frontiers | Research Topics

Databases and nutrition, volume II

Edited by
Alessandra Durazzo, Massimo Lucarini
and Igor Pravst

Published in
Frontiers in Nutrition

ISSN 1664-8714
ISBN 978-2-8325-4083-1
DOI 10.3389/978-2-8325-4083-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Databases and nutrition, volume II

Topic editors
Alessandra Durazzo — Research Centre for Food and Nutrition, Council for Agricultural Research and Economics, Italy
Massimo Lucarini — Research Centre for Food and Nutrition, Council for Agricultural Research and Economics, Italy
Igor Pravst — Institute of Nutrition, Slovenia

Citation
Durazzo, A., Lucarini, M., Pravst, I., eds. (2023). *Databases and nutrition, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4083-1

Table of contents

frontiers | Frontiers in Nutrition

TYPE Editorial
PUBLISHED 22 November 2023
DOI 10.3389/fnut.2023.1307370

Editorial: Databases and nutrition, volume II

Alessandra Durazzo[1]*, Igor Pravst[2]* and Massimo Lucarini[1]*

[1]CREA-Research Centre for Food and Nutrition, Rome, Italy, [2]Nutrition Institute, Ljubljana, Slovenia

KEYWORDS

food data, food groups, nutrients, natural substances, dietary supplements, classification, categorization, food composition databases

Editorial on the Research Topic
Databases and nutrition, volume II

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Angela M. Zivkovic,
University of California, Davis, United States

*CORRESPONDENCE
Alessandra Durazzo
✉ alessandra.durazzo@crea.gov.it
Igor Pravst
✉ igor.pravst@nutris.org
Massimo Lucarini
✉ massimo.lucarini@crea.gov.it

RECEIVED 04 October 2023
ACCEPTED 03 November 2023
PUBLISHED 22 November 2023

CITATION
Durazzo A, Pravst I and Lucarini M (2023) Editorial: Databases and nutrition, volume II. *Front. Nutr.* 10:1307370. doi: 10.3389/fnut.2023.1307370

Introduction

This Research Topic is dedicated to covering high-level aspects of "Databases and Nutrition" from a global and interdisciplinary perspective and interoperability as a tool toward improved populational health. Studies that examine the relationship between diet and health have led to increased interest in nutrients and other biologically active constituents in food, and data on these and other compounds are increasingly required in the database system.

Detailed knowledge of the presence and levels of nutrients and bioactive components in food is the basis for data storage and structuration into databases.

Databases around food and nutrition represent fundamental resources and tools in many fields, such as nutrition, food science, public health, and healthcare, by supporting human research studies, policymaking, and consumer education.

The development of databases, based on foods and beverages consumed by the population, is based on efforts dedicated to measuring, collecting, and integrating data with information and detailed documentation on food and methodologies, etc., taking into account the current needs and demands, i.e., environmental aspects, lifestyle changes, and global market trade.

A lot of efforts have been made to harmonize food nutritional databases and repositories throughout worldwide projects and networks, leading to standard methodologies and guidelines, and also to promote applications of food classification and description systems and advances in systems for ontology alignment.

The standardization, harmonization, and FAIRization of data are being reached throughout automatized technologies, innovative systems, and digital tools for organizing and exploiting food data into various applications. Integrating and linking information and data from different sources, i.e., food, environmental, nutrition, and health ones, lead to the development of a comprehensive, multidimensional resource -based on integrating modeling, a digital platform, and cloud space at the multidimensional and multisource level-that can be used across disciplines.

Thirteen articles are published in the collection of articles under the Research Topic "*Databases and Nutrition - Volume 2*".

Ahmed et al. presented the Food Label Information Program (FLIP), a comprehensive data approach for the evaluation of the Canadian food supply, and the latest methods used in the development of this database. Ferraz de Arruda et al. discussed vegetable oils as a case study of food composition databases in the era of big data.

Gilbert et al. presented an algorithm-based mapping of products in a Canadian branded food and beverage database to their equivalents in Health Canada's Canadian Nutrient File. The study of Balakrishna et al. addressed classifications of food items for health requirements and nutrition guidelines using Gaussian mixture models. The study by Liu et al. presented a novel food-components-target-function (FCTF) evaluation and prediction model for food efficacy based on association rule mining. Endaltseva et al. used an eater-oriented knowledge framework for reducing salt and dietary sodium intake and reviewed and presented an interdisciplinary documentary base of dietary sodium consumption factors.

It is also worth mentioning the study of Malcomson et al., which presented the operationalization of a standardized scoring system to assess adherence to the World Cancer Research Fund and American Institute for Cancer Research cancer prevention recommendations in the UK biobank.

Vlassopoulos et al. reported the performance of Nutri-Score in branded foods in Greece. Building on food composition data, Hribar et al. presented a validation of the food frequency questionnaire for the assessment of dietary vitamin D intake. The investigation by Shabnam focused on the assessment of non-linearity in the calorie–income relationship in Pakistan.

Davison et al. showed how lower energy-adjusted nutrient intakes occur among food energy under-reporters with poor mental health. Another study by Tang et al. focused on the intake of dietary fiber and femoral bone mineral density among middle-aged and older US adults from a cross-sectional study of the National Health and Nutrition Examination Survey 2013–2014.

Mognard et al. reviewed and explored "Eating Out", spatiality, temporality, and sociality, and presented a database for China, Indonesia, Japan, Malaysia, Singapore, and France.

This collection of research articles published as part of the Research Topic collection presented not only innovative approaches for the collection and management of big data in the area of nutrition but also demonstrated the use of such data for progress in nutrition research. The collection also demonstrated several opportunities that should be addressed with future studies. In the past, essential nutrients were the primary target for most nutrient and food datasets, with the result that databases have generally lacked detailed information about other food constituents, including bioactives and other components, and the effects of different processing techniques on their content in the resulting foods. Furthermore, the public health challenges of the highly populated and industrialized environment also highlight that other components in foods, including contaminants, should be taken into account.

Author contributions

AD: Writing—original draft. IP: Writing—original draft. ML: Writing—original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We would like to acknowledge the authors and reviewers of the publications in this Research Topic for their invaluable contributions and efforts. We are also grateful to the editorial board members and support staff of the journal for their kind support during the preparation of this Research Topic.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Nutrition

ORIGINAL RESEARCH
published: 03 February 2022
doi: 10.3389/fnut.2021.825050

Development of the Food Label Information Program: A Comprehensive Canadian Branded Food Composition Database

*Mavra Ahmed[1], Alyssa Schermel[1], Jennifer Lee[1], Madyson Weippert[1], Beatriz Franco-Arellano[1,2] and Mary L'Abbé[1]**

[1] Department of Nutritional Sciences, Temerty Faculty School of Medicine, University of Toronto, Toronto, ON, Canada,
[2] Faculty of Health Sciences, Ontario Tech University, Oshawa, ON, Canada

Objectives: Traditional methods for creating food composition databases struggle to cope with the large number of products and the rapid pace of turnover in the food supply. This paper introduces Food Label Information Program (FLIP), a big data approach to the evaluation of the Canadian food supply and presents the latest methods used in the development of this database.

Methods: The Food Label Information Program (FLIP) is a database of Canadian food and beverage package labels by brand name. The latest iteration of the FLIP, FLIP 2020, was developed using website "scraping" to collect food labeling information (e.g., nutritional composition, price, product images, ingredients, brand, etc.) on all foods and beverages available on seven major Canadian e-grocery retailer websites between May 2020 and February 2021.

Results: The University of Toronto's Food Label Information Program (FLIP) 2020 was developed in three phases: Phase 1, database development and enhancements; Phase 2, data capture and management of food products and nutrition information; Phase 3, data processing and food categorizing. A total of 74,445 products available on websites of seven retailers and 2 location-specific duplicate retailers were collected for FLIP 2020. Of 57,006 food and beverage products available on seven retailers, nutritional composition data were available for about 60% of the products and ingredients were available for about 45%. Data for energy, protein, carbohydrate, fat, sugar, sodium and saturated fat were present for 54–65% of the products, while fiber information was available for 37%. Food products were classified under multiple categorization systems, including Health Canada's Table of Reference Amounts, Health Canada's sodium categories for guiding benchmark sodium levels, sugar-focused categories and categories specific to various global nutrient profiling models.

Conclusions: FLIP is a powerful tool for evaluating and monitoring the Canadian food supply environment. The comprehensive sampling and granularity of collection provides

Edited by:
Massimo Lucarini,
Council for Agricultural Research and Economics, Italy

Reviewed by:
Maria Kapsokefalou,
Agricultural University of Athens, Greece
Aida Turrini,
Independent Researcher, Rome, Italy

****Correspondence:***
Mary L'Abbé
mary.labbe@utoronto.ca

Specialty section:
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

Received: *29 November 2021*
Accepted: *24 December 2021*
Published: *03 February 2022*

Citation:
Ahmed M, Schermel A, Lee J, Weippert M, Franco-Arellano B and L'Abbé M (2022) Development of the Food Label Information Program: A Comprehensive Canadian Branded Food Composition Database. Front. Nutr. 8:825050. doi: 10.3389/fnut.2021.825050

power for revealing analyses of the relationship between nutritional quality and marketing of branded foods, timely observation of product reformulation and other changes to the Canadian food supply.

Keywords: branded food database, food composition, packaged foods, artificial intelligence, Optical Character Recognition, nutrition facts table, web-scraping

INTRODUCTION

The study of nutritional epidemiology relies on understanding the association between nutrient consumption and health outcomes and usually involves monitoring the nutritional quality of food consumed by a population (1). Thus, such studies rely on the assessments of dietary intakes based on the collection of nutrition information from food composition tables or databases (1). Packaged foods and beverages represent a major segment of the food supply, providing approximately two-thirds of energy intake (2, 3). Despite the dominant role of these packaged foods and beverages in the diets of populations, existing food composition databases are limited in their ability to capture accurate nutrient content information for specific products due to the complex and dynamic nature of the national food supplies (4–6).

National food composition databases are expensive to develop, construct and maintain (7, 8). The packaged food and beverage sector includes a wide array of products and is characterized by continuous changes and turnover due to introduction of new products, reformulation or discontinuation of others (8, 9). Most national food composition databases include aggregate nutrition information for only a limited number of generic food items. For example, to assess dietary intakes of Canadians, researchers rely on the Canadian Nutrient File (10) to estimate the dietary intake component of dietary data, including 24-h recalls, food frequency questionnaires (11, 12), and other national nutrition surveys, [e.g., Canadian Community Health Survey (CCHS) (13)]. The CNF database is the standard reference food composition database developed and maintained by the Government of Canada, and is used by a number of Government of Canada agencies including Statistics Canada, Health Canada, Agriculture and Agri-Food Canada, and the Canadian Food Inspection Agency (10), industry and researchers. Nutrient information for all food and beverages reported by CCHS participants comes from the CNF, which is composed of nutrient profiles for about 6,000 products that are primarily generic representative composites, where more than half are recipe-based foods/beverages based on common preparation methods, rather than individual food/beverage items (14). However, using the CNF to analyze changes in the food supply or Canadian population dietary intakes poses several challenges due to its lack of scheduled comprehensive and systematic updating, the use of some non-Canadian food composition data, and aggregated data for packaged foods.

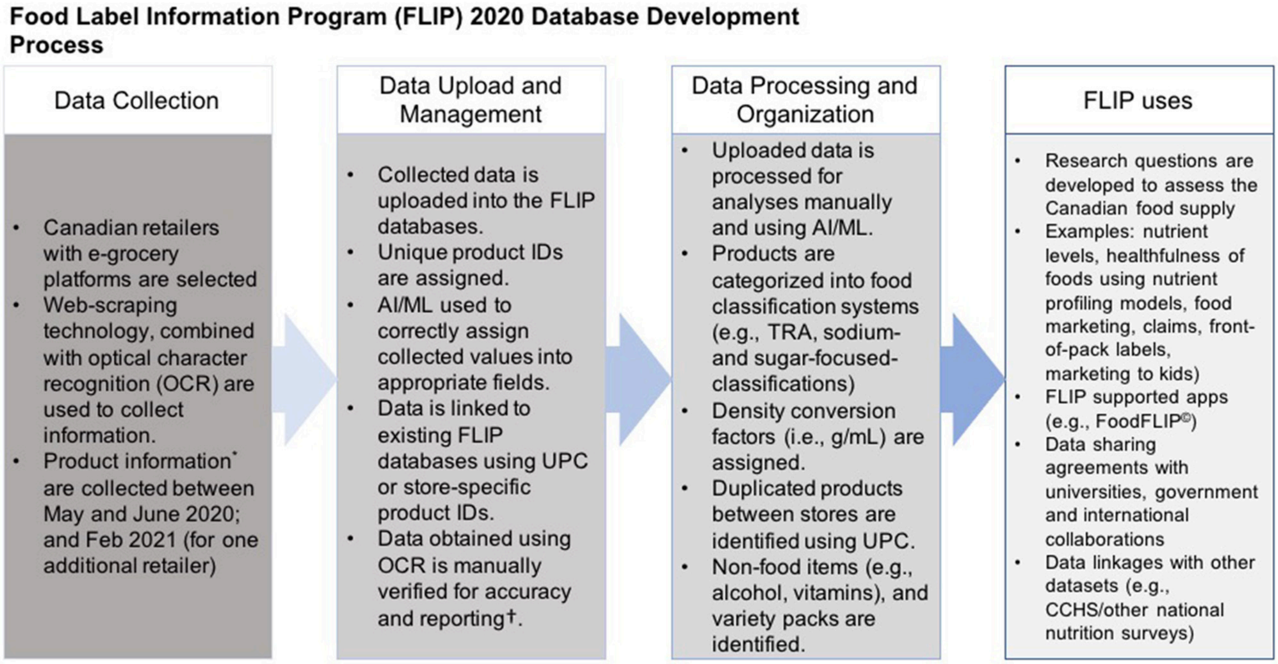
The packaged food and beverage industry is also characterized by fast-moving continuous turnover as new products are introduced and/or reformulated, some to replace less-favored or discontinued products (5, 6, 15–18). These continuous changes require food composition databases to be updated frequently; however, lack of resources limits the updating of the nutritional composition of all foods in many food databases, especially for foods found in the generic CNF. Globally, there have been various attempts to collate such large-scale up-to-date nutritional data on a comprehensive set of foods: for example, by crowdsourcing food label data using mobile phones [e.g., FoodSwitch in Australia (19)] or web applications [e.g., Open Food Facts in the United States (20)]; collecting data through contact with food manufacturers/industry (e.g., the USDA Global Branded Food Products Database) (21–23); or periodic audits of the foods on the market (e.g., FoodDB extracted weekly nutrition information on products using web-scraping in the United Kingdom) (24). However, there are limited developments of food composition databases that achieve comprehensive coverage of the Canadian supply system with brand-specificity and regularly up-to-date, extensive nutrition information of food products.

To address these research gaps, we aimed to develop a product- and brand-specific comprehensive database containing nutrition information for a diverse array of packaged foods and beverages in the Canadian food supply. Such a database allows for identifying important levers for promoting healthy diets, prioritizing nutritional interventions for public health policy, evaluating the impact of population-level policies such as Sodium Reformulation (25) or Trans Fat Ban (26) and effects of future policy interventions such as front-of-pack labeling (FOPL) or Marketing to Kids (M2K). It can also be used in national nutrition surveys to access the nutritional quality of diets of Canadians and assess the changes in nutritional composition of the food supply in response to policy or other changes such as the COVID-19 pandemic.

Previously, we have developed three versions of FLIP datasets in 2010, 2013, and 2017, described in detail elsewhere (26–28). However, in these previous versions, data were collected in stores using a smartphone application in 2013 and 2017, while in 2010, data from the products (e.g., nutrition information) were manually entered into the database/website. Given the common usage of big data techniques in collecting, storing, processing and analyzing data, now applied in many fields across non-profit, scientific, business and public sectors, this paper introduces FLIP 2020, an Artificial Intelligence (AI)-enhanced/powered Optical Character Recognition (OCR) (AI-enhanced OCR) approach to the collection and evaluation of the Canadian packaged food and beverage supply and presents the methods used in the development of this database.

METHODS

The University of Toronto's Food Label Information Program (FLIP) 2020 was developed in three phases: Phase 1, database



Food Label Information Program (FLIP) 2020 Database Development Process

Data Collection

- Canadian retailers with e-grocery platforms are selected
- Web-scraping technology, combined with optical character recognition (OCR) are used to collect information.
- Product information* are collected between May and June 2020; and Feb 2021 (for one additional retailer)

Data Upload and Management

- Collected data is uploaded into the FLIP databases.
- Unique product IDs are assigned.
- AI/ML used to correctly assign collected values into appropriate fields.
- Data is linked to existing FLIP databases using UPC or store-specific product IDs.
- Data obtained using OCR is manually verified for accuracy and reporting†.

Data Processing and Organization

- Uploaded data is processed for analyses manually and using AI/ML.
- Products are categorized into food classification systems (e.g., TRA, sodium- and sugar-focused-classifications)
- Density conversion factors (i.e., g/mL) are assigned.
- Duplicated products between stores are identified using UPC.
- Non-food items (e.g., alcohol, vitamins), and variety packs are identified.

FLIP uses

- Research questions are developed to assess the Canadian food supply
- Examples: nutrient levels, healthfulness of foods using nutrient profiling models, food marketing, claims, front-of-pack labels, marketing to kids)
- FLIP supported apps (e.g., FoodFLIP©)
- Data sharing agreements with universities, government and international collaborations
- Data linkages with other datasets (e.g., CCHS/other national nutrition surveys)

FIGURE 1 | Phases of development of Food Label Information Program (FLIP) 2020. FLIP 2010, 2013, and 2017 are previous versions of FLIP.

development and enhancements; Phase 2, data capture and management of food products and nutrition information (i.e., Nutrition Facts table (NFt); Phase 3, data processing and food categorization (**Figure 1**).

Overview of Food Label Information Program Database

The FLIP is a database of Canadian pre-packaged food and beverage package labels by brand name that was updated every 3–4 years at the University of Toronto (UofT), Toronto, ON, Canada. The purpose of the FLIP is to provide comprehensive food product nutrition information to allow for assessment and monitoring over time. To date, three previous versions of the FLIP datasets have been completed in 2010, 2013, and 2017 and are described in detail elsewhere (26–28). Briefly, data for FLIP 2010, FLIP 2013 and FLIP 2017 were collected in person in stores and stored on the FLIP website, a database (website) to collect, process, store and manage the data. FLIP 2010/11 (n = 10,487) (27) was collected in the Greater Toronto Area (GTA) and Calgary between March 2010 and April 2011; FLIP 2013 (n = 15,342) (28) data were collected in the GTA, Ottawa and Calgary during May to September 2013; and FLIP 2017 (n = 17,671) (26) data were collected during July to September 2017 in the GTA. These collections represented 56% (27), 75% (28), and 68% (29) of grocery retail sales in Canada, respectively for 2010, 2013, and 2017. All nationally available and private-label brands, but excluding seasonal products, were collected from major retailers (Loblaws, Metro, Sobeys and Safeway in 2010 and 2013; Loblaws, Metro and Sobeys in 2017). While inclusion criteria for foods and beverages changed little between each FLIP version, data were collected using a smartphone application in 2013 and 2017, while in 2010, data from the products (e.g., nutrition information) were manually entered into the database/website (**Table 1**).

The FLIP website enables users to generate data outputs and reports in Microsoft Excel for further analyses. The FLIP website contains a user tutorial, user guides, and a dashboard with the FLIP version number and details on the latest updates. The information captured on each product is described in **Table 2**.

Data Security

The FLIP is hosted on a cloud-based infrastructure located in Virginia, USA and Quebec, Canada. Raw data for each product page with date and time of data collection is stored separately for audit and data verification purposes, and to provide a mechanism for re-extracting data in the event that data was previously extracted incorrectly, or additional data points are required. At present, the FLIP website is available to the L'Abbé Lab nutritional sciences researchers at the University of Toronto, as well as national and international university and government researchers with whom the University of Toronto has set up data sharing agreements.

Phase 1: Food Label Information Program Database Development and Enhancements

FLIP 2020 Data Collection

The latest phase, FLIP 2020, is described in this manuscript. The FLIP 2020 contains nutrition information for 74,445 product listings, representing 48,829 unique universal product codes (UPC). Food information from the leading grocery retailers

TABLE 1 | History of the Food Label Information Program (FLIP) Database collections and updates.

FLIP database	Collection period	*n**	Number of stores	Collected product variables/information	Collection method
2010–11	March 2010–April 2011	10,487	4	Name, Brand, Company, Container size, NFt, UPC, Marketing information[‡]	• Food packages purchased for data collection. • Variables of interest were manually entered in Microsoft Excel. • Uploaded to FLIP cloud database following the 2013 collection.
2013	May–Sept	15,342	4	Name, Brand, Company, Container size, NFt, UPC, Marketing information[‡] Ingredients List, Photos of all sides of packages	• iPhone app development for digital collection of food package images in stores. • Database software development using Cloud storage. • OCR software development to automate NFt and ingredients list data entry. • Excel report generation capabilities added.
2017	July–Sept	19,267	3	Name, Brand, Company, Container size, NFt, UPC, Marketing information[‡], Ingredients List, Price (regular & sale), Photos of all sides of packages	• Upgraded technology capabilities, including ability to update databases using Excel. • Automated linking & matching products between databases using UPC or store-specific product codes • Development of algorithms for food categorization and nutrient profiling
2020–21	May 2020–Feb 2021	74,445	9[†]	Name, Brand, Container size, NFt, UPC, Ingredients List, Price (regular & sale), Photo of front of package (if available)	• Web scraping to collect all product information • AI-enhanced OCR technology to collect all product information. • Automated linking & matching UPCs / store-specific codes between FLIP 2020 and FLIP 2017, and between stores within FLIP 2020

**Sample sizes in the FLIP 2010–11, 2013 reflect unique products, while the sample size in FLIP 2017 also includes multiple package sizes and FLIP 2020 includes multiple package sizes and duplicates across stores.*

[†] *Data was collected from seven retailers plus two location-specific stores.*

[‡] *Marketing information included nutrient content claims, health claims, front-of-pack labeling, and children's marketing. Tabs and options can be and have been expanded over time, depending on research needs.*

FLIP, Food Label Information Program; NFt, Nutrition Facts table; OCR, optical character recognition; UPC, universal product code.

in Canada with online information were acquired from their respective websites and digitalized to enhance ease and efficiency of collection and analysis. Food composition database software (University of Toronto, Toronto, ON, Canada) (web and mobile) was developed for FLIP 2020, resulting in a shorter and more efficient food collection and data processing approach (**Table 1**).

