current screening tools for malnutrition in inpatients, for example, suffer a lack of agreement and poor adherence from hospital staff, suggesting that automatized approaches may be appreciated (93). The decision system of Yin et al. (94) sought to realize this by applying *k*-means to hospital record data to separate patients based on nutritional status. Well-nourished and mild, moderate, and severely malnourished clusters were identified, the characteristics of which formed the basis for a logistic regression classifier to assign unseen data points to 1 of the 4 clusters with perfect performance (AUROC: 1). Subramanian et al. (95) characterized a "healthy" microbiota index in children, from which RF could predict chronological age (AUROC: 0.73). Severe acute malnutrition could subsequently be predicted by deviation from the index for a given age, since malnourished children have a relatively immature microbiota when compared with healthy children of the same age. This approach allows targeted intervention in children at risk of malnutrition-induced growth stunting. Malnutrition can also be predicted with ML from demographic data in developing countries, which is attractive since such data are routinely collected and available to health organizations (96–98).

ML is helpful not only for generating PN outcomes but also for collecting PN data. Current dietary assessment methods have serious limitations such that estimated intakes can vary wildly from true intakes (99, 100). However, ML-assisted dietary intake monitoring could make more convenient and accurate the process of collecting intake data, the benefits of which would extend beyond PN to the broader nutrition domain. Indeed, examples in which ML has been used in dietary assessment include image-, smart watch–, piezoelectric-, and audio-based methods [see Table 4 in Kirk et al. (11)]. Although some of these instances are relatively primitive and often confined to controlled settings, it can be imagined that their successors will be refined and convenient in real-world settings. Natural language processing (NLP) can also be valuable in dietary assessment. NLP is a specific field in computer and linguistic sciences that has the goal

to interpret written and spoken text in such a way that its meaning is understandable by a computer. In PN, this can automate the processing of food diaries (101) and consolidate multiple data sets (data integration) and food tables (102). Eventually, NLP can communicate messages to users of health-tracking apps to offer personalized advice and provide support, omitting the need for such advice to come exclusively from health professionals, which is expensive and time-consuming. Activity tracking is relevant to PN outcomes by providing information on the exercise and sedentarism of individuals. ML has been used to classify activity patterns (103–106) and estimate energy expenditure (33, 34, 107) based on accelerometer data, which can improve the quality of activity data acquisition, thus increasing its value as a feature in PN approaches.

Omics is a discipline that derives its name from the suffix of components from which it is composed: genomics, transcriptomics, epigenomics, proteomics, metabolomics, and occasionally others (microbiome, lipidomics, etc.) (108). **Figure 3** shows the main omics components and their proximity to the genotype and phenotype. Data captured on any of these levels can be valuable in PN, and integrating their data (i.e., multi-omics) can provide a systems-level view capable of providing more information than the constituent parts independently (109). Genetic information has been used in tandem with ML for predicting obesity (110–112) and diabetes (113, 114), although, despite the wealth of information within the human genome, genetic information often explains little of the variance in complex health outcomes (115). Gene expression, transcripts, and the proteins that they encode can all be modified by environmental factors such as food, which can make the genetic sequence encoding them effectively redundant; hence, epigenomics, transcriptomics, and proteomics, respectively, exist in response to this. Further still, the microbiome is increasingly recognized as a key player in health and disease, at times being responsible for a significant portion of the variance in predictive health models (5, 20, 87, 90, 116, 117). The microbiome is of particular interest to PN because it can act as an input variable and a target variable to be modified in PN.

It should be emphasized that studies need not be designed with PN in mind to realize how responses can differ among individuals within a study. A prime example of this is seen in the weight loss study of Gardener et al. (118), where participants followed either a low-carbohydrate or low-fat diet for 12 months. Whereas the group means suggested that weight loss was similar, analyzing the data on an individual basis within the groups told a different story: although some lost a great deal of weight, others failed to lose or even gained weight. This highlights a pitfall of research in health sciences in that, although comparing groups means is convenient, it can mask individual differences that can be much more informative.

Despite its potential, ML in PN must still prove itself able to reduce disease burden when applied in real-world situations. Most of the aforementioned studies are descriptive, and indeed more experimental studies are required

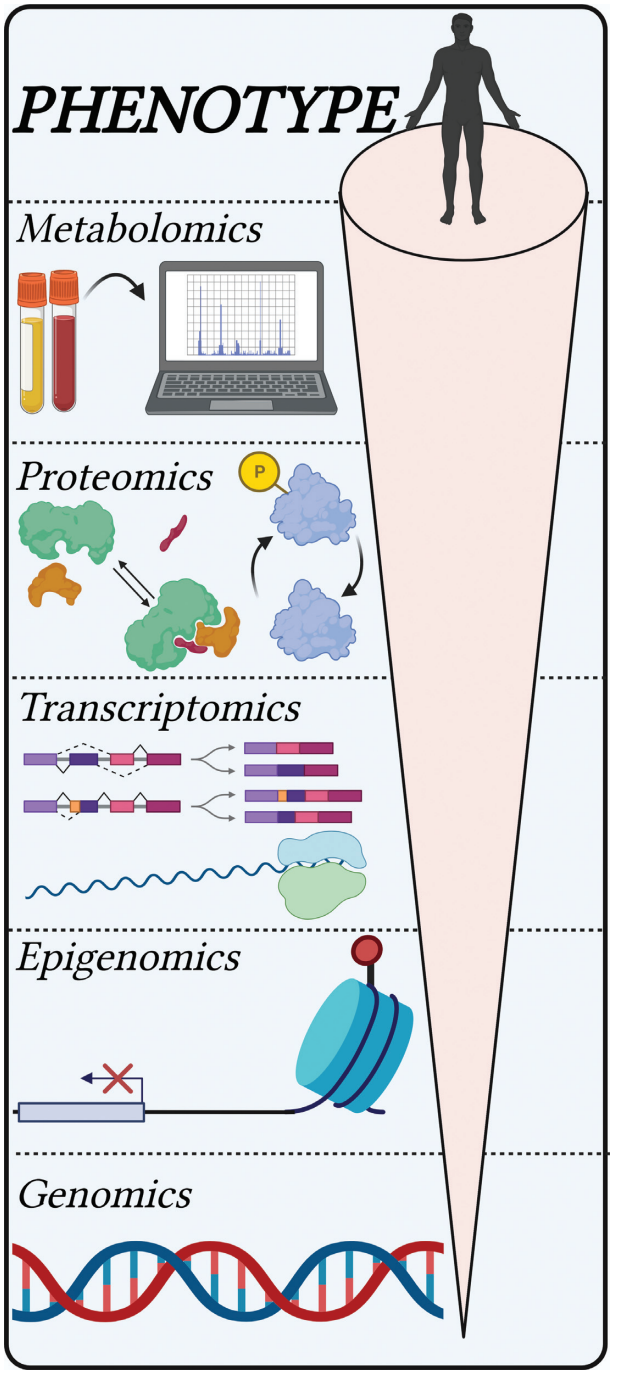

FIGURE 3 The major components of the omics field and their proximity to the phenotype.

to prove that PN is more effective than generic healthy eating recommendations when both are adhered to. Even if experimental studies can prove a theoretical role for PN in improving health, PN approaches must be practical. If the suggested dietary alterations are restrictive or infeasible, it is unlikely that they will be adhered to in the long term. For example, although the large-scale study Food4Me found that personalized advice favorably altered dietary habits in participants (119), there is insufficient evidence that

personalized approaches will lead to sufficient adherence to reap the potential benefits of PN. This was demonstrated in a recent Korean study where only those in the highest adherence group saw improvements in the markers of health that were measured, and in fact, the same markers deteriorated in the group of lowest adherence (120). Whereas ML has much potential to help with data generation and analysis in PN, these approaches must be able to demonstrate practical application and ultimately a reduction in clinical burden, both of which require many more studies for their verification.

Metabolomics

One area of nutrition that has received much attention in recent times is metabolomics. Metabolomics is closely related to PN, as a predictor of health outcomes and for data collection (11), though it also has functions that do not necessarily relate to PN. Modern technologies are now enabling the profiling of many metabolites all at once, within one or a few samples, followed by analysis of their interactions (121). The profiling of thousands of metabolites makes for noisy raw data, which requires preprocessing and analysis, two tasks for which ML is highly suited. Some examples of metabolomics research using ML in nutrition are presented in turn.

A popular application of metabolomics is phenotyping, which overlaps significantly with PN. Metabolites are generally considered to give a much more representative picture of a phenotype than other omics varieties since they more closely reflect the reactions that actually occur in a system (122). The simultaneous assessment of many metabolites in an individual enables a form of phenotyping specific to shared metabolite characteristics known as metabotyping (115). A randomized controlled trial found no effect of 15 μg of vitamin D supplementation on markers of metabolic syndrome (123). However, after metabotyping via *k*-means clustering, a vitamin D–responsive cluster was found where, in contrast to the population as a whole, vitamin D supplementation did improve markers of metabolic syndrome. Also utilizing *k*-means, O'Donovan et al. (124, 125) used a range of metabolites to identify healthy and unhealthy clusters in 2 cohorts, and on both occasions targeted advice was given based on the defining characteristics of the clusters. For example, a cluster composed of individuals with elevated cholesterol was administered personalized advice oriented toward lowering cholesterol (125). Given the diversity of metabolic alterations that people can experience despite having similar demographic or anthropometric characteristics, tailoring nutritional recommendations to the individual is a logical approach.

The concept of "metabolically healthy obese" and whether it actually exists is another example of phenotyping. It is indeed curious that approximately 1 in 3 obese individuals does not show metabolic alterations on commonly investigated clinical parameters (126) and that the exact metabolic consequences of the remaining two-thirds vary greatly among individuals (127). The drivers of this divergence remain unknown; hence, investigating metabolomic signature differences between the phenotypes may be revealing. Studies using ML techniques have found key differences between the metabolite profiles of healthy and unhealthy obese subjects (128–130). In a systematic review, BCAAs, aromatic amino acids, lipids, and acylcarnitines were all found to be elevated in the unhealthy obese phenotype as compared with the healthy (131). Due to the high dimensionality of the data sets in metabolomics, PCA is typically opted for. PCA-reduced feature sets can then be used to identify differences in metabolites, amino acids, and lipid patterns (132, 133). Using ML to understand how and which metabolic aberrations could develop among obese phenotypes can inform targeted treatment to minimize obesity-imposed harm.

ML has been applied in metabolomics when studying the microbiota. Microbiota-derived metabolomics data are complex and high-dimensional, which has motivated researchers to consider applying ML. Some notable examples include distinguishing healthy and unhealthy metabolite signatures following a red wine intervention (134), distinguishing women with and without food addiction from fecal samples alone (135), identifying pediatric IBS (136), and comparing metabolic activity of the microbiota between vegans and omnivores (137). The last of these studies is particularly relevant to the metabolomics field because, although differences in microbiota composition were minimal, metabolic activity differed significantly, and an RF classifier could distinguish the groups with 91.7% accuracy. This shows the importance of not only the composition but also the metabolic activity of the microbiota.

Metabolomics is interesting to the nutrition community as a free-living, objective dietary assessment tool (138, 139). Biomarkers of intake have been found for various foods, such as bread (140), coffee (141), citrus consumption (142), and meat and fish (143), as well as dietary components such as polyphenols (144, 145) and fermented foods (146). Such information can be used not only for simply monitoring food intake but also for associations with health and disease outcomes (141, 142, 147). This function of metabolomics permits estimating adherence to dietary patterns. Acar et al. (139) used metabolite profiles to identify participants potentially noncompliant to a particular dietary pattern through partial least squares discriminant analysis with a reduced feature set. Aside from identifying regular noncompliers, they observed that at any given time approximately 10% of the participants may have been deviating from their prescribed diets, which, if generalizable, has clear implications for nutrition intervention studies. The capabilities of ML make it suitable for finding associations in metabolomics, as well as identifying new phenotypes or markers of intake in untargeted approaches. Many processing steps are required to transform raw metabolomics data into a form from which information can be derived. This, however, is essentially feature selection and pattern recognition. To this end, the competency of deep artificial neural networks, such as convolutional neural networks in feature selection, could

be particularly useful, especially given the complexity of data used in metabolomics (148). Research investigating this in nutrition, though, is lacking but would be valuable.

In sum, ML can be a useful tool in the data preprocessing stages of metabolomics and in generating predictive models on the prepared data. Through clustering and classification, ML can analyze processed metabolomics data for applications such as disease prediction and understanding disease mechanisms, phenotyping, characterizing the metabolic environment, identifying biomarkers, and dietary assessment.

Framework for Applying Machine Learning in Nutrition Science

After understanding the advantages of using ML in research, researchers should be able to know when and how ML can be applied to a problem. The present section aims to support decision making in this process by providing a framework to guide researchers interested in using ML in their work. The framework takes inspiration from the concept of method engineering (149), though is adapted with nutrition research in mind.

Understanding the problem and the data

Whether ML can be used to solve a problem depends primarily on the problem itself and the data involved. If the problem is one concerned with predicting an outcome based on a given data set, ML can be considered. Data sets with many features and complex, nonlinear interactions suggest themselves suitable for application with ML because of its ability to identify patterns among the input variables that can then map the output variable, thus producing results that would otherwise go amiss. This is exemplified in cases where data are clustered without expectation yet new findings are discovered (52, 123).

ML could be considered in data preprocessing. This function of ML was evidenced on various occasions in the present article, such as the use of PCA for dimensionality reduction, deep learning for feature extraction, and ML approaches for data collection in PN and for processing noisy and complicated metabolomics data. "Missing data" is the often colloquial term denoting a data set in which not all the data entries that should have values are filled, regardless of the reason why. In nutrition, due to the practical challenges of longitudinal data gathering, missing data are a common issue. It is up to the researcher to understand the impact of values that are missing in the data set and how they are to be dealt with. Techniques exist for their imputation, and sometimes it is appropriate to remove entire variables or data entries (150). Since each of these approaches has advantages and disadvantages, the decision ultimately taken by the researcher should be done so after deliberation. Resources describing the missing data problem and its solutions exist (150–154). ML techniques can also be considered for imputation (155–158). The extent of the missing data can influence the modeling approach. Certain approaches in traditional statistics and ML can handle missing data well, including linear mixed models, decision trees, kNN, and XGBoost, sometimes even when missing data are as high as 20% (159, 160). Regardless of how this is done, the method of handling missing data should be explained and reported (161). Understanding the data will provide insight into how exactly ML can be applied.

Background research and existing solutions.

An understanding of the existing solutions to the problem at hand is crucial to knowing exactly how ML can be applied. If existing solutions are already suitable, additional benefits from ML may be marginal. ML is also incapable of replacing the human aspect involved in managing health and nutrition. For example, although an algorithm may make accurate personalized nutritional recommendations, it cannot deliver the same information in a way that a trained professional would, and this aspect may be important for inducing behavior change.

ML, instead, is better applied when existing methods are insufficient. This is observed in data with complex relationships where classical statistical methods are incapable or in situations where domain-specific techniques are inadequate. Examples of the latter include the prediction of PPGR (78, 90), for which existing methods have low accuracy. It may be that current solutions, although accurate, have other limitations, such as invasiveness, as is the case for NAFLD detection (39–41). Further still, existing solutions that have a higher human element naturally suffer from human limitations such as fatigue or calculation mistakes. This was exemplified by Kondrup et al. (162), who found that a major reason why patients in clinical care were not screened for nutritional assessment was that nurses "just forgot." ML in these cases can increase predictive capacity, improve efficiency, reduce patient risk, and mitigate human error.

Possible solutions

Of the options within the domain of ML, the eventual candidate solutions should be tailored to the needs of the problem, the data, and the ultimate goal of the project. Primarily, one must think about the task required, as this will naturally limit the options available since certain algorithms are capable of only certain tasks. Also important is the algorithm in relation to the data. For example, naive Bayes would be an unsuitable choice in data sets with classes that contain certain values of very low frequency due to the zero-frequency problem (see Supplemental Material) (163). Likewise, applying estimators that assume linearity to a data set that has predictors with nonlinear relationships with the dependent variable would lead to suboptimal performance. In this case, performing transformations on the data or choosing a model with nonlinear capabilities, such as RF or SVM with a nonlinear kernel, would be preferred.

Another possible solution might be to use dimensionality reduction techniques. Feature selection and engineering are valuable methods in ML and can be specifically helpful to nutrition data sets where collinearity is often present. Such

techniques must be chosen with care as their improper use can affect performance, but due to the breadth of their possible applications, it is not possible to state which should be used; instead, they should be tailored to the problem. On a general level, if dimensionality reduction is applied with collinearity in mind, then the method should be chosen and applied in a way that preserves or increases the score on the test set. When dimensionality is applied to reduce computational strain, one must have preconceptions about the degree to which error is allowed to increase to enable decreases in computational time.

The end goal must be kept in mind, which too will dictate the pool of candidate solutions. For example, if predictive performance was the primary goal in the detection of cancer based on medical images, high-performance convolutional neural networks could be used. Alternatively, when interpretability is needed, algorithms can be selected that provide coefficients (e.g., regression) or can be easily understood (e.g., decision trees). Black box models (e.g., RF, XGBoost, artificial neural networks) can be understood through xAI techniques (e.g., SHAP, LIME, partial dependence plots), but such techniques have their pros and cons and should be considered in the context of the entire problem scope. When ML approaches are to be deployed or incorporated into an application, such as using ML for food tracking on a mobile device, pragmatic factors such as computational time are relevant. For example, whereas RF would usually outperform naive Bayes on a classification test, naive Bayes is much faster. These practicalities must be kept in mind.

Testing the available solutions

One of the advantages of using ML in research is the capacity to test multiple options in multiple configurations and converge on an optimal one. This capacity should be exercised by trying various, if not all, candidate models. The interpretation of the test data results should be given more importance than the training data to reduce overfitting. If possible, hyperparameters of each algorithm should be optimized with techniques such as grid or random search (see Hyperparameter Optimization section in Supplemental Material), and nested CV should always be considered. Although this can be time-consuming, it will give a fairer representation of the quality of the possible solutions since some models perform better in their default parameters than others.

Trying many solutions is always advisable, even when the choice might seem obvious beforehand. For example, although ensemble methods consistently outperform logistic regression in classification, this is not always the case (164).

Indeed, one example of this was by Yin et al. (94), where logistic regression achieved perfect performance predicting malnutrition in a data set of 14, 000 patients with cancer, outperforming ensemble and deep learning techniques (94). Additionally, techniques such as stacking, which involve combining multiple learners into 1 meta-learner, should be explored. Packages such as SuperLearner (165) exist to facilitate this, but it can also be done manually. The process ensures a result at least as good as, if not better than, the best single learner alone. Stacking has been used in some nutrition research (166–168) but is typically underutilized.

Understanding and communicating the results

The evaluation process must be undertaken in the context of the solution (i.e., the algorithm) and the problem (see Evaluating Performance). Comparisons of the possible solutions should be made by the most suitable metrics to enable the optimal solution to be chosen. It is common to visualize results in various plots, such as AUROC plots for classification, R^2 for regression, silhouette score and cluster plots for clustering, PCA score plots for PCA, and heat plots for correlations, among others. Common libraries for achieving this include ggplot in R and Matplotlib, Seaborn, and Yellowbrick in Python. If xAI techniques are used, these results too can be communicated. Feature importance plots, plots of the SHAP library (waterfall, force plots, bee swarm, etc.), and partial dependence plots, among others, can allow visualization of the features most relevant to a decision.

Limitations of Machine Learning in Nutrition Research

Although the present article has focused on the promise that ML is demonstrating, it does not come without limitations. There is an apparent overoptimism in ML research that exists due to nonrigorous methodologies. Although we described methods for detecting this, such as CV, it is not uncommon to see circumstances where these are not made use of. For example, unless a data set is sufficiently large, different training-test splits can lead to different results when using a simple data split for validation. This opens the possibility to the generation of interesting results based solely on how the data were split. Further still, it is rare to see nested CV used in nutrition literature, the consequences of which were discussed in the Validation section. Another consideration is that ML algorithms are generally evaluated on homogeneous data collected in affluent societies; the performance of these models in distinct populations and with different data generation techniques is not guaranteed. Both these considerations compromise generalizability, meaning that if such models cannot be applied outside the setting in which they were tested, ultimately their utility is greatly diminished. Another often overlooked issue is flawed feature selection and derivation of importance. For example, feature importances from algorithms such as RF and XGBoost are readily available and often reported in studies that utilize them. However, the mechanism by which such methods estimate importance means that correlated features, though scoring similarly, appear less important than they are. This, relatively speaking, also means that the importance of less important features are inflated. Such similar phenomena occur with other algorithms and xAI techniques but are often not checked, for example, by corresponding with

other feature importance techniques for corroboration and, instead, are reported as is.

Finally, the application of ML in certain circumstances has practical drawbacks. There can be substantial costs for data collection, hardware, ML engineers, infrastructures (data storage, cloud computing), integration (pipeline development and documentation), and maintenance. In certain applications, data are generated from different sources by using different programming languages and arriving in different forms. Unifying this in a multimodal approach can be very challenging. Similarly, although the ability of ML to transform unstructured data into data suitable for use in models represents a stark advantage for ML over traditional methods, such techniques can be challenging even for those specialized in the area and thus may not be fruitful for researchers specialized in nutrition at the time of writing. In closing, ML is showing much potential in research in nutrition but still has much to prove to reduce the burden of nutrition-related ill-health in society.

Conclusion

In conclusion, there is much potential for ML to make progress in nutrition science. ML can capture the complex interactions that exist and are increasingly generated with modern technologies in nutrition and health data. The failure to be able to use techniques that can analyze complex data, such as ML, represents an unnecessary barrier to scientific progress. Although still relatively new, it is evident that AI approaches have much potential to supplant traditional and domain-specific methods in predictive capabilities, efficiency, costs, and convenience. ML can also be helpful in the data collection and data preprocessing stages in various fields of nutrition. To realize this potential, researchers must be familiar with ML concepts, be knowledgeable on when AI can be suitably applied to a problem and how to use it, and be willing to branch out of the techniques historically used in their disciplines. We hope that the intuitive explanation of ML and the examples of its application in nutrition science in the current article will facilitate this and be a useful reference guide to researchers of health and nutrition who would like to make use of ML in answering their research questions.

Acknowledgments

The authors' responsibilities were as follows — DK: performed the literature review, wrote and designed the article, and made the figures; GC: spawned the idea for the article, providing machine learning and nutritional expertise, and was involved in writing the article and editing; EJMF: involved in editing and providing nutritional expertise; BT: involved in editing and providing machine learning expertise; EK: provided valuable feedback and machine learning expertise and edited text; MT: provided valuable feedback and was involved in editing; and all authors: involved in brainstorming and structuring the article and read and approved the final manuscript.

References

1. Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. Nutrients 2020;12(8):2466.
2. Lecroy MN, Kim RS, Stevens J, Hanna DB, Isasi CR. Identifying key determinants of childhood obesity: a narrative review of machine learning studies. Child Obes 2021;17(3):153–9.
3. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv 2021;49:107739.
4. Li R, Li L, Xu Y, Yang J. Machine learning meets omics: applications and perspectives. Brief Bioinform 2022;23(1):bbab460.
5. Wang X, Liu J, Ma L. Identification of gut flora based on robust support vector machine. J Phys Conf Ser 2022;2171(1):012066.
6. Namkung J. Machine learning methods for microbiome studies. J Microbiol 2020;58(3):206–16.
7. Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nat Rev Gastroenterol Hepatol 2020;17(10):635–48.
8. Jorm LR. Commentary: towards machine learning-enabled epidemiology. Int J Epidemiol 2021;49(6):1770–3.
9. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health 2020;41:21–36.
10. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis 2018;66(1):149–53.
11. Kirk D, Catal C, Tekinerdogan B. Precision nutrition: a systematic literature review. Comput Biol Med 2021;133:104365.
12. Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. Cell 2020;181(1):92–101.
13. Vilne B, Ķibilds J, Siksna I, Lazda I, Valciņa O, Krūmiņa A. Could artificial intelligence/machine learning and inclusion of diet-gut microbiome interactions improve disease risk prediction? Case study: coronary artery disease. Front Microbiol 2022;13: 627892.
14. Chollet F. On the measure of intelligence [Internet]. 2019Nov 5 [cited 2022 Jul 29]. Available from: https://arxiv.org/abs/1911.01547v2
15. Wang H, Ma C, Zhou L. A brief review of machine learning and its application. In: 2009 International Conference on Information Engineering and Computer Science. New York (NY): IEEE; 2009. doi:10.1109/ICIECS.2009.5362936
16. Witten IH, Frank E, Hall MA, Pal CJ. Data mining: practical machine learning tools and techniques. Amsterdam (Netherlands): Elsevier; 2016.
17. Bzdok D, Altman N, Krzywinski M. Points of significance: statistics versus machine learning. Nat Methods 2018;15(4):233–4.
18. De Silva K, Jönsson D, Demmer RT. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. J Am Med Inform Assoc 2020;27(3):396–406.
19. Poss AM, Maschek JA, Cox JE, Hauner BJ, Hopkins PN, Hunt SC, et al. Machine learning reveals serum sphingolipids as cholesterol-independent biomarkers of coronary artery disease. J Clin Invest 2020;130(3):1363.
20. Gou W, Ling CW, He Y, Jiang Z, Fu Y, Xu F, et al. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. Diabetes Care 2021;44(2): 358.
21. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 2019;19(1):211.
22. Tap J, Derrien M, Törnblom H, Brazeilles R, Cools-Portier S, Doré J, et al. Identification of an intestinal microbiota signature associated with severity of irritable bowel syndrome. Gastroenterology 2017;152(1):111–23.e8.
23. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Gregory Hundley W, McClelland R, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res 2017;121(9):1092–101.