Data was acquired from the websites of seven Canadian retailers (Costco®, Costco Wholesale Canada Ltd., Nepean, ON, Canada; Grocery Gateway by Longo's, Longo's Brothers Fruit Markets Inc., Empire Company Ltd., Stellarton, NS, Canada; Loblaws®, Loblaws Companies Ltd., Brampton, ON, Canada; Metro, Metro Inc., Montreal, QB, Canada; No Frills®, Loblaws Companies Ltd., Brampton, ON, Canada; Voilà by Sobeys, Empire Company Ltd., Stellarton, NS, Canada; and Walmart, Walmart Canada Corp., Mississauga, ON, Canada), representing over 80% of the grocery retail market share (30, 31). Data were collected between May and June 2020, and in February 2021 (the latter for Voilá due to the lack of e-commerce availability during the initial scraping period). Two additional websites of two retailers (Loblaws and No Frills), located in a populated metropolitan Toronto area, were selected for additional data acquisition for further analysis of the e-commerce food environment.

Food and beverage product information was captured using a website "scraping" [webscraping, which is an automated process used for extracting data from websites implemented using a bot or a web crawler (24)] program developed in Python. Each e-retailer's online website was first scanned to get a general outline of how the product information is stored on the website, followed by a Python-based routine to locate the hyperlinks of product pages. Once the hyperlinks are located, each product page was loaded, and its data was extracted into FLIP. The scraping was customized for each website. Developers analyzed the structure of the webpages, looking for common patterns to the way the data was displayed for each product. Random pages were selected for manual comparison to the results. Data that didn't make sense once imported into FLIP was compared to detailed logs captured during the scraping process. In all cases where the websites displayed data that was inconsistent with the data captured during scraping, it was confirmed by viewing the detailed logs. The data was then further processed with algorithms developed in C# programming language. A set of core classes and helper

TABLE 2 | Information collected, managed and processed on the FLIP website.

FLIP website tabs	Description of information
Description	Product ID, company, brand, name, preparation required, variety pack, TRA categories, ingredients, sampling date, store code, container size and price
Barcode/UPC	Barcode/UPC, sample date, store code and linkages
Nutrition facts	Collected data: Serving size, weight/volume, nutrient contents as identified on the package Nutrition Facts table (amount and %DV) (Kcal, Fat, saturated fat, trans fat, cholesterol, sodium, potassium, carbohydrates, fiber, sugar, free sugar, protein, vitamin A, vitamin C, calcium, iron) Calculated data: weight/volume conversions (g/mL and mL/g e.g., density), as prepared nutrition information captured in 2013, 2017, nutrition information per 100 g
Marketing	Children's marketing, nutrient content claims, other claims, disease risk reduction claims, front of pack symbols, structure/function claims*
Nutrient profiling	Nutrient profiling and other information used in calculating nutrient profiling scores (e.g., FSANZ, Ofcom, UK TLL, Health Canada Surveillance Tool, FoodFlip (FoodFlip app related nutrient profiling models), added sugar, free sugar, PHO, sweeteners, NOVA Processing, added fats, whole grains*
Sodium [Categories]	Sodium-focused categories
Sugar [Categories]	Sugar-focused categories
Photos	Images of the product including front, back and sides, NFt, ingredient declaration (~8 photos per product as available)
Matches	Matched products with previous versions of FLIP
Log	User-inputted comments on validation and updates to the product

*Tabs and options can be and have been expanded over time, depending on research needs.

FLIP, Food Label Information Program; FSANZ, Food Standards Australia New Zealand; NFt, Nutrition Facts table; TLL, Traffic Light Label; TRA, Table of Reference Amounts; UPC, Universal Product Code.

libraries provided the main functionality for data collection and processing while custom routines were developed to handle each e-retailer's unique website layout, page structures, webpage loading mechanisms, and data formats.

Every food product was collected, including all available national and private label brands, multiple sizes, and all flavors and varieties of a product. Information collected for each product included, where available, the following: product name, UPC, brand, NFt information, ingredients, container size, product image(s) as available, price (regular and sale price), dietary or allergen information (e.g., suitable for vegetarians) (if available on packaging as part of the ingredient list), and date and location/store information of sampling (**Figure 2**). Each product's UPC is used for identification of and tracking unique products over time.

Phase 2: FLIP Data Capture and Management

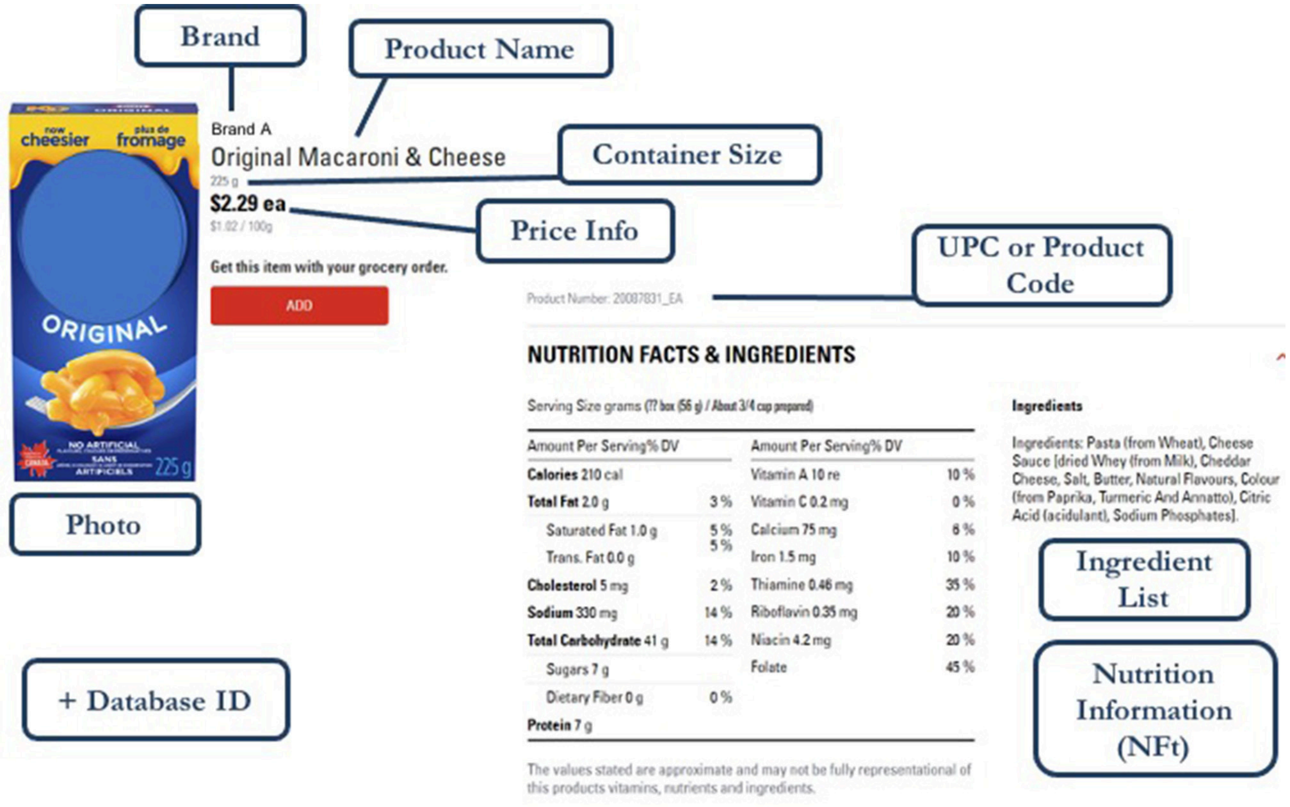
After web-scraping the product information, foods were automatically assigned a product ID (an internal unique identifier) and photos and web data were uploaded onto the FLIP website for data management and processing. Each product's ID is used for identifying and tracking unique products over time. Artificial Intelligence (AI)-enhanced/powered Optical Character Recognition (OCR) (AI-enhanced OCR) technology was used to automatically extract data available in photo format only (e.g., NFt and ingredients list from certain websites). In the AI-enhanced OCR process, each image of the product was scanned for text and a text parsing algorithm determined which image had text that resembled the NFt or ingredient list, followed by extraction of that particular text only. All NFt data extracted from OCR technology ($n = 7{,}400$ products) were manually validated by FLIP staff and students, for this version to determine accuracy of the AI-OCR technology.

Barcodes of food products from Metro, Walmart, and Grocery Gateway and store-specific product numbers from Loblaws and No Frills in FLIP 2020 were matched to those in FLIP 2017 barcodes and store-specific product codes, respectively ($n = 25{,}980$). The no change in barcodes and product codes were used as indicators of no significant product change, therefore, the matching process allowed for any empty data fields in FLIP 2020 to be populated by FLIP 2017 (e.g., company name was not available on websites, but was determined in 2017 from package photos). However, food products from Costco, Walmart, Grocery Gateway, No Frills and Voilá could not be linked to FLIP 2017 as the previous versions of FLIP did not contain any Costco, Walmart, Grocery Gateway and No Frills products and Voilá did not contain any barcodes on the website. All product matches were manually validated by two Research Assistants and the following information was transferred over for the matched products: Table of Reference Amount (TRA) categories, sodium and sugar categories, Company/Parent Company, As Prepared NFts (nutrition facts information as per preparation required as specified on the product packaging but only for products if their NFts were identical), container size, serving size g/mL conversion factors (only if the package information for products was identical) and free sugars.

Barcodes of foods from FLIP 2020 were also linked to identical barcodes from FLIP 2020 from different stores (e.g., Kellogg's Cornflakes Family Size was linked across all webscraped stores given the barcode was identical). If one of these products was missing NFt or ingredients information, its data fields would be populated using a linked product with the most complete data. The FLIP log tracks when data is transferred from one product to another and the source.

Phase 3: Data Processing and Food Categorization

In phase 3, food products were classified using Health Canada's Table of Reference Amounts (TRA) (32, 33), and two additional categories for therapeutic or supplemental products (e.g., meal replacements, nutritional supplements, vitamins) and variety packs (i.e., contain multiple products), followed by other multiple categorization systems (see **Tables 1**, **2**). Health Canada's TRA categories consist of 24 major and 172 sub-categories, as well as an "other" category. Details on TRA



Brand A

Original Macaroni & Cheese

225 g

$2.29 ea

$1.02 / 100g

Get this item with your grocery order.

Product Number: 20087831_EA

NUTRITION FACTS & INGREDIENTS

Serving Size grams (1/? box (56 g) / About 3/4 cup prepared)

Amount Per Serving% DV		Amount Per Serving% DV	
Calories 210 cal		Vitamin A 10 re	10 %
Total Fat 2.0 g	3 %	Vitamin C 0.2 mg	0 %
Saturated Fat 1.0 g	5 %	Calcium 75 mg	6 %
Trans. Fat 0.0 g	5 %	Iron 1.5 mg	10 %
Cholesterol 5 mg	2 %	Thiamine 0.46 mg	35 %
Sodium 330 mg	14 %	Riboflavin 0.35 mg	20 %
Total Carbohydrate 41 g	14 %	Niacin 4.2 mg	20 %
Sugars 7 g		Folate	45 %
Dietary Fiber 0 g	0 %		
Protein 7 g			

The values stated are approximate and may not be fully representational of this products vitamins, nutrients and ingredients.

Ingredients

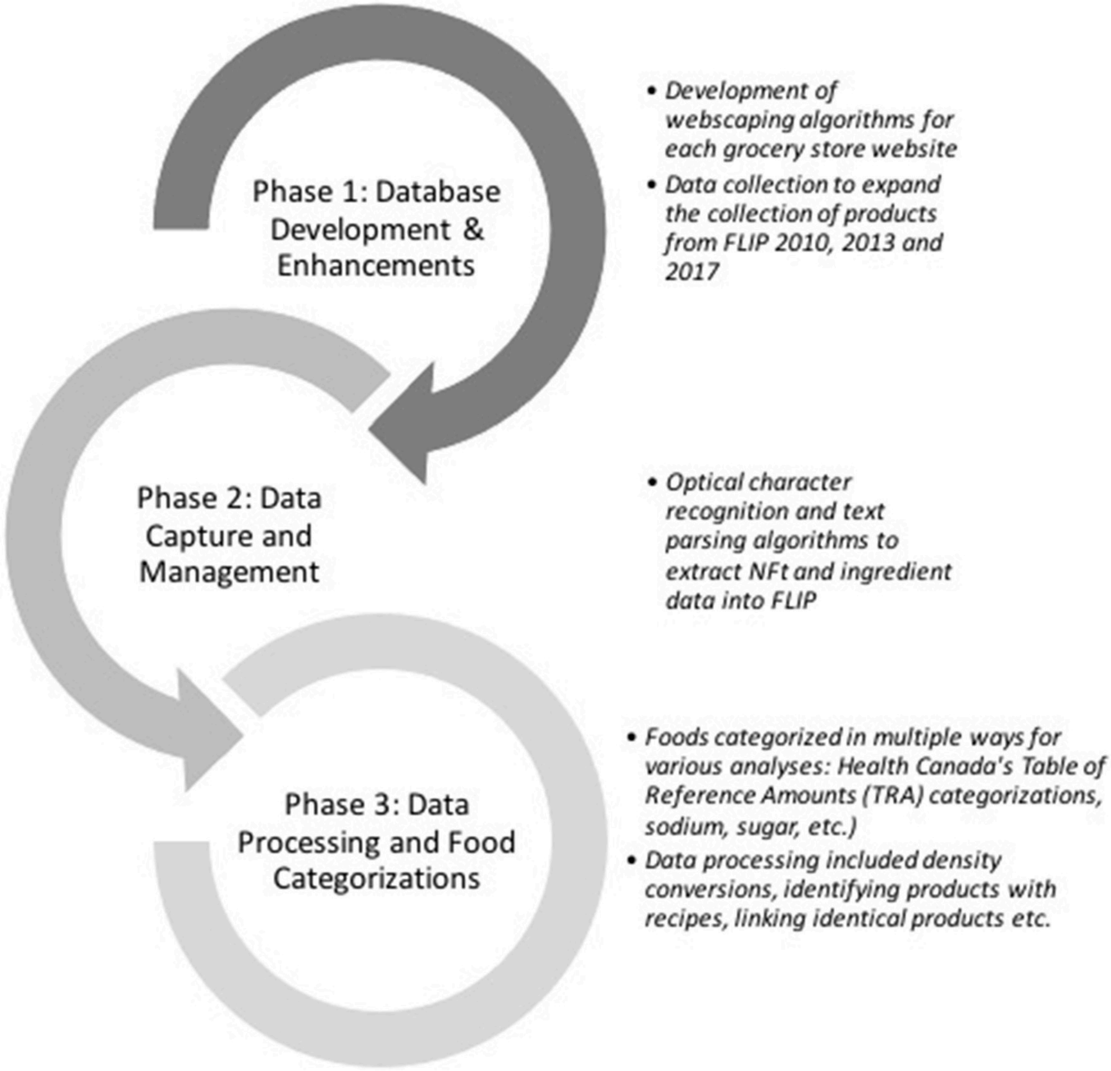
Ingredients: Pasta (from Wheat), Cheese Sauce [dried Whey (from Milk), Cheddar Cheese, Salt, Butter, Natural Flavours, Colour (from Paprika, Turmeric And Annatto), Citric Acid (acidulant), Sodium Phosphates].

FIGURE 2 | Example of the web-scraped information captured from the website of a major Canadian grocery retailer.

categories can be found on Health Canada's website (33). TRA categorizations for unmatched products were applied using predictive algorithms, a method of AI-based estimation. All products with identical product names and brands were grouped together and given a predicted TRA category, powered by Artificial Intelligence/Machine Learning (AI/ML) predictive algorithms. Each product was then manually validated by FLIP staff and students.

Additional automation algorithms were developed for classifying foods into sodium-focused (34) and sugar-focused categories. Sodium-focused food categories were as follows: 13 food group categories, 52 major subcategories and 171 minor subcategories, as published in Health Canada's Guidance for the Food Industry on Reducing Sodium in Processed Foods (34). Sugar-focused categories were created, as described earlier (28, 35), consisting of 19 major food groups, 87 major subcategories, and 252 minor categories. All sugar and sodium categories were mapped manually to the TRA categories using the FLIP 2017 database as a guide. As a second step, keywords for each sodium and sugar category were manually created to assign products to particular categories (e.g., Toaster Strudel or Pop-Tarts as keywords for the category Toaster Pastries). Additional categories specific to various nutrient profiling models were also applied (e.g., FSANZ, Nutri-Score, PAHO etc.) (36–39).

For analyses requiring application of nutrient profiling models [models used to classify foods based on their nutrient composition (40)], foods and beverages in FLIP were categorized using the criteria established by the respective nutrient profiling model, verified independently by two research assistants, and any discrepancy resolved by consensus. The classification of FLIP products into each model's categories was based upon using a combination of information from TRA categories and subcategories (described above), sodium/sugar-focused categories, and the ingredient list. Products were also used to generate a list of foods and beverages with nutrition information and front-of-pack symbols (based on nutrient profiling model) for a FoodFlip© smartphone application, as described in detail elsewhere (41). FoodFlip© app categories consist of categorizing the FLIP database into product specific major categories (n = 19), sub-categories (n = 101) and minor categories (n = 397) in order to allow consumers to easily locate products in consumer-friendly categories. Categorization of foods for FoodFlip© is based on merging Health Canada's TRA categories (33), Canada's sodium reformulation target categories [50], and more specific subsets of food categories [based on the iterative development process as described elsewhere (41)]. Categories were modified if found to be ambiguous or difficult for


Phase 1: Database Development & Enhancements
Phase 2: Data Capture and Management
Phase 3: Data Processing and Food Categorizations

- *Development of webscaping algorithms for each grocery store website*
- *Data collection to expand the collection of products from FLIP 2010, 2013 and 2017*

- *Optical character recognition and text parsing algorithms to extract NFt and ingredient data into FLIP*

- *Foods categorized in multiple ways for various analyses: Health Canada's Table of Reference Amounts (TRA) categorizations, sodium, sugar, etc.)*
- *Data processing included density conversions, identifying products with recipes, linking identical products etc.*

FIGURE 3 | Database development, data management and processing of FLIP 2020. *Product information included product images, serving size, price, nutrition information, and ingredient list. †For this phase, for the purpose of method development and reliability. FLIP, Food Label Information Program (FLIP); NFt; Nutrition facts table.

participants to understand or find during the reliability testing of the app.

For some products, serving sizes reported in milliliters were converted to grams for consistency across all products within a food category. Dependent upon specific research objectives and analyses, the database underwent quality control checks including verification of inputted nutrient contents using Atwater factors (i.e., checking for errors in nutrient declarations in the NFt, as determined by Atwater calculations where nutrients that were >20% from the declared caloric values were checked) and outliers to check for erroneous values.

Additional data extraction or processing, dependent on research objectives and analyses, are ongoing or will be conducted (e.g., application of nutrient profiling models, assessing marketing techniques, identification of nutrition claims and specific ingredients, calculation of free sugars content etc.) (**Figure 3**).

FLIP Database in Other Countries

The development of the smartphone data collector app and web-based software has supported the establishment of FLIP databases in other countries including Argentina, Costa Rica, Paraguay and Peru, called FLIP for Latin American Countries or FLIP-LAC. Data from Argentina (n = 3,724) were collected between August 2017 and May 2018 from three leading groceries stores located in the province of Buenos Aires and Buenos Aires city (42). Costa Rican packaged food label data (n = 6,835) were collected from two grocery stores located in San Jose during and January-August 2018 (43), in addition to pilot data collected in the Summer of 2017. Data in Paraguay (n = 4,091) and Peruvian

TABLE 3 | Frequency and percentage of occurrence of data for nutrients collected in FLIP 2020.

Nutrient	Number of products with nutrient data	% of products with specific nutrient data (n = 73,036 food downloads after removing non-foods)	% of products with specific nutrient data (n = 57,006 food downloads after No Frills and Loblaws duplicates removed)
Energy	47,057	64.4%	58.9%
Total fat	44,479	60.9%	54.4%
Saturated fat	44,405	60.8%	54.2%
Protein	46,200	63.3%	57.4%
Total Carbohydrates	46,436	63.6%	57.8%
Total sugars	50,154	68.7%	64.3%
Fiber	27,239	37.3%	35.2%
Sodium	46,154	63.2%	57.3%
Ingredients	25,196	34.5%	44.2%

data (n = 1,533) were collected during the Summer 2017 and December 2019. The Canadian FLIP or the FLIP-LAC have been used for research, food supply monitoring, policy evaluation and modeling (28, 44–48).

RESULTS

A total of 74,445 products were collected from Metro, Costco, Walmart, Grocery Gateway (Longo's), Loblaws, Loblaw's Maple Leaf Gardens (a specific location in a metropolitan Toronto area), No Frills, No Frills Joe's (a specific location in a metropolitan Toronto area) and Voila. The number of products for each store was as follows: Metro (n = 11,268), Costco (n = 735), Walmart (n = 8,153), Grocery Gateway (n = 9,621), Loblaws (9,428), Loblaw's Maple Leaf Gardens (n = 9,414), No Frills (n = 6,603), No Frills Joe's (n = 6,764), and Voila (n = 12,459). However, food products from Loblaw's Maple Leaf Gardens (n = 9,414) and No Frills Joe's (n = 6,764) were omitted from the current analysis as they are duplicate outlets of the same data discussed in this manuscript. There were 1,261 (from seven retailers) and 1,409 (from nine retailers) non-food products (e.g., food intended solely for children under 4 years of age, meal replacements and nutritional supplements, alcohol), which were removed from further analysis. In total, 25,980 of the FLIP 2020 products (across all stores) were matched to 8,646 of the FLIP 2017 products. FLIP 2020 products may have been matched to multiple 2017 products, and vice versa. Therefore, the total number of matches was 26,395.