24. De Silva K, Lim S, Mousa A, Teede H, Forbes A, Demmer RT, et al. Nutritional markers of undiagnosed type 2 diabetes in adults: findings of a machine learning analysis with external validation and benchmarking. PLoS One 2021;16(5):e0250832.
25. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author) [Internet]. 2001;16(3):199–231. Available from: https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling–The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full
26. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Stud Health Technol Inform 2004;107:736–40.
27. Stolfi P, Castiglione F. Emulating complex simulations by machine learning methods. BMC Bioinformatics 2021;22(Suppl 14):483.
28. Choi TY, Chang MY, Heo S, Jang JY. Explainable machine learning model to predict refeeding hypophosphatemia. Clin Nutr ESPEN 2021;45:213–9.
29. Khorraminezhad L, Leclercq M, Droit A, Bilodeau JF, Rudkowska I. Statistical and machine-learning analyses in nutritional genomics studies. Nutrients 2020;12(10):3140.
30. Ahmadi MN, Brookes D, Chowdhury A, Pavey T, Trost SG. Free-living evaluation of laboratory-based activity classifiers in preschoolers. Med Sci Sports Exercise 2020;52(5):1227–34.
31. Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG. Ensemble methods for classification of physical activities from wrist accelerometry. Med Sci Sports Exercise 2017;49(9):1965–73.
32. Pavey TG, Gilson ND, Gomersall SR, Clark B, Trost SG. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. J Sci Med Sport 2017;20(1):75–80.
33. Catal C, Akbulut A. Automatic energy expenditure measurement for health science. Comput Methods Programs Biomed 2018;157:31–7.
34. Ahmadi MN, Chowdhury A, Pavey T, Trost SG. Laboratory-based and free-living algorithms for energy expenditure estimation in preschool children: a free-living evaluation. PLoS One 2020;15(5):e0233229.
35. 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.
36. Sánchez-Cabo F, Rossello X, Fuster V, Benito F, Manzano JP, Silla JC, et al. Machine learning improves cardiovascular risk definition for young, asymptomatic individuals. J Am Coll Cardiol 2020;76(14):1674–85.
37. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. J Am Heart Assoc 2018;7(22):e009476.
38. Alaa AM, Bolton T, Angelantonio E Di, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One 2019;14(5):e0213653.
39. Sorino P, Campanella A, Bonfiglio C, Mirizzi A, Franco I, Bianco A, et al. Development and validation of a neural network for NAFLD diagnosis. Sci Rep 2021;11(1):20240.
40. Canbay A, Kälsch J, Neumann U, Rau M, Hohenester S, Baba HA, et al. Non-invasive assessment of NAFLD as systemic disease—a machine learning perspective. PLoS One 2019;14(3):e0214436.
41. Khusial RD, Cioffi CE, Caltharp SA, Krasinskas AM, Alazraki A, Knight-Scott J, et al. Development of a plasma screening panel for pediatric nonalcoholic fatty liver disease using metabolomics. Hepatol Commun 2019;3(10):1311–21.
42. Frantzi M, Gomez Gomez E, Blanca Pedregosa A, Valero Rosa J, Latosinska A, Culig Z, et al. CE-MS-based urinary biomarkers to distinguish non-significant from significant prostate cancer. Br J Cancer 2019;120(12):1120–8.
43. Cao R, Mohammadian Bajgiran A, Afshari Mirak S, Shakeri S, Zhong X, Enzmann D, et al. Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet. IEEE Trans Med Imaging 2019;38(11):2496–506.
44. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 2019;292(1):60–6.
45. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–61.
46. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2007;2:59–77.
47. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
48. Sorino P, Caruso MG, Misciagna G, Bonfiglio C, Campanella A, Mirizzi A, et al. Selecting the best machine learning algorithm to support the diagnosis of non-alcoholic fatty liver disease: a meta learner study. PLoS One 2020;15(10):e0240867.
49. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018;9:515.
50. Kirk D, Catal C, Tekinerdogan B. Predicting plasma vitamin C using machine learning. Applied Artificial Intelligence 2022;36(1):2042924.
51. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2021;2(3):160.
52. Ramyaa R, Hosseini O, Krishnan GP, Krishnan S. Phenotyping women based on dietary macronutrients, physical activity, and body weight using machine learning tools. Nutrients 2019;11(7):1681.
53. Basu S. Semi-supervised learning. In: Liu L, Özsu MT, editors. Encyclopedia of database systems [Internet]. Boston (MA): Springer; 2009 [cited 2022 Feb 28]. p. 2613–5. Available from: https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_609
54. Nguyen TP, Ho TB. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med 2012;54(1):63–71.
55. Davidson I. Clustering with constraints. In: Liu L, Özsu MT, editors. Encyclopedia of database systems [Internet]. Boston, MA: Springer; 2009 [cited 2022 Feb 28]. p. 393–6. Available from: https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_610
56. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science 2018;362(6419):1140–4.
57. Mulani J, Heda S, Tumdi K, Patel J, Chhinkaniwala H, Patel J, et al. Deep reinforcement learning based personalized health recommendations. In: Dash S, Acharya B, Mittal M, Abraham A, Kelemen A, editors. Deep learning techniques for biomedical and health informatics. Cham (Switzerland): Springer; 2020. p. 231–55.
58. Zhou M, Mintz Y, Fukuoka Y, Goldberg K, Flowers E, Kaminsky P, et al. Personalizing mobile fitness apps using reinforcement learning [Internet]. CEUR Workshop Proc 2018;2068. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7220419/
59. Breiman L. Random forests. Machine Learning 2001;45(1):5–32.
60. Awad M, Khanna R, Awad M, Khanna R. Support vector regression. In: Efficient learning machines. Cham (Switzerland): Springer; 2015. p. 67–80.
61. Zhou M, Fukuoka Y, Goldberg K, Vittinghoff E, Aswani A. Applying machine learning to predict future adherence to physical activity programs. BMC Med Inform Decis Mak 2019;19(1):169.
62. Wu WK, Panyod S, Liu PY, Chen CC, Kao HL, Chuang HL, et al. Characterization of TMAO productivity from carnitine challenge facilitates personalized nutrition and microbiome signatures discovery. Microbiome 2020;8(1):162.
63. Jones PJ, Catt M, Davies MJ, Edwardson CL, Mirkes EM, Khunti K, et al. Feature selection for unsupervised machine learning of accelerometer data physical activity clusters—a systematic review. Gait Posture 2021;90:120–8.
64. Kim J, Lin S, Ferrara G, Hua J, Seto E. Identifying people based on machine learning classification of foods consumed in order to offer tailored healthier food options. In: Advances in intelligent systems and computing. Cham (Switzerland): Springer; 2020. p. 190–4.

65. Sowah RA, Bampoe-Addo AA, Armoo SK, Saalia FK, Gatsi F, Sarkodie-Mensah B. Design and development of diabetes management system using machine learning. Int J Telemed Appl 2020;2020:8870141.
66. Mitchell EG, Heitkemper EM, Burgermaster M, Levine ME, Miao Y, Hwang ML, et al. From reflection to action: combining machine learning with expert knowledge for nutrition goal recommendations. Proc SIGCHI Conf Hum Factor Comput Syst 2021;2021:206.
67. Metwally AA, Leong AK, Desai A, Nagarjuna A, Perelman D, Snyder M. Learning personal food preferences via food logs embedding 2022;2281–6.Available from: https://doi.org/10.48550/arXiv.2110.15498.
68. Baek JW, Kim JC, Chun J, Chung K. Hybrid clustering based health decision-making for improving dietary habits. Technol Health Care 2019;27(5):459–72.
69. Hua J, Tembe WD, Dougherty ER. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognit 2009;42(3):409–24.
70. Waggoner PD. Modern dimension reduction. Cambridge (UK): Cambridge University Pres; 2021. doi:10.1017/9781108981767
71. Gavai AK. Bayesian networks for omics data analysis. Wageningen (Netherlands); Wageningen University; 2009.
72. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI Information Fusion 2020;58:82–115.
73. Maduri C, Sabrina Hsueh, PY, Li Z, Chen CH, Papoutsakis C. Applying contemporary machine learning approaches to nutrition care real-world evidence: findings from the national quality improvement data set. J Acad Nutr Diet 2021;121(12):2549–59.e1.
74. Wong RK, Pitino MA, Mahmood R, Zhu IY, Stone D, O'Connor DL, et al. Predicting protein and fat content in human donor milk using machine learning. J Nutr 2021;151(7):2075–83.
75. Lundberg SM, Lee SI. A unified approach to interpreting model predictions [Internet]. Adv Neural Inf Process Syst 2017;2017:4766–75. Available from: https://arxiv.org/abs/1705.07874v2
76. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189–232.
77. Ribeiro MT, Singh S, Guestrin C. Model-Agnostic interpretability of machine learning [Internet]. 2016Jun 16 [cited 2022 Sep 2]. Available from: https://arxiv.org/abs/1606.05386v1
78. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized nutrition by prediction of glycemic responses. Cell 2015;163(5):1079–94.
79. Davagdorj K, Li M, Ryu KH. Local interpretable model-agnostic explanations of predictive models for hypertension. In: Pan JS, Li J, Ryu KH, Meng Z, Klasnja-Milicevic A, editors. Advances in intelligent information hiding and multimedia signal processing. Singapore: Springer; 2021. p. 426–33.
80. Palacio-Niño JO, Berzal F. Evaluation metrics for unsupervised learning algorithms [Internet]. 2019May 14 [cited 2022 Feb 28]. Available from: https://arxiv.org/abs/1905.05667v2
81. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice Front Public Heal, 2017;5:307.doi: 10.3389/fpubh.2017.00307
82. National diabetes statistics report 2020: Estimates of diabetes and its burden in the United States. Atlanta (GA): US Department of Health and Human Services; 2020.
83. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract 2006;12(2):132–9.
84. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 2010;11:2079–107.
85. Mendes-Soares H, Raveh-Sadka T, Azulay S, Ben-Shlomo Y, Cohen Y, Ofek T, et al. Model of personalized postprandial glycemic response to food developed for an Israeli cohort predicts responses in midwestern american individuals. Am J Clin Nutr 2019;110(1):63–75.
86. Berry S, Valdes A, Davies R, Delahanty L, Drew D, Chan AT, et al. Predicting personal metabolic responses to food using multi-omics machine learning in over 1000 twins and singletons from the UK and US: the PREDICT I study (OR31-01-19). Curr Dev Nutr 2019;3(Suppl 1):nzz037.
87. Wu H, Tremaroli V, Schmidt C, Lundqvist A, Olsson LM, Krämer M, et al. The gut microbiota in prediabetes and diabetes: a population-based cross-sectional study. Cell Metab 2020;32(3):379–90.e3.
88. Chatelan A, Bochud M, Frohlich KL. Precision nutrition: hype or hope for public health interventions to reduce obesity? Int J Epidemiol 2019;48(2):332–42.
89. NIH awards $170 million for precision nutrition study [Internet]. 2022 [cited 2022 Mar 21]. Available from: https://www.nih.gov/news-events/news-releases/nih-awards-170-million-precision-nutrition-study
90. Berry SE, Valdes AM, Drew DA, Asnicar F, Mazidi M, Wolf J, et al. Human postprandial responses to food and potential for precision nutrition. Nat Med 2020;26(6):964–73.
91. Zellerbach K, Ruiz C. Machine learning to predict overeating from macronutrient composition. In: Proceedings—2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019. Piscataway (NJ): Institute of Electrical and Electronics Engineers Inc.; 2019. p. 1268–70.
92. Valdes AM, Walter J, Segal E, Spector TD. Role of the gut microbiota in nutrition and health. BMJ 2018;361:36–44.
93. Raphaeli O, Singer P. Towards personalized nutritional treatment for malnutrition using machine learning-based screening tools. Clin Nutr 2021;40(10):5249–51.
94. Yin L, Song C, Cui J, Lin X, Li N, Fan Y, et al. A fusion decision system to identify and grade malnutrition in cancer patients: machine learning reveals feasible workflow from representative real-world data. Clin Nutr 2021;40(8):4958–70.
95. Subramanian S, Huq S, Yatsunenko T, Haque R, Mahfuz M, Alam MA, et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature 2014;510(7505):417–21.
96. Fenta HM, Zewotir T, Muluneh EK. A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones. BMC Med Inform Decis Mak 2021;21(1):291.
97. Talukder A, Ahammed B. Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. Nutrition 2020;78:110861.
98. Islam MM, Rahman MJ, Islam MM, Roy DC, Ahmed N, Hussain S, et al. Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh. International Journal of Cognitive Computing in Engineering 2022;3:46–57.
99. Bingham SA. Limitations of the various methods for collecting dietary intake data. Ann Nutr Metab 1991;35(3):117–27.
100. Schoeller DA. Limitations in the assessment of dietary energy intake by self-report. Metabolism 1995;44(Suppl 2):18–22.
101. Mezgec S, Eftimov T, Bucher T, Koroušić Seljak B. Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment. Public Health Nutr 2019;22(7):1193–202.
102. van Erp M, Reynolds C, Maynard D, Starke A, Ibáñez Martín R, Andres F, et al. Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food. Front Artif Intell 2021;3:115.
103. Ahmadi MN, Pavey TG, Trost SG. Machine learning models for classifying physical activity in free-living preschool children. Sensors (Basel) 2020;20(16):4364.
104. Fridolfsson J, Arvidsson D, Doerks F, Kreidler TJ, Grau S. Workplace activity classification from shoe-based movement sensors. BMC Biomed Eng 2020;2:8.
105. Fergus P, Hussain AJ, Hearty J, Fairclough S, Boddy L, Mackintosh K, et al. A machine learning approach to measure and monitor physical activity in children. Neurocomputing 2017;228:220–30.

106. Kingsley MIC, Nawaratne R, O'Halloran PD, Montoye AHK, Alahakoon D, De Silva D, et al. Wrist-specific accelerometry methods for estimating free-living physical activity. J Sci Med Sport 2019;22(6):677–83.
107. O'Driscoll R, Turicchi J, Hopkins M, Horgan GW, Finlayson G, Stubbs JR. Improving energy expenditure estimates from wearable devices: a machine learning approach. J Sports Sci 2020;38(13):1496–505.
108. Perakakis N, Yazdani A, Karniadakis GE, Mantzoros C. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. Metabolism 2018;87:A1.
109. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinf Biol Insights 2020;14:117793221989905.
110. Montanez CAC, Fergus P, Hussain A, Al-Jumeily D, Abdulaimma B, Hind J, et al. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In: Proceedings of the International Joint Conference on Neural Networks. Piscataway (NJ): Institute of Electrical and Electronics Engineers Inc.; 2017. p. 2743–50.
111. Montañez CAC, Fergus P, Hussain A, Al-Jumeily D, Dorak MT, Abdullah R. Evaluation of phenotype classification methods for obesity using direct to consumer genetic data. In: Lecture notes in computer science. Berlin (Germany): Springer Verlag; 2017. p. 350–62.
112. Rodríguez-Pardo C, Segura A, Zamorano-León JJ, Martínez-Santos C, Martínez D, Collado-Yurrita L, et al. Decision tree learning to predict overweight/obesity based on body mass index and gene polymporphisms. Gene 2019;699:88–93.
113. López B, Torrent-Fontbona F, Viñas R, Fernández-Real JM. Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. Artif Intell Med 2018;85:43–9.
114. Wang Y, Zhang L, Niu M, Li R, Tu R, Liu X, et al. Genetic risk score increased discriminant efficiency of predictive models for type 2 diabetes mellitus using machine learning: cohort study. Front Public Heal 2021;9: 606711.
115. Holmes E, Wilson ID, Nicholson JK. Metabolic phenotyping in health and disease. Cell 2008;134(5):714–7.
116. Korem T, Zeevi D, Zmora N, Weissbrod O, Bar N, Lotan-Pompan M, et al. Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. Cell Metab 2017;25(6):1243–53.e5.
117. Nielsen RL, Helenius M, Garcia SL, Roager HM, Aytan-Aktug D, Hansen LBS, et al. Data integration for prediction of weight loss in randomized controlled dietary trials. Sci Rep 2020;10(1):20103.
118. Gardner CD, Trepanowski JF, Gobbo LCD, Hauser ME, Rigdon J, Ioannidis JPA, et al. Effect of low-fat VS low-carbohydrate diet on 12-month weight loss in overweight adults and the association with genotype pattern or insulin secretion the DIETFITS randomized clinical trial. JAMA 2018;319(7):667–79.
119. Celis-Morales C, Livingstone KM, Marsaux CFM, Macready AL, Fallaize R, O'Donovan CB, et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4Me European randomized controlled trial. Int J Epidemiol 2017;46(2):578–88.
120. An J, Yoon SR, Lee JH, Kim H, Kim OY. Importance of adherence to personalized diet intervention in obesity related metabolic improvement in overweight and obese Korean adults. Clin Nutr Res 2019;8(3):171.
121. Letertre MPM, Giraudeau P, de Tullio P. Nuclear magnetic resonance spectroscopy in clinical metabolomics and personalized medicine: current challenges and perspectives. Front Mol Biosci 2021;8:698337.
122. Fiehn O. Metabolomics—the link between genotypes and phenotypes. Plant Mol Biol 2002;48(1):155–71.
123. O'Sullivan A, Gibney MJ, Connor AO, Mion B, Kaluskar S, Cashman KD, et al. Biochemical and metabolomic phenotyping in the identification of a vitamin D responsive metabotype for markers of the metabolic syndrome. Mol Nutr Food Res 2011;55(5):679–90.
124. O'Donovan CB, Walsh MC, Woolhead C, Forster H, Celis-Morales C, Fallaize R, et al. Metabotyping for the development of tailored dietary advice solutions in a European population: the Food4Me study. Br J Nutr 2017;118(8):561–9.
125. O'Donovan CB, Walsh MC, Nugent AP, McNulty B, Walton J, Flynn A, et al. Use of metabotyping for the delivery of personalised nutrition. Mol Nutr Food Res 2015;59(3):377–85.
126. Rochlani Y, Pothineni NV, Kovelamudi S, Mehta JL. Metabolic syndrome: pathophysiology, management, and modulation by natural compounds. Ther Adv Cardiovasc Dis 2017;11(8):215–25.
127. Saklayen MG. The global epidemic of the metabolic syndrome. Curr Hypertens Reports 2018;20(2):12.
128. Korduner J, Nilsson PM, Melander O, Gerl MJ, Engström G, Bachus E, et al. Proteomic and metabolomic characterization of metabolically healthy obesity: a descriptive study from a Swedish cohort. J Obes 2021;2021 :6616983.
129. Cirulli ET, Guo L, Leon Swisher C, Shah N, Huang L, Napier LA, et al. Profound perturbation of the metabolome in obesity is associated with health risk. Cell Metab 2019;29(2):488–500.e2.
130. Chen HH, Tseng YJ, Wang SY, Tsai YS, Chang CS, Kuo TC, et al. The metabolome profiling and pathway analysis in metabolic healthy and abnormal obesity. Int J Obes 2015;39(8):1241–8.
131. Cheng D, Zhao X, Yang S, Cui H, Wang G. Metabolomic signature between metabolically healthy overweight/obese and metabolically unhealthy overweight/obese: a systematic review. Diabetes Metab Syndr Obes 2021;14:991.
132. Bagheri M, Farzadfar F, Qi L, Yekaninejad MS, Chamari M, Zeleznik OA, et al. Obesity-related metabolomic profiles and discrimination of metabolically unhealthy obesity. J Proteome Res 2018;17(4):1452–62.
133. Chashmniam S, Madani NH, Ghoochani B, Safari-Alighiarloo N, Khamseh ME. The metabolome profiling of obese and non-obese individuals: metabolically healthy obese and unhealthy non-obese paradox. Iran J Basic Med Sci 2020;23(2):186–94.
134. Vázquez-Fresno R, Llorach R, Perera A, Mandal R, Feliz M, Tinahones FJ, et al. Clinical phenotype clustering in cardiovascular risk patients for the identification of responsive metabotypes after red wine polyphenol intake. J Nutr Biochem 2016;28:114–20.
135. Dong TS, Mayer EA, Osadchiy V, Chang C, Katzka W, Lagishetty V, et al. A distinct brain-gut-microbiome profile exists for females with obesity and food addiction. Obes 2020;28(8):1477.
136. Hollister EB, Oezguen N, Chumpitazi BP, Luna RA, Weidler EM, Rubio-Gonzales M, et al. Leveraging human microbiome features to diagnose and stratify children with irritable bowel syndrome. J Mol Diagn 2019;21(3):449.
137. Prochazkova M, Budinska E, Kuzma M, Pelantova H, Hradecky J, Heczkova M, et al. Vegan diet is associated with favorable effects on the metabolic performance of intestinal microbiota: a cross-sectional multi-omics study. Front Nutr 2022;8:783302.
138. Wang DD, Hu FB. Precision nutrition for prevention and management of type 2 diabetes. Lancet Diabetes Endocrinol 2018;6:416–26.
139. Acar E, Gürdeniz G, Khakimov B, Savorani F, Korndal SK, Larsen TM, et al. Biomarkers of individual foods, and separation of diets using untargeted LC-MS-based plasma metabolomics in a randomized controlled trial. Mol Nutr Food Res 2019;63(1):1800215.
140. Garcia-Aloy M, Llorach R, Urpi-Sarda M, Tulipani S, Salas-Salvadó J, Martínez-González MA, et al. Nutrimetabolomics fingerprinting to identify biomarkers of bread exposure in a free-living population from the PREDIMED study cohort. Metabolomics 2015;11(1): 155–65.
141. Cornelis MC, Erlund I, Michelotti GA, Herder C, Westerhuis JA, Tuomilehto J. Metabolomic response to coffee consumption: application to a three-stage clinical trial. J Intern Med 2018;283(6):544–57.
142. Heinzmann SS, Brown IJ, Chan Q, Bictash M, Dumas ME, Kochhar S, et al. Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. Am J Clin Nutr 2010;92(2):436–43.
143. Cheung W, Keski-Rahkonen P, Assi N, Ferrari P, Freisling H, Rinaldi S, et al. A metabolomic study of biomarkers of meat and fish intake. Am J Clin Nutr 2017;105(3):600–8.

144. Cerdá B, Tomás-Barberán FA, Espín JC. Metabolism of antioxidant and chemopreventive ellagitannins from strawberries, raspberries, walnuts, and oak-aged wine in humans: identification of biomarkers and individual variability. J Agric Food Chem 2005;53(2):227–35.
145. Neveu V, Perez-Jiménez J, Vos F, Crespy V, du Chaffaut L, Mennen L, et al. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. Database (Oxford) 2010;2010:bap024.
146. Li KJ, Brouwer-Brolsma EM, Burton-Pimentel KJ, Vergères G, Feskens EJM. A systematic review to identify biomarkers of intake for fermented food products. Genes Nutr 2021;16(1):5.
147. Hang D, Zeleznik OA, He X, Guasch-Ferre M, Jiang X, Li J, et al. Metabolomic signatures of long-term coffee consumption and risk of type 2 diabetes in women. Diabetes Care 2020;43(10):2588–96.
148. Pomyen Y, Wanichthanarak K, Poungsombat P, Fahrmann J, Grapov D, Khoomrung S. Deep metabolome: applications of deep learning in metabolomics. Comput Struct Biotechnol J 2020;18:2818–25.
149. Brinkkemper S. Method engineering: engineering of information systems development methods and tools. Inf Softw Technol 1996;38(4):275–80.
150. Scheffer J. Dealing with missing data [Internet]. Research Letters in the Information and Mathematical Sciences 2002;3:153–60. Available from: https://mro.massey.ac.nz/handle/10179/4355
151. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013;64(5):402.
152. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. BMC Med Res Method 2017;17(1):162.
153. Harrell FE. Missing data. In: Regression modeling strategies. Cham (Switzerland); Springer Nature; 2015. p. 45–61.
154. Ware JH, Harrington D, Hunter DJ, D'Agostino RBS. Missing data. N Engl J Med 2012;367(14):1353–4.
155. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 2010;50(2):105–15.
156. Lakshminarayan K, Harp SA, Goldman R, Samad T. Imputation of missing data using machine learning techniques [Internet]. 1996 [cited 2022 Jul 30]. Available from: http://www.aaai.org
157. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Method 2020;20(1):199.
158. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–8.
159. Chen T, Guestrin C. XGBoost: a scalable tree boosting system [Internet]. Available from: https://arxiv.org/abs/1603.02754v3
160. Krueger C, Tian L. A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. Biol Res Nurs 2004;6(2):151–7.
161. Nijman SWJ, Leeuwenberg AM, Beekers I, Verkouter I, Jacobs JJL, Bots ML, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol 2022;142:218–29.
162. Kondrup J, Johansen N, Plum LM, Bak L, Larsen HI, Martinsen A, et al. Incidence of nutritional risk and causes of inadequate nutritional care in hospitals. Clin Nutr 2002;21(6):461–8.
163. Wu J, Cai Z, Zhu X. Self-adaptive probability estimation for naive Bayes classification. In: The 2013 International Joint Conference on Neural Networks. New York (NY): IEEE; 2013.
164. Levy JJ, O'Malley AJ. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. BMC Med Res Method 2020;20(1):171.
165. van der Laan MJ, Polley EC, Hubbard AE. Super learner [Internet]. Statistical Applications in Genetics and Molecular Biolog 2007;6(1). Available from: https://www.degruyter.com/document/doi/10.2202/1544-6115.1309/html
166. Taha AA, Malebary SJ. A hybrid meta-classifier of fuzzy clustering and logistic regression for diabetes prediction. Computers, Materials and Continua 2022;71(2):6089–105.
167. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. Eur J Epidemiol 2018;33(5):459.
168. Bodnar LM, Cartus AR, Kirkpatrick SI, Himes KP, Kennedy EH, Simhan HN, 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(6):1235–43.

See corresponding perspective on pages 401 and 405.

Food for thought: A natural language processing analysis of the 2020 Dietary Guidelines publice comments

Joseph Lindquist,[1] *Diana M Thomas,*[1] *Dusty Turner,*[2] *Jeanne Blankenship,*[3] *and Theodore K Kyle*[4]

[1]Department of Mathematical Sciences, United States Military Academy, West Point, NY, USA; [2]Center for Army Analysis, Fort Belvoir, VA, USA; [3]Academy of Nutrition and Dietetics, Chicago, IL, USA; and [4]ConscienHealth, Pittsburgh, PA, USA

ABSTRACT

Background: The Administrative Procedure Act of 1946 guarantees the public an opportunity to view and comment on the 2020 Dietary Guidelines as part of the policymaking process. In the past, public comments were submitted by postal mail or public hearings. The convenience of public comment through the Internet has generated increased comment volume, making manual analysis challenging.

Objectives: To apply natural language processing (NLP NLP is natural language processing.) to identify sentiment, emotion, and themes in the 2020 Dietary Guidelines public comments.

Methods: Written comments to the Scientific Report of the 2020 Dietary Guidelines Advisory Committee that were uploaded and visible at https://beta.regulations.gov/docket/FNS-2020-0015 were extracted using a computer program and retained for analysis. All comments were filtered, and duplicates were removed. A 2-round latent Dirichlet analysis (LDA) was used to identify 3 overarching topics as well as subtopics addressed in the comments. Sentiment analysis was applied to categorize emotion and overall positive and negative sentiment within each topic.

Results: Three different topics were identified by LDA. The first topic involved negative sentiment surrounding removing dairy from the guidelines because the commenters felt dairy is unnecessary. The second topic focused on positive sentiment involved in restricting added sugars. The third topic was too diverse to characterize under 1 theme. A second LDA within the third topic had 3 subtopics containing positive sentiment. The first subtopic valued the inclusion of dairy in the recommendations, the second involved the health benefits of consuming beef, and the third indicated that the recommendations lead to overall good health outcomes.

Conclusions: Public comments were diverse, held conflicting viewpoints, and often did not base comments on personal anecdotes or opinions without citing scientific evidence. Because the volume of public comments has grown dramatically, NLP has promise to assist in objective analysis of public comment input. *Am J Clin Nutr* 2021;114:713–720.

Keywords: 2020 Dietary Guidelines, natural language processing, machine learning, public comments, sentiment, emotion, topic modeling, latent Dirichlet allocation

Introduction

The 1946 Administrative Procedures Act gave the American public the right to be involved in the federal regulatory process by mandating public notice of proposed regulations by federal agencies and providing opportunities to receive public comment on the proposed regulations (1). Public notice and comment were primarily achieved through public hearings and mail-in comments. More than 50 y later, the E-Government Act of 2002 (2) sparked much greater public engagement because it supported the application of the Internet to provide more opportunities for participation. Nowhere is this increase more apparent than with the Dietary Guidelines for Americans. The 2010 Dietary Guidelines Scientific Report received 2000 comments (3), whereas the 2015 Dietary Guidelines for Americans received more than 29,000 comments (4). After publication of the 2015 Dietary Guidelines for Americans (5), significant questions arose about underlying process for developing the guidelines, including questions about public input and transparency (6). One reason for

The authors reported no funding received for this study.

Supplemental Data and Supplemental Figure 1 are available from the "Supplementary data" link in the online posting of the article and from the same link in the online table of contents at https://academic.oup.com/ajcn/.

Address correspondence to DT (e-mail: diana.thomas@westpoint.edu).

Abbreviations used: HHS, Department of Health and Human Services; LDA, latent Dirichlet analysis; NLP, natural language processing.

Received December 2, 2020. Accepted for publication March 17, 2021.

First published online June 16, 2021; doi: https://doi.org/10.1093/ajcn/nqab119.

those questions was the sheer volume of public comments on the scientific report that supported the final guidelines.

On the basis of the number of public comments, the US Congress requested a consensus report from the National Academies of Sciences, Engineering, and Medicine on making the process for producing the 2020 report more transparent, inclusive, and science driven (6). Despite this, the report did not address transparency involving the analysis of the public comments.