Of 73,036 food products, NFt were available for over 60% of products and data on ingredients were available for about 30% of the food and drinks. Data for energy, protein, carbohydrate, fat, sugars, sodium and saturated fat were present for about 65% of the products, data for fiber for 37%, while data for other nutrients were present for about 60% of the products (**Table 3**).

DISCUSSION

We developed a comprehensive product- and brand-specific database containing nutrition information for >70,000 foods and beverages sold by the largest Canadian food retailers, using web-scraping and OCR/AI capabilities. As consumers' eating patterns change toward an increased consumption of pre-packaged foods, branded food composition databases are a critical component for monitoring the packaged food supply, related to ongoing public health nutrition interventions and policy development (e.g., front of pack labeling, marketing to children, sodium reformulation, trans fat ban etc.).

Using automated techniques (e.g., webscraping, OCR with AI/ML) to collect data from e-grocery retailers can result in food composition databases with far greater coverage and temporality than have been achieved in the past (24), allowing for more detailed evaluation of the food supply system. Such large amounts of data require the development of automated procedures, but this level of granularity can also reveal insights about the constantly changing set of products available in the Canadian marketplace, including the rapid turnover and reformulation of products, and evaluating the real-time impact of food policies. The greater coverage allows for a comprehensive collection of the nutritional quality of foods available in the marketplace, and an assessment of the association between nutritional quality and other key variables that affect purchasing behavior, such as price. Analyses of this large and dynamic dataset can reveal insights such as the differences in fat, saturated fat, sugar and sodium between lower-priced and higher-priced ready products, and of the variability of available products, and changes in their composition over time. Such investigations have previously been conducted around the world and in Canada using past versions of the FLIP databases (i.e., FLIP 2010, 2013, and 2017) (26–28, 44, 46, 49–55).

The FLIP 2020 data collection via web-scraping showed that from about 73,000 foods, about 60% of products had NFt information, suggesting that automatically and repeatedly scraping data from online e-retailers websites can produce food composition databases with sufficient information on nutrients and ingredients with reliability to allow for monitoring/evaluating a highly dynamic food and beverage supply. In comparison, a study from UK on foodDB, with over 97,368 products, found data on specific nutrients were present for over 90% of nutrient declaration tables, with data on ingredients available for >80% of the foods and drinks (24). Considering that almost 30% of products had missing NFt/ingredient information in this study, this points to the need for policy or regulations on mandating retailers to provide food labeling information in the e-grocery retail environment in Canada to help consumers make healthy decisions when purchasing foods and beverages on these platforms.

The need for branded food databases as well as the challenges of creating such tools are recognized by researchers and policymakers (5, 6). These challenges include obtaining, updating and maintaining the database to accurately capture variation in product availability and formulation over time. Most importantly, once data collections and data input methods are automated, much more frequent collections become possible. New technology such as crowdsourcing, artificial intelligence and machine learning are the critical tools in addressing these challenges (24). These technologies enable a wider, more granular collection of food products, including capturing fresh and/or ready-to-eat foods. The AI/ML also has extended applications such as prediction of nutrients (e.g., added sugar, fiber) content in packaged foods using available nutrient, ingredient and food category information (56). Additionally, crowdsourced data allows for input of information on missing products and may provide a novel means for low-cost, real-time tracking of nutritional composition of the food supply, thereby enabling an expansion of the number of products captured (19). Notably, in many jurisdictions, e-retailers (e.g., online grocery/restaurants websites) are not required to provide nutrition information. Such discussions have begun at the CODEX Alimentarius Committee on Food Labeling (57). Given the impact of the COVID-19 pandemic on the uptake of online grocery shopping, which is likely to continue increasing in the upcoming years, it is essential for retailers to provide and for policymakers and researchers to be able to gather nutrition information at the (virtual) point of purchase. Therefore, collecting and maintaining current food nutrition information is a unique opportunity for health and nutrition researchers to collaborate with mathematicians and computer science specialists to develop faster and more reliable cloud-based databases. As an example, foodDB, using big data techniques, is a weekly updated database that collects data on a comprehensive sample of foods and beverages available for purchase in all major UK grocery stores (24). Another important evolution to gather data from the food supply is to request manufacturers provide digital food labels to centralized government databases. As an example, manufacturers in the United States already provide nutrition information to the USDA Food Branded Database, in text format (22). Adding copies or links to the digital food label from which nutrition information can be extracted using technology could enable ongoing data collection into the future.

LIMITATIONS

Some limitations to our approach are related to the continued evolution and changes to the e-retailers product availability and websites, in order to ensure the data on these products are up-to-date. Although, the use web-scraping with OCR and AI/ML for data collected in FLIP 2020 were key innovations of our database that provided up-to-date product-specific nutrient information in a systematic and comprehensive manner, it was also a key limitation. E-grocery retailers may detect web-scraping, enabling OCR blockers and other techniques to make it difficult to scrape the data. Furthermore, there are no e-grocery food labeling regulations to mandate and standardize the availability and presentation of product information resulting in poor availability and wide inconsistencies in food labeling information, including missing information, number and quality of images, NFt and ingredients in the e-grocery retail environment in Canada (58).

The current FLIP2020 does not capture local regional and geographical variability of food and drink availability within individual online grocery retailers nor does it capture regional and local ethnic supermarkets that once catered to immigrant communities but are serving non-immigrant consumers seeking new products. Furthermore, convenience stores and large drugstore chains are introducing new product ranges that often include foods, which are not currently captured by the FLIP database.

Some tasks needed for research or monitoring remain time- and labor-intensive. For example, creating scores for some nutrient profiling models, automatic mapping of categories and subcategories and parsing of ingredients in any database remain, although work is underway to apply AI/ML to such tasks.

STRENGTHS

The automation of FLIP 2020 is a first step in providing real time nutritional data on foods. Web-scraping coupled with AI-powered OCR technology are important tools in automating the collection of real-time foods and nutrition information. The automated data collect process, using AI-enhanced OCR, provides FLIP with distinct analytical advantages compared to previous versions of FLIP and the generic food composition database in Canada (CNF) and takes the burden off manual processing by staff and students. A systematic methodology was established, based on previous versions of FLIP, to validate and categorize information, thereby enhancing the collection, storage, processing and management of nutrition information for each product. The use of web-scraping and automation further lowers the cost for future collections and allows for regularity in data capture on products. These features can also be implemented for future collections of FLIP databases, such as the FLIP-LAC and can be useful for other nutrition databases. Automating the systematic and consistent data capture will ensure sustainability and feasibility of maintaining large-scale branded food composition databases as new products and other changes to product formulation are introduced and others discontinued.

CONCLUSION

FLIP 2020 is an automated methodological step forward for food composition databases, which are the bedrock of nutritional epidemiology. Web-scraping coupled with OCR technology (AI/ML) are important tools in automating the collection of real-time food and nutrition information. The FLIP 2020 data collection demonstrated that automatically scraping data from online supermarkets can produce a food composition database with sufficient accuracy, transparency, granularity and flexibility to regularly monitor a highly dynamic food and drink marketplace. Such information are important in understanding the relationships between the nutritional quality of food products and measurements of policy impacts and health over time.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because complete database for non-commercial use can be obtained from the corresponding author at mary.labbe@utoronto.ca through data sharing agreements. Requests to access the datasets should be directed to Mary L'Abbe, mary.labbe@utoronto.ca.

AUTHOR CONTRIBUTIONS

MA, AS, and ML'A conceived and designed the study. AS, JL, and MW collected the data. MA, AS, JL, MW, and BF-A processed the different phases of the FLIP database. All authors contributed to interpreting the data as well as writing and editing the manuscript and have read and agreed to the published version of the manuscript.

FUNDING

This research was supported by funds from the Canadian Institute of Health Research Project Grant - Policy Impact (2016PJT-378415)(ML'A). MA was supported by a Postdoctoral Fellowship from the Joannah and Brian Lawson Center for Child Nutrition at University of Toronto, Toronto, ON, Canada. JL was supported by CIHR Doctoral Fellowship and the Banting and Best Diabetes Center at University of Toronto, Toronto, ON, Canada. BF-A was supported by a Postdoctoral Fellowship at Ontario Tech University. MW was supported by a Banting and Best Graduate Studentship Award.

ACKNOWLEDGMENTS

We thank ICS Solutions (Mark Wilson, Leo Lam) for technical support. We would also like to acknowledge the support of Preveena Mani, Cristina Benea, and Jenna Campoli for assisting with various tasks required for FLIP2020.

REFERENCES

1. Willett W. *Nutritional Epidemiology*. Oxford; New York, NY: Oxford University Press (2013). Available from: http://myaccess.library.utoronto.ca/login?url=http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199754038.001.0001/acprof-9780199754038 (accessed September 21, 2021).
2. Poti JM, Popkin BM. Trends in energy intake among US children by eating location and food source, 1977-2006. *J Am Diet Assoc.* (2011) 111:1156–64. doi: 10.1016/j.jada.2011.05.007
3. Ng SW, Popkin BM. The Healthy Wight Commitment Foundation pledge: calories purchased by U.S. households with children, 2000-2012. *Am J Prev Med.* (2014) 47:520–30. doi: 10.1016/j.amepre.2014.05.030
4. Leclercq C, Valsta LM, Turrini A. Food composition issues–implications for the development of food-based dietary guidelines. *Public Health Nutr.* (2001) 4:677–82. doi: 10.1079/PHN2001153
5. Pennington JA, Stumbo PJ, Murphy SP, McNutt SW, Eldridge AL, McCabe-Sellers BJ, et al. Food composition data: the foundation of dietetic practice and research. *J Am Diet Assoc.* (2007) 107:2105–13. doi: 10.1016/j.jada.2007.09.004
6. Ng SW, Dunford E. Complexities and opportunities in monitoring and evaluating US and global changes by the food industry. *Obes Rev.* (2013) 14(Suppl. 2):29–41. doi: 10.1111/obr.12095
7. Greenfield H, Southgate DAT, Food and Agriculture Organization of the United Nations. *Food Composition Data: Production, Management, and Use*. Rome: FAO (2003). xiii, 288 p.
8. Poti JM, Yoon E, Hollingsworth B, Ostrowski J, Wandell J, Miles DR, et al. Development of a food composition database to monitor changes in packaged foods and beverages. *J Food Compost Anal.* (2017) 64(Pt 1):18–26. doi: 10.1016/j.jfca.2017.07.024
9. Weippert MV, Schermel A., L'Abbe M. Assessing the turnover rate of Canadian food and beverage labels – implications for the proposed front-of-package labelling transition period. *Can Nutr Soc Virtual* (2021) S41:1. https://cdnsciencepub.com/doi/pdf/10.1139/apnm-2021-0172
10. Government of Canada. *The Canadian Nutrient File.* (2015). Available from: https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/nutrient-data/canadian-nutrient-file-about-us.html (accessed December 21, 2015).
11. Fung C, McIsaac JL, Kuhle S, Kirk SF, Veugelers PJ. The impact of a population-level school food and nutrition policy on dietary intake and body weights of Canadian children. *Prev Med.* (2013) 57:934–40. doi: 10.1016/j.ypmed.2013.07.016
12. Wu XY, Ohinmaa A, Veugelers PJ. Diet quality, physical activity, body weight and health-related quality of life among grade 5 students in Canada. *Public Health Nutr.* (2012) 15:75–81. doi: 10.1017/S1368980011002412
13. Health Canada. *2015 Canadian Community Health Survey - Nutrition.* (2015). Available from: https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs/2015-canadian-community-health-survey-nutrition-food-nutrition-surveillance.html
14. Health Canada. *Canadian Nutrient File.* (2015). Available from: https://food-nutrition.canada.ca/cnf-fce/index-eng.jsp (accessed September 21, 2021).
15. Mozaffarian D, Jacobson MF, Greenstein JS. Food reformulations to reduce trans fatty acids. *N Engl J Med.* (2010) 362:2037–9. doi: 10.1056/NEJMc1001841
16. Jacobson MF, Havas S, McCarter R. Changes in sodium levels in processed and restaurant foods, 2005 to 2011. *JAMA Intern Med.* (2013) 173:1285–91. doi: 10.1001/jamainternmed.2013.6154
17. Otite FO, Jacobson MF, Dahmubed A, Mozaffarian D. Trends in trans fatty acids reformulations of US supermarket and brand-name foods from 2007 through 2011. *Prev Chronic Dis.* (2013) 10:E85. doi: 10.5888/pcd10.120198
18. Briguglio S CS, Spungen J, Rabbani PI, Hoffman-Pennesi D, Wirtz M. Sodium trends in selected US Total Diet Study foods, 2003–2011. *Proc Food Sci.* (2015) 4:27–38. doi: 10.1016/j.profoo.2015.06.006
19. Dunford E, Trevena H, Goodsell C, Ng KH, Webster J, Millis A, et al. FoodSwitch: a mobile phone app to enable consumers to make healthier food choices and crowdsourcing of national food composition data. *JMIR Mhealth Uhealth.* (2014) 2:e37. doi: 10.2196/mhealth.3230
20. Open Food Facts. *Open Food Facts Methodology*. Available from: https://world.openfoodfacts.org/ (accessed September 21, 2021).
21. Access to Nutrition Index (ATNI). *Harnessing the Power of the Private Sector to Tackle the World's Biggest Nutrition Challenges*. Available from: https://www.accesstonutrition.org/ (accessed September 23, 2021).
22. Kretser A, Murphy D, Starke-Reed P. A partnership for public health: USDA branded food products database. *J Food Comp Anal.* (2017) 64:10–2. doi: 10.1016/j.jfca.2017.07.019
23. Pehrsson PRH, David B, McKillop, Kyle A, Moore G, Finley, et al. *USDA Branded Food Products Database.* (2018). Available from: https://data.nal.usda.gov/dataset/usda-branded-food-products-database (accessed September 23, 2021).
24. Harrington RA, Adhikari V, Rayner M, Scarborough P. Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. *BMJ Open.* (2019) 9:e026652. doi: 10.1136/bmjopen-2018-026652
25. Health Canada. *Voluntary Sodium Reduction Targets for Processed Foods 2020-2025.* (2020). Available from: https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/sodium/sodium-reduced-targets-2020-2025.html (accessed July 15, 2021).
26. Franco-Arellano B, Arcand JA, Kim MA, Schermel A, L'Abbe M. Progress towards eliminating industrially-produced trans-fatty acids in the Canadian marketplace, 2013-2017. *Public Health Nutr.* (2019) 23:2257–67. doi: 10.1017/S1368980019004816

27. Schermel A, Emrich TE, Arcand J, Wong CL, L'Abbe MR. Nutrition marketing on processed food packages in Canada: 2010 Food Label Information Program. *Appl Physiol Nutr Metab.* (2013) 38:666–72. doi: 10.1139/apnm-2012-0386
28. Bernstein JT, Schermel A, Mills CM, L'Abbe MR. Total and free sugar content of canadian prepackaged foods and beverages. *Nutrients.* (2016) 8:E582. doi: 10.3390/nu8090582
29. United States Department of Agriculture Foreign Agricultural Services. *Canada Retail Sector Overview - 2018.* (2018). Available from: https://gain.fas.usda.gov/Recent%20GAIN%20Publications/Retail%20Foods_Ottawa_Canada_6-26-2018.pdf (accessed July 15, 2021).
30. Canadian Grocer. *Executive Report, Canadian Grocery Industry 2012–2013.* Toronto, ON (2012).
31. Statista. *Leading Canadian Food Retailers in Canada in FY 2019/20, by Grocery Sales Value.* (2019). Available from: https://www.statista.com/statistics/707809/leading-canadian-food-retailers-by-sales-value/ (accessed October 5, 2021).
32. Government of Canada. *Food and Drug Regulations (C.R.C., c. 870).* (2012). Available from: http://laws-lois.justice.gc.ca/PDF/C.R.C.,_c._870.pdf (accessed October 15, 2021).
33. Health Canada. *Table of Reference Amounts.* (2016). Available from: https://www.canada.ca/en/health-canada/services/technical-documents-labelling-requirements/table-reference-amounts-food.html (accessed April 18, 2021).
34. Health Canada. *Guidance for Food Industry on Reducing Sodium in Processed Food.* (2012). Available from: https://www.canada.ca/en/health-canada/services/food-nutrition/legislation-guidelines/guidance-documents/guidance-food-industry-reducing-sodium-processed-foods-2012.html (accessed November 12, 2020).
35. Weippert MV, L'Abbe M. Free sugars and sweeteners in the Canadian food supply: changes from 2013 to 2017. *Can Nutr Soc Virtual.* (2020). Available online at: https://cns-scn.ca/abstracts/view/1099
36. British Food Standards Agency. *Nutrient Profiling Technical Guidance.* London (2009).
37. Sante Publique France. *Nutri-Score Scientific and Technical Document.* (2019). Available from: https://www.santepubliquefrance.fr/determinants-de-sante/nutrition-et-activite-physique/articles/nutri-score (accessed June 14, 2020).
38. Food Standards Australia New Zealand. *Food Standards Australia New Zealand Nutrient Profiling.* (2016). Available from: http://www.foodstandards.gov.au/industry/labelling/Pages/Short-guide-for-industry-to-the-NPSC.aspx (accessed March 11, 2020).
39. Pan American Health Organization. *Pan American Health Organization Nutrient Profile Model.* (2016). Available from: http://www.paho.org/hq/index.php?option=com_content&view=article&id=11662%3Apaho-nutrient-profile-model&catid=1370%3Amicronutrients&Itemid=41739&lang=en (accessed November 18, 2021).
40. World Health Organization. *Nutrient Profiling : Report of a WHO/IASO Technical Meeting, London, United Kingdom.* Geneva: World Health Organization (2010).
41. Ahmed M, Oh, A, Vanderlee, L, Franco-Arellano, Schermel A, Lou W, et al. A randomized controlled trial examining consumers' perceptions and opinions on using different version of a FoodFlip© smartphone application for delivery of nutrition information. *International J Behav Nutr Phys Act.* (2019) 17:22–38. doi: 10.1186/s12966-020-0923-1
42. Allemandi L, Tiscornia MV, Guarnieri L, Castronuovo L, Martins E. Monitoring sodium content in processed foods in Argentina 2017-2018: compliance with national legislation and regional targets. *Nutrients.* (2019) 11:1474–86. doi: 10.3390/nu11071474
43. Vega-Solano J, Blanco-Metzler A, Benavides-Aguilar KF, Arcand J. An evaluation of the sodium content and compliance with the national sodium reduction targets among packaged foods sold in Costa Rica in 2015 and 2018. *Nutrients.* (2019) 11:2226–37. doi: 10.3390/nu11092226
44. Arcand J, Jefferson K, Schermel A, Shah F, Trang S, Kutlesa D, et al. Examination of food industry progress in reducing the sodium content of packaged foods in Canada: 2010 to 2013. *Appl Physiol Nutr Metab.* (2016) 41:684–90. doi: 10.1139/apnm-2015-0617
45. Vergeer L, Vanderlee L, Ahmed M, Franco-Arellano B, Mulligan C, Dickinson K, et al. A comparison of the nutritional quality of products offered by the top packaged food and beverage companies in Canada. *BMC Public Health.* (2020) 20:650. doi: 10.1186/s12889-020-08828-w
46. Labonté ME, Emrich TE, Scarborough P, Rayner M, L'Abbé MR. Traffic light labelling could prevent mortality from noncommunicable diseases in Canada: A scenario modelling study. *PLoS ONE.* (2019) 14:e0226975. doi: 10.1371/journal.pone.0226975
47. L'Abbe MR, Schermel A, Franco-Arellano B, Vega S J, Arcand J, Lam L. *FLIP-LAC User Guide.* (2018). Available from: https://idl-bnc-idrc.dspacedirect.org/handle/10625/58940 (accessed November 18, 2021).
48. Pan American Health Organization. *Updated PAHO Regional Sodium Reduction Targets.* (2021). Available from: https://iris.paho.org/handle/10665.2/54658 (accessed November 18, 2021).
49. Arcand J, Au JT, Schermel A, L'Abbe MR. A comprehensive analysis of sodium levels in the Canadian packaged food supply. *Am J Prev Med.* (2014) 46:633–42. doi: 10.1016/j.amepre.2014.01.012
50. Emrich TE, Cohen JE, Lou WY, L'Abbe MR. Food products qualifying for and carrying front-of-pack symbols: a cross-sectional study examining a manufacturer led and a non-profit organization led program. *BMC Public Health.* (2013) 13:846. doi: 10.1186/1471-2458-13-846
51. Emrich TE, Qi Y, Cohen JE, Lou WY, L'Abbe ML. Front-of-pack symbols are not a reliable indicator of products with healthier nutrient profiles. *Appetite.* (2015) 84:148–53. doi: 10.1016/j.appet.2014.09.017
52. Franco-Arellano B, Bernstein JT, Norsen S, Schermel A, L'Abbe MR. Assessing nutrition and other claims on food labels. *BMC Nutrition.* (2017) 3:74–90. doi: 10.1186/s40795-017-0192-9
53. Labonte ME, Poon T, Gladanac B, Ahmed M, Franco-Arellano B, Rayner M, et al. Nutrient profile models with applications in government-led nutrition policies aimed at health promotion and noncommunicable disease prevention: a systematic review. *Adv Nutr.* (2018) 9:741–88. doi: 10.1093/advances/nmy045
54. Poon T, Labonte ME, Mulligan C, Ahmed M, Dickinson KM, L'Abbe MR. Comparison of nutrient profiling models for assessing the nutritional quality of foods: a validation study. *Br J Nutr.* (2018) 120:567–82. doi: 10.1017/S0007114518001575
55. Vergeer L, Vanderlee L, Potvin Kent M, Mulligan C, L'Abbe MR. The effectiveness of voluntary policies and commitments in restricting unhealthy food marketing to Canadian children on food company websites. *Appl Physiol Nutr Metab.* (2019) 44:74–82. doi: 10.1139/apnm-2018-0528
56. Davies T, Louie JCY, Ndanuko R, Barbieri S, Perez-Concha O, Wu JHY. A machine learning approach to predict the added sugar content of packaged foods. *J Nutr.* (2021) 126. doi: 10.1093/jn/nxab341
57. Codex Alimentarius Commission. *Joint FAO/WHO Food Standards Programme Codex Committee on Food Labelling.* (2021). Available from: https://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FMeetings%252FCX-714-45%252Fdocuments%252Ffl45_06e_final.pdf
58. Lee JJ, Ahmed M, Zhang T, Weippert MV, Schermel A, L'Abbe MR. The availability and quality of food labelling components in the Canadian E-Grocery retail environment. *Nutrients.* (2021) 13:2611–22. doi: 10.3390/nu13082611

Conflict of Interest: ML'A and MA report receiving a competitive research grant from IAFNS formerly ILSI NA to analyze NHANES data to determine the intakes and sources of sodium in the diets of Americans 2021–2022. None of these companies/organizations had any involvement in the present research.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Nutrition

ORIGINAL RESEARCH
published: 14 March 2022
doi: 10.3389/fnut.2022.851820

Dietary Fiber Intake and Femoral Bone Mineral Density in Middle-Aged and Older US Adults: A Cross-Sectional Study of National Health and Nutrition Examination Survey 2013–2014

Yuchen Tang [1,2,3†], ***Jinmin Liu*** [1,2,3†], ***Xiaohui Zhang*** [1,2,3] ***and Bin Geng*** [1,2,3*]

[1] Department of Orthopaedics, Lanzhou University Second Hospital, Lanzhou, China, [2] Orthopaedics Key Laboratory of Gansu Province, Lanzhou, China, [3] Orthopaedic Clinical Research Center of Gansu Province, Lanzhou, China

Sufficient dietary fiber intake (DFI) is considered necessary for human health. However, the association between DFI and bone mineral density (BMD) remains unclear. Therefore, this study aimed to investigate the association between DFI and BMD and to determine whether sex modifies the association between DFI and BMD. Participants aged ≥ 40 years from the 2013–2014 National Health and Nutrition Examination Survey were included in the final analysis. The association between DFI and BMD was evaluated using a multivariate linear regression model. The non-linear relationship between DFI and BMD was characterized by smooth curve fittings and generalized additive models. Finally, 1,935 participants with a mean age of 58.12 ± 11.84 years were included in the final analysis. The results revealed that DFI was positively associated with femoral BMD in the unadjusted model. However, no correlation was observed between DFI and femoral BMD after adjusting for covariates. Moreover, the results showed an inverted U-shaped association between total DFI and femoral BMD among men but not women for the nonlinear relationship between DFI and femoral BMD. In conclusion, our results indicate that DFI might not follow a linear relationship with femoral BMD, and sex factors might modify the association between DFI and BMD. Particularly, high total DFI might contribute to lower femoral neck BMD. However, more studies are needed to investigate whether the negative effect of high DFI on femoral BMD does exist and whether high DFI has clear biological effects on bone metabolism, such as increasing the risk of osteoporosis.