USDA and Department of Health and Human Services (HHS) staff process public comments by reviewing, categorizing, and summarizing them. Those summaries are not available to the public until after the final guidelines are published. The 2015 review of comments was performed manually by senior staff at USDA with expertise in nutrition science and dietary guidelines (4). Although it is unclear exactly how duplicates were managed and how topics and themes were classified (4), manual review of comments largely involves subjective analysis.

Once again, the Scientific Report of the 2020 Dietary Guidelines Advisory Committee has stimulated a large number (38,368) of public comments. Here, we applied a web scraping script to collect the comments from the USDA website and apply natural language processing (NLP) methods to produce an objective summary of the themes and sentiment reflected in this large volume of content.

Materials and Methods

Dietary Guidelines comment extraction

The USDA and HHS solicited public comments on the Scientific Report of the 2020 Dietary Guidelines Advisory Committee, and they were uploaded to https://www.dietaryguidelines.gov/work-under-way/get-involved/submit-comment and visible at https://beta.regulations.gov/docket/FNS-2020-0015. A total of 38,368 comments were scraped from the site using the RSelenium (7) package in the statistical software R (Version 4.0.3; R Core Team).

Delineating unique comments

Because many of the comments were duplicates or nearly duplicates, a program was written in R (Version 4.0.3; R Core Team) to flag comments that had identical first sentences. These were then grouped and evaluated by whether the comments that had identical first sentences were truly identical, differed only because of signatures, or had some sentences at the end of the comment that personalized the duplication. Only the first of each duplicated comment, combined with the unique comments, was retained for topic modeling analysis. A workflow diagram of the process to identify duplicate comments and retain only one from the group appears in **Figure 1**. An additional detailed programming flowchart is also supplied in **Supplemental Figure 1**.

Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is an unsupervised method to identify distinct topics in a set of documents, which in our case is the set of all comments. Each word in entire data set is assigned a probability of belonging to one of k topics, where k is seeded by the user. The R package "topicmodels" (8) was used to implement the LDA algorithm in a program written in R (R Core Team). LDA was performed for $k = 2, \ldots, 4$, and a plot of words with the highest probability of belonging to each respective topic was evaluated by topic number. The probability that each comment belonged to topic k was calculated. The comment was then assigned to topic number k by the maximum probability.

There were 3 main topics identified by the LDA. Examination of the third topic revealed high variability in the comments within. Therefore, we performed a second LDA in the third topic to classify the different themes within.

Sentiment analysis

The "tidyverse" and "tm" packages for the programming language R (Version 4.0.3; R Core Team) (9) were applied to calculate sentiment within each of the topics identified by the LDA algorithm. Sentiment is calculated by matching words to a dictionary or lexicon that contains frequently used English words (9). The words in the lexicon are classified as positive or negative. Here, we applied the "nrc" lexicon, named for the National Research Council of Canada, because in addition to classifying words as positive and negative, the lexicon also classifies each word by emotions of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (10). Words that are not in the lexicon are assigned the value 0. The sentiment scores for each comment were calculated in the unique comments. Sentiment and a count of the words associated with each emotion within each LDA topic were determined.

Word frequency

Free-text response patterns can be tabulated by the frequency of words or phrases (n-grams) within a set of free-text responses. The topmost frequent trigrams (sets of 3 consecutive words) were tabulated within each LDA topic and plotted as a bar chart. Stop words such as "the," "of," and "to" were removed from the text (9). In addition, common nouns that were referencing the committee or the report, such as "Dietary Guidelines Committee" and "Scientific Report," were removed.

Results

Dietary guidelines comment extraction

A total of 38,368 comments were received at the 2020 Dietary Guidelines Committee Comment website (11). That total count included petitions with multiple signatures. The petitions with multiple signatures were counted as a comment. Also, a much smaller number of vulgar and inappropriate comments were removed from public view by the USDA. Thus, only 26,510 comments were available for this analysis. Some organizations submit comments on behalf of their membership (e.g., the Academy of Nutrition and Dietetics). Other comments represent the views of a single individual.

After removing comments that were only uploaded files, 14,689 comments were extracted from the comment site. Of

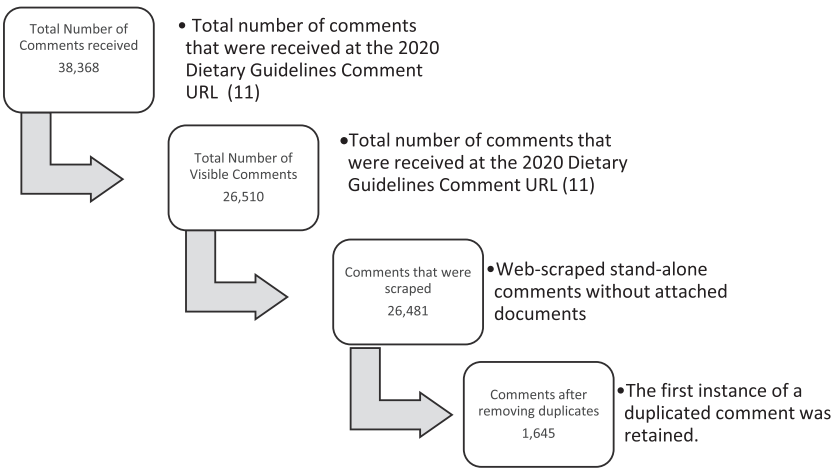

FIGURE 1 Workflow diagram describing retainment of final database of 2020 Dietary Guidelines public comments.

these, 13,004 were duplicated. There were 15 distinct groups of duplicated comments. **Table 1** contains information about each duplicated comment represented by the first sentence. Of the duplicated comments, 12,710 of the comments were linked to one duplication.

LDA topics

After visually reviewing words assigned with the highest probabilities to 2, 3, and 4 topics, 3 topics appeared to be the most distinct. **Figure 2** is a bar plot of the estimated probability a word belonged to topics 1, 2, and 3.

LDA topic classifications using top word bigrams

The top row of **Figure 3** depicts the most frequently appearing bigrams by topic. Top topic 1 bigrams such as "lactose intolerant" and "remove dairy" suggest that topic 1 involves comments associated with removing dairy from the Dietary Guidelines. Top topic 2 bigrams like "added sugars" and "lower limits" suggest that topic 2 classifies comments on added sugars. Finally, topic 3 bigrams like "health outcomes," "cardiovascular disease," and "public health" suggest that topic 3 groups together comments involving health. Because topic 3 had more wide-ranging words in the bigrams, we reran a LDA on topic 3.

Emotions by LDA topic

The second row of Figure 3 represents the counts of words classified by emotion in each topic. Emotions of anger were higher in the dairy (topic 1) group. Emotions of trust were the lowest and fear was the highest associated with added sugars. Emotions of anticipation were higher with health outcomes (topic 3).

Sentiment distribution by LDA topic

The distribution of summed positive and negative by comment in each LDA topic was plotted in **Figure 4**. The highest negative sentiment was observed in topic 1 associated with dairy. There is a left-skewed distribution for topic 2 associated with added sugar. Added sugar received positive sentiment because of support for restriction of added sugars in the Dietary Guidelines. Figure 4 demonstrates a right-skewed distribution for topic 3 associated with health outcomes.

Description of topic 3

Figure 5 is a bar plot of the estimated probability a word appeared in 1 of 3 subtopics identified by LDA within topic 3. The 3 words with the highest probability within each subtopic were "dairy," "beef," and "health," respectively.

Discussion

These results offer an objective and timely analysis of voluminous comments on the Scientific Report of the 2020 Dietary Guidelines Advisory Committee. Until 2015, manual review and summarization of this input to the guidelines process might have been reasonably complete and effective, but this situation has clearly changed. The present analysis demonstrates that NLP techniques can provide a practical means for a complete and objective summary.

The public feedback on this Scientific Report can be categorized into 3 major topics. The first and largest topic comprises comments suggesting that including dairy products as a broad recommendation is unwarranted and potentially harmful. The highest number of comments classified as angry fall into this topic. A typical comment in this cluster might say, "It's time for the Dietary Guidelines to protect the health of Americans by making it clear that dairy is unnecessary." At the same time, it is worth noting that this sentiment runs counter to the need for adequate calcium intake. In fact, some analyses suggest that it is quite difficult or impossible to provide adequate calcium intake while meeting other nutrient recommendations with dairy-free diets (12).

The second topic focuses on added sugars. Examples from this grouping include "added sugars are contributing to an obesity crisis among our youngest children" and "nearly 60% of infant and toddler food and drink advertising dollars

TABLE 1 First sentence of duplicate comment, number of times repeated, type of comment, and topic comment was associated with[1]

First sentence	Number	Type	Content topic
Dear Secretary Sonny Perdue, I call on the US Department of Agriculture and the US Department of Health and Human Services to follow the Dietary Guidelines Advisory Committees recommendations to support the long-term sustainability of the food system, as well as lowering limits on added sugars in the 2020–2025 Dietary Guidelines for Americans.	13,109	I	Added sugars and sustainability
Its time for the Dietary Guidelines to protect the health of Americans by making it clear that dairy is unnecessary.	5029	S & P	Dairy
To the USDA and HHS: For the health of our nation, I urge you to remove dairy as a recommended food group in the upcoming 2020–2025 Dietary Guidelines.	3062	S & P	Dairy
I'm writing to ask that the Departments of Agriculture and Health and Human Services ensure Americans get dietary advice based on sound science and common sense.	1571	P: some duplicates, some very personalized	Alcohol
Added sugars are contributing to an obesity crisis among our youngest children.	795	S & P	Added sugar
I am proud to raise cattle and feel good about serving beef to my family because I know that no other food delivers the same nutrient-rich package as a three-ounce serving of beef.	324	P	Beef
I want to commend the Dietary Guidelines Scientific Advisory Committee for its work on its report highlighting the latest nutrition science and offering recommendations for healthy diets.	73	I	Dairy
Thank you for the opportunity to comment on the development of the next dietary guidelines.	58	P: highly varied	Dairy
Dear Secretary Sonny Perdue, I urge the US Department of Agriculture and the US Department of Health and Human Services to follow the Dietary Guidelines Advisory Committees recommendations to support the long-term sustainability of the food system, as well as lowering limits on added sugars in the 2020–2025 Dietary Guidelines for Americans.	15	S: signed at end of text block	Added sugar
To the USDA and HHS: For the health of your nation, I urge you to remove dairy as a recommended food group in the upcoming 2020–2025 Dietary Guidelines.	11	P: very few slightly altered	Dairy
I support whole-food, vegan diet not only to end our exploitative relationship with animals but because of the array of health benefits they provide for individuals of all ages and lifestyles.	11	I	
A new report found that the food-related emissions in G20 countries, which make up two-thirds of the world's population, account for 75% of the world's carbon budget for food.	8	P	Sustainability
Dear Secretary Sonny Perdue, We call on the US Department of Agriculture and the US Department of Health and Human Services to follow the Dietary Guidelines Advisory Committees recommendations to support the long-term sustainability of the food system, as well as lowering limits on added sugars in the 2020–2025 Dietary Guidelines for Americans.	8	S: signed at the end of text block	Added sugars and sustainability
Dear Secretary Sonny Perdue, I ask the US Department of Agriculture and the US Department of Health and Human Services to follow the Dietary Guidelines Advisory Committees recommendations to support the long-term sustainability of the food system, as well as lowering limits on added sugars in the 2020–2025 Dietary Guidelines for Americans.	7	P: barelyS: signed at end of text block	Added sugars and sustainability
I urge the U.S. Departments of Agriculture and Health and Human Services to align the 2020 dietary guidelines with the planetary health diet.	6	P: highly varied	Sustainability

[1]HHS, Department of Health and Human Services; I, identical; P, personalized; S, signature.

promoted products not recommended for young children." Positive sentiment surrounding added sugars was found in this topic. The positive sentiment was associated with comments that viewed the Dietary Guidelines restriction on added sugars positively.

The third major topic used positive language associated with healthfulness and sustainability of the food supply. The comments in this topic were more variable than the other 2 topics. The LDA within topic 3 revealed 3 subtopics that can be represented as dairy, beef, and health. In contrast to the first

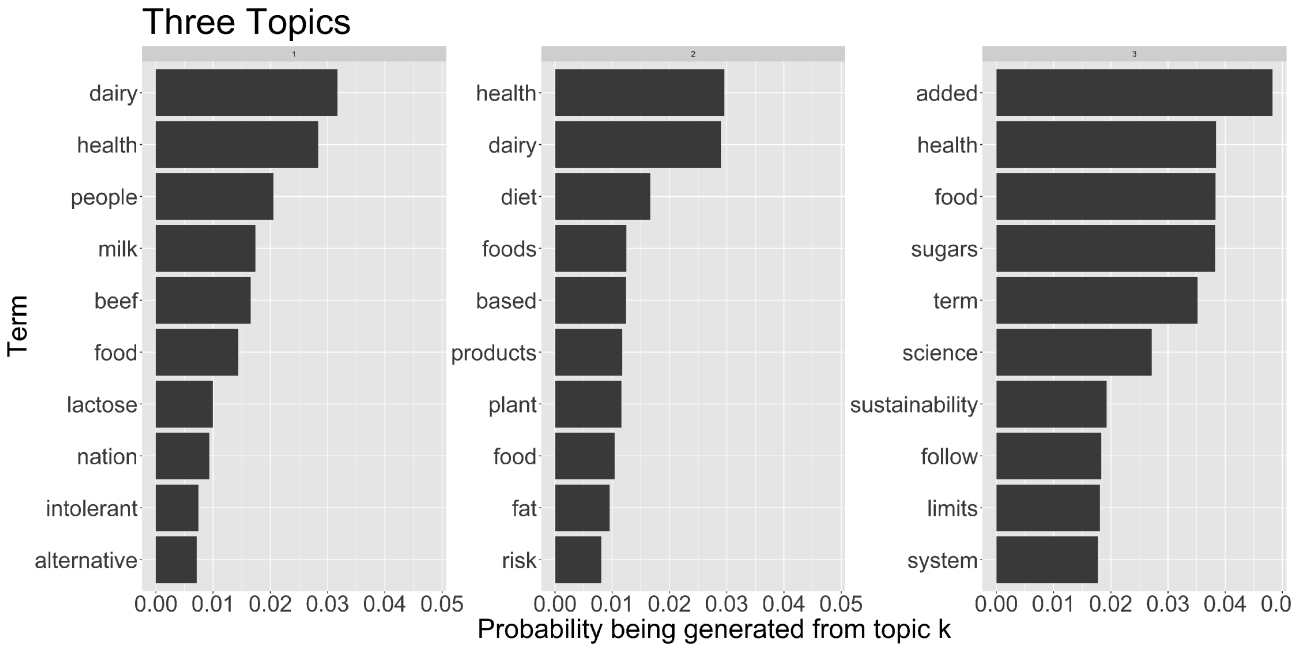

FIGURE 2 Bar plot of latent Dirichlet allocation estimated probability a word belongs to topic 1, 2, or 3.

topic, many comments in subtopic 1 were positive about the value of dairy products for health. Emotions of anticipation were highest in this group, perhaps reflecting a hope for guidelines that would promote better health. Three excerpts of comments from each subtopic in topic 3 illustrate the positive sentiment used in topic 3.

Excerpt 1

I want to commend the Dietary Guidelines Scientific Advisory Committee for its work on its report highlighting the latest nutrition science and offering recommendations for healthy diets. I especially appreciated its many positive findings on the importance of dairy, including the committee specifically pointing to the underconsumption of dairy as problematic when discussing the current landscape of the American diet.

Excerpt 2

I know that no other food delivers the same nutrient-rich package as a three-ounce serving of beef. Study after study shows that beef plays an important role in a balanced, healthy diet across the lifespan. Beef supports healthy pregnancies, ensures the healthy growth and development of children, and helps adults maintain strength and energy throughout adulthood so they can age vibrantly and have independent and active lifestyles.

Excerpt 3

The US Department of Agriculture and the US Department of Health and Human Services have a chance to make a difference in how Americans can grow healthful food and improve the health of our people in the future. I encourage you to follow the Dietary Guidelines Advisory Committees recommendations to support the long-term sustainability of the food system, as well as lowering limits on added sugars in the 2020–2025 Dietary Guidelines for Americans.

The comments supplied here include little or no supporting evidence, which should serve as an important criterion for inclusion in the regulatory process (13). With the increase in public comments, federal agencies could provide education on how to draft effective comments available at the commenting site (13).

How can the text analysis help the USDA review comments?

Duplicate comments were removed before the comments were posted (4). Despite this, the algorithm that identified duplicate comments found many duplicate comments that were undetected differed slightly because they contained different signatures or 1 personalized sentence. Our algorithm would be beneficial to the USDA not only because it pared down the comments to the truly unique ones but also because it automatically grouped together the duplicate comments by the level of replication.

As the number of comments increase, it will become more challenging to manually organize comments. The time involved to read the comments will be substantial, and classification is subject to the reviewer's bias. Moreover, conveying exactly how the decision was made to move each individual comment into a classification will be prohibitive. The LDA analysis automates this process using a well-defined algorithm to group comments. Once the comments are grouped by LDA, the comments within each topic can be manually evaluated.

It may be the case that comments are grouped under a topic that is wide-ranging like the third topic in our analysis. Understanding why the comments were grouped together will still require manual reading, but the number of comments in the topic is substantially less than the entire original set of comments, making this manual reading more manageable. It only took a few

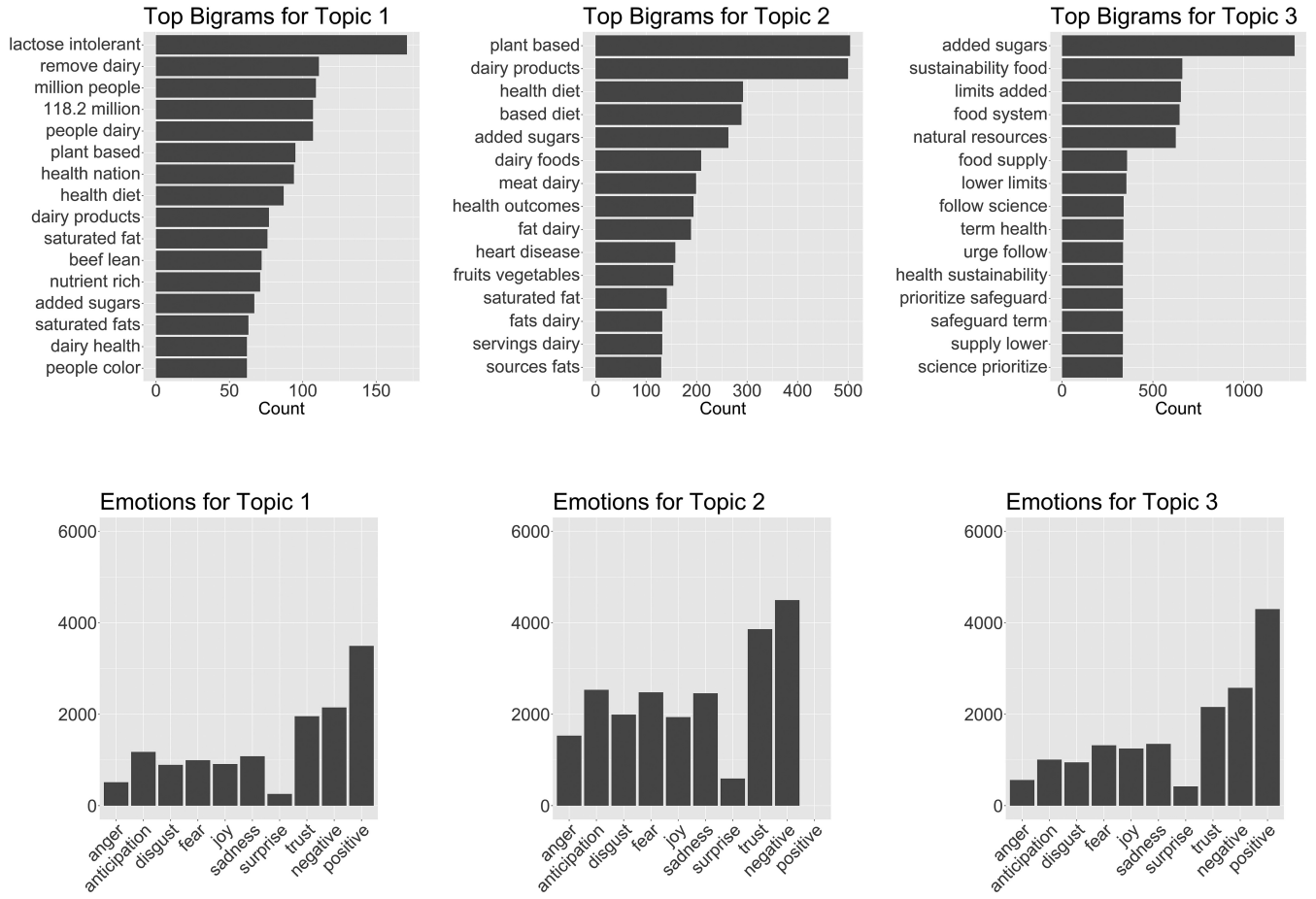

FIGURE 3 The top row represents the counts of word bigrams by topic. The second row represents the count of words classified by emotions for each topic.

minutes for our team to browse topic 3 comments and identify the overarching reason the comments were grouped together and the variety within these comments. This variety is what led to the second LDA within topic 3 to break up topic 3 into 3 further subgroups. Furthermore, applying sentiment analysis is not possible for a large set of comments manually. Using the precoded "nrc" dictionary in R that had words assigned to be positive or negative allows a user to compare sentiment across the topics as we have done here. Although analysis will take less time due to the NLP algorithms, conveying how the NLP

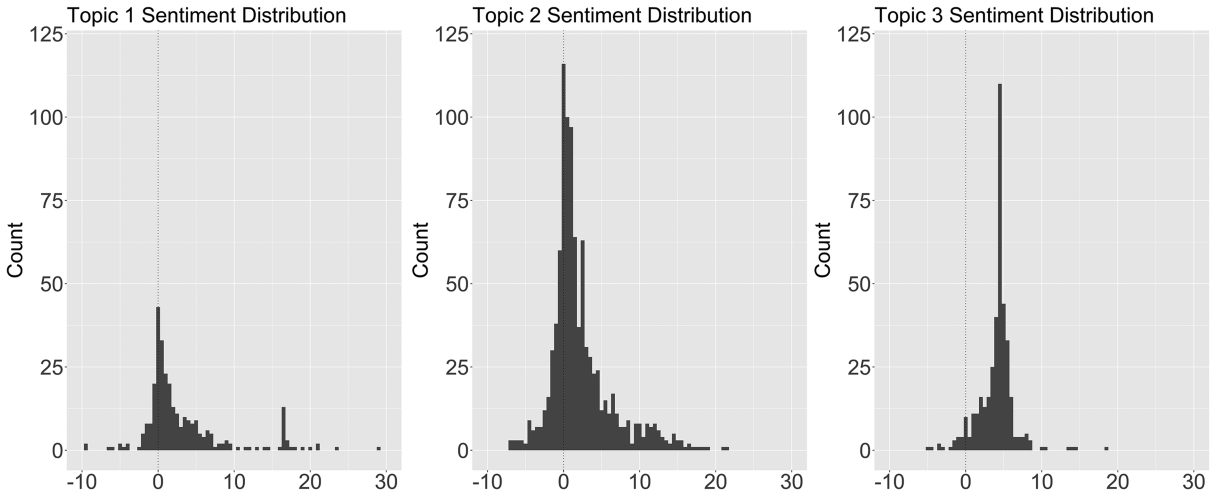

FIGURE 4 The distribution of sentiment by topic.

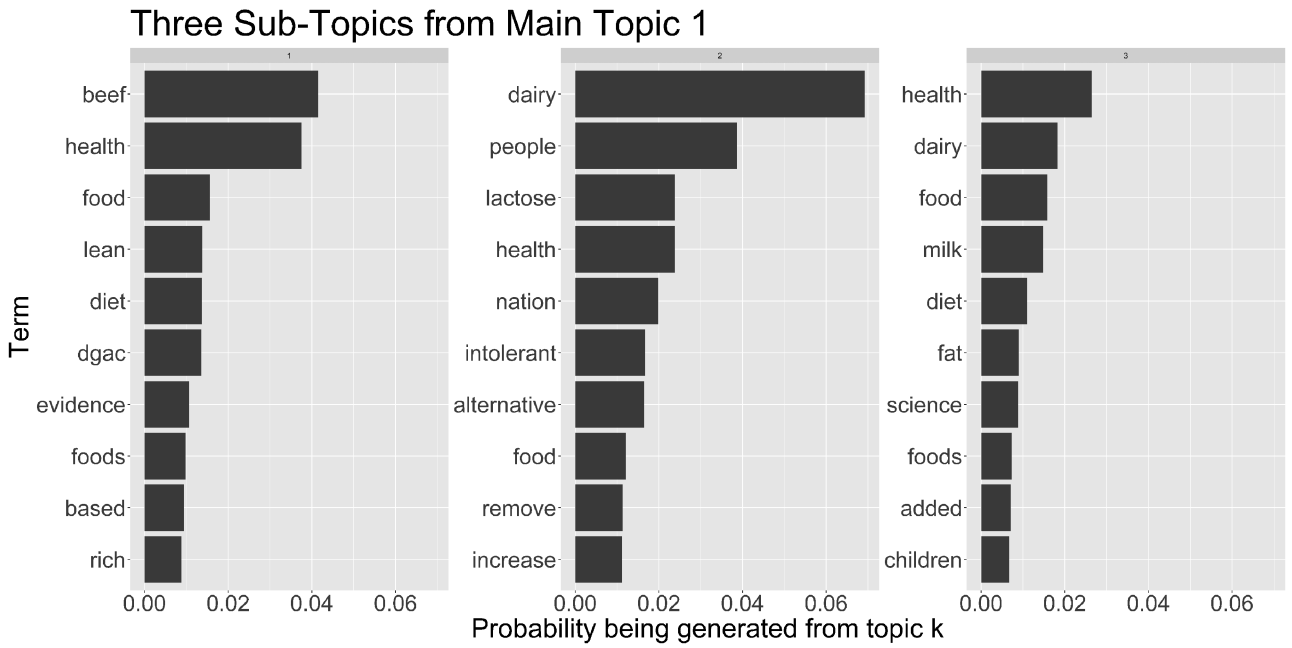

FIGURE 5 Bar plot of estimated latent Dirichlet allocation performed on topic 3 comments of the probability a word belongs to subtopic 1, 2 or 3 within topic 3.

algorithms work to a lay public will take more time but is a worthwhile effort.

A limitation of sentiment analysis is that we are relying on a general dictionary, namely, the "nrc" package that predetermined whether a word is positive or negative. The general dictionary is now being applied to a very specific list of words used to comment on the Dietary Guidelines. For example, the word "sugar" in the "nrc" lexicon has been assigned positive sentiment. This means every time a comment includes the word "sugar," the comment gets 1 count in the positive direction. However, it is not necessarily the case that all references to "sugar" are positive in the comments. It is possible to tune the lexicon to your specific analysis, but tuning is problematic because now you are fitting the lexicon to your data, and your findings may not be objective. We chose to not tune the lexicon and relied only on the general form of "nrc" in order to not introduce subjective and biased decisions into the analysis.

A second limitation is that topics in LDA need to be manually evaluated. Although the topics and the words that fall within each topic are determined under a Dirichlet distribution assumption, the theme of the topic still needs to be determined by a human.

Despite these limitations, using LDA and sentiment analysis reduces the burden on the analyst, is equipped to handle large numbers of comments, and is transparent with algorithms determining assignments of sentiment and topics.