Keywords: dietary fiber, dietary fiber intake, bone mineral density, sex, femoral neck

OPEN ACCESS

Edited by:
Massimo Lucarini,
Council for Agricultural Research and Economics, Italy

Reviewed by:
Victor L. Fulgoni III,
Independent Researcher,
Battle Creek, United States
Dongfeng Zhang,
Qingdao University, China

****Correspondence:***
Bin Geng
cxxxf@foxmail.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

Received: *10 January 2022*
Accepted: *27 January 2022*
Published: *14 March 2022*

Citation:
Tang Y, Liu J, Zhang X and Geng B (2022) Dietary Fiber Intake and Femoral Bone Mineral Density in Middle-Aged and Older US Adults: A Cross-Sectional Study of National Health and Nutrition Examination Survey 2013–2014. Front. Nutr. 9:851820. doi: 10.3389/fnut.2022.851820

INTRODUCTION

Osteoporosis, characterized by reduced bone mineral density (BMD) and bone tissue microstructure degradation, is a common chronic disease worldwide (1). Approximately one-third of women and one-fifth of men aged ≥ 50 years are at risk of osteoporosis globally (1–3). Moreover, osteoporotic fracture, the most serious complication of osteoporosis, is also an important cause of death in older adults (4, 5). The pathogenesis of osteoporosis is complex and it is generally accepted that osteoporosis is determined by numerous genes and environmental factors (1). In addition, lifestyle factors play essential roles in the pathogenesis of osteoporosis (1, 6). For example, sufficient calcium or vitamin D intake is considered a key factor in the maintenance of bone mass (6, 7). Additional evidence has demonstrated that intake of other nutritional elements also essentially contribute to maintaining normal BMD, except for calcium and vitamin D. Therefore, exploring the impact of nutritional element intake on bone metabolism is receiving increasing attention, and it is expected to open novel avenues to prevent bone loss.

Dietary fiber (DF) is a carbohydrate polymer with ten or more monomeric units, which are not hydrolyzed by endogenous enzymes in the small intestine of humans and are typically derived from whole-grain cereals, fruits, vegetables, and legumes (8, 9). Several previous studies have shown that adequate DF intake (DFI) is necessary for disease prevention. Tanaka et al. observed that increased DFI reduces the incidence of stroke (10). Fujii et al. demonstrated that increased DFI is associated with better glycemic control and a lower risk of chronic kidney disease in patients with type 2 diabetes (11). Ananthakrishnan et al. found that adequate long-term DFI is associated with the decreased risk of Crohn's disease (12). Although the number is limited, related studies on bone metabolism have found that DFI might be associated with BMD (13–16). Dai et al. observed that increased DFI was associated with less bone loss among males but not females (14). Lee and Suh found that DFI was positively associated with lumbar BMD in men aged 18–45 years, but this correlation was not observed among women regardless of age (15). Zhou et al. demonstrated that higher DFI was associated with higher heel BMD among individuals aged 40–69 years, regardless of sex (16). Conversely, Barron et al. observed that a higher DFI was associated with lower lumbar BMD among young female athletes with oligomenorrhea (13). These contradicting findings suggest that the relationship between DFI and BMD remains unclear. Moreover, there was no definite evidence of whether sex modified the association between DFI and BMD.

Therefore, this study aimed to investigate the association between DFI and BMD. Moreover, we also tried to determine whether sex modified the association between DFI and BMD.

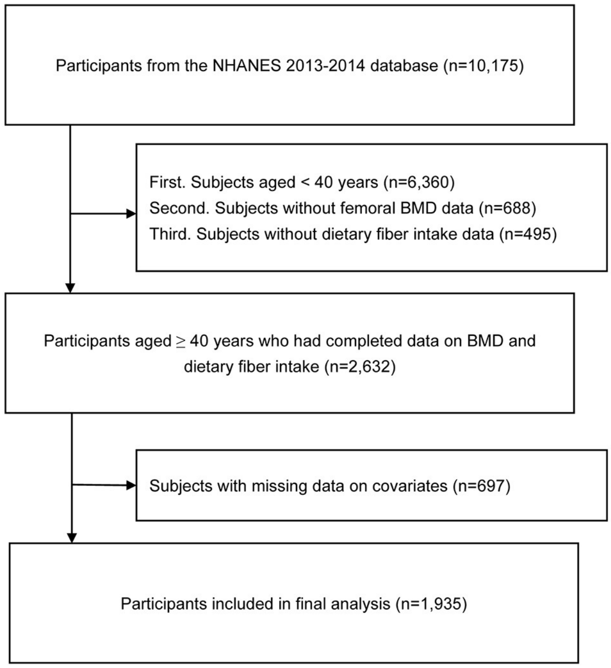

Participants from the NHANES 2013-2014 database (n=10,175)
First. Subjects aged < 40 years (n=6,360)
Second. Subjects without femoral BMD data (n=688)
Third. Subjects without dietary fiber intake data (n=495)
Participants aged ≥ 40 years who had completed data on BMD and dietary fiber intake (n=2,632)
Subjects with missing data on covariates (n=697)
Participants included in final analysis (n=1,935)

FIGURE 1 | Flow chart of participant selection. NHANES, National Health and Nutrition Examination Survey; BMD, bone mineral density.

MATERIALS AND METHODS

Study Population

We extracted data from the National Health and Nutrition Examination Survey (NHANES) database (2013–2014) (17). The NHANES database, affiliated with the Centers for Disease Control and Prevention (USA), aimed to assess the health and nutritional status of US residents and was updated biannually. Participants aged ≥ 40 years (the BMD test was only performed among participants aged ≥ 40 years in the NHANES 2013–2014) with complete data on BMD and DFI were enrolled in the present study. Moreover, subjects with missing covariate data (see details in the Covariates section) were excluded from the study. Each participant included in the present study obtained and signed the informed consent, and the Ethics Review Board of the National Center for Health Statistics approved the study (18).

Bone Mineral Density Testing

All participants underwent BMD testing, which was based on the dual energy X-ray absorptiometry scan and assessed the BMD of four femoral regions (total femur, femoral neck, trochanter, and intertrochanter). Moreover, certified radiologic technologists conducted the dual-energy X-ray absorptiometry examinations using Hologic QDR-4500A fan-beam densitometers (Hologic; Bedford, MA), and the data analysis was performed using the

Hologic APEX, version 4.0, software. Other details are available from the NHANES website (19).

Dietary Fiber Intake

NHANES assessed the types and amounts of foods and beverages (including all types of water) consumed during the 24 h before the interview and estimated the DFI from those foods and beverages. In this study, the DFI referred to total DFI from the above foods and beverages. Information on DFI was collected through in-person interviews and telephone surveys (3–10 days after the in-person interview). The dietary recall statuses were classified as follows (i) reliable and met the minimum criteria; (ii) not reliable or did not meet the minimum criteria; (iii) reported consuming breast milk (for infants); and (iv) not done. In the present study, we enrolled only participants with a dietary recall status that was "reliable and met the minimum criteria" in the final analysis. Moreover, to balance the errors in both methods (in person or by phone), we calculated the mean values between the two and used them as the final values of DFI. Other details about the measurement of DFI are listed on the NHANES website (20, 21).

Covariates

Considering that there were several factors that affected bone metabolism, we included covariates in the present study. Based on some previous studies (1, 22, 23), this study included the following covariates: age, sex, race, education level, income level, body mass index (BMI), smoking status, alcohol consumption, hypertension, diabetes, blood calcium level, serum 25-hydroxyvitamin D, rheumatoid arthritis (RA), cancer, use of glucocorticoid, family history of osteoporosis, previous fractures, physical activity level, calcium intake level, and vitamin D intake level. The specific information on the covariates is provided in **Supplementary Table 1**.

Statistical Analysis

The baseline characteristics were described using the mean (for continuous variables) or proportion (for categorical variables). The linear relationship between DFI and BMD was assessed by multivariate linear regression models, while the non-linear relationship between DFI and BMD was evaluated by smooth curve fitting and generalized additive models. Moreover, if the non-linear relationship shows that an inflection point might exist, the inflection point can be calculated using two-piecewise linear regression models by a recursive algorithm. All analyses were performed using R software (version 4.0.3; https://www.R-project.org) and EmpowerStats (version 2.0; http://www.empowerstats.com). Statistical significance was set at $P < 0.05$.

RESULTS

Participant Selection and Baseline Characteristics

We extracted data from 10,175 participants from the NHANES (2013–2014) database. First, subjects aged < 40 years ($n = 6,360$) were excluded from the present study. Second, subjects without femoral BMD data ($n = 688$) were also excluded. Third, subjects without dietary fiber intake data ($n = 495$) were excluded from

TABLE 1 | Baseline characteristics of included participants.

Characteristics		Mean or proportion
Age (year)		58.12 ± 11.84
Sex *n*, (%)	Male	949 (49.04%)
	Female	986 (50.96%)
Race *n*, (%)	Mexican American	249 (12.87%)
	Other hispanic	168 (8.68%)
	Non-hispanic white	937 (48.42%)
	Non-hispanic black	358 (18.50%)
	Other race	223 (11.52%)
Education level *n*, (%)	Under high school	371 (19.17%)
	High school or equivalent	421 (21.76%)
	Above high school	1,143 (59.07%)
Income level *n*, (%)	PIR < 1	300 (15.50%)
	PIR ≥ 1	1,635 (84.50%)
BMI *n*, (%)	Normal	535 (27.65%)
	Overweight	680 (35.14%)
	Obesity	720 (37.21%)
Smoking status *n*, (%)	Current smokers	327 (16.90%)
	Quit smoking	563 (29.10%)
	Never	1,045 (54.01%)
Alcohol consumption *n*, (%)	Yes	1,419 (73.33%)
	No	516 (26.67%)
Hypertension *n*, (%)	Yes	867 (44.81%)
	No	1,068 (55.19%)
Diabetes *n*, (%)	Yes	300 (15.50%)
	No	1,560 (80.62%)
	Borderline	75 (3.88%)
Blood calcium level *n*, (%)	Q1: 8.2–9.1 (mg/dL)	382 (19.74%)
	Q2: 9.2–9.3 (mg/dL)	396 (20.47%)
	Q3: 9.4–9.6 (mg/dL)	655 (33.85%)
	Q4: 9.7–12.0 (mg/dL)	502 (25.94%)
Serum 25-hydroxyvitamin D *n*, (%)	Q1: 9.37–50.90 (nmol/L)	484 (25.01%)
	Q2: 51.00–67.20 (nmol/L)	476 (24.60%)
	Q3: 67.30–85.60 (nmol/L)	488 (25.22%)
	Q4: 85.70–318.00 (nmol/L)	487 (25.17%)
RA *n*, (%)	Yes	119 (6.15%)
	No	1,816 (93.85%)
Cancer *n*, (%)	Yes	252 (13.02%)
	No	1,683 (86.98%)
Use of glucocorticoid *n*, (%)	Yes	109 (5.63%)

(Continued)

TABLE 1 | Continued

Characteristics		Mean or proportion
	No	1,826 (94.37%)
Family history of osteoporosis *n*, (%)	Yes	286 (14.78%)
	No	1,649 (85.22%)
Previous fractures *n*, (%)	Yes	532 (27.49%)
	No	1,403 (72.51%)
Physical activity level *n*, (%)	NMVPA	486 (25.12%)
	LMVPA	295 (15.25%)
	MMVPA	232 (11.99%)
	HMVPA	922 (47.65%)
Calcium intake level *n*, (%)	Q1: 39.50–580.00 (mg/day)	484 (25.01%)
	Q2: 580.50–829.00 (mg/day)	483 (24.96%)
	Q3: 829.50–1,107.50 (mg/day)	484 (25.01%)
	Q4: 1,108.00–4,022.00 (mg/day)	484 (25.01%)
Vitamin D intake level *n*, (%)	Q1: 0.00–1.85 (mcg/day)	484 (25.01%)
	Q2: 1.90–3.50 (mcg/day)	475 (24.55%)
	Q3: 3.55–6.00 (mcg/day)	483 (24.96%)
	Q4: 6.05–46.30 (mcg/day)	493 (25.48%)
Total femur BMD (g/cm^2)		0.95 ± 0.15
Femoral neck BMD (g/cm^2)		0.78 ± 0.14
Trochanter BMD (g/cm^2)		0.72 ± 0.12
Intertrochanter BMD (g/cm^2)		1.13 ± 0.18

BMI, body mass index; PIR, poverty-income ratio; RA, rheumatoid arthritis; HMVPA, high moderate-to-vigorous physical activity (≥1,200 MET-mins/week); MMVPA, medium moderate-to-vigorous physical activity (600–1,199 MET-mins/week); LMVPA, low moderate-to-vigorous physical activity (1–599 MET-mins/week); NMVPA, no moderate-to-vigorous physical activity (0 MET-mins/week).

this study. In addition, we excluded 697 subjects with missing data (missing data; refused to answer; or answered "do not know") on covariates (**Supplementary Figure 1**). Finally, 1,935 participants were included in the final analysis. A flowchart of participant selection is shown in **Figure 1**.

The mean age of included participants was 58.12 ± 11.84 years. Moreover, most participants were females (50.96%), non-Hispanic whites (48.42%), had above high school education (59.07%), and were with $\geq$ 1 of poverty-income ratio (84.50%). In addition, the ratios of cases who were obese, current smoker, consumed at least 12 alcoholic drinks in the previous year, and were with diabetes, hypertension were 37.21, 16.90, 73.33, 44.81, 15.50%, respectively. Besides, the mean total femur BMD, femoral neck BMD, trochanter BMD, and intertrochanter BMD were 0.95 ± 0.15 g/cm^2, 0.78 ± 0.14 g/cm^2, 0.72 ± 0.12 g/cm^2, 1.13 ± 0.18 g/cm^2, respectively. Other details of the baseline characteristics are listed in **Table 1**.

Association Between DFI and BMD

The results of multivariate linear regression models showed that DFI was positively associated with total femur (β: 0.0011; 95% CI: 0.0004–0.0019), trochanter (β: 0.0007; 95% CI: 0.0001–0.0013), and intertrochanter (β: 0.0013; 95% CI: 0.0005–0.0022) BMD in Model 1 (unadjusted model). However, no correlation was observed between DFI and femoral BMD after adjusting for covariates (Model 2 and Model 3). The specific results are shown in **Table 2**.

When the variable of DFI was converted into a categorical variable, the results of multivariate linear regression models revealed that participants with the higher quartile of DFI (Q3 and Q4) had higher femoral BMD than those with the lowest quartile of DFI in Model 1 (unadjusted model). After adjusting for age, sex, and race (Model 2), the results revealed that participants with the third quartile of DFI showed higher total femur (β: 0.0187; 95% CI: 0.0023–0.0352) and trochanter (β: 0.0183; 95% CI: 0.0044–0.0323) BMD compared with those with the lowest quartile of DFI. When all covariates were adjusted (Model 3), participants with the third quartile of DFI still showed higher trochanter (β: 0.0147; 95% CI: 0.0013–0.0281) BMD than those with the lowest quartile of DFI, and no significant differences were observed in other groups. The specific results are listed in **Table 3**.

Association Between DFI and BMD Stratified by Sex

The subgroup analysis stratified by sex is shown in **Table 4**. The results of multivariate linear regression models revealed that DFI was not associated with femoral BMD ($P > 0.05$) regardless of sex. Moreover, further analysis of the non-linear relationship between DFI and femoral BMD showed an inverted U-shaped association between DFI and femoral BMD among men but not women, and the inflection points of DFI observed were about 25 gm/day (**Figure 2**). In addition, the two-piecewise linear regression models demonstrated the inverted U-shaped association between DFI and femoral BMD among men. In particular, DFI was negatively associated with femoral neck BMD (β: −0.0017; 95% CI: −0.0032 to −0.0002) among men when DFI was >25 gm/day. The details are listed in **Table 5**.

DISCUSSION

Osteoporosis in middle-aged and older individuals has become a global issue in the past decade. Currently, there is an increasing awareness that dietary changes or lifestyle modifications might be an effective mean of preventing osteoporosis. This study found that DFI was positively associated with femoral BMD in the unadjusted model. However, no correlation was observed between DFI and femoral BMD after adjusting for covariates. For the non-linear relationship between DFI and femoral BMD, the

TABLE 2 | Association between dietary fiber intake and femoral BMD.

Index	Model 1	Model 2	Model 3
	β (95% CI)	**β (95% CI)**	**β (95% CI)**
Total femur BMD	**0.0011 (0.0004, 0.0019)**	0.0002 (−0.0004, 0.0009)	0.0003 (−0.0003, 0.0010)
Femoral neck BMD	0.0006 (−0.0000, 0.0013)	0.0003 (−0.0003, 0.0009)	0.0002 (−0.0004, 0.0009)
Trochanter BMD	**0.0007 (0.0001, 0.0013)**	0.0001 (−0.0005, 0.0006)	0.0002 (−0.0004, 0.0007)
Intertrochanter BMD	**0.0013 (0.0005, 0.0022)**	0.0002 (-0.0006, 0.0010)	0.0003 (−0.0005, 0.0012)

Bold variables indicate P-value < 0.05. Model 1: unadjusted model; Model 2: age, sex, and race were adjusted; Model 3: age, sex, race, education level, income level, BMI, smoking status, alcohol consumption, hypertension, diabetes, blood calcium level, serum 25-hydroxyvitamin D, RA, cancer, use of glucocorticoid, family history of osteoporosis, previous fractures, physical activity level, calcium intake level, and vitamin D intake level were adjusted. BMD, bone mineral density; CI, confidence interval; BMI, body mass index; RA, rheumatoid arthritis.

TABLE 3 | Association between dietary fiber intake and femoral BMD.

Index	Group	Model 1	Model 2	Model 3
		β (95% CI)	**β (95% CI)**	**β (95% CI)**
Total femur BMD	Q1: 0.15–11.25 gm/day	Reference (0)	Reference (0)	Reference (0)
	Q2: 11.30–15.90 gm/day	0.0153 (−0.0038, 0.0343)	0.0106 (−0.0057, 0.0270)	0.0021 (−0.0129, 0.0172)
	Q3: 15.95–22.05 gm/day	**0.0199 (0.0009, 0.0389)**	**0.0187 (0.0023, 0.0352)**	0.0141 (−0.0016, 0.0297)
	Q4: 22.10–95.20 gm/day	**0.0340 (0.0150, 0.0530)**	0.0132 (−0.0037, 0.0300)	0.0131 (−0.0041, 0.0304)
Femoral neck BMD	Q1: 0.15–11.25 gm/day	Reference (0)	Reference (0)	Reference (0)
	Q2: 11.30–15.90 gm/day	0.0018 (−0.0161, 0.0197)	0.0022 (−0.0132, 0.0176)	−0.0058 (−0.0206, 0.0090)
	Q3: 15.95–22.05 gm/day	0.0059 (−0.0120, 0.0237)	0.0129 (−0.0026, 0.0284)	0.0072 (−0.0082, 0.0227)
	Q4: 22.10–95.20 gm/day	**0.0198 (0.0019, 0.0377)**	0.0131 (−0.0028, 0.0289)	0.0093 (−0.0077, 0.0263)
Trochanter BMD	Q1: 0.15–11.25 gm/day	Reference (0)	Reference (0)	Reference (0)
	Q2: 11.30–15.90 gm/day	0.0147 (−0.0007, 0.0302)	0.0117 (−0.0021, 0.0256)	0.0042 (−0.0087, 0.0170)
	Q3: 15.95–22.05 gm/day	**0.0189 (0.0035, 0.0343)**	**0.0183 (0.0044, 0.0323)**	**0.0147 (0.0013, 0.0281)**
	Q4: 22.10–95.20 gm/day	**0.0247 (0.0092, 0.0401)**	0.0111 (−0.0032, 0.0253)	0.0117 (−0.0030, 0.0265)
Intertrochanter BMD	Q1: 0.15–11.25 gm/day	Reference (0)	Reference (0)	Reference (0)
	Q2: 11.30–15.90 gm/day	0.0148 (−0.0079, 0.0375)	0.0084 (−0.0113, 0.0281)	−0.0006 (−0.0189, 0.0177)
	Q3: 15.95–22.05 gm/day	0.0200 (−0.0026, 0.0427)	0.0170 (−0.0029, 0.0368)	0.0114 (−0.0076, 0.0305)
	Q4: 22.10–95.20 gm/day	**0.0379 (0.0152, 0.0606)**	0.0110 (−0.0093, 0.0314)	0.0113 (−0.0097, 0.0323)

Bold variables indicate P-value < 0.05. Model 1: unadjusted model; Model 2: age, sex, and race were adjusted; Model 3: age, sex, race, education level, income level, BMI, smoking status, alcohol consumption, hypertension, diabetes, blood calcium level, serum 25-hydroxyvitamin D, RA, cancer, use of glucocorticoid, family history of osteoporosis, previous fractures, physical activity level, calcium intake level, and vitamin D intake level were adjusted. BMD, bone mineral density; CI, confidence interval.

TABLE 4 | Association between dietary fiber intake and femoral BMD stratified by sex.

Sex	Index	Model 1	Model 2	Model 3
		β (95% CI)	**β (95% CI)**	**β (95% CI)**
Male	Total femur BMD	0.0000 (−0.0009, 0.0009)	0.0002 (−0.0006, 0.0011)	−0.0001 (−0.0010, 0.0008)
	Femoral neck BMD	0.0000 (−0.0009, 0.0009)	0.0003 (−0.0005, 0.0012)	−0.0002 (−0.0011, 0.0008)
	Trochanter BMD	−0.0001 (−0.0008, 0.0006)	0.0001 (−0.0006, 0.0008)	−0.0000 (−0.0008, 0.0008)
	Intertrochanter BMD	−0.0001 (−0.0011, 0.0009)	0.0002 (−0.0009, 0.0012)	−0.0002 (−0.0013, 0.0009)
Female	Total femur BMD	0.0000 (−0.0010, 0.0011)	0.0002 (−0.0007, 0.0012)	0.0008 (−0.0002, 0.0018)
	Femoral neck BMD	−0.0002 (−0.0012, 0.0008)	0.0002 (−0.0007, 0.0011)	0.0005 (−0.0005, 0.0015)
	Trochanter BMD	−0.0002 (−0.0010, 0.0007)	0.0000 (−0.0008, 0.0008)	0.0004 (−0.0004, 0.0012)
	Intertrochanter BMD	0.0001 (−0.0012, 0.0014)	0.0003 (−0.0009, 0.0015)	0.0010 (−0.0002, 0.0022)

Model 1: unadjusted model; Model 2: age and race were adjusted; Model 3: age, race, education level, income level, BMI, smoking status, alcohol consumption, hypertension, diabetes, blood calcium level, serum 25-hydroxyvitamin D, RA, cancer, use of glucocorticoid, family history of osteoporosis, previous fractures, physical activity level, calcium intake level, and vitamin D intake level were adjusted. BMD, bone mineral density; CI, confidence interval; BMI, body mass index; RA, rheumatoid arthritis.

results showed an inverted *U*-shaped association between DFI and femoral BMD among males but not females. In addition, DFI was negatively associated with femoral neck BMD among males when DFI was >25 gm/day.

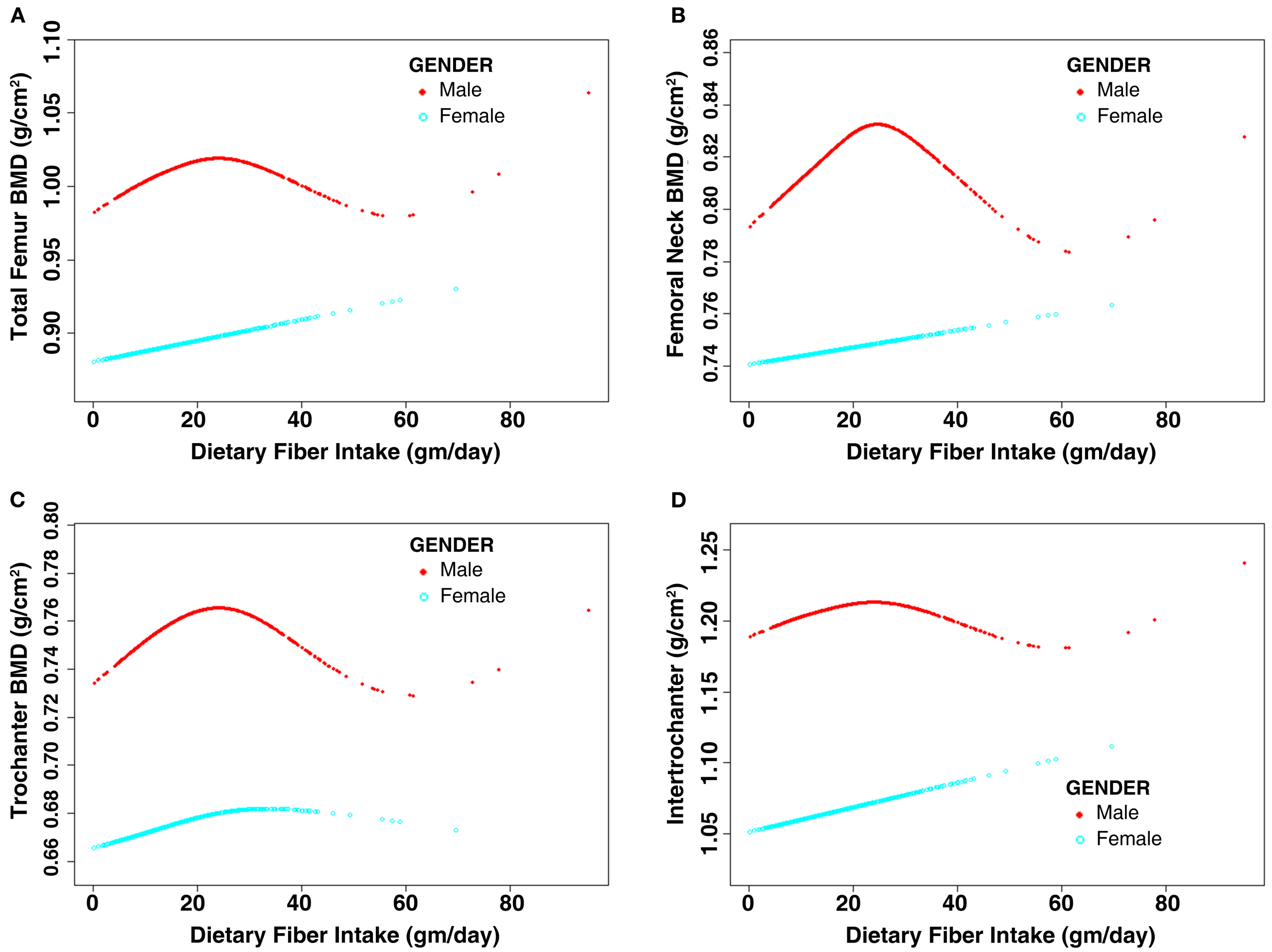


FIGURE 2 | Non-linear relationship between dietary fiber intake and femoral BMD stratified by sex. Age, race, education level, income level, BMI, smoking status, alcohol consumption, hypertension, diabetes, blood calcium level, serum 25-hydroxyvitamin D, RA, cancer, use of glucocorticoid, family history of osteoporosis, previous fractures, physical activity level, calcium intake level, and vitamin D intake level were adjusted. **(A)** Total femur BMD; **(B)** Femoral neck BMD; **(C)** Trochanter BMD; **(D)** Intertrochanter BMD. BMD, bone mineral density; BMI, body mass index; RA, rheumatoid arthritis.

TABLE 5 | Two-piecewise linear regression models of dietary fiber intake on BMD in males.

Index	Total femur BMD	Femoral neck BMD	Trochanter BMD	Intertrochanter BMD
Fitting by the standard linear model	−0.0001 (−0.0010, 0.0008)	−0.0002 (−0.0011, 0.0008)	−0.0000 (−0.0008, 0.0008)	−0.0002 (−0.0013, 0.0009)
Fitting by the two-piecewise linear model				
Inflection point (gm/day)	25	25	25	25
Dietary fiber intake < Infection point	0.0011 (−0.0005, 0.0027)	0.0015 (−0.0001, 0.0031)	0.0012 (−0.0002, 0.0025)	0.0007 (−0.0012, 0.0026)
Dietary fiber intake > Infection point	−0.0011 (−0.0026, 0.0004)	−0.0017 (−0.0032, −0.0002)	−0.0012 (−0.0025, 0.0002)	−0.0010 (−0.0028, 0.0008)

Age, race, education level, income level, BMI, smoking status, alcohol consumption, hypertension, diabetes, blood calcium level, serum 25-hydroxyvitamin D, RA, cancer, use of glucocorticoid, family history of osteoporosis, previous fractures, physical activity level, calcium intake level, and vitamin D intake level were adjusted. BMD, bone mineral density; BMI, body mass index; RA, rheumatoid arthritis.

This study found that DFI was positively associated with femoral BMD in the unadjusted model, but no correlation was observed between DFI and femoral BMD after adjusting for covariates. This finding seemed to be the important differences compared with the existing literature (14–16). Dai et al. observed that increased DFI was associated with reduced bone loss in men (14). Moreover, Lee and Suh found that DFI was positively associated with lumbar BMD in men aged 18–45 years (15). Zhou et al. demonstrated that a higher DFI was associated with higher heel BMD among individuals aged 40–69 years (16). There are several possible explanations for the discrepancy between our study and previous study. First, DFI may not be associated with BMD. DF is derived from whole-grain cereals, fruits, vegetables, and legumes, and these foods also contain other nutritional elements such as calcium and vitamin D, which are considered to play essential roles in maintaining bone mass (6, 7). High DFI might also be associated with high calcium or vitamin D intake, which might be a potential reason for the discrepancy between the present and previous studies. Therefore, we initially included the variables of calcium and vitamin D intake as covariates to avoid potential bias. However, our study found that no correlation was observed between DFI and femoral BMD after adjusting for all covariates. Second, DFI might be associated with BMD, but our study did not observe this because of the limitations of the present study. On the one hand, the DFI data were collected based on short-term intake, using short-term dietary intake as usual intake, to assess the association between DFI and femoral BMD; this might lead to a biased estimate of the association. Therefore, these findings also suggest that further studies on the relationship between DFI and BMD need to consider the influence of exposure time. Meanwhile, the information on DFI was collected based on self-report in the present study, which is a subjective parameter and might not reflect the actual DFI. Third, DFI might be associated with BMD, but the relationship between DFI and BMD was influenced by other factors, such as age, sex, or anatomical sites. In the present study, the association between DFI and BMD seemed to be modified by sex. Similarly, the association between DFI and BMD may be modified by other factors. For example, Lee and Suh found that DFI was associated with BMD in men aged 18–45 years but not in those aged over 65 years. Dai et al. observed that total DFI was correlated with femoral neck BMD but not lumbar BMD (14). Considering the limited number of related studies, additional studies are needed to confirm our hypothesis.

This study also observed sex differences in the association between DFI and femoral BMD. We considered that the sex differences might have resulted from hormone levels, especially sex hormones. Barron et al. observed that higher DFI was associated with lower lumbar BMD among young female athletes with oligomenorrhea (13), which is a symptom possibly caused by disorders of sex hormones. These findings combined with the results of our study suggested that DF might play various roles in different sex hormone levels. In addition, the impact of DF on the gut microbiota may have sex differences. Zhang et al. observed sex differences in the gut microbiome in response to DF supplementation in experimental animals (24). Similarly, Morrison et al. found a sex-specific effect of DFI on the gut microbiota community composition in animal experiments (25). However, there is no direct evidence supporting our hypotheses, and the mechanisms remain ambiguous. Therefore, further studies are needed to investigate this intriguing observation.

Interestingly, this study observed that DFI was negatively associated with femoral neck BMD among men when DFI was >25 gm/day, suggesting that high DFI might be unfavorable to prevent bone loss or even contribute to lower femoral BMD. We considered that there were some underlying mechanisms of high DFI leading to low BMDs. First, a high DFI might contribute to low femoral BMDs by altering the composition of the intestinal microbiota. Actually, high DFI could indeed alter the composition of the intestinal microbiota (26, 27). Moreover, cumulative evidence indicates that the gut microbiota is linked to bone metabolism (28, 29). However, further studies on the impact of DF on bone metabolism are needed to support our hypotheses because direct proof has been missing. Second, high DFI might contribute to low femoral BMDs by affecting hormone levels, such as estrogen levels. Wayne et al. found that high DFI was associated with low serum estradiol levels among postmenopausal breast cancer survivors (30). Similarly, Zengul et al. observed an inverse association between DFI and estradiol levels in postmenopausal women with breast cancer (31). However, these studies did not prove that DFI could directly affect estrogen metabolism, and no evidence has demonstrated that the inverse association between DFI and estrogen levels exists among middle-aged and older men. Therefore, whether high DFI might contribute to low femoral BMDs by reducing estrogen levels is an interesting topic for further study. Third, high DFI might contribute to low femoral BMDs by enhancing intestinal inflammation and affecting calcium and vitamin D absorption. Grabitske and Slavin suggested that a higher or excessive fiber intake might cause gastrointestinal effects, such as diarrhea and abdominal discomfort (32). Miles et al. demonstrated that the addition of inulin, a DF, exacerbated the severity of colitis induced by dextran sulfate sodium in mice (33). However, these studies did not directly prove our hypotheses, and the number of related studies is limited. Moreover, there were also several studies demonstrated that high DFI might be a protective factor for inflammatory bowel disease (34, 35). In addition, it remains unclear whether the negative correlation between DFI and femoral BMD has clear biological effects, such as increasing the risk of osteoporosis. Therefore, additional research is needed to explore whether high DFI contributes to lower femoral BMD or whether the negative effect of high DFI on femoral BMD only applies to specific populations.

The findings of the present study could also provide references or guidelines for daily routine practice and future research. Specifically, the findings of this study might provide a reference for the recommended intake of DF, especially in high-risk population. According to the 2020–2025 Dietary Guidelines for Americans (36), individuals aged 31–50 years and those aged over 50 years should consume at least 31 and 28 g of DF per day, respectively. In the present study, we observed that DFI was negatively associated with femoral neck BMD among men when DFI was >25 gm/day. Therefore, to prevent bone loss, excess DFI might not be appropriate for middle-aged and older men. However, high DFI might also be a protective factor against other diseases, such as coronary artery disease, cancer, and diabetes (37, 38). The number of studies on the impact of DF on bone metabolism was limited. Therefore, additional prospective studies are needed to determine the optimal threshold of DF intake. On the other hand, the findings of the present study might also provide a reference for future research on the relationship between DFI and bone metabolism. Except for the negative association between DFI and femoral BMD among men with high DFI (>25 gm/day), this study also observed sex differences in the relationship of DFI with femoral BMD between men and women. Although more studies are needed to investigate whether the negative correlation between DFI and femoral BMD has clear biological effects, such as increasing the risk of osteoporosis, this study offers a new perspective on the potential impact of DF on bone metabolism.

This study had some limitations. First, the DFI data were collected based on short-term intake, using short-term dietary intake as usual intake to assess the association between DFI and femoral BMD, which might lead to a biased estimate of the association. Second, the final analysis was based on individuals with complete data. Subjects with missing data were excluded from the present study, which might have produced bias. Third, the DFI data were collected based on subjective self-reports. Therefore, there might be some discrepancy between self-reported DFI and actual DFI. Fourth, the participants included in the final analysis were based on the general US population. Considering the differences in culture, lifestyle, and diet among different countries and regions, more studies are needed to investigate whether the conclusions of the present study are generally applicable. Finally, some unmeasured confounding variables (such as bone turnover markers), which are also considered important factors for bone metabolism, were not assessed in the present study because these variables were not available in the NHANES database, and the lack of adjustment for these potential factors may have biased the results.

In conclusion, our results indicate that DFI might not follow a linear relationship with femoral BMD, and sex factors might modify the association between DFI and BMD. In particular, high DFI (>25 gm/day) might contribute to lower femoral neck BMDs among males aged ≥ 40 years. However, more studies are needed to investigate whether the negative effect of high DFI on femoral BMD does exist and whether high DFI has clear biological effects on bone metabolism, such as increasing the risk of osteoporosis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by NCHS Research Ethics Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YT and JL conceived the study, data curation, data analysis, and draft writing. YT and XZ completed images and tables preparation. BG conceived the study, funding acquisition, and writin—review and editing. All authors participated in critical revision of the manuscript, contributed to the article, and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (81960403), Cuiying Scientific and Technological Innovation Program of Lanzhou University Second Hospital (CY2021-MS-A07), and Innovation Star Project for Excellent Graduate Students of the Education Department of Gansu Province (2021CXZX-143).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2022.851820/full#supplementary-material

REFERENCES

1. Compston JE, McClung MR, Leslie WD. Osteoporosis. *Lancet.* (2019) 393:364–76. doi: 10.1016/S0140-6736(18)32112-3
2. Chen P, Li Z, Hu Y. Prevalence of osteoporosis in China: a meta-analysis and systematic review. *BMC Public Health.* (2016) 16:1039. doi: 10.1186/s12889-016-3712-7
3. Cheng X, Zhao K, Zha X, Du X, Li Y, Chen S, et al. Opportunistic screening using low-dose CT and the prevalence of osteoporosis in China: a Nationwide, multicenter study. *J Bone Miner Res.* (2021) 36:427–35. doi: 10.1002/jbmr.4187
4. Johnell O, Kanis J. Epidemiology of osteoporotic fractures. *Osteoporos Int.* (2005) 16(Suppl. 2):S3–7. doi: 10.1007/s00198-004-1702-6
5. Lu J, Ren Z, Liu X, Xu YJ, Liu Q. Osteoporotic fracture guidelines and medical education related to the clinical practices: a Nationwide survey in China. *Orthop Surg.* (2019) 11:569–77. doi: 10.1111/os.12476
6. Klibanski A, Adams-Campbell L, Bassford T, Blair SN, Boden SD, Dickersin K, et al. Osteoporosis prevention, diagnosis, and therapy. *JAMA.* (2001) 285:785–95. doi: 10.1001/jama.285.6.785
7. Warensjö E, Byberg L, Melhus H, Gedeborg R, Mallmin H, Wolk A, et al. Dietary calcium intake and risk of fracture and osteoporosis: prospective longitudinal cohort study. *BMJ.* (2011) 342:d1473. doi: 10.1136/bmj.d1473
8. Weickert MO, Pfeiffer AF. Metabolic effects of dietary fiber consumption and prevention of diabetes. *J Nutr.* (2008) 138:439–42. doi: 10.1093/jn/138.3.439
9. Jones JM. CODEX-aligned dietary fiber definitions help to bridge the 'fiber gap'. *Nutr J.* (2014) 13:34. doi: 10.1186/1475-2891-13-34
10. Tanaka S, Yoshimura Y, Kamada C, Tanaka S, Horikawa C, Okumura R, et al. Intakes of dietary fiber, vegetables, and fruits and incidence of cardiovascular disease in Japanese patients with type 2 diabetes. *Diabetes Care.* (2013) 36:3916–22. doi: 10.2337/dc13-0654
11. Fujii H, Iwase M, Ohkuma T, Ogata-Kaizu S, Ide H, Kikuchi Y, et al. Impact of dietary fiber intake on glycemic control, cardiovascular risk factors and chronic kidney disease in Japanese patients with type 2 diabetes mellitus: the Fukuoka Diabetes Registry. *Nutr J.* (2013) 12:159. doi: 10.1186/1475-2891-12-159
12. Ananthakrishnan AN, Khalili H, Konijeti GG, Higuchi LM, de Silva P, Korzenik JR, et al. A prospective study of long-term intake of dietary fiber and risk of Crohn's disease and ulcerative colitis. *Gastroenterology.* (2013) 145:970–7. doi: 10.1053/j.gastro.2013.07.050
13. Barron E, Cano Sokoloff N, Maffazioli GDN, Ackerman KE, Woolley R, Holmes TM, et al. Diets high in fiber and vegetable protein are associated with low lumbar bone mineral density in young athletes with oligoamenorrhea. *J Acad Nutr Diet.* (2016) 116:481–9. doi: 10.1016/j.jand.2015.10.022