Conclusions

After issuing the Scientific Report of the Dietary Guidelines Committee and receiving public comments, USDA and HHS have developed and issued the final guidelines. This analysis offers an objective means to enhance the transparency of factoring public comments into the final guidelines. In addition, the analysis provides a scalable and feasible method for analysis in anticipation of even more future public commenting.

The 2020 Dietary Guidelines will reflect a full range of comments received throughout the yearlong process of analysis, deliberations, public hearings, input, and review. The final product must strike a balance between public sentiment and scientific evidence while being sensitive to the realities of political factors. Transparency in objectively reviewing these comments remains both challenging and essential.

We thank COL Raymond Kimball for coming up with the title for the manuscript.

The authors' contributions were as follows—TKK: conceived the study; JL: developed the program to scrape the data from the site; JL, DT, and DMT: performed the natural language processing analysis; TKK and JB: interpreted the natural language processing findings; and all authors: participated in writing several drafts of the manuscript and read and approved the final manuscript. DMT is an associate editor at *AJCN* and had no role in the journal's evaluating this manuscript. TKK has received consulting fees unrelated to this work from Gelesis, Johnson & Johnson, Novo Nordisk, and Nutrisystem. All other authors report no conflicts of interest.

Data Availability

Data described in the manuscript, code book, and analytic code are made publicly and freely available without restriction at https://beta.regulations.gov/document/FNS-2020-0015-0001/comment.

References

1. Bureau of National Affairs. Administrative Procedure Act; summary and analysis. Washington (DC): The Bureau; 1946.

2. Seifert JW, Relyea HC. E-Government Act of 2002 in the United States. In: Anttiroiko A-V, Malkia M, eds. Encyclopedia of Digital Government. Hershey (PA): Idea Group Reference; 2007:476–81.
3. USDA. 2010 Dietary Guidelines for Americans: background, history and process. Washington (DC): USDg; 2010.
4. Public comments on the Scientific Report of the 2015 Dietary Guidelines Advisory Committee. Washington (DC): US Departments of Health and Human Services (HHS) and Agriculture (USDA); 2015.
5. 2015–2020 Dietary Guidelines for Americans. 8th ed. Washington (DC): U.S. Department of Health and Human Services (HHS) and the U.S. Department of Agriculture (USDA); 2015.
6. Redesigning the process for establishing the Dietary Guidelines for Americans. Washington (DC): National Academies of Sciences, Engineering, and Medicine; 2017.
7. Harrison J. RSelenium: R bindings for 'Selenium WebDriver'. [Internet]. 2019. Available from: https://cran.r-project.org/web/packages/RSelenium/index.html.
8. Grün B, Hornik K. Topic models: an R package for fitting topic models. J Stat Softw 2011;40(13):1–30.
9. Silge J, Robinson D. Text mining with R: a tidy approach. Boston (MA): O'Reilly; 2017.
10. Mohammad SM. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. Presented at: 56th Annual Meeting of the Association for Computational Linguistics 20 July 2018; Melbourne, Australia; 2018.
11. [Internet] [cited 2020 Oct 14]. Available from: https://beta.regulations.gov/document/FNS-2020-0015-0001/comment
12. Gao X, Wilde PE, Lichtenstein AH, Tucker KL. Meeting adequate intake for dietary calcium without dairy foods in adolescents aged 9 to 18 years (National Health and Nutrition Examination Survey 2001–2002). J Am Diet Assoc 2006;106(11):1759–65.
13. Looney A. How to effectively comment on regulations. Brookings Institute, Washington D.C.; 2018.

Biostatistics (2020) **21**, 2, *pp.* 339–344
doi:10.1093/biostatistics/kxz040
Advance Access publication on November 19, 2019

Teaching yourself about structural racism will improve your machine learning

WHITNEY R. ROBINSON*

Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill; Carolina Population Center, University of North Carolina at Chapel Hill, CB #7345 McGavran-Greenberg, Chapel Hill, NC 27599-7435, USA

whitney_robinson@unc.edu

AUDREY RENSON

Department of Epidemiology, University of North Carolina at Chapel Hill, 135 Dauer Dr., Chapel Hill, NC 27599, USA

ASHLEY I. NAIMI

Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, 5131 Public Health Building, Pittsburgh, PA 15261-3100, USA

Summary

In this commentary, we put forth the following argument: Anyone conducting machine learning in a health-related domain should educate themselves about structural racism. We argue that structural racism is a critical body of knowledge needed for generalizability in almost all domains of health research.

Keywords: Causal inference; Directed acyclic graphs; Machine learning; Structural racism.

1. Motivation

In this commentary, we put forth the following argument: anyone conducting machine learning in a health-related domain should educate themselves about structural racism. As Domingos and others have argued, "Every learner must embody some knowledge or assumptions beyond the data it is given in order to generalize beyond it" (Domingos, 2012). We argue here that structural racism is a critical body of knowledge needed for generalizability in almost all domains of health research. We believe that this is especially true when inference relies on algorithms ("learners") to choose statistical models.

We make the recommendation to incorporate structural racism based on our experiences as epidemiologists. Epidemiologists are fundamentally interested in causing changes that result in improved individual and population health and that reduce health disparities (Glass *and others*, 2013). Quantifying statistical associations are central to these objectives, yet they are not sufficient. As we observe the challenges facing applied machine learning, we see echoes of debates that epidemiology has encountered. For instance, epidemiology's disease screening literature has imparted the intuition that even highly sensitive and specific

*To whom correspondence should be addressed.

screening algorithms can produce more false positives (analogous to the high "false discovery rate" in machine learning terminology) than true positives when an outcome is rare.

In particular, we are inspired by our field's 1980s- and 1990s-era debates about "black box epidemiology" (Weed, 1998). In many respects, this debate mirrors debates in machine learning about the trade-offs between improved prediction versus greater model interpretability (Seligman *and others*, 2018). The earlier epidemiology debate contrasted the use of multivariable-adjusted regression models to identify behavioral risk factors for cancer incidence to target in prevention efforts ("black box") versus research elucidating biological pathways of cancer development, particularly at the level of molecular biology ("mechanistic") (Weed, 1998). However, the "mechanistic" side's focus on molecular mechanisms ignored all the parts of the causal structure that were "above" the molecular level. A key insight of this debate was that, even a mechanistically oriented research orientation has its blind spots. Specifically, the integration of sociopolitical forces with a consideration of biology and behavior was missing in early debates in the field (Weed, 1998). The need to integrate factors from across the full breadth of the causal structure is likely even more crucial when making causal inference.

2. Why structural racism is critical knowledge

Structural racism refers to "the totality of ways in which societies foster [racial] discrimination, via mutually reinforcing [inequitable] systems...(e.g., in housing, education, employment, earnings, benefits, credit, media, health care, criminal justice, etc.) that in turn reinforce discriminatory beliefs, values, and distribution of resources," reflected in history, culture, and interconnected institutions (Bailey *and others*, 2017). Below we present two reasons why expanding one's knowledge base of structural racism will improve the quality of work produced by machine learning applications to health.

First, even when racial categorization is not a topic of interest in an analysis, structural racism will shape associations of health-related processes. For instance, in a U.S. context, the sociodemographic variable White/Black race is a variable that is frequently available in health-related datasets and often highly predictive of health outcomes. For instance, Seligman *and others* (2018) attempted to predict body mass index (BMI), blood pressure, and waist circumference in a population of 15,784 longitudinally assessed participants using four machine learning methods (linear regression, penalized regressions, random forests, and neural networks). With access to 458 variables, "Black/African-American Race" was one of the top five variables selected by all four machine learning models Seligman *and others* (2018). In non-U.S. contexts, other markers of social stratification, such as caste, ethnicity, religion, social class, country of birth, home village, gender, or sexual identity, will be similarly predictive in health processes that involves human agency and social organization. An analyst striving to produce the most predictive and illuminating models ignores structural racism (and other axes of inequality) at his or her own peril. Ignoring structural racism is a decision to ignore structures that give rise to many and varied associations with health.

A second motivation for educating oneself about structural racism is the common occurrence of "algorithmic bias." When "race-neutral" approaches are employed in model development, prediction will tend to be poorer for racial minority populations. Greater error rates, and even failure of algorithms to perform at all, for racial minorities have been widely reported. Examples include facial recognition software that misidentifies gender and even species when presented with dark-skinned women of African descent (Buolamwini and Gebru, 2016) and proprietary formulas used in criminal sentencing that misclassify defendants as high risk for recidivism at a greater rate for Black versus White defendants (Rudin *and others*, 2018). Two explanations for differentially poorer model performance can be addressed by collecting more data: too few observations of members of racial minority groups and unrepresentative sampling that can differentially limit generalizability (Kreatsoulas and Subramanian, 2018).

However, an additional cause of algorithmic bias is not well appreciated and cannot be overcome simply by adding more of the same kind of data to a learner. The problem is the data generation process

itself. The data generation process "[is] an inherently subjective enterprise in which a discipline's norms and conventions help to reinforce existing racial (and other) hierarchies" (Ford and Airhihenbuwa, 2010). As explicated by the Public Health Critical Race Praxis, without an explicit focus on social equity, the concerns of the most privileged members of society are overrepresented in data and research (Ford and Airhihenbuwa, 2010). One empirical demonstration of this phenomenon is Tehranifar *and others*' (2009) investigation of racial disparities in survival ranked by the degree of knowledge that had been generated about a cancer. Ranked from "nonamenble" (little knowledge, uniformly high mortality rates) to "mostly amenable" (well-studied, high survival), "the hazard ratios (95% confidence intervals) for African Americans versus Whites from nonamenable, partly amenable, and mostly amenable cancers were 1.05 (1.03–1.07), 1.38 (1.34–1.41), and 1.41 (1.37–1.46), respectively" (Tehranifar *and others*, 2009). Absent much knowledge about prevention or treatment, Blacks and Whites had similar mortality outcomes. The kind of knowledge that was produced disproportionately benefited the health of the White populations. One factor is the ability of more privileged groups to access knowledge and leverage financial and medical resources (Phelan *and others*, 2010). But a more fundamental factor is the research enterprise itself tends to collect more data and advance more quickly on problems that disproportionately affect those in society with more power and resources. For instance, over the past decades, medical treatment for the "triple-negative" subtype of breast cancer that disproportionately affects Black women has advanced much more slowly than for the hormone-receptor positive subtypes that are disproportionately diagnosed among White women in the Unites States (Foulkes *and others*, 2010). Similarly, in the case of multiple myeloma wherein Black patients make up 20% of all people diagnosed, in a review of 21 clinical trials in patients with multiple myeloma, the median proportion of Black patients enrolled was only 4.5% (Bhatnagar *and others*, 2017).

One solution to algorithmic bias is to follow Doug Weed's suggestion from the "black box" epidemiology debates: devote special attention to causal structures that act at different levels. A tool that we find valuable for this effort are directed acyclic graphs, or DAGs. In a DAG, arrows between nodes represent our beliefs about the presence of causal relationships among factors under study. D-separation, a set of rules for drawing and analyzing the relationships in the DAG (http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html), results in concrete guidance for model building and interpretation. For instance, DAG analysis can identify biases, such as uncontrolled confounding ("omitted variable bias"), unaddressed in an analysis or even biases introduced when problematic variables are included in the adjustment set (e.g., instruments or mediators) (Keil *and others*, 2018; Hernán *and others*, 2001), which is not the same as, but related to, the concept of bias from "overfitting" (Seligman *and others*, 2018).

3. Empirical example: algorithms and lung function

Here we present an example from the pulmonary health literature. Lung function is typically assessed using a tool called a spirometer (Braun, 2015). Internationally, most commercially available spirometers require the operator to "select the race of an individual, as well as indicate their age, sex/gender and height. These data are fed into an algorithm that "corrects" for each factor, based on the assumption that normal levels of lung function differ by age, sex, height, and race. The first formulas were produced in the United States in the 1920s, "during a period when eugenic policies rooted in hereditarianism were popular" (Braun, 2015). The formulas have been updated over time, most recently based on data from the 1988–1994 NHANES exam (Braun, 2015). Today, for example, in the United States, compared to a "Caucasian" population, correction factors for individuals labeled "black" range from 10% to 15%, on the assumption that Black people have constitutionally poorer lung function; for people labeled "Asian," correction factors are between 4% and 6% (Braun, 2015).

Figure 1 shows a DAG depicting how the spirometer manufacturers and most lung function researchers tacitly conceptualize the relationship between race and normal lung function. In addition, White workers

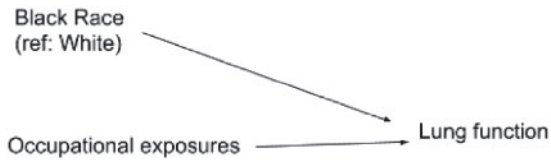

Fig. 1. Directed acyclic graph (DAG) depicting naive conceptualization of causal relationships among race, occupational exposures, and lung function.

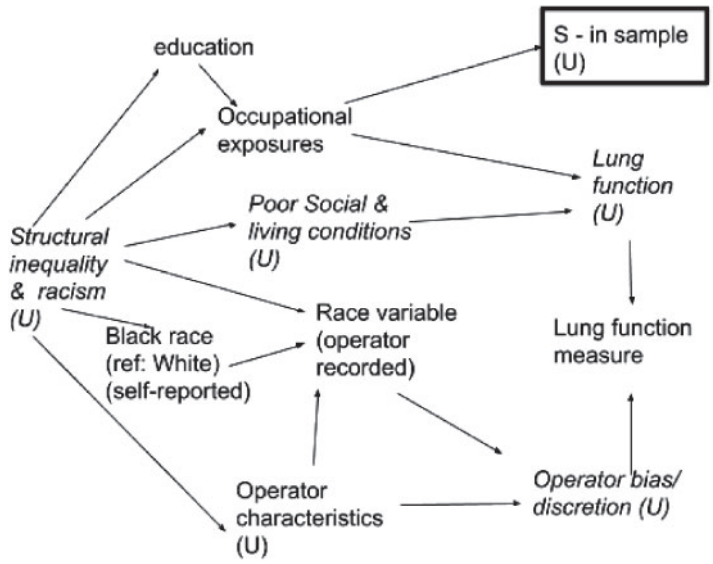

Fig. 2. Directed acyclic graph (DAG) of relationships among race, occupational exposures, and lung function, incorporating structural racism and theories from the Public Health Critical Race Praxis.

are centered in this conceptualization (Ford and Airhihenbuwa, 2010). Their lung function is considered to be the norm from which non-White people deviate. Race is being conceptualized here as a characteristic of a person that independently and innately depresses lung function. In addition, this DAG encodes the prediction that the effect of harmful occupational exposure on lung function would differ for Black and White people. Therefore, it would makes sense to correct for race when predicting lung function based on occupational exposures. The result is that a given level of poor lung function would be considered abnormally low for a White person but normal for a Black person. The implications are profound: Black and Asian people must exhibit lower levels of actual lung function than White people to cross clinical thresholds for qualifying for therapeutic services or disability benefits. Further, the model predictions could easily justify allowing greater levels of exposures to occupational hazards for Black workers who do not exhibit the depressed levels of lung function associated with harm.

Figure 2 presents a similar DAG informed by knowledge of structural racism. This DAG explicitly incorporates possible operator bias, interrogates the source of the "race" variable and its meaning (e.g., self-reported forced choice among categories, operator perception). Further, the DAG makes the data generation process and variations in data quality more explicit. Finally, by acknowledging that observed associations between race and lung function could be entirely explained by racial disparities in living and working conditions, the DAG animates the analyst to collect or include additional, more proximally causal,

and likely more statistically predictive, variables, such as living conditions, for which race is operating as a surrogate. Crucially, incorporating more proximal and predictive variables into models, rather than relying on race variables to act as proxies, will improve transportability of algorithms across contexts.

4. Conclusion

"The fundamental goal of machine learning is to generalize beyond the algorithm training set" (Kreatsoulas and Subramanian, 2018). In particular, when applied to health and health care, even models intended only for prediction will have causal impacts. Model results will be used for decision-making about the allocation of health care, access to social welfare and disability systems, and acceptable limits of medical exposures for vulnerable populations. We have argued that grounding one's work in an understanding of structural racism will improve model accuracy and help avoid the pitfalls of limited application to racial minority populations, algorithmic bias, limited transportability and reinforcing racial inequities.

Acknowledgments

Conflict of Interest: None declared.

Funding

Whitney R. Robinson is supported by the National Institute of Minority Health and Health Disparities (NIMHD) of the National Institutes of Health (NIH) under award number NIH-R01MD011680.

Audrey Renson is supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) of the National Institutes of Health (NIH) under award number T32HD091058-01.

Ashley I. Naimi is supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) of the National Institutes of Health (NIH) under award number NIH-R01HD093602.

References

Bailey, Z. D., Krieger, N., Agénor, M., Graves, J., Linos, N. and Bassett, M. T. (2017). Structural racism and health inequities in the USA: evidence and interventions. *The Lancet* **389**, 1453–1463.

Bhatnagar, V., Gormley, N., Kazandjian, D., Goldberg, K., McKee, A. E., Blumenthal, G., Farrell, A. T. and Pazdur, R. (2017). FDA analysis of racial demographics in multiple myeloma trials. *Blood* **130**(Suppl 1), 4352.

Braun L. (2015). Race, ethnicity and lung function: a brief history. *Canadian Journal of Respiratory Therapy: CJRT* **51**, 99.

Buolamwini, J. and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler, S. A. and Wilson, C. (editors), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in Proceedings of Machine Learning Research, New York University, NYC, pp. 1–15.

Domingos, P. M. (2012). A few useful things to know about machine learning. *Communications of the ACM* **55**, 7887.

Ford, C. L. and Airhihenbuwa, C. O. (2010). The public health critical race methodology: praxis for antiracism research. *Social Science & Medicine* **71**, 1390–1398.

Foulkes, W. D., Smith, I. E. and Reis-Filho, J. S. (2010). Triple-negative breast cancer. *New England Journal of Medicine* **363**, 1938–1948.

Glass, T. A., Goodman, S. N., Hernán, M. A. and Samet, J. M. (2013). Causal inference in public health. *Annual Review of Public Health* **34**, 1–75.

HERNÁN, M. A., HERNÁNDEZ-DÍAZ, S., ROBINS, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15**, 615–625.

KEIL, A. P., MOONEY, S. J., JONSSON FUNK, M., COLE, S. R., EDWARDS, J. K., WESTREICH, D. (2018). Resolving an apparent paradox in doubly robust estimators. *American Journal of Epidemiology* **187**, 891–892.

KREATSOULAS, C. AND SUBRAMANIAN, S. V. (2018). Machine learning in social epidemiology: learning from experience. *SSM-Population Health* **4**,347.

PHELAN, J. C., LINK, B. G. AND TEHRANIFAR, P. (2010). Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *Journal of Health and Social Behavior* **51**(Suppl 1), S28–S40.

RUDIN, C., WANG, C. AND COKER B. (2018). The age of secrecy and unfairness in recidivism prediction. arXiv preprint arXiv:1811.00731.

SELIGMAN, B., TULJAPURKAR, S. AND REHKOPF, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. *SSM-Population Health* **4**, 95–99.

TEHRANIFAR, P., NEUGUT, A. I., PHELAN, J. C., LINK, B. G., LIAO, Y., DESAI, M. AND TERRY, M. B. (2009). Medical advances and racial/ethnic disparities in cancer survival. *Cancer Epidemiology and Prevention Biomarkers* **18**, 2701–2708.

WEED, D. L. (1998). Beyond black box epidemiology. *American Journal of Public Health* **88**, 12–14.

[*Received September 25, 2019; revised September 25, 2019; accepted for publication September 25, 2019*]

HHS Public Access

Author manuscript

Nat Mach Intell. Author manuscript; available in PMC 2022 May 20.

Published in final edited form as:

Nat Mach Intell. 2019 May ; 1(5): 206–215. doi:10.1038/s42256-019-0048-x.

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Cynthia Rudin

Duke University

Abstract

Black box machine learning models are currently being used for high stakes decision-making throughout society, causing problems throughout healthcare, criminal justice, and in other domains. People have hoped that creating methods for explaining these black box models will alleviate some of these problems, but trying to *explain* black box models, rather than creating models that are *interpretable* in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society. There is a way forward – it is to design models that are inherently interpretable. This manuscript clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare, and computer vision.

1 Introduction

There has been an increasing trend in healthcare and criminal justice to leverage machine learning (ML) for high-stakes prediction applications that deeply impact human lives. Many of the ML models are black boxes that do not explain their predictions in a way that humans can understand. The lack of transparency and accountability of predictive models can have (and has already had) severe consequences; there have been cases of people incorrectly denied parole [1], poor bail decisions leading to the release of dangerous criminals, ML-based pollution models stating that highly polluted air was safe to breathe [2], and generally poor use of limited valuable resources in criminal justice, medicine, energy reliability, finance, and in other domains [3].

Rather than trying to create models that are inherently interpretable, there has been a recent explosion of work on "Explainable ML," where a second (posthoc) model is created to explain the first black box model. This is problematic. Explanations are often not reliable, and can be misleading, as we discuss below. If we instead use models that are inherently interpretable, they provide their own explanations, which are faithful to what the model actually computes.

cynthia@cs.duke.edu .

In what follows, we discuss the problems with Explainable ML, followed by the challenges in Interpretable ML. This document is mainly relevant to high-stakes decision making and troubleshooting models, which are the main two reasons one might require an interpretable or explainable model. Interpretability is a domain-specific notion [4, 5, 6, 7], so there cannot be an all-purpose definition. Usually, however, an interpretable machine learning model is *constrained in model form* so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity [e.g., 8], causality, structural (generative) constraints, additivity [9], or physical constraints that come from domain knowledge. Interpretable models could use case-based reasoning for complex domains. Often for structured data, sparsity is a useful measure of interpretability, since humans can handle at most 7±2 cognitive entities at once [10, 11]. Sparse models allow a view of how variables interact *jointly* rather than individually. We will discuss several forms of interpretable machine learning models for different applications below, but there can never be a single definition; e.g., in some domains, sparsity is useful, and in others is it not. There is a spectrum between fully transparent models (where we understand how all the variables are *jointly* related to each other) and models that are lightly constrained in model form (such as models that are forced to increase as one of the variables increases, or models that, all else being equal, prefer variables that domain experts have identified as important, see [12]).

A preliminary version of this manuscript appeared at a workshop, entitled “Please Stop Explaining Black Box Machine Learning Models for High Stakes Decisions” [13].

2 Key Issues with Explainable ML

A black box model could be either (i) a function that is too complicated for any human to comprehend, or (ii) a function that is proprietary (see Appendix A). Deep learning models, for instance, tend to be black boxes of the first kind because they are highly recursive. As the term is presently used in its most common form, an explanation is a separate model that is supposed to replicate most of the behavior of a black box (e.g., “the black box says that people who have been delinquent on current credit are more likely to default on a new loan”). Note that the term “explanation” here refers to an understanding of how a model works, as opposed to an explanation of how the world works. The terminology “explanation” will be discussed later; it is misleading.

I am concerned that the field of interpretability/explainability/comprehensibility/ transparency in machine learning has strayed away from the needs of real problems. This field dates back to the early 90’s at least [see 4, 14], and there are a huge number of papers on interpretable ML in various fields (that often do not have the word “interpretable” or “explainable” in the title, as the recent papers do). Recent work on explainability of black boxes – rather than interpretability of models – contains and perpetuates critical misconceptions that have generally gone unnoticed, but that can have a lasting negative impact on the widespread use of machine learning models in society. Let us spend some time discussing this before discussing possible solutions.

(i) It is a myth that there is necessarily a trade-off between accuracy and interpretability.

There is a widespread belief that more complex models are more accurate, meaning that a complicated black box is necessary for top predictive performance. However, this is often not true, particularly when the data are structured, with a good representation in terms of naturally meaningful features. When considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing. (Appendix B discusses this further.) In data science problems, where structured data with meaningful features are constructed as part of the data science process, there tends to be little difference between algorithms, assuming that the data scientist follows a standard process for knowledge discovery [such as KDD, CRISP-DM, or BigData, see 15, 16, 17].

Even for applications such as computer vision, where deep learning has major performance gains, and where interpretability is much more difficult to define, some forms of interpretability can be imbued directly into the models without losing accuracy. This will be discussed more later in the Challenges section. Uninterpretable algorithms can still be useful in high-stakes decisions as part of the knowledge discovery process, for instance, to obtain baseline levels of performance, but they are not generally the final goal of knowledge discovery.

Figure 1, taken from the DARPA Explainable Artificial Intelligence program's Broad Agency Announcement [18], exemplifies a blind belief in the myth of the accuracy-interpretability trade-off. This not a "real" figure, in that it was not generated by any data. The axes have no quantification (there is no specific meaning to the horizontal or vertical axes). The image appears to illustrate an experiment with a static dataset, where several machine learning algorithms are applied to the same dataset. However, this kind of smooth accuracy/interpretability/explainability trade-off is atypical in data science applications with meaningful features. Even if one were to quantify the interpretability/explainability axis and aim to show that such a trade-off did exist, it is not clear what algorithms would be applied to produce this figure. (Would one actually claim it is fair to compare the 1984 decision tree algorithm CART to a 2018 deep learning model and conclude that interpretable models are not as accurate?) One can always create an artificial trade-off between accuracy and interpretability/explainability by removing parts of a more complex model to reduce accuracy, but this is not representative of the analysis one would perform on a real problem. It is also not clear why the comparison should be performed on a static dataset, because any formal process for defining knowledge from data [15, 16, 17] would require an iterative process, where one refines the data processing after interpreting the results. Generally, in the practice of data science, the small difference in performance between machine learning algorithms can be overwhelmed by the ability to interpret results and process the data better at the next iteration [19]. In those cases, the accuracy/interpretability tradeoff is reversed – more interpretability leads to better overall accuracy, not worse.

Efforts working within a knowledge discovery process led me to work in interpretable machine learning [20]. Specifically, I participated in a large-scale effort to predict electrical grid failures across New York City. The data were messy, including free text documents

(trouble tickets), accounting data about electrical cables from as far back as the 1890's, inspections data from a brand new manhole inspections program; even the structured data were not easily integrated into a database, and there were confounding issues and other problems. Algorithms on a static dataset were at most 1% different in performance, but the ability to interpret and reprocess the data led to significant improvements in performance, including correcting problems with the dataset, and revealing false assumptions about the data generation process. The most accurate predictors we found were sparse models with meaningful features that were constructed through the iterative process.

The belief that there is always a trade-off between accuracy and interpretability has led many researchers to forgo the *attempt* to produce an interpretable model. This problem is compounded by the fact that researchers are now trained in deep learning, but not in interpretable machine learning. Worse, toolkits of machine learning algorithms offer little in the way of useful interfaces for interpretable machine learning methods.

To our knowledge, all recent review and commentary articles on this topic imply (implicitly or explicitly) that the trade-off between interpretability and accuracy generally occurs. It could be possible that there are application domains where a complete black box is required for a high stakes decision. As of yet, I have not encountered such an application, despite having worked on numerous applications in healthcare and criminal justice [e.g., 21], energy reliability [e.g., 20], and financial risk assessment [e.g., 22].

(ii) Explainable ML methods provide explanations that are not faithful to what the original model computes.

Explanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would be interpretable.) This leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space. [See also for instance, 23, among others.]

An inaccurate (low-fidelity) explanation model limits trust in the explanation, and by extension, trust in the black box that it is trying to explain. An explainable model that has a 90% agreement with the original model indeed explains the original model most of the time. However, an explanation model that is correct 90% of the time is wrong 10% of the time. If a tenth of the explanations are incorrect, one cannot trust the explanations, and thus one cannot trust the original black box. If we cannot know for certain whether our explanation is correct, we cannot know whether to trust either the explanation or the original model.

A more important misconception about explanations stems from the terminology "explanation," which is often used in a misleading way, because explanation models do not always attempt to mimic the calculations made by the original model. Even an explanation model that performs almost identically to a black box model might use completely different features, and is thus not faithful to the computation of the black box. Consider a black box

model for criminal recidivism prediction, where the goal is to predict whether someone will be arrested within a certain time after being released from jail/prison. Most recidivism prediction models depend explicitly on age and criminal history, but do not explicitly depend on race. Since criminal history and age are correlated with race in all of our datasets, a fairly accurate explanation model could construct a rule such as "This person is predicted to be arrested because they are black." This might be an accurate explanation model since it correctly mimics the predictions of the original model, but it would not be faithful to what the original model computes. This is possibly the main flaw identified by criminologists [24] in the ProPublica analysis [25, 26] that accused the proprietary COMPAS recidivism model of being racially biased. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a proprietary model that is used widely in the U.S. Justice system for parole and bail decisions. ProPublica created a linear explanation model for COMPAS that depended on race, and then accused the black box COMPAS model of depending on race, conditioned on age and criminal history. In fact, COMPAS seems to be nonlinear, and it is entirely possible that COMPAS does not depend on race (beyond its correlations with age and criminal history) [27]. ProPublica's linear model was not truly an "explanation" for COMPAS, and they should not have concluded that their explanation model uses the same important features as the black box it was approximating. (There will be a lot more discussion about COMPAS later in this document.)

An easy fix to this problem is to change terminology. Let us stop calling approximations to black box model predictions *explanations*. For a model that does not use race explicitly, an automated explanation "This model predicts you will be arrested because you are black" is not an explanation of what the model is actually doing, and would be confusing to a judge, lawyer or defendant. Recidivism prediction will be discussed more later, as it is a key application where interpretable machine learning is necessary. In any case, it can be much easier to detect and debate possible bias or unfairness with an interpretable model than with a black box. Similarly, it could be easier to detect and avoid data privacy issues with interpretable models than black boxes. Just as in the recidivism example above, many of the methods that claim to produce *explanations* instead compute useful *summary statistics of predictions* made by the original model. Rather than producing explanations that are faithful to the original model, they show trends in how predictions are related to the features. Calling these "summaries of predictions," "summary statistics," or "trends" rather than "explanations" would be less misleading.

(iii) Explanations often do not make sense, or do not provide enough detail to understand what the black box is doing.

Even if both models are correct (the original black box is correct in its prediction and the explanation model is correct in its approximation of the black box's prediction), it is possible that the explanation leaves out so much information that it makes no sense. I will give an example from image processing, for a low-stakes decision (not a high-stakes decision where explanations are needed, but where explanation methods are often demonstrated). Saliency maps are often considered to be explanatory. Saliency maps can be useful to determine what part of the image is being omitted by the classifier, but this leaves out all information about how relevant information *is* being used. Knowing where

the network is looking within the image does not tell the user what it is doing with that part of the image, as illustrated in Figure 2. In fact, the saliency maps for multiple classes could be essentially the same; in that case, the explanation for why the image might contain a Siberian husky would be the same as the explanation for why the image might contain a transverse flute.

An unfortunate trend in recent work is to show explanations only for the observation's *correct* label when demonstrating the method (e.g., Figure 2 would not appear). Demonstrating a method using explanations only for the correct class is misleading. This practice can instill a false sense of confidence in the explanation method and in the black box. Consider, for instance, a case where the explanations for multiple (or all) of the classes are identical. This situation would happen often when saliency maps are the explanations, because they tend to highlight edges, and thus provide similar explanations for each class. These explanations could be identical even if the model is *always* wrong. Then, showing only the explanations for the image's correct class misleads the user into thinking that the explanation is useful, *and* that the black box is useful, even if neither one of them are.

Saliency maps are only one example of explanations that are so incomplete that they might not convey why the black box predicted what it did. Similar arguments can be made with other kinds of explanation methods. Poor explanations can make it very hard to troubleshoot a black box.

(iv) Black box models are often not compatible with situations where information outside the database needs to be combined with a risk assessment.

In high stakes decisions, there are often considerations outside the database that need to be combined with a risk calculation. For instance, what if the circumstances of the crime are much worse than a generic assigned charge? There are often circumstances whose knowledge could either increase or decrease someone's risk. But if the model is a black box, it is very difficult to manually calibrate how much this additional information should raise or lower the estimated risk. This issue arises constantly; for instance, the proprietary COMPAS model used in the U.S. Justice System for recidivism risk prediction does not depend on the seriousness of the current crime [27, 29]. Instead, the judge is instructed to somehow manually combine current crime with COMPAS. Actually, it is possible that many judges do not know this fact. If the model were transparent, the judge could see directly that the seriousness of the current crime is not being considered in the risk assessment.

(v) Black box models with explanations can lead to an overly complicated decision pathway that is ripe for human error.

Typographical errors seem to be common in computing COMPAS, and these typographical errors sometimes determine bail decision outcomes [1, 27]. This exemplifies an important drawback of using overly complicated black box models for recidivism prediction – they may be incorrectly calculated in practice. The computation of COMPAS requires 130+ factors. If typographical errors by humans entering these data into a survey occur at a rate of 1%, then more than 1 out of every 2 surveys on average will have at least one typographical error. The multitude of typographical errors has been argued to be a type of *procedural*

unfairness, whereby two individuals who are identical might be randomly given different parole or bail decisions. These types of errors have the potential to reduce the in-practice accuracy of these complicated models.

On the separate topic of model troubleshooting, an overly complicated black box model may be flawed but we do not know it, because it is difficult to troubleshoot. Having an (incomplete) explanation of it may not help, and now we must troubleshoot two models rather than one (the black box model and the explanation model).

In the next section, we completely switch gears. We will discuss reasons why so many people appear to advocate for black box models with separate explanation models, rather than inherently interpretable models – even for high-stakes decisions.

3 Key Issues with Interpretable ML

There are many cases where black boxes with explanations are preferred over interpretable models, even for high-stakes decisions. However, for most applications, I am hopeful that there are ways around some of these problems, whether they are computational problems, or problems with training of researchers and availability of code. The first problem, however, is currently a major obstacle that I see no way of avoiding other than through policy, as discussed in the next section.

(i) Corporations can make profits from the intellectual property afforded to a black box.

Companies that charge for individual predictions could find their profits obliterated if an interpretable model were used instead.

Consider the COMPAS proprietary recidivism risk prediction tool discussed above that is in widespread use in the U.S. Justice System for predicting the probability that someone will be arrested after their release [29].

The COMPAS model is equally accurate for recidivism prediction as the very simple three rule interpretable machine learning model involving only age and number of past crimes shown in Figure 3 below. However, there is no clear business model that would suggest profiting from the simple transparent model. The simple model in Figure 3 was created from an algorithm called Certifiably Optimal Rule Lists (CORELS) that looks for if-then patterns in data. Even though the model in Figure 3 looks like a rule of thumb that a human may have designed without data, it is instead a full-blown machine learning model. A qualitative comparison of the COMPAS and CORELS models is in Table 1. Standard machine learning tools and interpretable machine learning tools seem to be approximately equally accurate for predicting recidivism, even if we define recidivism in many different ways, for many different crime types [30, 31]. This evidence, however, has not changed the momentum of the justice system towards proprietary models. As of this writing, California has recently eliminated its cash bail system, instead enforcing that decisions be made by algorithms; it is unclear whether COMPAS will be the algorithm used for this, despite the fact that it is not known to be any more accurate than other models, such as the simple CORELS model in Figure 3.

COMPAS is not a machine learning model – it was not created by any standard machine learning algorithm. It was designed by experts based on carefully designed surveys and expertise, and it does not seem to depend heavily on past criminal history [27]. Interestingly, if the COMPAS model were not proprietary, its documentation [29] indicates that it would actually be an interpretable predictive model. (It is a black box of the second type – proprietary – but not the first type – complicated – discussed above.) Revealing this model, however, would be revealing a trade secret.

Let us switch examples to consider the proprietary machine learning model by BreezoMeter, used by Google during the California wildfires of 2018, which predicted air quality as "good – ideal air quality for outdoor activities," when air quality was dangerously bad according to multiple other models [2], and people reported their cars covered in ash. The Environmental Protection Agency's free, vigorously-tested air quality index would have provided a reliable result [33]. How could BreezoMeter's machine learning method be so badly wrong and put so many in danger? We will never find out, but BreezoMeter, who has probably made a profit from making these predictions, may not have developed this new technology if its models were forced to be transparent.

In medicine, there is a trend towards blind acceptance of black box models, which will open the door for companies to sell more models to hospitals. For instance, radiology and in-hospital patient monitoring are areas of medicine that stand to gain tremendously by automation; humans cannot process data fast enough or rapidly enough to compete with machines. However, in trusting these automated systems, we must also trust the full database on which they were trained, the processing of the data, along with the completeness of the database. If the database does not represent the full set of possible situations that can arise, then the model could be making predictions in cases that are very different from anything it was trained on. An example of where this can go wrong is given by Zech et al. [34], who noticed that their neural network was picking up on the word "portable" within an x-ray image, representing the type of x-ray equipment rather than the medical content of the image. If they had used an interpretable model, or even an explainable model, this issue would never have gone unnoticed. Zech et al. [34] pointed out the issue of confounding generally; in fact, the plague of confounding haunts a vast number of datasets, and particularly medical datasets. This means that proprietary models for medicine can have serious errors. These models can also be fragile, in that if the model is used in practice in a slightly different setting than how it was trained (e.g., new x-ray equipment), accuracy can substantially drop.

The examples of COMPAS, Breezometer, and black box medical diagnosis all illustrate a problem with the business model for machine learning. In particular, there is a conflict of responsibility in the use of black box models for high-stakes decisions: *the companies that profit from these models are not necessarily responsible for the quality of individual predictions*. A prisoner serving an excessively long sentence due to a mistake entered in an overly-complicated risk score could suffer for years, whereas the company that constructed this complicated model is unaffected. On the contrary, the fact that the model was complicated and proprietary allowed the company to profit from it. In that sense, the model's designers are not incentivized to be careful in its design, performance, and ease

of use. These are some of the same types of problems affecting the credit rating agencies who priced mortgages in 2008; that is, these are the same problems that contributed to the financial crisis in the United States at that time.

One argument favoring black boxes is that keeping these models hidden prevents them from being gamed or reverse-engineered. It is not clear that this argument generally makes sense. In fact, the reason a system may be gamed is because it most likely was not designed properly in the first place, leading to a form of Goodhart's law if it were revealed. Quoting from Chang et al. [35] about product rating systems: "If the ratings are accurate measures of quality, then making the ratings more transparent could have a uniformly positive impact: it would help companies to make better rated products, it would help consumers to have these higher quality products, and it would encourage rating companies to receive feedback as to whether their rating systems fairly represent quality." Thus, transparency could help improve the quality of the system, whereby attempting to game it would genuinely align with the overall goal of improvement. For instance, improving one's credit score should actually correspond to an improvement in creditworthiness.

Another argument favoring black boxes is the belief that "counterfactual explanations" of black boxes are sufficient. A counterfactual explanation describes a minimal change to the input that would result in the opposite prediction. For instance, a possible counterfactual explanation might be "your loan application was denied, but if you had $1000 less debt, you would have qualified for the loan." This type of explanation can suffer from key issue (iv) discussed above, about combining information outside the database with the black box. In particular, the "minimal" change to the input might be different for different individuals. Appendix C discusses in more depth why counterfactual explanations generally do not suffice for high stakes decisions of black boxes.

(ii) Interpretable models can entail significant effort to construct, in terms of both computation and domain expertise.

As discussed above, interpretability usually translates in practice to a set of application-specific constraints on the model. Solving constrained problems is generally harder than solving unconstrained problems. Domain expertise is needed to construct the definition of interpretability for the domain, and the features for machine learning. For data that are unconfounded, complete, and clean, it is much easier to use a black box machine learning method than to troubleshoot and solve computationally hard problems. However, for high-stakes decisions, analyst time and computational time are less expensive than the cost of having a flawed or overly complicated model. That is, it is worthwhile to devote extra effort and cost into constructing a high-quality model. But even so, many organizations do not have analysts who have the training or expertise to construct interpretable models at all.

Some companies have started to provide interpretable ML solutions using proprietary software. While this is a step in the right direction, it is not clear that the proprietary software is better than publicly available software. For instance, claims made by some companies about performance of their proprietary algorithms are not impressive (e.g., Interpretable AI, whose decision tree performance using mixed integer programming

software in 2017 is reported to be often beaten by or comparable to the 1984 Classification and Regression Tree algorithm, CART).

As discussed earlier, interpretability constraints (like sparsity) lead to optimization problems that have been proven to be computationally hard in the worst case. The theoretical hardness of these problems does not mean we cannot solve them, though in real cases, these optimization problems are often difficult to solve. Major improvements have been made in the last decade, and some are discussed later in the Challenges section. Explanation methods, on the other hand, are usually based on derivatives, which lead to easier gradient-based optimization.

(iii) Black box models seem to uncover "hidden patterns."

The fact that many scientists have difficulty constructing interpretable models may be fueling the belief that black boxes have the ability to uncover subtle hidden patterns in the data that the user was not previously aware of. A transparent model may be able to uncover these same patterns. If the pattern in the data was important enough that a black box model could leverage it to obtain better predictions, an interpretable model might also locate the same pattern and use it. Again, this depends on the machine learning researcher's ability to create accurate-yet-interpretable models. The researcher needs to create a model that has the capability of uncovering the types of patterns that the user would find interpretable, but also the model needs to be flexible enough to fit the data accurately. This, and the optimization challenges discussed above, are where the difficulty lies with constructing interpretable models.

4 Encouraging Responsible ML Governance

Currently the European Union's revolutionary General Data Protection Regulation and other AI regulation plans govern "right to an explanation," where only an explanation is required, not an interpretable model [36], in particular "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" (Article 22 of GDPR regulations from http://www.privacy-regulation.eu/en/22.htm). If one were to provide an explanation for an automated decision, it is not clear whether the explanation is required to be accurate, complete, or faithful to the underlying model [e.g., see 37]. Less-than-satisfactory explanations can easily undermine these new policies.

Let us consider a possible mandate that, for certain high-stakes decisions, *no black box should be deployed when there exists an interpretable model with the same level of performance*. If such a mandate were deployed, organizations that produce and sell black box models could then be held accountable if an equally accurate transparent model exists. It could be considered a form of false advertising to sell a black box model if there is an equally-accurate interpretable model. The onus would then fall on organizations to produce black box models only when no transparent model exists for the same task.

This possible mandate could produce a change in the business model for machine learning. Opacity is viewed as essential in protecting intellectual property, but it is at odds with

the requirements of many domains that involve public health or welfare. However, the combination of opacity and explainability is not the only way to incentivize machine learning experts to invest in creating such systems. Compensation for developing an interpretable model could be provided in a lump sum, and the model could be released to the public. The creator of the model would not be able to profit from licensing the model over a period of time, but the fact that the models are useful for public good applications would make these problems appeal to academics and charitable foundations.

This proposal will not solve all problems, but it could at least rule out companies selling recidivism prediction models, possibly credit scoring models, and other kinds of models where we can construct accurate-yet-interpretable alternatives. If applied too broadly, it could reduce industrial participation in cases where machine learning might benefit society.

Consider a second proposal, which is weaker than the one provided above, but which might have a similar effect. Let us consider the possibility that organizations that introduce black box models would be mandated to report the accuracy of interpretable modeling methods. In that case, one could more easily determine whether the accuracy/interpretability trade-off claimed by the organization is worthwhile. This also forces the organization to try using interpretable modeling methods. It also encourages the organization to use these methods carefully, otherwise risking the possibility of criticism.

As mentioned earlier, I have not yet found a high-stakes application where a fully black box model is necessary, despite having worked on many applications. As long as we continue to allow for a broad definition of interpretability that is adapted to the domain, we should be able to improve decision making for serious tasks of societal importance. However, in order for people to design interpretable models, the technology must exist to do so. As discussed earlier, there is a formidable computational hurdle in designing interpretable models, even for standard structured data with already-meaningful features.

5 Algorithmic Challenges in Interpretable ML

What if every black box machine learning model could be replaced with one that was equally accurate but also interpretable? If we could do this, we would identify flaws in our models and data that we could not see before. Perhaps we could prevent some of the poor decisions in criminal justice and medicine that are caused by problems with using black box models. We could also eliminate the need for explanations that are misleading and often wrong.

Since interpretability is domain-specific, a large toolbox of possible techniques can come in handy. Below we expand on three of the challenges for interpretable machine learning that appear often. All three cases have something in common: people have been providing interpretable predictive models for these problems for decades, and the human-designed models look just like the type of model we want to create with machine learning. I also discuss some of our current work on these well-known problems.

Each of these challenges is a representative from a major class of models: modeling that uses logical conditions (Challenge 1), linear modeling (Challenge 2), and case-based reasoning

(Challenge 3). By no means is this set of challenges close to encompassing the large number of domain-specific challenges that exist in creating interpretable models.

Challenge #1: Constructing optimal logical models

A logical model consists of statements involving "or," "and," "if-then," etc. The CORELS model in Figure 3 is a logical model, called a *rule list*. Decision trees are logical models, as well as conjunctions of disjunctions ("or's" of "and's" – for instance, IF condition A is true OR conditions B AND C are true, THEN predict yes, otherwise predict no).

Logical models have been crafted by hand as *expert systems* as far back as the 1970's. Since then, there have been many heuristics for creating logical models; for instance, one might add logical conditions one by one (greedily), and then prune conditions away that are not helpful (again, greedily). These heuristic methods tend to be inaccurate and/or uninterpretable because they do not choose a globally best choice (or approximately best choice) for the logical conditions, and are not designed to be optimally sparse. They might use 200 logical conditions when the same accuracy could be obtained with 5 logical conditions. [C4.5 and CART 38, 39, decision trees suffer from these problems, as well as a vast number of models from the associative classification literature]. An issue with algorithms that do not aim for optimal (or near-optimal) solutions to optimization problems is that it becomes difficult to tell whether poor performance is due to the choice of algorithm or the combination of the choice of model class and constraints. (Did the algorithm perform poorly because it did not optimize its objective, or because we chose constraints that do not allow enough flexibility in the model to fit the data well?) The question of computing optimal logical models has existed since at least the mid 1990's [40].

We would like models that look like they are created by hand, but they need to be accurate, full-blown machine learning models. To this end, let us consider the following optimization problem, which asks us to find a model that minimizes a combination of the fraction of misclassified training points and the size of the model. Training observations are indexed from $i = 1,..,n$, and $\mathcal{F}$ is a family of logical models such as decision trees. The optimization problem is:

$$\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[\text{training observation } i \text{ is misclassified by } f]} + \lambda \times \text{size}(f) \right). \quad (1)$$

Here, the size of the model can be measured by the number of logical conditions in the model, such as the number of leaves in a decision tree. The parameter λ is the classification error one would sacrifice in order to have one fewer term in the model; if λ is 0.01, it means we would sacrifice 1% training accuracy in order to reduce the size of the model by one. Another way to say this is that the model would contain an additional term only if this additional term reduced the error by at least 1%.

The optimization problem in (1) is generally known to be computationally hard. Versions of this optimization problem are some of the fundamental problems of artificial intelligence. The challenge is whether we can solve (or approximately solve) problems like this in practical ways, by leveraging new theoretical techniques and advances in hardware.

The model in Figure 3 is a machine learning model that comes from the CORELS algorithm [32]. CORELS solves a special case of (1), for the special choice of $\mathcal{F}$ as the set of rule lists, and where the size of the model is measured by the number of rules in the list. Figure 3 has three "if-then" rules so its size is 3. In order to minimize (1), CORELS needs to avoid enumerating all possible models, because this would take an extremely long time (perhaps until the end of the universe on a modern laptop for a fairly small dataset). The technology underlying the CORELS algorithm was able to solve the optimization problem to optimality in under a minute for the Broward County, FL, dataset discussed above. CORELS' backbone is: (i) a set of theorems allowing massive reductions in the search space of rule lists, (ii) a custom fast bit-vector library that allows fast exploration of the search space, so that CORELS does not need to enumerate all rule lists, and (iii) specialized data structures that keep track of intermediate computations and symmetries. This set of ingredients proved to be a powerful cocktail for handling these tough computational problems.

The example of CORELS enforces two points discussed above, which are, first, that interpretable models sometimes entail hard computational problems, and second, that these computational problems can be solved by leveraging a combination of theoretical and systems-level techniques. CORELS creates one type of logical model; however, there are many more. Formally, *the first challenge is to create algorithms that solve logical modeling problems in a reasonable amount of time, for practical datasets*.

We have been extending CORELS to more complex problems, such as Falling Rule Lists [41, 42], and optimal binary-split decision trees, but there is much work to be done on other types of logical models, with various kinds of constraints.

Note that it is possible to construct interpretable logical models for which the global model is large, and yet each explanation is small. This is discussed in Appendix D.

Challenge #2: Construct optimal sparse scoring systems

Scoring systems have been designed by hand since at least the Burgess criminological model of 1928 [43]. The Burgess model was designed to predict whether a criminal would violate bail, where individuals received points for being a "ne'er do well" or a "recent immigrant" that increased their predicted probability of parole violation. (Of course, this model was not created using machine learning, which had not been invented yet.) A scoring system is a sparse linear model with integer coefficients – the coefficients are the point scores. An example of a scoring system for criminal recidivism is shown in Figure 4, which predicts whether someone will be arrested within 3 years of release. Scoring systems are used pervasively throughout medicine; there are hundreds of scoring systems developed by physicians. Again, the challenge is whether scoring systems – which look like they could have been produced by a human in the absence of data – can be produced by a machine learning algorithm, and be as accurate as any other model from any other machine learning algorithm.

There are several ways to formulate the problem of producing a scoring system [see, e.g., 46, 47]. For instance, we could use a special case of (1), where the model size is the number of terms in the model. (Figure 4 is a machine learning model with 5 terms.) Sometimes, one

can round the coefficients of a logistic regression model to produce a scoring system, but that method does not tend to give accurate models, and does not tend to produce models that have particularly nice coefficients (such as 1 and −1 used in Figure 4). However, solving (1) or its variants is computationally hard, because the domain over which we solve the optimization problem is the integer lattice. (To see this, consider an axis for each of $\{b_1, b_2, \ldots, b_p\}$, where each b_j can take on integer values. This is a lattice that defines the feasible region of the optimization problem.)

The model in Figure 4 arose from the solution to a very hard optimization problem. Let us discuss this optimization problem briefly. The goal is to find the coefficients $b_j, j = 1 \ldots p$ for the linear predictive model $f(\mathbf{z}) = \sum_j b_j z_j$ where z_j is the jth covariate of a test observation $\mathbf{z}$. In Figure 4, the b_j's are the point scores, which turned out to be 1, −1, and 0 as a result of optimization, where only the nonzero coefficients are displayed in the figure. In particular, we want to solve:

$$\min_{b_1, b_2, .., b_p \in \{-10, -9, \ldots, 9, 10\}} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-\sum_{j=1}^{p} b_j x_{i,j}\right)\right) + \lambda \sum_j 1[b_j \neq 0],$$

where the point scores b_j are constrained to be integers between −10 and 10, the training observations are indexed by $i = 1, \ldots, n$, and p is the total number of covariates for our data. Here the model size is the number of non-zero coefficients, and again λ is the trade-off parameter. The first term is the logistic loss used in logistic regression. The problem is hard, specifically it is a mixed-integer-nonlinear program (MINLP) whose domain is the integer lattice.

Despite the hardness of this problem, new cutting plane algorithms have been able to solve this problem to optimality (or near-optimality) for arbitrarily large sample sizes and a moderate number of variables within a few minutes. The latest attempt at solving this problem is the RiskSLIM (Risk-Supersparse-Linear-Integer-Models) algorithm, which is a specialized cutting plane method that adds cutting planes only whenever the solution to a linear program is integer-valued, and otherwise performs branching [44].

This optimization problem is similar to what physicians attempt to solve manually, but without writing the optimization problem down like we did above. Because physicians do not use optimization tools to do this, accurate scoring systems tend to be difficult for physicians to create themselves from data. One of our collaborators spent months trying to construct a scoring system himself by adding and removing variables, rounding, and using other heuristics to decide which variables to add, remove, and round. RiskSLIM was useful for helping him with this task [48]. Formally, *the second challenge is to create algorithms for scoring systems that are computationally efficient*. Ideally we would increase the size of the optimal scoring system problems that current methods can practically handle by an order of magnitude.

Challenge #3 Define interpretability for specific domains and create methods accordingly, including computer vision

Since interpretability needs to be defined in a domain-specific way, some of the most important technical challenges for the future are tied to specific important domains. Let us start with computer vision, for classification of images. There is a vast and growing body of research on posthoc explainability of deep neural networks, but not as much work in designing *interpretable neural networks*. My goal in this section is to demonstrate that even for classic domains of machine learning, where latent representations of data need to be constructed, there could exist interpretable models that are as accurate as black box models.

For computer vision in particular, there is not a clear definition of interpretability, and the sparsity-related models discussed above do not apply – sparsity in pixel space does not make sense. There can be many different ideas of what constitutes interpretability, even between different computer vision applications. However, if we can define interpretability somehow for our particular application, we can embed this definition into our algorithm.

Let us define what constitutes interpretability by considering *how people explain to each other* the reasoning processes behind complicated visual classification tasks. As it turns out, for classification of natural images, domain experts often direct our attention to different parts of the image and explain why these parts of the image were important in their reasoning process. The question is whether we can construct network architectures for deep learning that can also do this. The network must then make decisions by reasoning about parts of the image so that the explanations are real, and not posthoc.

In a recent attempt to do this, Chen, Li, and colleagues have been building architectures that append a special prototype layer to the end of the network [49, 55]. During training, the prototype layer finds parts of training images that act as prototypes for each class. For instance, for bird classification, the prototype layer might pick out a prototypical head of a blue jay, prototypical feathers of a blue jay, etc. The network also learns a similarity metric between parts of images. Thus, during testing, when a new test image needs to be evaluated, the network finds parts of the test image that are similar to the prototypes it learned during training, as shown in Figure 5. The final class prediction of the network is based on the weighted sum of similarities to the prototypes; this is the sum of evidence throughout the image for a particular class. The explanations given by the network are the prototypes (and the weighted similarities to them). These explanations are the actual computations of the model, and these are not posthoc explanations. The network is called “This look like that” because its reasoning process considers whether “this” part of the image looks like “that” prototype.

Training this prototype network is not as easy as training an ordinary neural network; the tricks that have been developed for regular deep learning have not yet been developed for the prototype network. However, so far these prototype networks have been trained to be approximately as accurate as the original black box deep neural networks they were derived from, before the prototype layer was added.

Discussion on Interpretability for Specific Domains

Let us finish this short discussion on challenges to interpretability for specific domains by mentioning that there are vast numbers of papers that have imbued interpretability in their methodology. Interpretability is not mentioned in the title of these papers, and often not in the body of the text. This is why it is almost impossible to create a review article on interpretability in machine learning or statistics without missing the overwhelming majority of it.

It is not clear why review articles for interpretability and explainability make sense to create. We do not normally have reviews of performance/accuracy measures, despite the fact that there are many of them – accuracy, area under the ROC curve, partial AUC, sensitivity, specificity, discounted cumulative gain, F-score, G-means, and many other domain-specific measures. Interpretability/explainability is just as domain-specific as accuracy performance, so it is not clear why reviews of interpretability make any more sense than reviews of accuracy/performance. I have yet to find even a single recent review that recognized the chasm between interpretability and explainability.

Let us discuss very briefly some of examples of work on interpretability that would not have been covered by recent review articles, and yet are valuable contributions to interpretability in their respective domains. Gallagher et al. [56] analyze brain-wide electrical spatiotemporal dynamics to understand depression vulnerability and find interpretable patterns in a low dimensional space. Dimension reduction to interpretable dimensions is an important theme in interpretable machine learning. Problems residing in *applied statistics* are often interpretable because they embed the physics of the domain; e.g., Wang et al. [57] create models for recovery curves for prostatectomy patients whose signal and uncertainty obey specific constraints in order to be realistic. Constraints on the uncertainty of the predictions make these models interpretable.