14. Dai Z, Zhang Y, Lu N, Felson DT, Kiel DP, Sahni S. Association between dietary fiber intake and bone loss in the Framingham offspring study. *J Bone Miner Res.* (2018) 33:241–9. doi: 10.1002/jbmr.3308
15. Lee T, Suh HS. Associations between dietary fiber intake and bone mineral density in adult Korean population: analysis of National Health and Nutrition Examination Survey in 2011. *J Bone Metab.* (2019) 26:151–60. doi: 10.11005/jbm.2019.26.3.151
16. Zhou T, Wang M, Ma H, Li X, Heianza Y, Qi L. Dietary fiber, genetic variations of gut microbiota-derived short-chain fatty acids, and bone health in UK Biobank. *J Clin Endocrinol Metab.* (2021) 106:201–10. doi: 10.1210/clinem/dgaa740
17. CDC. *National Health and Nutrition Examination Survey 2013-2014.* (2022). Available online at: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013 (accessed January 1, 2022).
18. CDC. *NCHS Research Ethics Review Board (ERB) Approval.* (2022). Available online at: https://www.cdc.gov/nchs/nhanes/irba98.htm (accessed January 1, 2022).
19. CDC. *Body Composition Procedures Manual.* (2022). Available online at: https://wwwn.cdc.gov/nchs/data/nhanes/2013-2014/manuals/2013_Body_Composition_DXA.pdf (accessed January 1, 2022).
20. CDC. *MEC In-Person Dietary Interviewers Procedures Manual.* (2022). Available online at: https://wwwn.cdc.gov/nchs/data/nhanes/2013-2014/manuals/mec_in_person_dietary_procedures_manual_jan_2014.pdf (accessed January 1, 2022).
21. CDC. *Phone Follow-Up Dietary Interviewer Procedures Manual.* (2022). Available online at: https://wwwn.cdc.gov/nchs/data/nhanes/2013-2014/manuals/Phone_Follow-up_Dietary_Interviewers_Manual.pdf (accessed January 1, 2022).
22. Kanis JA, Cooper C, Rizzoli R, Reginster JY. European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporos Int.* (2019) 30:3–44. doi: 10.1007/s00198-018-4704-5
23. Ensrud KE, Crandall CJ. Osteoporosis. *Ann Intern Med.* (2017) 167:Itc17-itc32. doi: 10.7326/AITC201708010
24. Zhang Z, Hyun JE, Thiesen A, Park H, Hotte N, Watanabe H, et al. Sex-specific differences in the gut microbiome in response to dietary fiber supplementation in IL-10-deficient mice. *Nutrients.* (2020) 12:2088. doi: 10.3390/nu12072088
25. Morrison KE, Jašarević E, Howard CD, Bale TL. It's the fiber, not the fat: significant effects of dietary challenge on the gut microbiome. *Microbiome.* (2020) 8:15. doi: 10.1186/s40168-020-0791-6
26. Makki K, Deehan EC, Walter J, Bäckhed F. The impact of dietary fiber on gut microbiota in host health and disease. *Cell Host Microbe.* (2018) 23:705–15. doi: 10.1016/j.chom.2018.05.012
27. So D, Whelan K, Rossi M, Morrison M, Holtmann G, Kelly JT, et al. Dietary fiber intervention on gut microbiota composition in healthy adults: a systematic review and meta-analysis. *Am J Clin Nutr.* (2018) 107:965–83. doi: 10.1093/ajcn/nqy041
28. Ozaki D, Kubota R, Maeno T, Abdelhakim M, Hitosugi N. Association between gut microbiota, bone metabolism, and fracture risk in postmenopausal Japanese women. *Osteoporos Int.* (2021) 32:145–56. doi: 10.1007/s00198-020-05728-y
29. Li C, Huang Q, Yang R, Dai Y, Zeng Y, Tao L, et al. Gut microbiota composition and bone mineral loss-epidemiologic evidence from individuals in Wuhan, China. *Osteoporos Int.* (2019) 30:1003–13. doi: 10.1007/s00198-019-04855-5
30. Wayne SJ, Neuhouser ML, Ulrich CM, Koprowski C, Baumgartner KB, Baumgartner RN, et al. Dietary fiber is associated with serum sex hormones and insulin-related peptides in postmenopausal breast cancer survivors. *Breast Cancer Res Treat.* (2008) 112:149–58. doi: 10.1007/s10549-007-9834-y
31. Zengul AG, Demark-Wahnefried W, Barnes S, Morrow CD, Bertrand B, Berryhill TF, et al. Associations between dietary fiber, the fecal microbiota and estrogen metabolism in postmenopausal women with breast cancer. *Nutr Cancer.* (2021) 73:1108–17. doi: 10.1080/01635581.2020.1784444
32. Grabitske HA, Slavin JL. Gastrointestinal effects of low-digestible carbohydrates. *Crit Rev Food Sci Nutr.* (2009) 49:327–60. doi: 10.1080/10408390802067126
33. Miles JP, Zou J, Kumar MV, Pellizzon M, Ulman E, Ricci M, et al. Supplementation of low- and high-fat diets with fermentable fiber exacerbates severity of dss-induced acute colitis. *Inflamm Bowel Dis.* (2017) 23:1133–43. doi: 10.1097/MIB.0000000000001155
34. Hou JK, Abraham B, El-Serag H. Dietary intake and risk of developing inflammatory bowel disease: a systematic review of the literature. *Am J Gastroenterol.* (2011) 106:563–73. doi: 10.1038/ajg.2011.44
35. Milajerdi A, Ebrahimi-Daryani N, Dieleman LA, Larijani B, Esmaillzadeh A. Association of dietary fiber, fruit, and vegetable consumption with risk of inflammatory Bowel disease: a systematic review and meta-analysis. *Adv Nutr.* (2021) 12:735–43. doi: 10.1093/advances/nmaa145
36. Swanson CM, Kohrt WM, Buxton OM, Everson CA, Wright KP, Jr., et al. The importance of the circadian system & sleep for bone health. *Metabolism.* (2018) 84:28–43. doi: 10.1016/j.metabol.2017.12.002
37. Anderson JW, Baird P, Davis RH, Jr., Ferreri S, Knudtson M, et al. Health benefits of dietary fiber. *Nutr Rev.* (2009) 67:188–205. doi: 10.1111/j.1753-4887.2009.00189.x
38. Veronese N, Solmi M, Caruso MG, Giannelli G, Osella AR, Evangelou E, et al. Dietary fiber and health outcomes: an umbrella review of systematic reviews and meta-analyses. *Am J Clin Nutr.* (2018) 107:436–44. doi: 10.1093/ajcn/nqx082

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

BRIEF RESEARCH REPORT
published: 08 August 2022
doi: 10.3389/fnut.2022.833354

Lower Energy-Adjusted Nutrient Intakes Occur Among Food Energy Under-Reporters With Poor Mental Health

Karen M. Davison [1]*, *Vanessa Araujo Almeida* [2] *and Lovedeep Gondara* [3]

[1] Health Science, Kwantlen Polytechnic University, Richmond, BC, Canada, [2] College of Tropical Agriculture & Human Resources, University of Hawai'i at Mānoa, Honolulu, HI, United States, [3] Computing Science, Simon Fraser University, Burnaby, BC, Canada

Background: Food energy under-reporting is differentially distributed among populations. Currently, little is known about how mental health state may affect energy-adjusted nutrient intakes among food energy under-reporters.

Methods: Stratified analysis of energy-adjusted nutrient intake by mental health (poor vs. good) and age/sex was conducted using data from Canadian Community Health Survey (CCHS) respondents (14–70 years; $n = 8,233$) who were deemed as under-reporters based on Goldberg's cutoffs.

Results: Most were experiencing good mental health (95.2%). Among those reporting poor mental health, significantly lower energy-adjusted nutrient intakes tended to be found for fiber, protein, vitamins A, B_2, B_3, B_6, B_9, B_{12}, C, and D, and calcium, potassium, and zinc (probability measures (p) < 0.05). For women (51–70 years), all micronutrient intakes, except iron, were significantly lower among those reporting poor mental health ($p < 0.05$). For men (31–50 years), B vitamin and most mineral intakes, except sodium, were significantly lower among those reporting poor mental health ($p < 0.05$). Among women (31–50 years) who reported poor mental health, higher energy-adjusted intakes were reported for vitamin B_9 and phosphorus ($p < 0.05$).

Conclusions: Among food energy under-reporters, poor mental health tends to lower the report of specific energy-adjusted nutrient intakes that include ones critical for mental health. Future research is needed to discern if these differences may be attributed to deviations in the accurate reports of food intakes, measurement errors, or mental health states.

Keywords: mental health, under-reporting, nutrition, measurement error, dietary intakes

OPEN ACCESS

Edited by:
Igor Pravst,
Institute of Nutrition, Slovenia

Reviewed by:
Aida Turrini,
Independent Researcher, Rome, Italy
Rajko Vidrih,
University of Ljubljana, Slovenia

****Correspondence:***
Karen M. Davison
Karen.davison@kpu.ca

Specialty section:
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

Received: *11 December 2021*
Accepted: *20 June 2022*
Published: *08 August 2022*

Citation:
Davison KM, Araujo Almeida V and Gondara L (2022) Lower Energy-Adjusted Nutrient Intakes Occur Among Food Energy Under-Reporters With Poor Mental Health. Front. Nutr. 9:833354. doi: 10.3389/fnut.2022.833354

INTRODUCTION

A longstanding criticism of self-reported dietary intake data is the underestimation of dietary energy intake (EI) in relation to requirements, commonly referred to as food energy under-reporting (1, 2). This measurement issue that appears to occur non-randomly (1–3) can lead to an inaccurate assessment of the relationships between diet and health (4, 5). Adjustments for EI in the evaluation of nutrient intakes may produce more valid findings as it controls for confounding and removes extraneous variation resulting from factors such as metabolic efficiency (6). A recent study reported that estimates of EI in the 2015 Canadian Community Health Survey (CCHS) were lower than those reported in 2004. The authors suggested that increased misreporting of dietary intake may explain part of this difference (7).

Food energy under-reporting is differentially distributed among populations. Factors associated with food energy under-reporting include female sex, older age, income, body weight status and history, diet composition (e.g., macronutrients), eating behaviors, social desirability, body image, and physical activity (8–10). These characteristics, which contribute to differences in nutrient analysis results, have not been reported in those identified as food energy under-reporters and plausible reporters. Furthermore, although behavior-related issues are relevant, studies of mental health state at the time of dietary intake data collection and its potential effects on nutrient analysis results, particularly among food energy under-reporters, have not been investigated.

The limited research related to mental health state and food energy under-reporting has mainly focused on individuals with a diagnosed condition. A small study, which compared food energy under-reporting in women with schizophrenia and controls, found that food energy under-reporting was more prevalent among those with the mental health condition (77%) vs. those without (50%) (11). In another study that examined individuals with mood disorders, it was found that food energy under-reporting was associated with diet quality, a history of weight change after taking psychiatric medication, and female sex (12). Depending on the type of regression models analyzed, women with probable major depressive episodes (13) or individuals with prior depression diagnosis (14) may have increased odds of food energy under-reporting. Further research is needed about those who report poor mental health, not necessarily those with a diagnosed condition, as this state of mind, which can impact overall functioning, is more common among different populations (15).

To help address gaps in knowledge about the effects of mental health state on nutrient intake analysis results among food energy under-reporters, data from a large, national sample from the CCHS were analyzed. The objective of the analysis was to examine if there are differences in energy-adjusted nutrient intakes among food energy under-reporters experiencing good and poor mental health by age and sex categories. It is hypothesized that the energy-adjusted nutrient intakes among food energy under-reporters will be significantly lower among those experiencing poor mental health when compared to those who report good mental health. The results from the analysis of this national survey may help to determine if mental health state is a factor to account for in studies that include dietary intakes.

METHODS

Sample of Food Energy Under-Reporters

The sample was derived from Statistics Canada's CCHS – Cycle 2.2 (2004), which provides the only Canadian national data to date that includes both detailed nutrient intake data and a measure of mental health (16). This survey included 35,107 respondents who were living in private residences in all of Canada's 10 provinces. It excluded full-time members of the Canadian Forces and individuals who lived on First Nation Reserves or Crown Lands, in prisons or care facilities, or in some remote areas due to resource limitations or that the health services delivered differ from the general population. Approval for the use of the de-identified dataset was granted by Statistics Canada. All data were vetted by a Statistics Canada analyst prior to release to ensure that respondent privacy was maintained. Institutional Review Board ethics approval was not required.

The sample included CCHS respondents between the age of 14 and 70 years (22,709) who were considered food energy under-reporters as defined by Goldberg's cutoffs for EI-to-basal metabolic rate (BMR) (5). EI plausibility was based on the ratio of self-reported EI from 24-h dietary intake recalls (EIrep) to BMR. Subjects with an EIrep:BMR ratio less than 1.36 were categorized as under-reporters (2). Estimated energy requirements (EERs) (17) were based on respondents' sex, age, self-reported physical activity level, and the self-reported or measured height and weight. The physical activity coefficients used in the EER equation were based on three levels: active, moderately active, or inactive (16).

Energy under-reporting is an important challenge in nutrition epidemiology as it affects the estimation of EI and consequently of other nutrients, which then may lead to a mis-estimation of nutrient inadequacy and bias in the associations between diet and diseases. Given that key characteristics of under-reporters are being women, younger age, and having non-favorable self-reported health perception status (14), the focus of this study was on characterizing energy-adjusted nutrient intakes in energy under-reporters by sex, age, and mental health state. This would enable quantification of the problem, identification of key nutrient intakes that are impacted, and help to identify strategies of how energy under-reporting may be mitigated in future studies.

Dietary Intake

Dietary intake data were based on 24-h dietary intake recalls that were conducted in-person and included the use of the multi-pass method. For a subset of CCHS respondents, a follow-up 24-h recall was done by telephone between 3 and 10 days after the first interview and this data helped to adjust for day-to-day variability. Energy-adjusted nutrient intakes were derived using the density method where values are reported per 1,000 kcal (18). The Canadian Nutrient File (CNF) was used as the nutrient analysis database. The CNF only had complete values of vitamin E (alpha-tocopherol) for 46% of the foods; therefore, vitamin E intakes were not reported.

Perceived Mental Health

Perceived mental health, a variable that captures the various dimensions of mental health experiences, was used to stratify the sample by mental health status. The variable is based on responses to the question "How would you say your mental health is: excellent? very good? good? fair? poor?". The variable was dichotomized as poor mental health (poor/fair responses) and good mental health (good/very good/excellent responses) as has been commonly done in various studies (19–21). Perceived mental health is an indicator for some forms of mental disorder, mental or emotional problems, or distress (22, 23). It has been associated with mental morbidity measures, such as non-specific psychological distress, depressive symptoms, activity limitations, and physical and emotional role functioning (24–27). A recent epigenome-wide association study (EWAS) assessed the predictive value of methylation beta values of EWAS that identified CpGs (5'-C-phosphate-G-3') for incidence of depressive symptoms in later life and found that subjective mental health and hypomethylation at cg27115863 are predictive of depressive symptoms, which are thought to be due to activation of the inflammatory signaling pathway (28).

Stratified Analysis

For those who were food energy under-reporters, stratified analysis was conducted according to perceived mental health and sex/age categories (14–19, 20–30, 31–50, and 51–70 years). The secured data were analyzed in the Statistics Canada Research Data Center at the University of British Columbia using SAS (version 9.1, 2003, SAS Institute) and Software for Intake Distribution Estimation in IML language (SIDE-IML, version 1.11, 2001, Iowa State University). Survey weights provided by Statistics Canada were incorporated into the calculations to provide national representation, and the bootstrap re-sampling technique was used (16). Nutrient intake values were stratified by age/sex categories and reported using the median and inter-quartile range. Given that the normality assumption is untenable for most nutrient intake distributions (29), statistical comparisons by mental health status within age/sex categories were done using Mann-Whitney *U* tests.

RESULTS

Of those who participated in the CCHS, between 14 and 70 years (8,233/22,709), 36.3% were considered as food energy under-reporters and formed the basis of the sample used in this investigation. Based on weighted frequencies, 8.9% were between 14 and 19 years, 21.2% were between 20 and 30 years, 41.8% were between 31 and 50 years, and 28.0% were between 51 and 70 years. Within this sample ($n = 8,233$), 95.2% reported good mental health and 51.3% were women.

Energy, Fiber, and Macronutrients

Among men between 31 and 50 years, energy-adjusted fiber and protein intakes were significantly lower in those reporting poor mental health state (probability measures (p) < 0.05; **Supplementary Figure S1**); conversely, carbohydrate intakes were significantly higher among those reporting poor mental health (**Supplementary Figure S1a**). For women, significantly lower intakes for protein (31–50 years) and fiber (31–70 years) were reported among those experiencing poor mental health ($p < 0.05$; **Supplementary Figure S1b**).

Micronutrients

Among men 20–30 years who reported poor mental health, significantly lower energy-adjusted intakes for vitamins B_2 and C were found ($p < 0.05$; **Supplementary Figure S2a**). Similar results were found for intakes of all B vitamins ($p < 0.05$) for men between 31 and 50 years (**Supplementary Figure S2b**) and vitamins A and D ($p < 0.05$) for men between 51 and 70 years (**Supplementary Figure S2c**). Among women between 14 and 19 years, energy-adjusted vitamin A intakes were lower among those with poor mental health (**Supplementary Figure S2d**). Across other age groups for women, energy-adjusted vitamin B_6 and C intakes (20–30 years; $p < 0.05$; **Supplementary Figure S2d**) and intakes of vitamins A and B_3 (31–50 years; $p < 0.05$) were significantly lower among those reporting poor mental health (**Supplementary Figure S2e**). For women between 51 and 70 years and reporting poor mental health, all vitamin intakes ($p < 0.05$) were significantly lower as compared to those reporting good mental health (**Supplementary Figure S2f**). Interestingly, among women between 31 and 50 years, vitamin B_9 intakes were significantly higher among the group with poor mental health ($p < 0.05$; **Supplementary Figure S2e**).

For mineral intakes, several significant differences by mental health state were also found. Among men 20–30 years, significantly lower energy-adjusted intakes of calcium and zinc were found for those reporting poor mental health (**Supplementary Figure S3a**). Among men 31–50 years, similar results were indicated for all minerals except sodium (**Supplementary Figures S3b,c**). For men between 51 and 70 years, calcium intakes were significantly lower among those reporting poor mental health ($p < 0.05$; **Supplementary Figure S3a**). Among women, significantly lower intakes of energy-adjusted calcium were found for those between 20 and 30 years (**Supplementary Figure S3d**), and lower calcium, phosphorus, potassium, and sodium intakes were found for those between 51 and 70 years (**Supplementary Figure S3e**). Among women of 51–70 years, magnesium and zinc intakes were also significantly lower among those reporting poor mental health ($p < 0.05$; **Supplementary Figure S3f**).

Overall, reported energy-adjusted nutrient intake differences tended to be significantly lower in those reporting poor mental health. Exceptions to this included reported carbohydrate intakes in men 31–50 years as well as vitamin B_9 and phosphorus in women 31–50 years, where energy-adjusted nutrient intakes were significantly higher among those reporting poor mental health.

DISCUSSION

Given that most energy-adjusted nutrient intakes were significantly lower among most groups reporting poor mental health, our hypothesis that significantly lower energy-adjusted nutrient intakes would be observed among those with poor mental health was supported. This, however, was not the case for carbohydrate intakes among men 31–50 years, as well as vitamin B_9 and phosphorus intakes among women 31–50 years, where significantly higher intakes were reported among those reporting poor mental health. Poor mental health state appeared to lower reported energy-adjusted nutrient intakes for protein, fiber, most of the B vitamins, and the majority of minerals, particularly among women and those between 31 and 70 years.

Although it appears that mental health state significantly impacts the report of energy-adjusted nutrient intakes, it is unclear whether those reporting poor mental health state are more prone to under-report food intakes due to reasons such as impairments in recall of food intake (30) or that they are simply consuming less food. In a study, which explored perceived mental health and dietary intakes in the same dataset analyzed for this study, it was reported that those reporting poorer mental health consumed diets of lower quality based on the Canadian Healthy Eating Index (20). In another study, it was indicated that intakes of vitamins B_1, B_2, B_6, B_9, B_{12}, phosphorus, and zinc were significantly lower among individuals with verified mood disorders when compared to a healthy population sample (31). Individuals with poor mental health status who are taking psychiatric medications may experience alterations in usual dietary intakes (32), which could contribute to differences in nutrient intake by mental health state. This would suggest that during data collection, mental health state and medication use should be accounted for and validation approaches, such as the multi-pass method, should be used to help to ensure the reliability of the recorded information. Given the potential impact that mental health state has on reporting of energy-adjusted nutrient intakes, it is questioned whether the results of studies that indicate differences in the reporting of dietary patterns and their associations with mental health outcomes are accurate (33). Our results have highlighted issues related to processes that may cause people to under-report their food intakes. Thus, multidisciplinary approaches, that could include psychology and pathophysiology, are needed to advance the understanding of mental health state and the under-reporting of dietary intake (34).

The findings of significantly higher intakes of carbohydrates among men 31–50 years, as well as vitamin B_9 and phosphorus among women 31–50 years who reported poor mental health, were surprising. Results of observational studies indicate that recurrent hypoglycemia is associated with poor mental health (35) and this may contribute to increased cravings for carbohydrates and intakes of the macronutrient. Previous studies have shown a positive association between the consumption of soft drinks, which contain high levels of phosphate additives and mental health concerns (36). Intakes of foods with high amounts of folate, have been reported to improve mental health and mood (37). Individuals who are experiencing poor mental health and trying to improve their symptoms may increase intakes of foods which are rich sources of folate.

Implications

The findings of this study are consistent with others that suggest that energy under-reporting is an issue in research that examines trends in food intakes (38). In particular, our results suggest that dietary intake assessments should utilize the most accurate methods to assess dietary exposures and account for mental health state that is measured by valid tools. If mental health improvements are part of a dietary intervention's goals, particular attention should be made to ensure foods, which are sources of nutrients critical to mental health, such as the omega 3 fatty acids, folate, and iron (39), are accurately recorded. Previous investigations indicate that under-reporting of food intakes tends to occur during afternoon snacks, dinner, and breakfast (40), suggesting intakes reported at these times of the day require additional attention during a dietary assessment. It has been identified that factors, such as lack of physical exercise and substance use may impact dietary recall (41). For individuals with severe mental health symptoms, food-frequency questionnaires, brief dietary assessment instruments, food image assessments, and wearable cameras may be helpful (41). However, further research is needed to ascertain how accurate these alternatives are in populations with mental health concerns. Ongoing investigations of under-reporting related to mental health status are needed to examine whether the findings observed in this study occur across different subpopulations that include those at different life stages, such as children (42). Furthermore, predictive modeling that can examine a number of factors will better ascertain the relationship between perceived mental health and energy under-reporting. Finally, it is recommended that in large-scale nutrition epidemiology studies, a proportion of the participants experiencing good and poor mental health should be selected and their dietary intake results validated by employing methods such as alternative dietary assessment, examining nutritional status (e.g., anthropometric measures), and measuring nutrition-related biomarkers (43–45).

Limitations

Although the Goldberg cutoffs are less accurate than objective methods, such as the use of doubly labeled water biomarkers to reference EI, they are considered appropriate for energy under-reporting classification (5). To better identify food energy under-reporters, detailed information on occupation and leisure activity to derive subject-specific physical activity levels to evaluate individual EI should be used. The inflation of the type I error rate from multiple statistical testing may have overestimated the impact that poor mental health has on reporting of energy-adjusted nutrient intakes. Due to limited sample size within groups stratified by age and sex and limitations of variables available in the CCHS dataset, other factors, such as eating behavior (e.g., eating restraint), social desirability, dieting, body image, and race/ethnicity (4, 5, 46), which may mediate or moderate the relationships between mental health state and dietary intakes, could not be assessed. Finally, it has been reported elsewhere that individuals experiencing depression have lower total energy expenditure (47), which raises questions about how food energy under-reporting may be defined in those with poor mental health.