The setup of the recent 2018 FICO Explainable ML Challenge exemplified the blind belief in the myth of the accuracy/interpretability tradeoff for a specific domain, namely credit scoring. Entrants were instructed to create a black box to predict credit default and explain the model afterwards. *However, there was no performance difference between interpre table models and explainable models for the FICO data*. A globally interpretable model [22] won the FICO Recognition Prize for the competition. This is a case where the organizers and judges had not expected an interpretable model to be able to be constructed and thus did not ask entrants to try to construct such a model. The model of [22] was an additive model, which is a known form of interpretable model [see also 9, 58, where additive models are used for medical data]. Additive models could be optimized using similar techniques to those introduced in Challenge 2 above.

A Technical Reason Why Accurate Interpretable Models Might Exist in Many Domains

Why is it that accurate interpretable models could possibly exist in so many different domains? Is it really possible that many aspects of nature have simple truths that are waiting to be discovered by machine learning? Although that would be intriguing, I will not make this kind of Occham's-Razor-style argument, in favor of a technical argument about function

classes, and in particular, Rashomon Sets. The argument below is fleshed out more formally in [59]. This is related to (but different from) the notation of "flat minima," for which a nice example is given by Hand [19].

Here is the *Rashomon set* argument: Consider that the data permit a large set of reasonably accurate predictive models to exist. Because this set of accurate models is large, it often contains at least one model that is interpretable. This model is thus both interpretable and accurate.

Unpacking this argument slightly, for a given data set, we define the *Rashomon set* as the set of reasonably accurate predictive models (say within a given accuracy from the best model accuracy of boosted decision trees). Because the data are finite, the data could admit many close-to-optimal models that predict differently from each other: a large Rashomon set. I suspect this happens often in practice because sometimes many different machine learning algorithms perform similarly on the same dataset, despite having different functional forms (e.g., random forests, neural networks, support vector machines). As long as the Rashomon set contains a large enough set of models with diverse predictions, it probably contains functions that can be approximated well by simpler functions, and so the Rashomon set can also contain these simpler functions. Said another way, uncertainty arising from the data leads to a Rashomon set, a larger Rashomon set probably contains interpretable models, thus interpretable accurate models often exist.

If this theory holds, we should expect to see interpretable models exist across domains. These interpretable models may be hard to find through optimization, but at least there is a reason we might expect that such models exist.

If there are many diverse yet good models, it means that algorithms may not be *stable*; an algorithm might choose one model, and a small change to that algorithm or to the dataset may yield a completely different (but still accurate) model. This is not necessarily a bad thing, in fact, the availability of diverse good models means that domain experts may have more flexibility in choosing a model that they find interpretable. Appendix E discusses this in slightly more detail.

6 Conclusion

If this commentary can shift the focus even slightly from the basic assumption underlying most work in Explainable ML – which is that a black box is necessary for accurate predictions – we will have considered this document a success.

If this document can encourage policy makers not to accept black box models without significant attempts at interpretable (rather than explainable) models, that would be even better.

If we can make people aware of the current challenges right now in interpretable machine learning, it will allow policy-makers the mechanism to demand that more effort should be made in ensuring safety and trust in our machine learning models for high-stakes decisions.

If we do not succeed at these efforts, it is possible that black box models will continue to be permitted when it is not safe to use them. Since the definition of what constitutes a viable explanation is unclear, even strong regulations such as “right to explanation” can be undermined with less-than-satisfactory explanations. Further, there will continue to be problems combining black box model predictions with information outside the database, and continued miscalculations of black box model inputs. This may continue to lead to poor decisions throughout our criminal justice system, incorrect safety guidance for air quality disasters, incomprehensible loan decisions, and other widespread societal problems.

Acknowledgments

I would like to thank Fulton Wang, Tong Wang, Chaofan Chen, Oscar Li, Alina Barnett, Tom Dietterich, Margo Seltzer, Elaine Angelino, Nicholas Larus-Stone, Elizabeth Mannshart, Maya Gupta, and several others who helped my thought processes in various ways, and particularly Berk Ustun, Ron Parr, Rob Holte, and my father, Stephen Rudin, who went to considerable efforts to provide thoughtful comments and discussion. I would also like to thank two anonymous reviewers for their suggestions that improved the manuscript. I would like to acknowledge funding from the Laura and John Arnold Foundation, NIH, NSF, DARPA, the Lord Foundation of North Carolina, and MIT-Lincoln Laboratory.

A: On the Two Types of Black Box

Black box models of the first type are too complicated for a human to comprehend, and black box models of the second type are proprietary. Some models are of both types. The consequences of these two types of black box are different, but related. For instance, for a black box model that is complicated but not proprietary, we at least know what variables it uses. We also know the model form and could use that to attempt to analyze the different parts of the model. For a black box model that is proprietary but not complicated [we have evidence that COMPAS is such a model, 27], we may not even have access to query it in order to study it. If a proprietary model is too sparse, there is a risk that it could be easily reverse-engineered, thus there is an incentive to make proprietary models complicated in order to preserve their secrecy.

B: Performance Comparisons

For most problems with meaningful structured covariates, machine learning algorithms tend to perform similarly, with no algorithm clearly dominating the others. The variation due to tuning parameters of a single algorithm can often be higher than the variation between algorithms. This lack of single dominating algorithm for structured data is arguably why the field of machine learning focuses on image and speech recognition, whose data are represented in raw features (pixels, sound files); these are fields for which the choice of algorithm impacts performance. Even for complex domains such as medical records, it has been reported in some studies that logistic regression has identical performance to deep neural networks [e.g. 60].

If there is no dominating algorithm, the Rashomon Set argument discussed above would suggest that interpretable models might perform well.

Unfortunately the culture of publication within machine learning favors selective reporting of algorithms on selectively chosen datasets. Papers are often rejected if small or no

performance gains are reported between algorithms. This encourages omission of accurate baselines for comparison, as well as omission of datasets on which the method does not perform well, and encourages authors to poorly tune the parameters of baseline methods, or equivalently, place more effort into tuning the parameters of the author's own method. This creates an illusion of large performance differences between algorithms, even when such performance differences do not truly exist.

C: Counterfactual Explanations

Some have argued that counterfactual explanations [e.g., see 37] are a way for black boxes to provide useful information while preserving secrecy of the global model. Counterfactual explanations, also called inverse classification, state a change in features that is sufficient (but not necessary) for the prediction to switch to another class (e.g., "If you reduced your debt by $5000 and increased your savings by $50% then you would have qualified for the loan you applied for"). This is important for recourse in certain types of decisions, meaning that the user could take an action to reverse a decision [61].

There are several problems with the argument that counterfactual explanations are sufficient. For loan applications, for instance, we would want the counterfactual explanation to provide the *lowest cost* action for the user to take, *according to the user's own cost metric*. [See 35, for an example of lowest-cost counterfactual reasoning in product rankings]. In other words, let us say that there is more than one counterfactual explanation available (e.g., the first explanation is "If you reduced your debt by $5000 and increased your savings by $50% then you would have qualified for the loan you applied for" and the second explanation is "If you had gotten a job that pays $500 more per week, then you would have qualified for the loan"). In that case, the explanation shown to the user should be the easiest one for the user to actually accomplish. However, it is unclear in advance which explanation would be easier for the user to accomplish. In the credit example, perhaps it is easier for the user to save money rather than get a job or vice versa. In order to determine which explanation is the lowest cost for the user, we would need to elicit cost information for the user, and that cost information is generally very difficult to obtain; worse, the cost information could actually change as the user attempts to follow the policy provided by the counterfactual explanation (e.g., it turns out to be harder than the user thought to get a salary increase). For that reason it is unclear that counterfactual explanations would suffice for high stakes decisions. Additionally, counterfactual explanations of black boxes have many of the other pitfalls discussed throughout this paper.

D: Interpretable Models that Provide Smaller-Than-Global Explanations

It is possible to create a global model (perhaps a complicated one) for which explanations for any given individual are very sparse. In other words, even if the global model would take several pages of text to write, the prediction for a given individual can be very simple to calculate (perhaps requiring only 1–2 conditions). Let us consider the case of credit risk prediction. Assume we do not need to justify to the client why we would grant a loan, but we would need to justify why we would deny a loan.

Let us consider a disjunctive normal form model, which is a collection of "or's" of "and's." For instance, the model might deny a loan if "(credit history too short AND at least one bad past trade) OR (at least 4 bad past trades) OR (at least one recent delinquency AND high percentage of delinquent trades)." Even if we had hundreds of conjunctions within the model, only one of these needs to be shown to the client; if any conjunction is true, that conjunction is a defining reason why the client would be denied a loan. In other words, if the client had "at least one recent delinquency AND high percentage of delinquent trades," then regardless of any other aspects of her credit history, she could be shown that simple explanation, and it would be a defining reason why her loan application would be denied.

Disjunctive normal form models are well-studied, and are called by various names, such as "or's of and's," as well as "decision rules," "rule sets" and "associative classifiers." There has been substantial work in being able to generate such models over the past few years so that the models are globally interpretable, not just locally interpretable (meaning that the global model consists of a small number of conjunctions) [e.g., see 62, 63, 64, 65, 66].

There are many other types of models that would provide smaller-than-global explanations. For instance, falling rule lists [41, 42] provide shorter explanations for the decisions that are most important. For instance, a falling rule list for predicting patient mortality would use few logical conditions to categorize whether a patient is in a high-risk group, but use several additional logical conditions to determine which low-risk group a patient falls into.

E: Algorithm Stability

A common criticism of decision trees is that they are not stable, meaning that small changes in the training data lead to completely different trees, giving no guidance as to which tree to choose. In fact, the same problem can happen in *linear* models when there are highly correlated features. This can happen even in basic least squares, where correlations between features can lead to very different models having precisely the same levels of performance. When there are correlated features, the lack of stability happens with most algorithms that are not strongly regularized.

I hypothesize this instability in the learning algorithm could be a side-effect of the Rashomon effect mentioned earlier – that there are many different almost-equally good predictive models. Adding regularization to an algorithm increases stability, but also limits flexibility of the user to choose which element of the Rashomon set would be more desirable.

For applications where the models are purely predictive and not causal (e.g., in criminal recidivism where we use age and prior criminal history to predict future crime), there is no assumption that the model represents how outcomes are actually generated. The importance of the variables in the model does not reflect a causal relationship between the variables and the outcomes. Thus, without additional guidance from the domain expert, there is no way to proceed further to choose a single "best model" among the set of models that perform similarly. As discussed above, regularization can act as this additional input.

I view the lack of algorithmic stability as an advantage rather than a disadvantage. If the lack of stability is indeed caused by a large Rashomon effect, it means that domain experts can add more constraints to the model to customize it without losing accuracy.

In other words, while many people criticize methods such as decision trees for not being stable, I view that as a strength of interpretability for decision trees. If there are many equally accurate trees, the domain expert can pick the one that is the most interpretable.

Note that not all researchers working in interpretability agree with this general sentiment about the advantages of instability [67].

References

[1]. Wexler R When a Computer Program Keeps You in Jail: How Computers are Harming Criminal Justice. New York Times. 2017 June 13;.

[2]. McGough M How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say. Sacramento Bee. 2018 August 7;.

[3]. Varshney KR, Alemzadeh H. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. Big Data. 2016 10;5.

[4]. Freitas AA. Comprehensible classification models: a position paper. ACM SIGKDD Explorations Newsletter. 2014 Mar;15(1):1–10.

[5]. Kodratoff Y. The comprehensibility manifesto. KDD Nugget Newsletter. 1994;94(9).

[6]. Huysmans J, Dejaeger K, Mues C, Vanthienen J, Baesens B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems. 2011;51(1):141–154.

[7]. Rüping S Learning Interpretable Models. Universität Dortmund; 2006.

[8]. Gupta M, Cotter A, Pfeifer J, Voevodski K, Canini K, Mangylov A, et al. Monotonic calibrated interpolated look-up tables. Journal of Machine Learning Research. 2016;17(109):1–47.

[9]. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate Intelligible Models with Pairwise Interactions. In: Proceedings of 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM; 2013..

[10]. Miller G The magical number seven, plus or minus two: Some limits on our capacity for processing information. The Psychological Review. 1956;63:81–97. [PubMed: 13310704]

[11]. Cowan N The magical mystery four how is working memory capacity limited, and why? Current directions in psychological science. 2010;19(1):51–57. [PubMed: 20445769]

[12]. Wang J, Oh J, Wang H, Wiens J. Learning Credible Models. In: Proceedings of 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM; 2018. p. 2417–2426.

[13]. Rudin C Please Stop Explaining Black Box Models for High Stakes Decisions. In: Proceedings of NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning; 2018..

[14]. Holte RC. Very simple classification rules perform well on most commonly used datasets. Machine Learning. 1993;11(1):63–91.

[15]. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996;17:37–54.

[16]. Chapman P, et al. CRISP-DM 1.0 - Step-by-step data mining guide. SPSS; 2000.

[17]. Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, et al. Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association; 2012. Available from: http://cra.org/ccc/resources/ccc-led-whitepapers/.

[18]. Defense Advanced Research Projects Agency. Broad Agency Announcement, Explainable Artificial Intelligence (XAI), DARPA-BAA-16-53; 2016. Published August 10. Available from https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

[19]. Hand D Classifier Technology and the Illusion of Progress. Statist Sci. 2006;21(1):1–14.

[20]. Rudin C, Passonneau R, Radeva A, Dutta H, Ierome S, Isaac D. A Process for Predicting Manhole Events In Manhattan. Machine Learning. 2010;80:1–31.

[21]. Rudin C, Ustun B. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. Interfaces. 2018;48:399–486. Special Issue: 2017 Daniel H. Wagner Prize for Excellence in Operations Research Practice September-October 2018.

[22]. Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T. An Interpretable Model with Globally Consistent Explanations for Credit Risk. In: Proceedings of NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy; 2018..

[23]. Mittelstadt B, Russell C, Wachter S. Explaining Explanations in AI. In: In Proceedings of Fairness, Accountability, and Transparency (FAT*); 2019..

[24]. Flores AW, Lowenkamp CT, Bechtel K. False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals”. Federal probation. 2016 September;80(2):38–46.

[25]. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. ProPublica; 2016. Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[26]. Larson J, Mattu S, Kirchner L, Angwin J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica; 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[27]. Rudin C, Wang C, Coker B. The age of secrecy and unfairness in recidivism prediction. arXiv e-prints 1811 00731 [applied statistics]. 2018 Nov;.

[28]. Checkermallow. Canis lupus winstonii (Siberian Husky); 2016.
Public domain image. https://www.flickr.com/photos/132792051@N06/28302196071/in/photolist-K7Y9RM-utZTV9-QWJmHo-QAEdSE-QAE3pL-TvjNJu-%20tziyrj-EWFwEx-DWb7T4-DTRAWu-CYLBpP-DMUVn2-dUbgLG-ccuabw-57nNvJ-UpDv4D-eNyCQP-q8aWpJ-86gced-QLBwiG-QP7k6v-aNxiRc-rmTdLW-oeTM8i-d1rkCG-ueSwz4-%20dYKwJx-7PxAPF-KFUqKN-TkarEj-7X5FZ2-7WS6Z2-7X5Gwa-7X5GkT-7Z8w5s-s4St8A-%20qsa12b-7X8Vqs-7X8VLy-7X5Gm6-7X5Gjp-PTy69W-7X8VQ3-7X8VEy-7X5GqD-iaMjUN-7X8VgE-odbiWy-TkacgQ-7X5Gk4/

[29]. Brennan T, Dieterich W, Ehret B. Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System. Criminal Justice and Behavior. 2009 January;36(1):21–40.

[30]. Zeng J, Ustun B, Rudin C. Interpretable classification models for recidivism prediction. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2017;180(3):689–722.

[31]. Tollenaar N, van der Heijden PGM. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2013;176(2):565–584.

[32]. Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C. Certifiably optimal rule lists for categorical data. Journal of Machine Learning Research. 2018;19:1–79.

[33]. Mannshardt E, Naess L. Air quality in the USA. Significance. 2018 Oct;15:24–27.

[34]. Zech JR, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15(e1002683).

[35]. Chang A, Rudin C, Cavaretta M, Thomas R, Chou G. How to Reverse-Engineer Quality Rankings. Machine Learning. 2012 September;88:369–398.

[36]. Goodman B, Flaxman S. EU regulations on algorithmic decision-making and a ‘right to explanation’. AI Magazine. 2017;38(3).

[37]. Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology. 2018;1(2).

[38]. Quinlan JR. C4. 5: programs for machine learning. vol. 1. Morgan Kaufmann; 1993.

[39]. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

[40]. Auer P, Holte RC, Maass W. Theory and Applications of Agnostic PAC-Learning with Small Decision Trees. In: Machine Learning Proceedings 1995. San Francisco (CA): Morgan Kaufmann; 1995. p. 21–29.

[41]. Wang F, Rudin C. Falling Rule Lists. In: Proceedings of Machine Learning Research Vol. 38: Artificial Intelligence and Statistics (AISTATS); 2015. p. 1013–1022.

[42]. Chen C, Rudin C. An optimization approach to learning falling rule lists. In: Proceedings of Machine Learning Research Vol. 84: Artificial Intelligence and Statistics (AISTATS); 2018. p. 604–612.

[43]. Burgess EW. Factors determining success or failure on parole; 1928. Illinois Committee on Indeterminate-Sentence Law and Parole Springfield, IL.

[44]. Ustun B, Rudin C. Optimized Risk Scores. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2017..

[45]. Ustun B, Rudin C. Supersparse linear integer models for optimized medical scoring systems. Machine Learning. 2015;p. 1–43.

[46]. Carrizosa E, Martín-Barragán B, Morales DR. Binarized support vector machines. INFORMS Journal on Computing. 2010;22(1):154–167.

[47]. Sokolovska N, Chevaleyre Y, Zucker JD. A Provable Algorithm for Learning Interpretable Scoring Systems. In: Proceedings of Machine Learning Research Vol. 84: Artificial Intelligence and Statistics (AISTATS); 2018. p. 566–574.

[48]. Ustun B, et al. The World Health Organization Adult Attention-Deficit/Hyperactivity Disorder Self-Report Screening Scale for DSM-5. JAMA Psychiatry. 2017;74(5):520–526. [PubMed: 28384801]

[49]. Chen C, Li O, Tao C, Barnett A, Su J, Rudin C. *This* Looks Like *that:* Deep Learning for Interpretable Image Recognition. In: Neural Information Processing Systems (NeurIPS); 2019..

[50]. O'Malley D Clay-colored Sparrow; 2014. Public domain image. https://www.flickr.com/photos/62798180@N03/11895857625/.

[51]. ksblack99. Clay-colored Sparrow; 2018. Public domain image. https://www.flickr.com/photos/ksblack99/42047311831/.

[52]. Schmierer A Clay-colored Sparrow; 2017. Public domain image. https://flic.kr/p/T6QVkY.

[53]. Schmierer A Clay-colored Sparrow; 2015. Public domain image. https://flic.kr/p/rguC7K.

[54]. Schmierer A Clay-colored Sparrow; 2015. Public domain image. https://www.flickr.com/photos/sloalan/16585472235/.

[55]. Li O, Liu H, Chen C, Rudin C. Deep Learning for Case-based Reasoning through Prototypes: A Neural Network that Explains its Predictions. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI); 2018. p. 3530–3537.

[56]. Gallagher N, et al. Cross-Spectral Factor Analysis. In: Proceedings of Advances in Neural Information Processing Systems 30 (NeurIPS). Curran Associates, Inc.; 2017. p. 6842–6852.

[57]. Wang F, Rudin C, Mccormick TH, Gore JL. Modeling recovery curves with application to prostatectomy. Biostatistics. 2018;p. kxy002. Available from: 10.1093/biostatistics/kxy002.

[58]. Lou Y, Caruana R, Gehrke J. Intelligible Models for Classification and Regression. In: Proceedings of Knowledge Discovery in Databases (KDD). ACM; 2012..

[59]. Semenova L, Parr R, Rudin C. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning; 2018. In progress.

[60]. Razavian N, et al. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data. 2015;3(4).

[61]. Ustun B, Spangher A, Liu Y. Actionable Recourse in Linear Classification. In: ACM Conference on Fairness, Accountability and Transparency (FAT*); 2019..

[62]. Su G, Wei D, Varshney KR, Malioutov DM. Interpretable Two-Level Boolean Rule Learning for Classification. In: Proceedings of ICML Workshop on Human Interpretability in Machine Learning; 2016. p. 66–70.

[63]. Dash S, Gunluk O, Wei D. Boolean Decision Rules via Column Generation. In: 32nd Conference on Neural Information Processing Systems (NeurIPS); 2018..

[64]. Wang T, Rudin C, Doshi-Velez F, Liu Y, Klampfl E, MacNeille P. A Bayesian Framework for Learning Rule Sets for Interpretable Classification. Journal of Machine Learning Research. 2017;18(70):1–37.

[65]. Rijnbeek PR, Kors JA. Finding a Short and Accurate Decision Rule in Disjunctive Normal Form by Exhaustive Search. Machine Learning. 2010 Jul;80(1):33–62.

[66]. Goh ST, Rudin C. Box Drawings for Learning with Imbalanced Data. In: Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD); 2014..

[67]. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Interpretable machine learning: definitions, methods, and applications. arXiv e-prints: 1901 04592 [statistical machine learning]. 2019 Jan;.

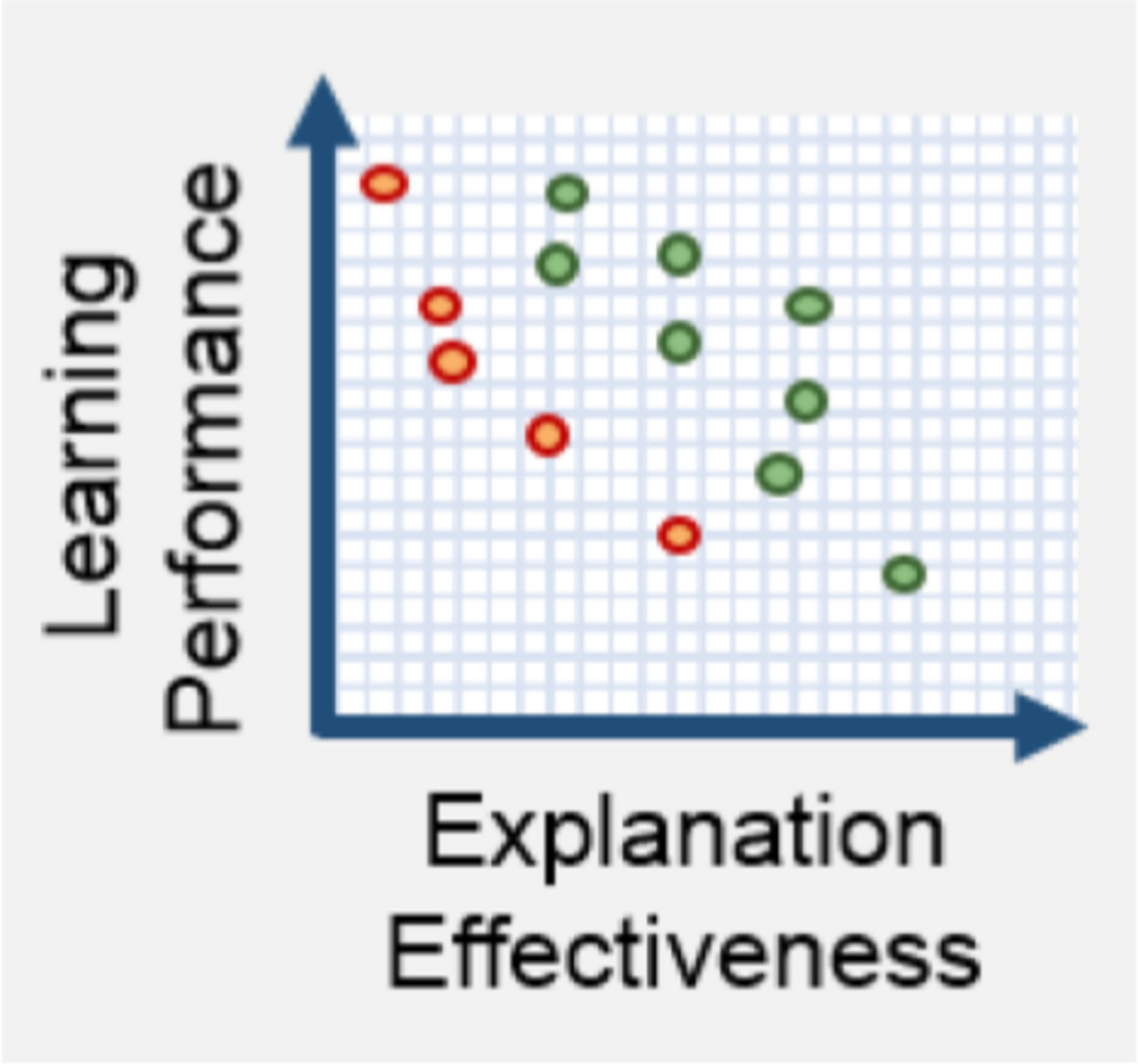

Figure 1:
A fictional depiction of the "accuracy-interpretability trade-off," taken from the DARPA XAI (Explainable Artificial Intelligence) Broad Agency Announcement [18].

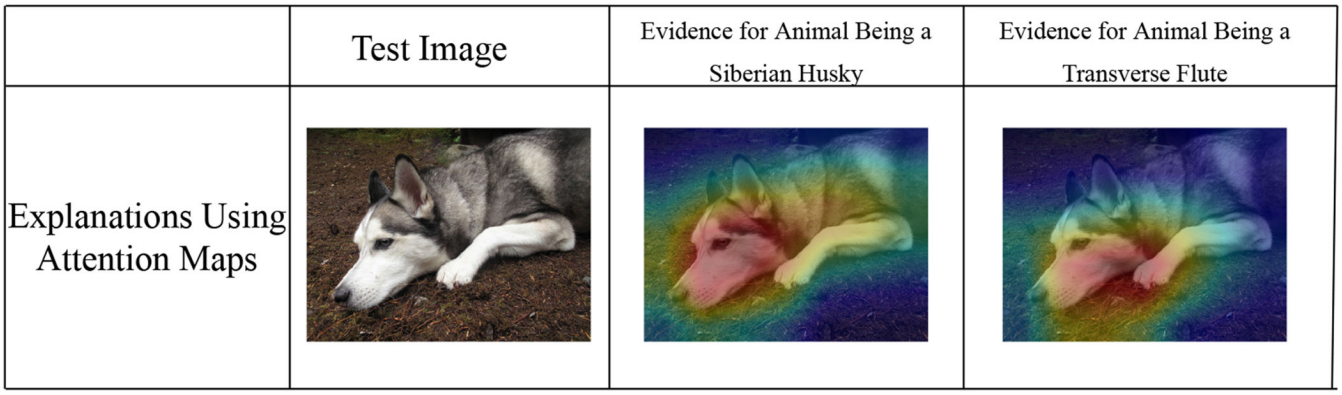

Figure 2:
Saliency does not explain anything except where the network is looking. We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Figure credit: Chaofan Chen and [28].

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Figure 3:
This is a machine learning model from the Certifiably Optimal Rule Lists (CORELS) algorithm [32]. This model is the minimizer of a special case of Equation 1 discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.