CONCLUSIONS

The report of energy-adjusted nutrient intakes tends to differ among those defined as food energy under-reporters reporting poor and good mental health. This suggests that the mental health state needs to be accounted for when dietary intake assessments are undertaken. This is particularly critical given that diet is becoming increasingly recognized as both a prevention and an intervention target to support mental health (48–50). Future research is needed to discern if deviations in energy-adjusted nutrient intake by mental health state among food energy under-reporters may be attributed to differences in the accurate reports of food intakes or a function of measurement error.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The data used is from Statistics Canada. Data is secured by Statistics Canada and permission to share data publicly cannot be granted as it might compromise patient confidentiality or participant privacy. The Canadian Community Health Survey data for this analysis was collected by Statistics Canada (third party data). Details about how to access Statistics Canada data are available at: https://www.statcan.gc.ca/eng/rdc/index. Researchers who have been sworn in as 'deemed employees' of Statistics Canada can access the confidential microdata files for approved projects through Statistics Canada's Research Data Centres (RDCs). The confidential microdata files contain information collected during the survey, derived variables, and the Bootstrap weights used to calculate the exact variance. Requests to access these datasets should be directed to Statistics Canada: https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs.html.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

KD submitted the project proposal that included the research plan to Statistics Canada for approval to access the secure data. KD and LG analyzed the data. KD and VA drafted the manuscript. All authors read and provided edits on manuscript drafts and approved the final manuscript.

FUNDING

This project was partially supported by Fulbright Canada funds.

ACKNOWLEDGMENTS

Gratitude is extended to Didier Garriguet of Statistics Canada who provided tools that helped with the use of SIDE-IML.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2022.833354/full#supplementary-material

REFERENCES

1. Livingstone MBE, Black AE. Markers of the validity of reported energy intake. *J Nutr.* (2003) 133:895S–920S. doi: 10.1093/jn/133.3.895S
2. Tooze JA, Subar AF, Thompson FE, Troiano R, Schatzkin A, Kipnis V. Psychosocial predictors of energy underreporting in a large doubly labelled water study. *Am J Clin Nutr.* (2004) 79:795–804. doi: 10.1093/ajcn/79.5.795
3. Mattisson I, Wirfalt E, Aronsson CA, Wallstrom P, Sonestedt E, Gullberg B et al. Misreporting of energy: prevalence, characteristics of misreporters and influence on observed risk estimates in the Malmo Diet and Cancer cohort. *Br J Nutr.* (2005) 94:832–42. doi: 10.1079/BJN20051573
4. Jakes RW, Day NE, Luben R, Welch A, Bingham S, Mitchell J et al. Adjusting for energy intake-What measure to use in nutritional epidemiological studies? *Int J Epidemiol.* (2004) 33:1382–6. doi: 10.1093/ije/dyh181
5. Black AE, Goldberg GR, Jebb SA, Livingstone MBE, Cole TF, Prentice AM. Critical evaluation of energy intake data using fundamental principles of energy physiology: 2. Evaluating the results of published surveys. *Eur J Clin Nutr.* (1991) 45:583–99.
6. Willett WC, Howe GR, Kushi LH. Adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr.* (1997) 65:1220S−1228S. doi: 10.1093/ajcn/65.4.1220S
7. Garriguet D. Accounting for misreporting when comparing energy intake across time in Canada. *Health Rep.* (2018) 29:3–12.
8. Murakami K, Miyake Y, Sasaki S, Tanaka K, Arakawa M. Characteristics of under- and over-reporters of energy intake among Japanese children and adolescents: The Ryukyus Child Health Study. *Nutr.* (2012) 28:532–8. doi: 10.1016/j.nut.2011.08.011
9. Maurer J, Douglas DL, Teixeira PJ, Thomson CA, Lohman TG, Going SB, et al. The psychosocial and behavioral characteristics related to energy misreporting. *Nutr Rev.* (2006) 64:53–66. doi: 10.1111/j.1753-4887.2006.tb00188.x
10. Poppitt SD, Swann D, Black AE, Prentice AM. Assessment of selective underreporting of food intake by both obese and non-obese women in a metabolic facility. *Int J Obes Relat Metab Disord.* (1998) 22:303–11. doi: 10.1038/sj.ijo.0800584
11. Khazaal Y, Rothen S, Morinière Trombert N, Frésard E, Zullino DF. Dietary underreporting in women with schizophrenia requiring dietary intervention: a case control study. *Eat Weight Disord.* (2007) 12:83–5. doi: 10.1007/BF03327600
12. Davison KM. Energy under-reporting in adults with mood disorders: prevalence and associated factors. *Eat Weight Disord.* (2013) 18:323–7. doi: 10.1007/s40519-013-0046-7
13. Lutomski JE, van den Broeck J, Harrington J, Shiely F, Perry IJ. Sociodemographic, lifestyle, mental health and dietary factors associated with direction of misreporting of energy intake. *Public Health Nutr.* (2011) 14:532–41. doi: 10.1017/S1368980010001801
14. Magalhães V, Severo M, Torres D, Ramos E, Lopes C. Characterizing energy intake misreporting and its effects on intake estimations, in the Portuguese adult population. *Public Health Nutr.* (2020) 23:1031–40. doi: 10.1017/S1368980019002465
15. Chiu M, Amartey A, Wang X, Vigod S, Kurdyak P. Trends in objectively measured and perceived mental health and use of mental health services: a population-based study in Ontario, 2002-2014. *CMAJ.* (2020) 192:E329–E337. doi: 10.1503/cmaj.190603
16. Statistics Canada. Canadian Community Health Survey (CCHS) Cycle 2.2 (2004). Nutrition: General Health and 24-h Dietary Recall Components User Guide (2006). Ottawa, ON, Canada, Statistics Canada. (accessed October 30, 2021).

17. Rennie K, Coward A, Jebb S. Estimating under-reporting of energy intake in dietary surveys using an individualised method. *Br J Nutr.* (2007) 97:1169–76. doi: 10.1017/S0007114507433086
18. Murakami K, Sasaki S, Takahashi Y, Uenishi K, Yamasaki M, Hayabuchi H et al. Misreporting of dietary energy, protein, potassium and sodium in relation to body mass index in young Japanese women. *Eur J Clin Nutr.* (2008) 62:111–8. doi: 10.1038/sj.ejcn.1602683
19. Puyat JH. Is the influence of social support on mental health the same for immigrants and non-immigrants? *J Immigr Minor Health.* (2013) 15:598–605. doi: 10.1007/s10903-012-9658-7
20. Davison KM, Gondara L, Kaplan BJ. Food insecurity, poor diet quality, and suboptimal intakes of folate and iron are independently associated with perceived mental health in Canadians. *Nutrients*. (2017) 9:274. doi: 10.3390/nu9030274
21. Manor O, Matthews S, Power C. Dichotomous or categorical response? Analysing self-rated health and lifetime social class. *Int J Epidemiol.* (2000) 29:149–57. doi: 10.1093/ije/29.1.149
22. Mawani FN, Gilmour H. Validation of self-rated mental health. *Health Rep.* (2010) 21:61–75.
23. Rust KF, Rao JN. Variance estimation for complex surveys using replication techniques. *Stat Methods Med Res.* (1996) 5:283–310. doi: 10.1177/096228029600500305
24. Statistics Canada. Perceived Mental Health (2006). Ottawa, ON, Canada, Statistics Canada. Available online at: http://www.statcan.gc.ca/pub/82-229-x/2009001/status/pmh-eng.htm (accessed October 30, 2021).
25. Fleishman JA, Zuvekas SH. Global self-rated mental health: Associations with other mental health measures and with role functioning. *Med Care.* (2007) 45:602–9. doi: 10.1097/MLR.0b013e31803bb4b0
26. Health Profile-Definitions. *Sources and Symbols Statistics* Canada, Ottawa, ON. (2014)
27. Ahmad F, Jhajj AK, Stewart DE, et al. Single item measures of self-rated mental health: a scoping review. *BMC Health Serv Res.* (2014) 14:398. doi: 10.1186/1472-6963-14-398
28. Perna L, Zhang Y, Matias-Garcia PR, Ladwig KH, Wiechmann T, Wild B, et al. Subjective mental health, incidence of depressive symptoms in later life, and the role of epigenetics: results from two longitudinal cohort studies. *Transl Psychiatry.* (2020) 10:323. doi: 10.1038/s41398-020-00997-x
29. Carriquiry AL. Estimation of usual intake distributions of nutrients and foods. *J Nutr.* (2003) 133:601S—8S. doi: 10.1093/jn/133.2.601S
30. Dillon DG, Pizzagalli DA. Mechanisms of memory disruption in depression. *Trends Neurosci.* (2018) 41:137–49. doi: 10.1016/j.tins.2017.12.006
31. Davison KM, Kaplan BJ. Vitamin and mineral intakes in adults with mood disorders: comparisons to nutrition standards and associations with sociodemographic and clinical variables. *J Am Coll Nutr.* (2011) 30:547–58. doi: 10.1080/07315724.2011.10720001
32. Davison KM. The relationships among psychiatric medications, eating behaviors, and weight. *Eat Behav.* (2013) 14:87–91. doi: 10.1016/j.eatbeh.2013.01.001
33. Markussen MS, Veierød MB, Ursin G, Andersen LF. The effect of under-reporting of energy intake on dietary patterns and on the associations between dietary patterns and self-reported chronic disease in women aged 50-69 years. *Br J Nutr.* (2016) 116:547–58. doi: 10.1017/S000711451600218X
34. Macdiarmid J, Blundell J. Assessing dietary intake: Who, what and why of under-reporting. *Nutr Res Rev.* (1998) 11:231–53. doi: 10.1079/NRR19980017
35. Haghighatdoost F, Azadbakht L, Keshteli AH, Feinle-Bisset C, Daghaghzadeh H, Afshar H, et al. Glycemic index, glycemic load, and common psychological disorders. *Am J Clin Nutr.* (2016) 103:201–9. doi: 10.3945/ajcn.114.105445
36. Shi Z, Taylor AW, Wittert G, Goldney R, Gill TK. Soft drink consumption and mental health problems among adults in Australia. *Public Health Nutr.* (2010) 13:1073–9. doi: 10.1017/S1368980009993132
37. Aucoin M, LaChance L, Cooley K, Kidd S. Diet and psychosis: A scoping review. *Neuropsychobiology.* (2020) 79:20–42. doi: 10.1159/000493399
38. Funtikova AN, Gomez SF, Fitó M, Elosua R, Benítez-Arciniega AA, Schröder H. Effect of energy under-reporting on secular trends of dietary patterns in a mediterranean population. *PLoS ONE.* (2015) 10:e0127647. doi: 10.1371/journal.pone.0127647
39. Vancampfort D, Stubbs B, Mitchell AJ, De Hert M, Wampers M, Ward PB, et al. Risk of metabolic syndrome and its components in people with schizophrenia and related psychotic disorders, bipolar disorder and major depressive disorder: A systematic review and meta-analysis. *World Psychiatry.* (2015) 14:339–347. doi: 10.1002/wps.20252
40. Gemming L, Ni Mhurchu C. Dietary under-reporting: what foods and which meals are typically under-reported? *Eur J Clin Nutr.* (2016) 70:640–1. doi: 10.1038/ejcn.2015.204
41. Mueller-Stierlin AS, Teasdale SB, Dinc U, Moerkl S, Prinz N, Becker T, et al. Feasibility and acceptability of photographic food record, food diary and weighed food record in people with serious mental illness. *Nutrients.* (2021) 13:2862. doi: 10.3390/nu13082862
42. Hébert JR, Hurley TG, Steck SE, Miller DR, Tabung FK, Peterson KE et al. Considering the value of dietary assessment data in informing nutrition-related health policy. *Adv Nutr.* (2014) 5:447–55. doi: 10.3945/an.114.006189
43. Lioret S, Touvier M, Balin M, Huybrechts I, Dubuisson C, Dufour A et al. Characteristics of energy under-reporting in children and adolescents. *Br J Nutr.* (2011) 105:1671–80. doi: 10.1017/S0007114510005465
44. Huang TT, Roberts SB, Howarth NC, McCrory MA. Effect of screening out implausible energy intake reports on relationships between diet and BMI. *Obes Res.* (2005) 13:1205–17. doi: 10.1038/oby.2005.143
45. Freedman LS, Kipnis V, Schatzkin A, Tasevska N, Potischman N. Can we use biomarkers in combination with self-reports to strengthen the analysis of nutritional epidemiologic studies? *Epidemiol Perspect Innov*. (2010) 7:1–9. doi: 10.1186/1742-5573-7-2
46. Subar AF, Freedman LS, Tooze JA, Kirkpatrick SI, Boushey C, Neuhouser ML et al. Addressing current criticism regarding the value of self-report dietary data. *J Nutr.* (2015) 145:2639–45. doi: 10.3945/jn.115.219634
47. Bel S, De Ridder KAA, Lebacq T, Ost C, Teppers E, Cuypers K et al. Habitual food consumption of the Belgian population in 2014-2015 and adherence to food-based dietary guidelines. *Arch Public Health.* (2019) 77:14. doi: 10.1186/s13690-019-0343-3
48. Wielopolski J, Reich K, Clepce M, Fischer M, Sperling W, Kornhuber J et al. Physical activity and energy expenditure during depressive episodes of major depression. *J Affect Disord.* (2015) 174:310–6. doi: 10.1016/j.jad.2014.11.060
49. Davison KM, D'Andreamatteo C, Mitchell S, Vanderkooy P. The development of a national nutrition and mental health research agenda with comparison of priorities among diverse stakeholders. *Public Health Nutr.* (2017) 20:712–25. doi: 10.1017/S1368980016002056
50. Sarris J, Logan AC, Akbaraly TN, Paul Amminger G, Balanzá-Martínez V, Freeman MP et al. International Society for Nutritional Psychiatry Research consensus position statement: nutritional medicine in modern psychiatry. *World Psychiatry.* (2015) 14:370–1. doi: 10.1002/wps.20223

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fnut.2022.950874

Validation of the food frequency questionnaire for the assessment of dietary vitamin D intake

Maša Hribar[1,2], Katarina Žlavs[1,2], Igor Pravst[1,2,3]* and Katja Žmitek[1,3]

[1]Nutrition and Public Health Research Group, Nutrition Institute, Ljubljana, Slovenia, [2]Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia, [3]VIST–Faculty of Applied Sciences, Ljubljana, Slovenia

Check for updates

OPEN ACCESS

EDITED BY
Christophe Matthys,
KU Leuven, Belgium

REVIEWED BY
Rosalind Fallaize,
University of Hertfordshire,
United Kingdom
Irena Keser,
University of Zagreb, Croatia

*CORRESPONDENCE
Igor Pravst
igor.pravst@nutris.org

SPECIALTY SECTION
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

RECEIVED 23 May 2022
ACCEPTED 26 August 2022
PUBLISHED 23 September 2022

CITATION
Hribar M, Žlavs K, Pravst I and Žmitek K (2022) Validation of the food frequency questionnaire for the assessment of dietary vitamin D intake. *Front. Nutr.* 9:950874. doi: 10.3389/fnut.2022.950874

Vitamin D and its adequate status are related to many aspects of human health; therefore, an appropriate tool is needed for the valid assessment of vitamin D status. The main contributor to vitamin D status is endogenous synthesis after cutaneous exposure to ultraviolet B light (UVB), but in the absence of UVB radiation, vitamin D intake becomes an important source of vitamin D. Various tools are available for vitamin D intake assessments, with the Food Frequency Questionnaire (FFQ) being among the fastest, cheapest, and most convenient; however, until now, these tools have not been adapted for the Slovenia (SI). To enable valid vitamin D intake estimation, we developed a simple one-page semi-quantitative FFQ (sqFFQ/SI) and tested its validity using a 5-day dietary record (DR) as a reference method. The reproducibility was tested with the second sqFFQ/SI (sqFFQ/SI2) 6 weeks after the first (sqFFQ/SI1). The validity and reproducibility of this method were tested on 54 participants using Bland–Altman plots, Spearman's correlation, and Kappa analyses of tertiles. The mean daily vitamin D intake was 3.50 ± 1.91 μg according to the 5-day DR, and 2.99 ± 1.35 and 3.31 ± 1.67 μg according to the sqFFQ/SI1 and repeated sqFFQ/SI (sqFFQ/SI2), respectively. When analyzing for validity, the sqFFQ/SI1 was found to be significantly correlated ($p < 0.05$) with the 5-day DR, with an acceptable correlation coefficient of 0.268 and a Bland–Altman index of 3.7%. For reproducibility, the correlation between the sqFFQ/SI1 and sqFFQ/SI2 was highly significant ($p < 0.001$), with a good correlation coefficient of 0.689 and a Bland–Altman index of 3.7%. Kappa analyses of tertiles showed a poor validity and acceptable reproducibility. Overall, we observed a higher reproducibility than validity. Validation and reproducibility analyses demonstrated that the proposed sqFFQ/SI is acceptable and is, therefore, an appropriate tool for the effective assessment of habitual vitamin D intake on an individual level. With this consideration, this tool will be used in further population studies to assess vitamin D intake and for the development of a screening tool for the assessment of the risk for vitamin D deficiency, which will be used as a foundation for evidence-based policy-making decisions.

KEYWORDS

vitamin D, nutrient intake, validation, reproducibility, Slovenia, Food Frequency Questionnaire (FFQ), dietary record

Introduction

Vitamin D is a fat-soluble vitamin that is, due to its many functions in the body, crucial for the growth and maintenance of health in all life stages (1–3). For humans, the sources of vitamin D are endogenous synthesis in the skin when exposed to ultraviolet B (UVB) radiation and dietary intake (either with foods that are naturally rich in vitamin D, fortified foods, or pharmaceutical preparations) (4). Although endogenous synthesis is the main source of vitamin D for most people, in the absence of sufficient UVB exposure, vitamin D becomes an essential nutrient and sufficient dietary intake is required (5–7).

The dietary vitamin D intake is usually well below recommendations (5), mainly because very few foods are rich in vitamin D, and, at the same time, they are seldom consumed. The recommended dietary vitamin D intake for the adult population is 5 μg/day (19–50 years), 10 μg/day (51–65 years), and 15 μg/day (>65 years) according to the recommendations of the World Health Organization (WHO) (8), and 15 and 20 μg/day (in the absence of endogenous synthesis) according to the recommendations of the European Food Safety Authority (EFSA) and the Nutrition Societies of Germany, Austria, and Switzerland (D-A-CH), respectively (9, 10). On the other hand, the nutrient reference value (NRV), as defined in the European union food labeling regulation, is 5 μg (11), while the threshold of 2.5 μg is sometimes used as a lower reference nutrient intake (LRNI) (12).

Studies have reported a high prevalence of inadequate dietary intakes of vitamin D in European populations and around the world (6, 13, 14), including Slovenia (15, 16). In most European countries, the daily intake of vitamin D is lower than 5 μg (6, 14, 16, 17); exceptions are Scandinavian countries, where oil-rich fish consumption is relatively high, and both fortification and supplementation policies have also been implemented (6, 18). Only a few studies have evaluated the dietary intake of vitamin D in the Slovenian population, using various methods to record the dietary intake (15, 16).

Due to insufficient UVB exposure and simultaneous inadequate vitamin D intake, an important public health task is the rapid identification of individuals exposed to the risk of inadequate vitamin D status. Furthermore, the accurate assessment of dietary vitamin D intake is important for the application of evidence-based public health measures in order to prevent poor vitamin D status in different population groups. To achieve this, a suitable screening procedure is necessary (19).

The optimal and most objective method for evaluating vitamin D status is a laboratory determination of serum 25-hydroxyvitamin D [25(OH)D] (20), but this method is invasive and not recommended for screening in large populations (19, 21). Because vitamin D status is also affected by dietary intake of vitamin D, a valid dietary assessment method that is easy to use is needed (22). Determining dietary vitamin D intake with the 24-h recall or the dietary record (DR) method, a gold standard for dietary intake assessments, is not the most well suited (23); because of large day-to-day variations in vitamin D intake (dependent on, e.g., fish intake and the diversity of fortified foods), an extended time period is necessary for data collection (22, 24). On the other hand, the Food Frequency Questionnaire (FFQ) is less useful for measuring the absolute dietary intake, but it can better reflect one's typical diet (25). Additionally, the FFQ may be more reliable for the estimation of micronutrient intake, such as vitamin D, as it covers a longer period and can focus on specific foods that are relevant to vitamin D intake (26).

When assessing dietary intake, the research method must be simple and fast, for both the subject and the researcher, and, at the same time, it must be valid and reproducible (25, 27). Various FFQs for the assessment of vitamin D intake were designed and validated in several countries and studies around the world (23, 26, 28–34). However, such tools need to be tailored for use in specific regions; country-specific food consumption patterns and foods need to be considered (25, 27, 35). The typical reference methods for the validation of the FFQ are DR or the 24-h recall method, and biomarkers are sometimes also used (25).

The objective of this study was to assess the validity and reproducibility of a semi-quantitative FFQ on the Slovenian population (sqFFQ/SI) for the assessment of the dietary intake of vitamin D, using 5-day DR as a reference method. The sqFFQ/SI was developed by the Nutrition Institute

(Slovenia) in cooperation with the National Institute of Public Health (Slovenia) within the national research project Nutri-D "Challenges in achieving adequate vitamin D status in the adult population" (L7-1849).

Materials and methods

Study design and data collection

The study protocol was approved by the Nutrition Research Ethics Committee (Biotechnical Faculty, University of Ljubljana), under the identification number KEP-1-2/2020 on 10 February 2020. The study was conducted in full compliance with the principles laid out in the Declaration of Helsinki. Participation in the study was voluntary. All of the subjects signed a written informed consent form before participating. They were informed that they can withdraw from the study at any time with no consequences. The study was conducted between February and April 2020 and included data collection using the FFQ and a 5-day DR. The participants received all the required information (and instructions) in oral and written format at individual meetings. To assess reproducibility, the participants were asked to fill out the sqFFQ/SI two times: the first one (sqFFQ/SI1) was filled out at the beginning of the study, and the second one (sqFFQ/SI2) was filled out approximately 6 weeks later (**Figure 1**). The participants were asked not to alter dietary habits between sqFFQ/SI1 and sqFFQ/SI2 if possible. It should be noted that the second one was conducted during the SARS-CoV-2 epidemic. To evaluate the validity of the questionnaire, the participants were requested to complete a 5-day DR during the time between both administered sqFFQ/SI. The participants were free to choose any 3 week/2 weekend days in that time period.