1.	Prior Arrests $\geq$ 2	1 point		···
2.	Prior Arrests $\geq$ 5	1 point	+	···
3.	Prior Arrests for Local Ordinance	1 point	+	···
4.	Age at Release between 18 to 24	1 point	+	···
5.	Age at Release $\geq$ 40	-1 points	+	···
		SCORE	=	···

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

Figure 4:
Scoring system for risk of recidivism from [21] [which grew out of 30, 44, 45]. This model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

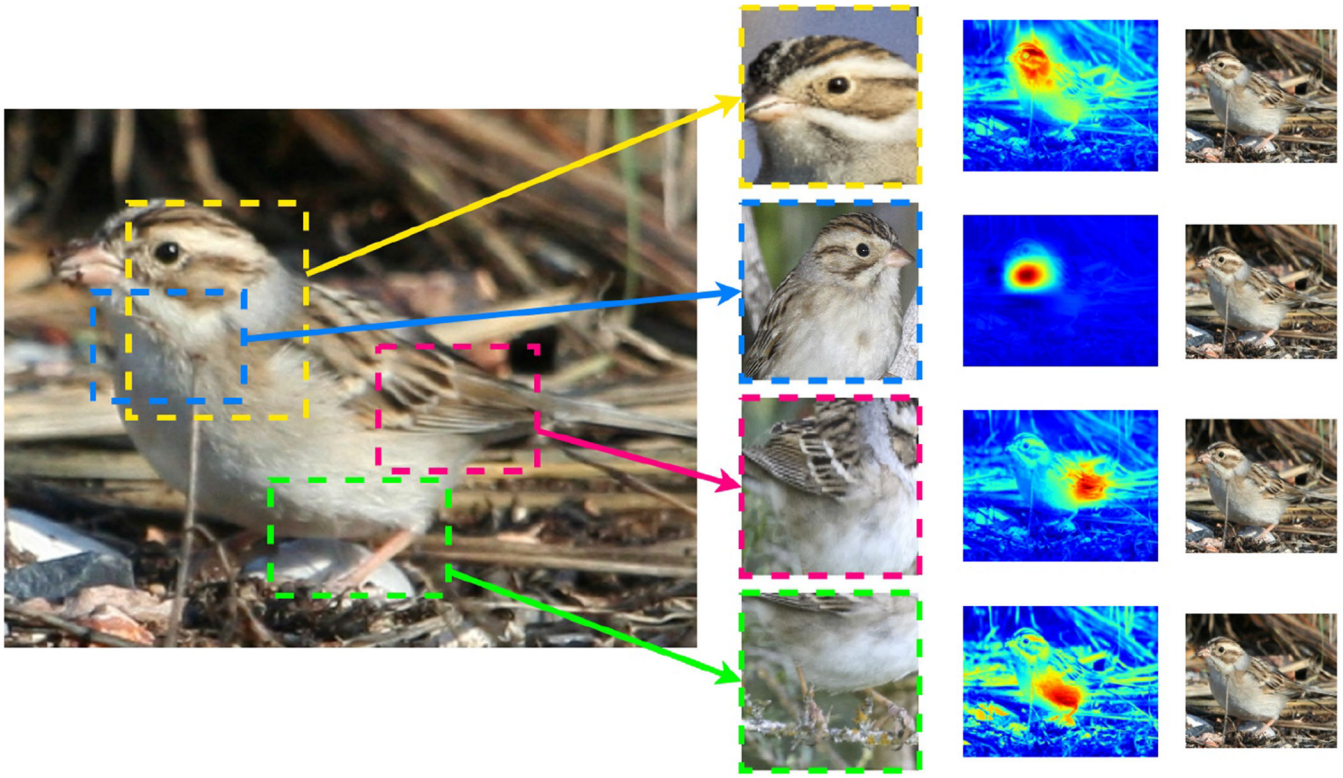

Figure 5:
Image from the authors of [49], indicating that parts of the test image on the left are similar to prototypical parts of training examples. The test image to be classified is on the left, the most similar prototypes are in the middle column, and the heatmaps that show which part of the test image is similar to the prototype are on the right. We included copies of the test image on the right so that it is easier to see what part of the bird the heatmaps are referring to. The similarities of the prototypes to the test image are what determine the predicted class label of the image. Here, the image is predicted to be a clay-colored sparrow. The top prototype seems to be comparing the bird's head to a prototypical head of a clay-colored sparrow, the second prototype considers the throat of the bird, the third looks at feathers, and the last seems to consider the abdomen and leg. Test image from [50]. Prototypes from [51, 52, 53, 54]. Image constructed by Alina Barnett.

Table 1:

Comparison of COMPAS and CORELS models. Both models have similar true and false positive rates and true and false negative rates on data from Broward County, Florida.

COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software license), within software used in U.S. Justice System	full model is in Figure 3 only age, priors, (optional) gender no other information free, transparent

ARTICLE OPEN

Check for updates

Machine learning modeling practices to support the principles of AI and ethics in nutrition research

Diana M. Thomas[1]✉, Samantha Kleinberg[2], Andrew W. Brown[3,4], Mason Crow[1], Nathaniel D. Bastian[5], Nicholas Reisweber[1], Robert Lasater[1], Thomas Kendall[1], Patrick Shafto[6], Raymond Blaine[7], Sarah Smith[7], Daniel Ruiz[7], Christopher Morrell[7] and Nicholas Clark[1]

BACKGROUND: Nutrition research is relying more on artificial intelligence and machine learning models to understand, diagnose, predict, and explain data. While artificial intelligence and machine learning models provide powerful modeling tools, failure to use careful and well-thought-out modeling processes can lead to misleading conclusions and concerns surrounding ethics and bias.
METHODS: Based on our experience as reviewers and journal editors in nutrition and obesity, we identified the most frequently omitted best practices from statistical modeling and how these same practices extend to machine learning models. We next addressed areas required for implementation of machine learning that are not included in commercial software packages.
RESULTS: Here, we provide a tutorial on best artificial intelligence and machine learning modeling practices that can reduce potential ethical problems with a checklist and guiding principles to aid nutrition researchers in developing, evaluating, and implementing artificial intelligence and machine learning models in nutrition research.
CONCLUSION: The quality of AI/ML modeling in nutrition research requires iterative and tailored processes to mitigate against potential ethical problems or to predict conclusions that are free of bias.

Nutrition and Diabetes (2022)12:48 ; https://doi.org/10.1038/s41387-022-00226-y

INTRODUCTION

Complex, large, and multimodal nutrition datasets are being aggregated for the purpose of advancing personalized nutrition, such as the Personalized Responses to Dietary Composition Trial-1 (PREDICT) study [1], a study focused on nutritional prediction of glycemic responses [2], and the new Nutrition for Precision Health program [3]. Such studies and programs highlight a critical need and growing desire to implement machine learning (ML) in nutrition research. For nutrition researchers new to ML but well-versed in statistical methods, using ML models will require adhering to best practices from statistical methods while establishing new approaches that address the complexities of ML models.

The availability of AI/ML capabilities in commercial software packages has made AI/ML algorithms accessible to the wider nutrition research community. However, the high accessibility of AI/ML models through "click and play programs" belies their complexity, which, when overlooked, can lead to myriad unanticipated ethical problems that violate published AI principles [4, 5]. Standardized procedures for the appropriate implementation of ML models often do not exist. Deceptively simple questions, such as whether the sample size is adequate for model fitting, often require iterative evaluation by the modeler that cannot be built into standardized software. Failure to follow a reflective thoughtful approach to AI/ML modeling can lead to errors and biased conclusions that can have deleterious results [6].

Herein we define ML as computer algorithms that improve automatically through experience [7, 8]. The closely related term "artificial intelligence" (AI) is often interchanged with ML. AI refers to an algorithm that can learn insights, adapt through feedback, be dynamic, respond to its environment, and problem solve independently with minimal human supervision [8, 9]. ML is sometimes considered a subset of AI and vice versa, and the terms are frequently used interchangeably [8]. We, therefore, refer to both types of algorithms as AI/ML because many of the ethical concerns discussed herein apply regardless of distinction.

The Alignment Problem by Brian Christian [6] and landmark studies like those of Buolamwini and Gebru [10] highlight many unfortunate consequences of launching ML models without careful examination of the data used for modeling, without application of more than one modeling approach, and without a thorough review and surveillance of model predictions and conclusions. Such negative consequences can range from racial or other discriminatory predictions, wasted time or opportunity, negative health outcomes, or even death. Many detrimental consequences of AI/ML applications covered in Christian's book

[1]Department of Mathematical Sciences, United States Military Academy, West Point, NY 10996, USA. [2]Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA. [3]Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA. [4]Arkansas Children's Research Institute, Little Rock, AR 72202, USA. [5]Army Cyber Institute, United States Military Academy, West Point, NY 10996, USA. [6]Department of Mathematics and Computer Science, Rutgers University, Newark, NJ 07102, USA. [7]Department of Electrical Engineering and Computer Science, United States Military Academy, West Point, NY 10996, USA.
✉email: diana.thomas@westpoint.edu

Received: 22 October 2022 Revised: 28 October 2022 Accepted: 15 November 2022
Published online: 02 December 2022

can be summarized as resulting from poor modeling practices. In addition, a recent review of 62 studies that used machine learning to detect and predict COVID-19 from chest radiographs and CT scans found that every single study had a methodological flaw [11]. These flaws ranged from lack of transparency regarding how key modeling decisions were made to not including model validation experiments [11].

With many and varied approaches available for evaluating AI/ML models, how can nutrition modelers, manuscript reviewers, and journal editors ensure that the models are complete, minimize predictions or conclusions that can cause patient harm, avoid bias, and minimize ethical violations [12]? While we cannot address every possible situation and scenario that could arise, we address common considerations that nutrition researchers may encounter when developing and/or evaluating AI/ML models. The considerations we address herein came from our experience as AI/ML modelers in nutrition, serving as reviewers of AI/ML modeling articles, and our service as editors for top nutrition research journals. We frame the discussion for an audience of nutrition researchers who are familiar with statistical and ML methods in nutrition research but may be new to or have limited experience with developing, evaluating, or implementing AI/ML models.

The description and recommendations here build upon an existing body of literature. The Findable, Accessible, Interoperable, and Reusable (FAIR) Data Principles [13] involve stewardship and management of data which have some overlap with AI/ML best modeling practices. There have been several articles on best AI/ML modeling practices which draw upon and integrate with FAIR principles [14, 15]. Articles that provide overviews of machine learning also include some best modeling practices [16, 17] and articles that are specific to an application like image analysis [18] include best modeling practices that scale to other disciplines. In addition, discipline-specific checklists are now being applied for publications such as the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [19], the machine learning checklist for Neural Information Processing Systems [20], and the machine learning reproducibility checklist produced by the Computer Vision and Pattern Recognition Conference [21]. The guidelines and checklist presented here focus on the viewpoint of a nutrition researcher who has a background in statistics and wishes to build on that background to include AI/ML models to describe, predict and explain nutrition data.

We begin with some well-known modeling practices derived from statistical methods that extend to AI/ML modeling. We next move to two important areas specific to AI/ML model development: appropriate sample sizes and balanced datasets. Next, we address the need for simultaneous development of models and specifically explainable AI/ML models. Finally, we emphasize the need for increased data literacy. With the application of new and complex AI/ML approaches in nutrition research, we as a community need to learn more about the underlying properties, requirements, capabilities, and limitations of AI/ML model development. Because AI/ML approaches are relatively new [1, 2] in nutrition, many of the examples of bias and error arising from poor development and evaluation of AI/ML models are drawn from other disciplines. These examples, while not specifically in nutrition, provide can raise our awareness of potential pitfalls as a higher dependence on AI/ML models in nutrition research advances. Table 1 serves as a Table of Contents, and Table 2 is a checklist that summarizes our tutorial. The checklist in Table 2 is presented in order of AI/ML execution starting with study design and ending with model evaluation. While every step in the checklist is important as a best practice, the most important result of the checklist is reproducibility. If we consider the AI/ML modeling process analogous to the methods behind the experiment, the checklist provides clear, rigorous, and transparent guidelines for the methods that ensure the results are reproducible.

Table 1. The Table of Contents is hyperlinked to ease navigation to sections within the article.

Hyperlinked Article Sections
Introduction to Extensions of Statistical Methods to AI/ML
Measurement Error
Selection Bias
Sample Size Calculations for AI/ML
Missing Data in AI/ML
Data Imbalance
Explainable AI
Data Literacy: The AI User Responsibility

EXTENSIONS TO AI/ML FROM STATISTICAL MODELING

Statistical modeling has well-developed methods for identifying, mitigating, and transparently reporting bias and error. We distinguish "bias" in the statistical sense from "bias" in the social sense. When we discuss bias in a model, we are indicating that the expectation of the model does not match the true value; that is, we reliably come to inaccurate conclusions. More specifically, we are referring to bias that comes from the statistic being used to estimate a parameter, or we are discussing bias that arises from using data that is not representative of our population. In either case, the result of the bias is a parameter estimate that is not accurate. However, we should note that all bias is not bad; statisticians will often use a biased estimator if it results in a lower mean squared error such as what is used in the popular LASSO algorithm. Biased data, or sampling data that is not reflective of our population, on the other hand, is rarely a good idea and can lead to disastrous results if not properly accounted for. This is different than the social aspects of bias, such as prejudice. Unfortunately, some forms of bias discussed herein (attrition bias, selection bias) may result in or result from socially biased research approaches, which in turn can create a model that inherits those biases, and ultimately creates a statistically biased model. Many of the statistically-based quality assurance checks still apply and are even more important to consider when developing machine learning models. Unfortunately, these statistical best practices are oft "forgotten" [22] and are not standard or routine when reporting the results of machine learning predictive models. Identifying whether the characteristics of participants who dropped out were different than completers, whether missing data were missing at random, or expressing limitations on extending predictions beyond the sample are common omissions [23, 24].

Statistical modeling best practices that ensure data are collected in manners that reduce bias and errors exist and are also relevant for AI/ML model development. It is not our intent to provide a comprehensive statistical tutorial on the statistical methods. Instead, we provide a summary of bias and error that is often observed in nutrition research and address how statistical mitigation strategies also prevail for AI/ML models. Some methods are "best (but oft-forgotten) practices" [25] and we recommend the statistical series at the *American Journal of Clinical Nutrition* for an in-depth tutorial into statistical practices frequently applied in nutrition research [22].

Measurement error

Take home message. Controlled data with minimal measurement error are needed as a gold standard to compare clinically relevant data that the models will be used on. Explainable AI/ML models are key to understanding the propagation of measurement error.

What is it? There is a wide range of measurements in clinical nutrition. Measurements of glycated hemoglobin (HbA1c) are

Table 2. Checklist for ethical and effective application of AI/ML modeling in nutrition research.

Considerations for Ethical & Effective Application of Machine Learning for Nutrition Research		
Item No.	**Item**	**Recommendations**
Study Design		
1	Describe Overall Goal	Is the purpose to understand new data, select the most informative features from data, inform other researchers/clinicians/public, develop a model that makes predictions or diagnoses, or something else?
2	Describe Data	Clearly identify
		1. The data source
		2. When data were collected
		3. Over what time period data were collected
		4. Whether data will continue to be collected
		5. The size of the dataset
		6. The collection methodology
		Disclose in the manuscript if any of the above are unknown.
		Describe
		1. The approaches for minimizing measurement error during data collection
		2. The approaches for minimizing collection procedure bias
		3. Warning labels regarding the representation of the data
		If existing data sources are used, why were these particular source(s) used?
		Do the data represent the target population(s) (the population that your AI/ML models will be used to predict) accurately?
4	Discuss AI/ML Suitability	Are the modeling approach(es) supervised or unsupervised? Will the models be updated with additional data and, if so, how?
		Explain the suitability of AI/ML to answering the question. For example, is there an abundance of complex data? What were the results of traditional approaches such as regression? Do you suspect the data contain underlying patterns or correlations that a computer could learn?
5	Establish Evaluation Criteria	What evaluation criteria will you use to assess the performance of your model(s)?
		Why did you settle on these criteria? If using categorical classification, report a confusion matrix in the results. In the discussion, explain what is the impact is of the false positive rate and false negative rate on your application. If you are predicting continuous outcomes, what is the cost of over or under-estimating?
Data Pre-Processing		
6	Handle Missing Data	What data are missing? What techniques were employed to account for missing data? If multiple techniques were used, how were they evaluated against each other?
		In the discussion, describe the potential cause for missing data. Are the data MCAR (Missing Completely At Random), MAR (Missing At Random), MNAR (Missing Not At Random)?
		Why did you settle on a particular method for handling missing data?
7	Classify Outliers	How were outliers defined? Were outliers removed? What is the impact of including the outliers on your modeling? Why did you settle on this method to identify and classify outliers? Simulate the potential comparison of model performance with/ without outliers or outliers defined by different approaches.
8	Balance Classes	Did you balance the subgroups used as inputs or the classes you are predicting? What was the method used to balance the dataset (e.g., up-sampling previously untapped populations)? Justify the choices of balancing classes. For example, did the initial class distribution fail to match the distribution of the population for which you are applying the AI/ML model? Describe your balancing methodology, including justifying why not balancing would be appropriate if you choose not to balance your dataset.
9	Select Features	Which features from your dataset did you select for AI/ML model training? Were all available features used or a subset? Explain why the features were selected.
10	Evaluate Dataset Size & Augmentation	Was the dataset reduced/expanded through resampling or augmentation?
		Is the dataset of an appropriate size for the AI/ML modeling methods? Why was the dataset reduced or expanded? Are you targeting a particular AI/ML algorithm (i.e., neural network)?
		How will the size of the dataset, after pre-processing, inform the choice AI/ML algorithm(s) (see next section)?

Table 2. continued

Considerations for Ethical & Effective Application of Machine Learning for Nutrition Research		
Item No.	**Item**	**Recommendations**
Algorithm Construction		
11	Select Algorithm(s)	List all AI/ML modeling approaches that were trained and evaluated.
		Justify why the approaches were selected. If only one approach is used, explain why it was not feasible or not desirable to test more than one model. Is at least one AI/ML approach explainable? If not, why?
		Clearly assess assumptions of the AI/ML models and describe whether they hold.
12	Algorithm Explainability	Describe approaches to enhance or select for explainability of models.
		Describe the level of explainability of the selected models. Can the model's decision-making be understood or interpreted?
		If selecting a non-explainable model, justify the choice (e.g., far superior model performance when non-explainable). If an explainable model was not paired with the non-explainable example, provide justification.
13	Model Reproducibility	Provide the exact hardware, software, and hyperparameter specifications used to train AI/ML model(s). Supply the algorithms, data, and code to reproduce the model. Explain the steps required to reproduce the results. If appropriate, explain why data, code, software, or other artifacts necessary to reproduce the work are not publicly available.
Algorithm Evaluation		
14	Determine Baseline Performance	Determine Comparison Performance. What is the state of the model accuracy in the literature? How are you improving understanding or accuracy beyond what already exists?
15	Internal Validation	Describe how the training and validation/test set were divided and why. Use evaluation approaches like k-fold cross-validation to capture internal variance. Include multiple runs on any machine learning model that relies on random initiation of weights or other model parameters (e.g. neural networks).
		Do the results suggest the model was overfitting or underfit? Did you use internal validation approaches like k-fold cross-validation? If so, what were the results?
16	Determine Best Model	Identify and explain which AI/ML model performed the best in accordance with your chosen evaluation criteria.
17	External Validation	Describe any approaches used to externally validate the model (i.e., model validated on an independent sample).
		If not externally validating, why not? If externally validating, explain why that external data source and approach are being used.
		What are the results of external validation? Alternatively, what are the ramifications of not externally validating?
18	Model Deployment Considerations	Describe how the model will be deployed and who the end user would be.
		Describe the use cases for the model. What are the limitations of the model? How often should it be reevaluated or retrained? What is the shelf life of the data?
19	Considerations	Offer considerations for future research. What additional techniques or data could be tested?

objective and correlate to a patient's diabetes status [26]. On the other hand, measurements that are obtained from accelerometers are also objective, but can be extremely noisy and are not able to estimate physical activity expenditure well in comparison to gold standard methods [27]. However, the largest source of measurement in nutrition research, self-reported energy intake, is not objective and sometimes not reliable without triangulating with other methods [28] for deriving scientific conclusions [29–31]. There are numerous additional diverse measurements in nutrition research such as clinical energy balance measurements [32, 33], body composition [34], anthropometry [35], and biomarkers [36]. Within these measurements, some of the measurement errors occur at random while some are systematic or idiosyncratic.

Statistical modeling has long included discussions of error, including assumptions about the nature of the error (e.g., normally distributed with zero mean) that have to be satisfied in order to make statistical inferences and methods that assume that the true values are measured with error (e.g. Bayesian error models) [37, 38]. Because measurement error can render the results of a study or model meaningless, imprecise, or unreliable [39], there is a vast literature on handling measurement error [40, 41] in the context of statistical modeling.

What should we do about it? While we cannot eliminate all measurement errors, there are best practices to reduce measurement errors during data collection. Some best practices to minimize measurement error is to take multiple measurements of the same variable when possible and to collect the data with precision. For example, body weights should be collected under similar conditions, such as first thing in the morning, on the same scale, and in a hospital gown. To obtain information on the variation in measurements, the measurement should be taken multiple times (e.g., three times for body weight). How much measurement error is in the input data needs to be conveyed, not just in peer-reviewed publications, but also as "warning labels" in data repositories that will include AI/ML prediction tools. An exemplar for including warning labels within a data repository is the All of Us Research Program [42], which alerts data users to the

quality and distribution of the data during access. A robust list of resources for tagging data for reuse and reproducibility appears on the Go FAIR website for the Findable, Accessible, Interoperable, and Reusable (FAIR) principles [13, 43].

In the case of non-objective measurement error, it has been suggested that self-reported dietary intake should not be used as true dietary intake to derive scientific conclusions [29, 30]. This does not mean that self-reported dietary intake data is not valuable during interventions. There are examples of self-reported dietary intake data being used in tandem with other tools such as energy intake wearables [44, 45] and mathematical models that predict weight loss to guide intake [46] improving dietary adherence even more than any of the dietary assessment methods used alone [28]. The danger of using data like self-reported dietary intake as true intake to train AI/ML models is that the models will identify patterns that are artifacts of error from the input data which will then be used to make erroneous predictions that inform decision-making. For example, intake has been found to be underreported in individuals with obesity [31, 47], which has led to erroneous predictions and conclusions that people with obesity gain weight while eating less [48]. It is important to note that if we knew the bias in the self-reported data this could easily be corrected. Future research should focus on identifying the magnitude and direction of biases in the data using proxy or alternate datasets. Multilevel models also serve as potential tools that should further be studied to determine how they can potentially be leveraged to correct self-reporting biases [49].

We also need to be concerned about the measurement and its error under conditions of research versus conditions of use. Using body weight as an example: if a model is trained on body weight collected under exacting conditions, multiple times, at the same time of day, the model may not perform as well when using body weights taken at the clinic once, at any time of day, often without removing excess clothing. The measurement for the model thus does not match the measurement for use.

Extension to AI/ML modeling Errors in measurement have the potential to result in erroneous decisions. Simple models allow us to track how error propagates from the initial variable to the final output. In comparison to simpler explainable models like linear regression, it is often challenging to track error propagation in AI/ML models when they contain nonlinearities and interconnections between variables that are not immediately apparent, also known as "black boxes" [50]. Furthermore, AI/ML methods often incorporate nonlinear aspects which tend to exacerbate error [51]. Specific methods to address individual AI/ML models exist, but there does not exist a one size fits all solution to generally characterize error propagation within AI/ML models [51]. The reliability of a model where the error propagation is unknown cannot be properly characterized; however, model developers can look to the literature for the specific model to find methods to quantify error propagation [52].

Selection bias

Take home message. Characteristics of the dataset, such as demographics, need to be summarized and explored for limitations prior to training algorithms. Justification should be provided for why the AI/ML model is appropriate for the sample size. Approaches such as up-sampling and down-sampling can be cautiously applied using an iterative process to mitigate concerns about selection bias.

What is it? One of the most well-known examples of selection bias in artificial intelligence occurred when a Google Photos image classifier incorrectly identified people of color as gorillas [6]. Google attempted to fix the artificial intelligence model from a top-down approach relying on various strategies; however, the underlying problem was that the model training dataset did not contain enough people of color. This is known as "selection bias". Selection bias occurs when the individuals or groups in a dataset differ from the population of interest in a systematic way [53]. In the Google Photos example, the data on which the model was trained did not fully represent the population the models were applied for. As summarized by Brian Christian, the problem with "a system that can, in theory, learn just about anything from a set of examples is that it finds itself, then, at the mercy of the examples from which it is taught" [6].

What should we do about it? Selection bias awareness is required in both study design and in reporting model capabilities. When recruiting, investigators should focus on the population they hope to generalize to and then recruit participants that meet those criteria. Recruiting a population that aligns with the target population for study outcomes will minimize selection bias. However, such recruitment may require creative ways to reach previously untapped populations [6].

Extension to AI/ML modeling Recruiting representative populations for training datasets may not always be possible. For instance, large datasets may consist of convenience samples like electronic health records [54]. One method to account for this limitation is to weigh the data for key characteristics between the sample and population of interest. Weighting the data for regression applications is straightforward, but does not extend to AI/ML models that are often nonlinear. An extension of the statistical weighting approach to AI/ML models is to "up-sample" or "down-sample" the data according to weights. For example, if the dataset contains a sample of 20% females and 80% males, "up-sample" by repeating the 20% observations until the dataset female:male ratio matches the population of interest (e.g., ~50%). Conversely, a random sample of male subjects can be selected to down-sample or develop a dataset that contains the target female:male ratio. While this concrete example addresses female:male imbalance, it does not address other potential imbalances. For example, the female sample may have a BMI distribution different from the population (e.g., the sample is all below 25 kg/m^2). AI/ML models may therefore incorrectly learn that females will have BMI below 25 kg/m^2 without appropriately addressing imbalance. In all cases, the limitations of the data used to train the model should be made explicit in publications and any software application or tools used to disseminate the model should warn the user of limitations such as the characteristics of the training dataset.

CONSIDERATIONS SPECIFIC TO AI/ML MODELING

Sample sizes calculations

Take home message. No one-size-fits-all approach exists to calculate sample sizes for AI/ML models. Adequate sample size depends on the application and model complexity. Sample size calculations for specific AI/ML models often require an iterative process. For reproducibility, the justification for the sample size always should be articulated.

What is it? Having a large enough sample to train and test AI/ML models is critical to avoid overfitting or underfitting models. The definition of model overfitting is when the model fits too closely to the training dataset [55], thereby capturing idiosyncrasies of the observed data rather than generalizing true data properties. Ethical issues with overfitting occur when models perform well on the training dataset, but do not translate well to new data. For example, an overfit model that uses biomarkers to predict patient health will predict accurately the patient's health used in the sample to develop the model, but misdiagnose patients not used in model development as being healthy when they actually require treatment [56]. There are several ways to mitigate potential overfitting and sample size can play a role. In general, the more complex the model (e.g., more weights, input variables, and layers in a neural network), the

more data required to avoid situations like overfitting. Underfitting, on the other hand, can occur when there is not enough complexity in the model to match the supplied data [57]. In both cases, selecting the right sample size depends on the complexity of the model, tests for goodness of fit in independent data, and iterative evaluation of the model design versus model's outcomes. In addition, in AI/ML models that are used for feature selection or identifying which variables are relevant, too small of a training dataset may result in lower data variability and, consequently, degrade the identification of important features [58].

What should we do about it? Power is the probability of detecting a difference when one really exists (that is, one minus the probability of making a type 2 error). In statistical analyses, it is used to determine the sample size required to make appropriate and corresponding statistical inferences. Although well-studied in the area of AI/ML modeling [59], a similar systematic and tractable method to determine sample sizes for AI/ML models cannot be provided. The nonlinearity and complexities of AI/ML models and the multiple models that fall into the category of AI/ML do not lend well to a uniform process for calculating sample sizes when compared to more simplistic analyses like a t-test. Despite these challenges, several published "rules of thumb" exist [60]. For classification models (e.g., decision trees or neural network classifiers), a rule-of-thumb is that the sample size needs to be at least 50–1000 times the number of classes being predicted [61]. For example, if you are predicting categories of obesity (BMI ≥ 30 versus BMI < 30), this is a binary classifier and your sample size would need to be between 100 and 2000. Similar rules of thumb exist relating sample sizes to the number of input variables or features, or sample size to number of weights used in the model. These rules ultimately relate the sample size to the complexity of the model (e.g., number of classes predicted, number of variables used as inputs, the number of hidden layers, or number of weights) and range widely as demonstrated with the 100–2000 range for a binary classifier. Thus, an iterative process is required to determine the appropriate sample size tailored for each individual problem and model. In publications or other forms of model dissemination, the sample size choice must be justified and clearly articulated.

For exploratory modeling when the number of covariates is high compared to the number of data points, regularization techniques such as LASSO regression or, more generally, Elastic Net regression offer ways to fit data. Here the resulting parameters will be biased, however, more complex models can be fit [62]. Whether these techniques are appropriate depends on the overall goal of modeling, but they are often good tools if practitioners are attempting to both diagnose a root cause as well as build a predictive model.

Missing data

Take home message. Nutrition research frequently includes missing data, such as from incomplete self-reported habits or missed clinical visits. How we handle missing data can influence AI/ML model predictions and conclusions. In addition to traditional statistical approaches for handling missing data such as imputation, methods using AI/ML models have been developed to handle missing data. In some cases, missingness can be treated as a model feature. Lack of adherence to prescribed interventions and other reasons for missingness can be captured using this approach.

What is it? Missing data are pervasive in healthcare and especially common in nutrition research. Missing data can occur in multiple ways. Nutrition research often relies on logs kept by human subjects or surveys (such as the food frequency questionnaire (FFQ), food diaries, or 24-hour recalls) [63]. Individuals may forget to record a specific meal, selectively omit information due to desirability bias [64], or fail to complete the dietary instrument altogether. Objective measures, too, may have missing data, such as missed samples for biomarkers or user and technological errors failing to record behaviors. Datasets may therefore be missing individual data points (e.g., a meal), entire variables (e.g., no blood glucose data), or specific time windows (e.g., losing a day of data due to technology failures).

There are three main types of missing data and each has different implications for data analysis [65]. The first is missing completely at random (MCAR). An example of this is if a researcher is out sick and misses follow-up appointments with some subjects. The probability of a data point being missing is then independent of any characteristics of the participants. MCAR data reduces the sample size (and study power) depending on the proportion of missing information. In some cases, information for some missing data can be inferred from other information in the dataset. However, many models can use only complete records, but in the case of MCAR, ignoring missing data will not lead to biased results. This type of missingness is unlikely. A more common scenario is data that is missing at random (MAR), which is when the likelihood of a variable being missing depends on other variables [66]. For example, if someone leaves out snacks in their meal logs only on days when they do not exercise, data on snacks would be MAR. Similarly, if people are more likely to answer survey questions based on their age or gender, those data would also be MAR. If we use only complete records with MAR data, we may get a biased estimate of how prevalent something is in the population (e.g., 100% of people who snack exercise). For some types of analysis, such as likelihood-based methods, this type of missingness is considered ignorable, though this terminology is a misnomer. We cannot ignore that missingness depends on other observed variables and cannot use only complete records without introducing bias. For causal inference, using only complete records can mean we fail to discover causal relationships (e.g., without any variation in reported snack behavior we cannot find what causes it). Finally, when the presence of data depends on the variable of interest itself, data is missing not at random (MNAR). An example of this is if people only self-report their weight when it falls in certain ranges if doctors measure HbA1c when they suspect it is high, or if an individual with diabetes tests their blood glucose only when they suspect it is too high or too low. Ignoring incomplete records will lead to biased results. For example, ignoring times without glucose values will give the impression that glucose is always at an extreme. Predictive models trained on datasets with data that are MNAR will fail when used in the real world, since they will have few examples of glucose values outside of the extremes. Finally, note that statistical tests to distinguish whether missing data are MCAR, MAR, or MNAR are often highly limited.

What should we do about it? Ignoring subjects who dropped out of a clinical trial can bias results [66], and the same is true for AI/ML methods. Failing to account for missing data can lead to incorrect results and models that fail when applied to new populations. The primary strategies for handling missing data are imputation or modeling the missingness. The majority of imputation methods are designed for data that is MAR, and use observed values to reconstruct missing ones. The simplest approach, using the mean (or mode) value in the observed data to replace missing values, has been used widely, but has significant limitations and is not recommended for use in nutrition studies. The mean recorded bodyweight or calorie intake in a dataset is simply not representative of missing instances. Similarly, carrying forward the last observation (e.g., assuming someone's bodyweight is the same until it is next recorded) requires assumptions about the stability of these variables that are not justified. More advanced approaches, such as k-nearest neighbor (kNN), aim to find similar observed instances to missing ones, and have been applied to FFQ data [67]. Rather than using a

population average, kNN finds the most similar subjects to one with missing data, and uses a function of their values to replace missing ones. Note that this approach is only appropriate for MAR data, where there is a relationship between observed values and missing ones. A limitation is that accuracy declines as more variables are missing for an instance, and it cannot be used when all data is missing (e.g., for time series data, if all variables are absent at one-time point). Multiple imputation [68] allows modeling of uncertainty in missing data. Rather than fill in gaps with a single value, these methods create multiple imputed datasets. Combining results on each enables estimates of error due to the missing data. This approach has been used on FFQ [69], 24-h recall [70], and food log data [71]. For data that are MNAR, fewer methods exist, though some have been introduced to model data with variables that may be MNAR or MAR [72].

Notably, missingness can be informative and has been used as a feature to improve prediction. Intuitively, if a doctor chooses not to run a test or a person decides not to record a specific meal, those events are likely to be different from the ones that are observed. Thus, if we impute values for missing data, but do not capture the fact that data was not originally recorded, we may lose valuable information. Lin and Huang [73] showed that including indicators representing missing data improved predictions from electronic health record data. This has been repeated using other methods such as recurrent neural networks [74, 75].

Data imbalance

Take home message. Datasets used for training must be balanced so models learn what and how input features are important to the application of the AI/ML model. The definition of balance will depend on the model type and intended application, but should consider the distribution of classes in a dataset. There are methods to "balance" a dataset that should be applied cautiously. For reproducibility and transparency, the percentage of different classes available in the training data as well as steps taken to balance the data need to be articulated.

What is it? Data imbalance occurs when most instances in a dataset belong to a single or small subset of the total classes. For example, if females represent only 20% of a training dataset and males are 80% of the dataset, then we would say the dataset is imbalanced. Similarly, if a specific outcome of interest occurs at lower rates than all other outcomes, such as pregnancies complicated by gestational diabetes, and we are developing an AI/ML model to predict which pregnancies result in gestational diabetes, the dataset is also referred to as imbalanced.

In the case where a sub-group is smaller in size than other groups, AI/ML models "see" the subgroup less when learning. The lack of exposure can result in poor performance when restricted to the subgroup. This is exactly what occurred in the Google Photo example described in the Selection Bias section. While people of color were contained in the large dataset, the learning models did not see enough examples of people's faces to be able to recognize faces of people of color when presented with a new photo.

In the second case, where the outcome occurs less frequently, such as gestational diabetes mellitus (GDM), failure to balance the dataset could result in flawed or non-informative models. It is estimated that GDM prevalence is between 4 and 10% of all pregnancies in the United States [76]. An AI/ML model that classifies GDM pregnancies would need more than 90% accuracy to outperform the model that assumes that GDM does not occur. This is because in the worst-case estimate of 10% prevalence of GDM pregnancies, the model that assumes GDM never occurs is already 90% accurate.

What should we do about it? In the section on Selection Bias, up-sampling and down-sampling were already discussed and represent the most frequently applied method to mitigate problems with data imbalance. However, sampling up or down should remain an alternative to the original collection of balanced data. As mentioned earlier, up-sampling can result in AI/ML models learning artifacts of up-sampled observations that are not true features. Similarly, down-sampling the other classifications or subgroups reduces the size of the dataset to the smallest-sized subgroup.

APPLICATION OF EXPLAINABLE MODELS

Goals of explainable AI

The challenge with modern AI/ML models is that oftentimes the complexity of the modeling approach comes at a cost of explainability. This becomes an issue when practitioners attempt to draw causal or suggest causal relationships between predictors and response variables in the model. Because there are many AI/ML modeling approaches, one of the most important best practices is to use more than one AI/ML method and specifically to combine non-explainable with explainable models. For example, neural network classifiers are sometimes referred to as "black boxes" because while neural networks may have high accuracy for prediction, their complexity results in loss of explainability. However, using neural networks in tandem with an explainable method like logistic regression can circumvent the black box and provide explainability.

In general, to understand what elements of a model should be explainable it is useful to think of the Generalized Linear Models (GLM) framework. In this commonly used methodology a practitioner specifies a linear predictor that captures covariates of interest, a link function that maps the linear predictor to function of parameters in the statistical model, and a distribution function that captures the unexplainable parts of the model. The covariates, in this case, are the explainable part of the model. The practitioner may never explain why the uncertainty in the data follow, say, a gamma distribution, but they can explain the meaning behind how the explanatory variables are related to the response. Uncertainty then can further be partitioned through the use of Generalized Linear Mixed Effects Models (GLMM) that allow additional model-based uncertainty to be specified, therefore partitioning the uncertainty into model-based and data-based uncertainty. An interpretable AI algorithm should seek to behave similarly, where some key aspects of the model can be captured as a meaningful part of the parameter. In the machine learning literature tools such as Gaussian Process Regression have recently been used to model more complex data patterns than can be done using GLMMs but in an interpretable manner.

Explainable AI

What is it?. AI/ML models have improved prediction beyond what was previously possible; however, due to model complexity, AI/ML models often lose internal model interpretability [77]. This loss of interpretability can eventually lead to unexpected and problematic model conclusions [6]. For example, deep convolution neural networks were trained using images of skin lesions, and they classified malignant versus benign melanomas with a high degree of accuracy when compared to the diagnosis of board-certified dermatologists [78]. However, it was later found that images of lesions that included rulers were classified as malignant because the model "learned" that when a ruler was included in the image, the lesion was more likely to be malignant. This artifact was introduced because rulers were included in images when the clinician already thought the lesion was more likely to be malignant [79]. If this artifact was not detected (that is, if the model was not explained), the model would have a high false-negative rate for new images. Explainable AI was promoted to preserve the high level of desirable accuracy that is provided by complex AI/ML models while retaining interpretation.

Explainable AI (XAI) [80], is a collection of methods to extract knowledge from opaque or "black box" machine learning methods like deep learning. XAI systems have been developed

to meet this challenge, primarily motivated by image classification concerns like the erroneous classifications with the ruler in the image problem [79]. One example of an XAI method that opens the AI black box for interpretability is a saliency map [81]. A saliency map reveals information on the degree that each feature in the image explain and contribute to predictions [82]. Saliency maps applied in tandem with a deep convolution neural network can leverage the high degree of accurate predictions while retaining interpretable and explainable aspects of the underlying model. Another similar example of XAI used in tandem with a less explainable model occurs with random forests where one can compare the "variable importance" resulting from a comparison of the number of decision trees in which the variable appears, normalized by the associated node impurity decrease.

What are the available tools and how can they be used to model in nutrition?. XAI methods in nutrition are just beginning to advance [50, 83]. For example, XAI has been recently applied to automatic identification of food from images [84]. Food imaging and classification have been used in the Remote Food Photography Method [85] and in eating sensors [86, 87] and represent a novel objective method to estimate food intake in free-living humans.

DATA LITERACY: THE AI USER RESPONSIBILITY

An issue that is rarely addressed is the accountability of AI/ML consumers regarding data literacy. Because of our increasing reliance on AI/ML in nutrition, a certain level of data literacy and data standards needs to be embraced by all nutrition stakeholders. A critical component of data literacy is properly specifying a data-driven question and analyzing whether the question can be answered through descriptive analytics, diagnostic analytics, or predictive analytics. Further, as practitioners increase their data literacy they are better postured to combine the techniques given above. Indeed, many of the methods that fall under AI/ML are diverse and require specialized training. Even trained mathematical modelers cannot be experts in all possible methods and areas – just like any other discipline that interfaces with nutrition. Therefore, we advocate for more articles like the one presented here with checklists and summaries that help the nutrition research community address the right questions that will require models to be transparent, reproducible, and ethically applied.

CONCLUSIONS

The quality of AI/ML modeling requires iterative and tailored processes to mitigate against potential ethical problems or to predict conclusions that are free of bias. Some of these feasibility checks may require a background in AI/ML training and including research team members with expertise will provide support for these analyses. Providing some basic best practice AI/ML modeling principles provides a path for researchers interested in using AI/ML models to understand and implement in nutrition applications.

REFERENCES

1. Berry SE, Valdes AM, Drew DA, Asnicar F, Mazidi M, Wolf J, et al. Human postprandial responses to food and potential for precision nutrition. Nat Med. 2020;26:964–73.
2. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized nutrition by prediction of glycemic responses. Cell. 2015;163:1079–94.
3. Li Z, Wang H, Zhang Y, Zhao X. Random forest–based feature selection and detection method for drunk driving recognition. Int J Distrib Sens Netw. 2020;16:1550147720905234.
4. World Health Organization. WHO Consultation Towards the Development of guidance on ethics and governance of artificial intelligence for Health Meeting report Geneva, Switzerland, 2–4 October 2019. World Health Organization; 2021.
5. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, concepts, and implementation practices of FAIR (findable, accessible, interoperable, and reusable) data principles in health data stewardship practice: protocol for a scoping review. JMIR Res Protoc. 2021;10:e22505.
6. Christian B. The alignment problem: machine learning and human values, First edition. W.W. Norton & Company: New York, NY; 2020.
7. Mitchell TM. Machine Learning: a guide to current research. In: Carbonell JG, Michalski RS, (eds).
8. Campesato O. Artificial intelligence, machine learning and deep learning.
9. Russell SJ. Artificial intelligence: a modern approach. In: Norvig P, (ed). 3rd ed. ed. Upper Saddle River, N.J.: Prentice Hall; 2010.
10. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Sorelle AF, Christo W, (eds). *Proceedings of the 1st conference on fairness, accountability and transparency*. Proceedings of Machine Learning Research 2018. p. 77–91.
11. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3:199–217.
12. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–53.
13. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
14. Artrith N, Butler KT, Coudert F-X, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13:505–8.
15. Makarov VA, Stouch T, Allgood B, Willis CD, Lynch N. Best practices for artificial intelligence in life sciences research. Drug Discov Today. 2021;26:1107–10.
16. Rajkomar A, Dean J, Kohane I. Machine Learning in medicine. 2019;380:1347–58.
17. DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, et al. A review of machine learning in obesity. Obes Rev. 2018;19:668–85.
18. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. Am J Roentgenol. 2019;212:513–9.
19. Mongan J, Moy L, Charles E Kahn J. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. 2020;2:e200029.
20. NeurIPS 2022 Paper Checklist Guidelines. In 2022.
21. The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020). In 2020.
22. Bier DM, Allison DB, Alpers DH, Astrup A, Cashman KD, Coates PM, et al. Introduction to the series "Best (but Oft-Forgotten) Practices". Am J Clin Nutr. 2015;102:239–40.
23. Shilo S, Godneva A, Rachmiel M, Korem T, Kolobkov D, Karady T, et al. Prediction of personal glycemic responses to food for individuals with type 1 diabetes through integration of clinical and microbial data. Diabetes Care. 2022;542:502–511.
24. Gallardo M, Munk MR, Kurmann T, De Zanet S, Mosinska A, Karagoz IK, et al. Machine learning can predict anti-VEGF treatment demand in a treat-and-extend regimen for patients with neovascular AMD, DME, and RVO associated macular edema. Ophthalmol Retin. 2021;5:604–24.
25. Ludwig DS, Ebbeling CB, Wong JMW, Wolfe RR, Wong WW. Methodological error in measurement of energy expenditure by the doubly labeled water method: much ado about nothing? Am J Clin Nutr. 2019;110:1253–4.
26. Sherwani SI, Khan HA, Ekhzaimy A, Masood A, Sakharkar MK. Significance of HbA1c test in diagnosis and prognosis of diabetic patients. Biomark Insights. 2016;11:95–104.
27. Murakami H, Kawakami R, Nakae S, Yamada Y, Nakata Y, Ohkawara K, et al. Accuracy of 12 wearable devices for estimating physical activity energy expenditure using a metabolic chamber and the doubly labeled water method: validation study. JMIR mHealth uHealth. 2019;7:e13938.
28. Goldstein CM, Goldstein SP, Thomas DM, Hoover A, Bond DS, Thomas JG. The Behavioral Intervention with Technology for E-Weight Loss Study (BITES): incorporating energy balance models and the bite counter into an online behavioral weight loss program. J Technol Behav Sci. 2020;6:406–18.
29. Dhurandhar NV, Schoeller D, Brown AW, Heymsfield SB, Thomas D, Sorensen TI, et al. Energy balance measurement: when something is not better than nothing. Int J Obes. 2015;39:1109–13.
30. Schoeller DA, Thomas D, Archer E, Heymsfield SB, Blair SN, Goran MI, et al. Self-report-based estimates of energy intake offer an inadequate basis for scientific conclusions. Am J Clin Nutr. 2013;97:1413–5.
31. Lichtman SW, Pisarska K, Berman ER, Pestone M, Dowling H, Offenbacher E, et al. Discrepancy between self-reported and actual caloric intake and exercise in obese subjects. N Engl J Med. 1992;327:1893–8.
32. Heymsfield SB, Peterson CM, Thomas DM, Hirezi M, Zhang B, Smith S, et al. Establishing energy requirements for body weight maintenance: validation of an intake-balance method. BMC Res Notes. 2017;10:220.
33. Hall KD, Guo J, Chen KY, Leibel RL, Reitman ML, Rosenbaum M, et al. Methodologic considerations for measuring energy expenditure differences between diets varying in carbohydrate using the doubly labeled water method. Am J Clin Nutr. 2019;109:1328–34.

34. Baracos V, Caserotti P, Earthman CP, Fields D, Gallagher D, Hall KD, et al. Advances in the science and application of body composition measurement. J Parenter Enter Nutr. 2012;36:96–107.
35. Barber J, Palmese L, Chwastiak LA, Ratliff JC, Reutenauer EL, Jean-Baptiste M, et al. Reliability and practicality of measuring waist circumference to monitor cardiovascular risk among community mental health center patients. Community Ment Health J. 2014;50:68–74.
36. Schoeller DA. A novel carbon isotope biomarker for dietary sugar. J Nutr. 2013;143:763–5.
37. Taguchi YI, Ki T. *Tosa nikki yōkai*, Shohan. edn Yūseidō: Tōkyō, 1955.
38. Lennox KP, Glascoe LG. A Bayesian measurement error model for misaligned radiographic data. Technometrics. 2013;55:450–60.
39. Fitzmaurice GMJN. Measurement error and reliability. 2002; 18 1: 112–4.
40. Buonaccorsi JP. Measurement error models, methods, and applications. In. Boca Raton: CRC Press; 2010.
41. Viswanathan M. Measurement error and research design. In. Thousand Oaks: Sage Publications; 2005.
42. Baller D, Thomas DM, Cummiskey K, Bredlau C, Schwartz N, Orzechowski K, et al. Gestational growth trajectories derived from a dynamic fetal-placental scaling law. J R Soc Interface. 2019;16:20190417–20190417.
43. FAIR principles. In: GO FAIR. Berlin, Germany; 2022.
44. Sazonov E, Schuckers S, Lopez-Meyer P, Makeyev O, Sazonova N, Melanson EL, et al. Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. Physiol Meas. 2008;29:525–41.
45. Alex J, Turner D, Thomas DM, McDougall A, Halawani MW, Heymsfield SB, et al. Bite count rates in free-living individuals: new insights from a portable sensor. BMC Nutr. 2018;4:23.
46. Martin CK, Miller AC, Thomas DM, Champagne CM, Han H, Church T. Efficacy of SmartLoss, a smartphone-based weight loss intervention: results from a randomized controlled trial. Obes (Silver Spring). 2015;23:935–42.
47. Fisher JO, Johnson RK, Lindquist C, Birch LL, Goran MI. Influence of body composition on the accuracy of reported energy intake in children. Obes Res. 2000;8:597–603.
48. Ford ES, Dietz WH. Trends in energy intake among adults in the United States: findings from NHANES. Am J Clin Nutr. 2013;97:848–53.
49. Hoffman L. Multilevel models for examining individual differences in within-person variation and covariation over time. Multivar Behav Res. 2007;42:609–29.
50. Gianfagna L, Di Cecco A. Explainable AI: needs, opportunities, and challenges. In: Gianfagna L, Di Cecco A (eds). *Explainable AI with Python*. Springer International Publishing: Cham; 2021, p. 27–46.
51. Li G, Hari SKS, Sullivan M, Tsai T, Pattabiraman K, Emer J, et al. Understanding error propagation in deep learning neural network (DNN) accelerators and applications. In: *Proc international conference for high performance computing, networking, storage and analysis*. Denver, Colorado: Association for Computing Machinery, 2017. Article 8.
52. Bharathi R, Selvarani R. A machine learning approach for quantifying the design error propagation in safety critical software system. IETE J Res. 2022;68:467–81.
53. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. In. 3rd_Edition ed: Lippincott Williams & Wilkins; 2008. p. 1–758.
54. Kight CE, Bouche JM, Curry A, Frankenfield D, Good K, Guenter P, et al. Consensus recommendations for optimizing electronic health records for nutrition care. Nutr Clin Pract: Off Publ Am Soc Parenter Enter Nutr. 2020;35:12–23.
55. Everitt B. The Cambridge dictionary of statistics. In. 3rd ed. ed. Cambridge, UK;: Cambridge University Press; 2006.
56. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods. 2016;13:703–4.
57. Gollapudi S. Practical machine learning: tackle the real-world complexities of modern machine learning with innovative and cutting-edge techniques. In: Laxmikanth V, (ed).
58. Löffler-Wirth H, Willscher E, Ahnert P, Wirkner K, Engel C, Loeffler M, et al. Novel anthropometry based on 3D-bodyscans applied to a large population-based cohort. PLoS One. 2016;11:e0159887.
59. Anthony M. Neural network learning: theoretical foundations. In: Bartlett PL, (ed). Cambridge, U.K.: Cambridge University Press; 1999.
60. Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. J Choice Model. 2018;28:167–82.
61. Cho J, Lee K, Shin E, Choy G, Do SJAL. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. 2015.
62. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Springer: New York, NY; 2009.
63. Delisle Nyström C, Henriksson H, Alexandrou C, Bergström A, Bonn S, Bälter K, et al. Validation of an online food frequency questionnaire against doubly labelled water and 24 h dietary recalls in pre-school children. *Nutrients*. 2017;9:66.
64. Cordeiro F, Epstein DA, Thomaz E, Bales E, Jagannathan AK, Abowd GD, et al. Barriers and negative nudges: exploring challenges in food journaling. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. Seoul, Republic of Korea: Association for Computing Machinery, 2015. p. 1159–62.
65. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
66. Li P, Stuart EA. Best (but oft-forgotten) practices: missing data methods in randomized controlled nutrition trials. Am J Clin Nutr. 2019;109:504–8.
67. Parr CL, Hjartåker A, Scheel I, Lund E, Laake P, Veierød MB. Comparing methods for handling missing values in food-frequency questionnaires and proposing k nearest neighbours imputation: effects on dietary intake in the Norwegian Women and Cancer study (NOWAC). Public Health Nutr. 2008;11:361–70.
68. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30:377–99.
69. Ichikawa M, Hosono A, Tamai Y, Watanabe M, Shibata K, Tsujimura S, et al. Handling missing data in an FFQ: multiple imputation and nutrient intake estimates. Public Health Nutr. 2019;22:1351–60.
70. Kupek E, de Assis MA. The use of multiple imputation method for the validation of 24-h food recalls by part-time observation of dietary intake in school. Br J Nutr. 2016;116:904–12.
71. Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. Stat Med. 2009;28:3657–69.
72. Rahman SA, Huang Y, Claassen J, Heintzman N, Kleinberg S. Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. J Biomed Inform. 2015;58:198–207.
73. Lin J-H, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform. 2008;41:1–14.
74. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8:6085.
75. Modeling Missing Data in Clinical Time Series with RNNs. *Machine Learning for Healthcare*; Saban Research Institute; 2016.
76. DeSisto CL, Kim SY, Sharma AJ. Prevalence estimates of gestational diabetes mellitus in the United States, Pregnancy Risk Assessment Monitoring System (PRAMS), 2007–2010. Prev Chronic Dis. 2014;11:E104–E104.
77. Setzu M, Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F. GLocalX—from local to global explanations of black box AI models. Artif Intell. 2021;294:103457.
78. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
79. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. J Investig Dermatol. 2018;138:2108–10.
80. Holzinger A, Kieseberg P, Tjoa AM, Weippl E (eds). Explainable AI: the new 42? *Machine Learning and Knowledge Extraction*. Cham. Springer International Publishing; 2018.
81. Nawaz M, Yan H. Saliency detection using deep features and affinity-based robust background subtraction. IEEE Trans Multimed. 2021;23:2902–16.
82. Kumar LA, Jayashree LS, Manimegalai R (eds). Visual importance identification of natural images using location-based feature selection saliency map. *Proceedings of international conference on artificial intelligence, smart grid and smart city applications*. Cham. Springer International Publishing; 2020.
83. Gianfagna. Explainable AI with Python. In: Gianfagna L (ed).
84. Tahir GA, Loo CK. Explainable deep learning ensemble for food image analysis on edge devices. Comput Biol Med. 2021;139:104972.
85. Martin CK, Han H, Coulon SM, Allen HR, Champagne CM, Anton SD. A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method. Br J Nutr. 2009;101:446–56.
86. Hossain D, Imtiaz MH, Ghosh T, Bhaskar V, Sazonov E. Real-time food intake monitoring using wearable egocnetric camera. In: *Proceedings of the annual international conference of the IEEE engineering in medicine and biology society. IEEE Engineering in Medicine and Biology Society*. Annual International Conference. 2020;2020:4191–5.
87. Doulah A, Ghosh T, Hossain D, Imtiaz MH, Sazonov E. "Automatic Ingestion Monitor Version 2"—a novel wearable device for automatic food intake detection and passive capture of food images. IEEE J Biomed Health Inform. 2021;25:568–76.

ACKNOWLEDGEMENTS

DMT, SK, MC, NC, RB, and CM were supported by NIH U54TR004279. AB was supported by NIH R25DK099080, NIH R25GM141507, and NIH R25HL124208. PS was supported by DARPA XAI FA8750-17-2-0146. Statements in the manuscript do not necessarily represent the official views of, or imply endorsement by, the National Institute of Health, AHRQ, or the US Department of Health and Human Services. The views expressed in this work are those of the authors and do not reflect the official policy or position of the United States Military Academy, Department of the Army, or the Department of Defense. The authors do not have any conflicts of interest to declare.

AUTHOR CONTRIBUTIONS

DMT conceived this study and prepared the original and final manuscript draft for this study. SK drafted the section on missing data. AWB co-drafted the introduction and the statistical section. MC reviewed the statistical rigor of the statistical sections. NDB reviewed the rigor of the machine learning sections and the checklist. NR, RL, and TK drafted the statistical sections of the manuscript. PS drafted the XAI section. RB, SS, DR, and CM drafted the machine learning sections and developed the first draft of the checklist. Nicholas Clark reviewed all sections of the manuscript and added missing content. All authors reviewed multiple versions of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Diana M. Thomas.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.