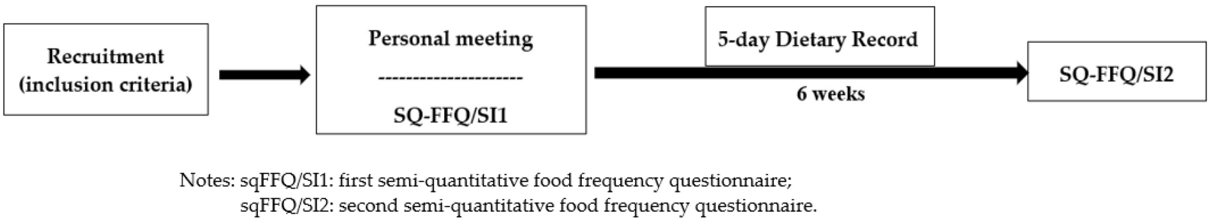
Study population

The sqFFQ/SI was validated among a group of Slovenian adults, aged between 18 and 65 years, mainly from central Slovenia and the Savinja statistical region. The subjects were enrolled with the use of invitations *via* social media profiles from the official Nutrition Institute profile, and personal invitations. The exclusion criteria were diagnosis of chronic disease, pregnant or breastfeeding women, and specific diets (vegan diet, ketogenic diet, energy-restricted diets, and diets due to medical reasons). It should be noted that vegetarians were not excluded. All the required information regarding inclusion/exclusion criteria was presented before the beginning of the study.

Semi quantitative food frequency questionnaire for Slovenian population

For this study, a semi-quantitative FFQ adapted for the Slovenian population was used (sqFFQ/SI), in which the frequency of food consumption and the size of portions are defined (36). The tool included food products that were previously identified as important sources of vitamin D in Slovenia (15). Although Slovenia does not have a mandatory vitamin D fortification of foods, some food groups are commonly fortified (37) and were therefore included. The final sqFFQ/SI consisted of 22 food items that contain at least 0.03 µg of vitamin D per 100 g, according to the reviewed literature (38) and the selected food composition databases: Slovenian Open Platform for Clinical Nutrition (OPEN) (39), McCance and Widdowson's The Composition of Foods (38), and the United States Department of Agriculture (USDA) database (40). The included food groups are presented in **Table 1**. For each food group, we identified all the relevant food records in the abovementioned food composition datasets and calculated the category average content of vitamin D. We did not include the use of pharmaceutical preparations.

The subjects were asked to rank their consumption frequencies during the past year. Previously reported (17) frequency options were implemented: multiple times a day, daily, 4–6 times per week, 1–3 times per week, 1–3 times per month, and rarely or never. Further, subjects were asked to rank their usual portion sizes (in comparison to the indicated reference portion size): (a) as indicated, (b) less than indicated (specified as at least one-half smaller than the normal portion size), and (c) more than indicated (specified as at least one-half


Recruitment (inclusion criteria)
Personal meeting

SQ-FFQ/SI1
5-day Dietary Record
6 weeks
SQ-FFQ/SI2
Notes: sqFFQ/SI1: first semi-quantitative food frequency questionnaire;
sqFFQ/SI2: second semi-quantitative food frequency questionnaire.

FIGURE 1
Study design.

larger than the normal portion size). The complete sqFFQ/SI is provided in the **Supplementary material**. The sqFFQ/SI was carried out online and took approximately 10 min to complete.

Five-day dietary record

In line with the previously reported approach (31), the 5-day DR was conducted on five typical random non-consecutive days (3 weekdays and 3 days during the weekend). At the first meeting, the participants were given detailed instructions on how to complete the DR. Participants were asked to maintain their usual eating habits and record all consumed foods and beverages in as much detail as possible (describing the type/brand of food, the amount of food, the method of preparation, and the recipes of composited dishes where applicable). The amounts were preferably weighted and written down in grams when participants had access to a kitchen scale. Exceptionally, the amounts were estimated using illustration material for different portion sizes of typical foods using a previously developed nationally adapted picture book (41). The participants returned their completed 5-day DR *via* a pre-paid postal service or in person.

Data processing and statistical analysis

The data collected by both sqFFQ/SI were used to calculate the daily vitamin D intake (μg/day) based on the method described in detail by Biro and Gee (42). The calculations were performed using the selected serving size and average vitamin D contents in 100 g of foods, as shown in **Table 1**.

Vitamin D intake (μg/day) was further determined using a 5-day DR using the online nutrition analysis software OPEN, which is linked to the food composition database (43). Due to some missing information regarding the vitamin D content in some foods, the OPEN database was updated in cooperation with the software owner, the Jožef Stefan Institute (JSI). The missing data were updated with data available in the USDA database (40), the National Food Composition Database in Finland (Fineli) (44), and McCance and Widdowson's The Composition of Foods (38).

The obtained data were statistically analyzed with the IBM SPSS version 27, Statistics program (IBM SPSS, IBM Corp., Armonk, NY, USA) (45). We investigated the validity (external validation compared with the results of the 5-day DR) and reproducibility of the method (internal validation comparing results obtained two times: sqFFQ/SI1 and sqFFQ/SI2) (46). Descriptive characteristics (means, median, and proportions) for the daily vitamin D intakes were calculated.

The estimated daily vitamin D intakes were grouped for cross-classification according to tertiles. In the analyses, we regarded the estimations as good if less than 10% of the participants were grossly misclassified into the opposite tertiles and at least 50% of the participants were correctly classified (47). In the Kappa analyses, we considered Kappa values below 0.20 to have a poor agreement, between 0.20 and 0.60 as having an acceptable agreement, and over 0.60 as having good agreement (48).

The normality of distribution was tested with the Shapiro–Wilk test. The analysis of correlations between the results obtained in the assessment of validity (sqFFQ/SI1 compared with a 5-day DR) and the assessment of reproducibility (comparison between sqFFQ/SI1 and sqFFQ/SI2) was used, where Spearman's correlation was applied. Correlation coefficients of less than 0.20 were a poor outcome; those between 0.20 and 0.49 were acceptable, and those of 0.50 or higher was considered a good outcome (48).

TABLE 1 Reference serving sizes and vitamin D content in 100 g of the foods used in the semi-quantitative Food Frequency Questionnaire (sqFFQ/SI).

Food group	Reference serving size (g/ml)	Vitamin D (μg/100 g)
Sardines, trout, salmon, and carp	120	7.84
Sea bass, tuna, cod, common sole, blue tilapia, and other fish	120	3.23
Canned fish	80	4.31
Plant-based milk alternatives: rice milk, soy milk, etc.	250	0.47
Semi-skimmed milk (1.5% milkfat), cocoa drink, and milk drinks	200	0.03
Whole milk (3.5% milkfat), a cocoa drink containing whole milk, milk drinks	200	0.09
Semi-skimmed (1.5% milkfat) flavored or plain yogurt	150	0.03
Whole milk (3.5% milkfat) flavored or plain yogurt	150	0.06
Hard cheese: Gouda cheese, Edam cheese, etc.	30	0.9
Blue cheese	20	0.39
Cottage cheese, mozzarella, other types of processed cheese	50	0.28
Ice cream	40	0.25
Butter	6	1.66
Margarine	6	2.5
Eggs	50	2.9
Egg pasta	100	0.28
Red meat	100	0.48
Poultry	100	0.26
Meat products	40	0.86
Calf's liver	60	1.2
Mushrooms	100	0.18
Cakes, pastry, and muffins	70	0.31

In all of the comparisons, significance was considered at $p < 0.05$. A Bland–Altman plot was further used for the validation and reproducibility assessment. Since the data were not normally distributed, we used log transformation, as previously proposed (49). A Bland–Altman index below 5% was interpreted as good, as suggested before in similar research (23, 28, 49–51).

Results

A total of 55 participants volunteered to participate. The final sample included 54 participants, as one of the individuals withdrew from the study (due to lack of time). The sample was represented by 37 women (69%) and 17 men (31%). The average age was 32.7 years (±13.6 years). Other characteristics of the population [age and body mass index (BMI)] are shown in **Table 2**. The participants completed two sqFFQ/SIs on average 46 days apart (sqFFQ/SI1 and sqFFQ/SI2, respectively), and a 5-day DR according to a study design, presented in **Figure 1**. The mean daily vitamin D intake was 3.50 ± 1.91 µg according to the 5-day DR, and 2.99 ± 1.35 and 3.31 ± 1.67 µg according to the sqFFQ/SI1 and sqFFQ/SI2, respectively (**Table 3**). Since none of the participants achieved the nationally recommended daily intake of vitamin D (20 µg), we analyzed the data with a cut-off value for the LRNI (2.5 µg) and NRV (5 µg). Overall, the NRV threshold was not met by 87.0% of subjects according to the 5-day DRs, 90.7% according to the sqFFQ/SI1, and 83.3% according to the sqFFQ/SI2. On the other hand, the lower LRNI threshold was not met by 35.2, 42.6, and 40.7%, respectively.

TABLE 2 Study population description.

Parameter	Criteria	Number (%)
Participants (total)		54 (100)
Sex	Men	17 (31.5)
	Women	37 (69.5)
Age	19–24	28 (52.9)
	25–65	26 (48.1)
Body mass index categories	<18.5 kg/m^2 (underweight)	2 (3.7)
	18.5–24.9 kg/m^2 (normal weight)	37 (68.5)
	25.0–29.9 kg/m^2 (overweight)	10 (18.5)
	>30.0 kg/m^2 (obese)	5 (9.3)

TABLE 3 Daily vitamin D intake estimated with a 5-day dietary record (DR) and semi-quantitative Food Frequency Questionnaires (sqFFQ/SI) administered 6 weeks apart.

	sqFFQ/SI1	sqFFQ/SI2	5-day DR
Mean ± SD (µg)	2.99 ± 1.35	3.31 ± 1.67	3.50 ± 1.91
Median (µg)	2.61	2.94	3.04
Minimum (µg)	0.44	0.58	0.97
Maximum (µg)	7.08	8.19	10.31
<2.5 µg (%)	42.6	40.7	35.2
<5 µg (%)	90.7	83.3	87

Validity

The validity of the sqFFQ/SI1 for the estimation of daily vitamin D intake was analyzed with comparison to the 5-day DR. The estimated intakes were analyzed with Spearman's rank correlation for the sqFFQ/SI1 and 5-day DR (**Figure 2**). The sqFFQ/SI1 was significantly correlated ($p < 0.05$) with the 5-day DR, with a correlation coefficient of 0.268; the mean difference between both methods was 0.514 µg (SD: 0.318 µg). Due to the non-normal distribution, further comparison of the 5-day DR and sqFFQ/SI1 using Bland–Altman plots were carried out with log-transformed data (**Figure 3**). The Bland–Altman index of the logarithmic model was good (3.70%). Further, we analyzed the percentages of subjects classified into the same vitamin D intake tertile (**Table 4**). When comparing the sqFFQ/SI1 and the 5-day DR, 42.6% of the participants were categorized into the same tertile, and 16.7% into the opposite tertile. This indicates a low agreement; the Kappa coefficient was 0.139.

Reproducibility

To investigate the reproducibility of the sqFFQ/SI, we compared the daily vitamin D intake as estimated with the sqFFQ/SI1 and sqFFQ/SI2, which were administered approximately 6 weeks apart. The correlation between the sqFFQ/SI1 and sqFFQ/SI2 was highly significant ($p < 0.001$), with a correlation coefficient of 0.689 (**Figure 4**). The mean difference between measurements was 0.318 µg (SD: 0.291 µg). Furthermore, the log-transformed Bland–Altman plot showed good reproducibility with an index of 3.70% (**Figure 5**). When testing the sqFFQ/SI1 for reproducibility, the analysis of the tertiles showed acceptable agreement; 59.3% of the subjects were categorized into the same tertile, and there were no classifications into opposite tertile, while the Kappa value was acceptable (0.389) (**Table 4**).

Discussion

Vitamin D is a crucial micronutrient for optimal human health in all life stages, and we should strive to achieve optimal status across all populations. Besides UVB-induced cutaneous synthesis, food intake is an important source of vitamin D (5–7). Vitamin D intake can be estimated using various methods, with the FFQ being one of the less burdensome methods. The FFQ is user friendly and time/cost efficient (25). Convenient tools for intake estimation are important for the efficient assessment

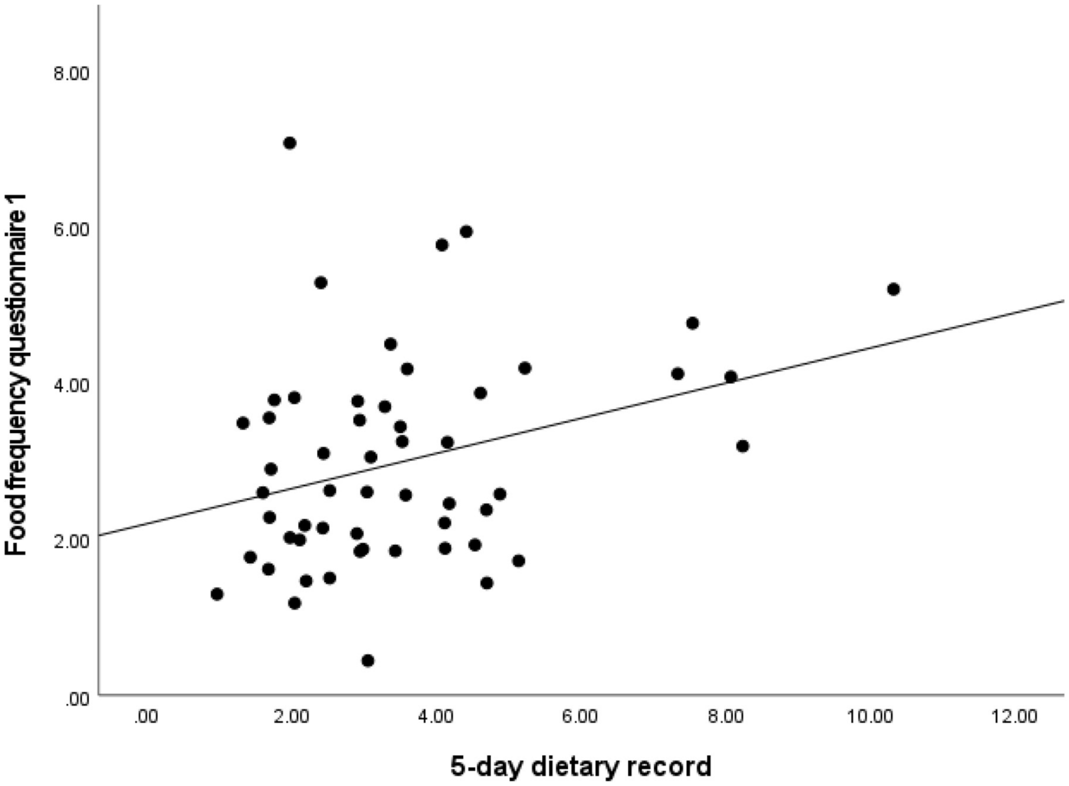


FIGURE 2
Analysis of correlation for daily vitamin D intake estimated with semi-quantitative Food Frequency Questionnaire 1 (sqFFQ/SI1) and 5-day dietary record (correlation coefficient = 0.268; $p < 0.05$).

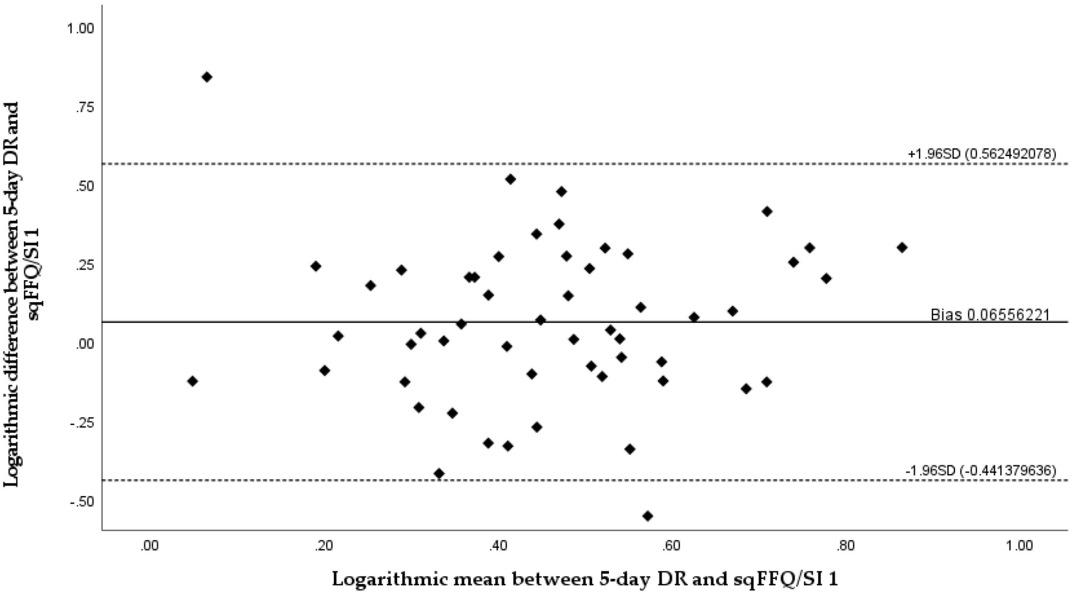


FIGURE 3
Bland–Altman plot comparing daily vitamin D intake estimated with semi-quantitative Food Frequency Questionnaire 1 (sqFFQ/SI1) and a 5-day dietary record (Bland–Altman index: 3.70%).

TABLE 4 Count and percentages of subjects classified into the same/opposite vitamin D intake tertile.

Category		sqFFQ/SI1 vs. 5-day DR	sqFFQ/SI1 vs. sqFFQ/SI2
Subjects classified into the same tertile	*N*	23	32
	%	42.6	59.3
Subjects misclassified into the opposite tertile	*N*	9	0
	%	16.7	0

DR, dietary record; sqFFQ/SI1, semi-quantitative Food Frequency Questionnaire 1; sqFFQ/SI2, semi-quantitative Food Frequency Questionnaire 2.

of the risk of vitamin D deficiency, particularly in the absence of endogenous synthesis. To accurately assess the dietary intake of vitamin D in the Slovenian population we developed a semi-quantitative FFQ and tested its validity and reproducibility using 5-day DR and repeated sqFFQ/SI, respectively. The estimated mean daily vitamin D intakes in our study were 3.50, 2.99, and 3.31 µg for the 5-day DR, sqFFQ/SI1, and sqFFQ/SI2, respectively. We did not observe a higher mean intake with the FFQ (in comparison to the 5-day DR), unlike some other validation studies (29, 52).

The validity and reproducibility were tested using Bland–Altman plots, a recommended "gold-standard" approach by which to compare results from different methods observing the same variable (53). Our results show that the developed sqFFQ/SI is fairly valid and reproducible; only 3.70% of the data points were outside the 95% limits of agreements for both validity and reproducibility. Other research investigating similar a topic reported from 2.7 to 6.3% of data points outside the 95% limits of agreement using Bland–Altman plot (23, 28). Additionally, Spearman's correlation was significant both for validity (<0.05) and reproducibility (<0.001). The correlation coefficients were acceptable and good (0.268 and 0.689, respectively). In similar studies comparing multiple day dietary vitamin D intake with FFQ, significant correlation coefficients ranged from 0.21 to 0.83 for validity and from 0.62 to 0.82 for reproducibility (23, 26, 28, 29, 52). It should be noted that due to the complexity of the estimation of micronutrient intakes, correlation coefficients above 0.2 are considered acceptable, and coefficients above 0.7 are rarely reported (32). However, the thresholds for acceptable correlations are not well harmonized (25, 48), and we should take caution when evaluating the outcomes. Analyses of terciles in our case showed less agreement than in some other studies. Altogether, in the validity study, 42.6% of subjects were classified in the same tercile (Kappa coefficient: 0.139; poor agreement), while some other studies reported up to 64% (28, 29), but we must note that the cross-classification is a relatively crude measurement (29). On the hand, we observed better results in analyses of terciles in the reproducibility study (59.3%; Kappa coefficient: 0.389; acceptable). Other studies also reported lower differences in the reproducibility of FFQs, in comparison to validity testing with DRs (23, 54), which might be affected by the limited ability of DRs to capture dietary patterns, related to vitamin D intake.

In a recent study, it was shown that in Slovenia vitamin D deficiency is highly prevalent, particularly in the wintertime when dietary intake becomes the main source of vitamin D. In the winter months, ca. 80% of adults and elderly people were shown to be vitamin D deficient (55), and the mean daily vitamin D intakes were 2.9 and 2.5 µg, respectively (15). Globally, various FFQs were developed and regionally adapted to estimate vitamin D intakes (23, 26, 28–34); however, to the best of our knowledge, there is no such tool available for use in the Slovenian population.

The intake of nutrients can be estimated using a variety of methods that have different levels of accuracy for different nutrients. For nutrients that are found in a limited number of foods, the use of short-period DRs can pose a risk of not capturing a typical dietary pattern, and it is therefore recommended to follow food intake over a period of several days. On the contrary, although the FFQ is much simpler to use, this method can better capture food consumption patterns over a longer period (26). In the case of vitamin D, the intake estimation is particularly challenging due to notable day-to-day variations as vitamin D-rich foods (i.e., fish) are seldom consumed (24, 26). This, of course, affects the estimation of daily vitamin D intake when different methods are used. For example, in a nationally representative Slovenian SI. Menu study, 72.8% of adults were recognized as sea fish consumers when two 24 h dietary recalls were used, while the Food Propensity Questionnaire method identified 80.8% as true consumers (15).We developed an FFQ that covers the most important contributors to vitamin D intake in Slovenia, including eggs, fish, and fish products, meat and meat products, milk and milk products, and commonly fortified foods, such as plant-based milk alternatives (15, 37). The validation of the FFQ (sqFFQ/SI1) was conducted on 54 participants using a 5-day DR as a reference method. Despite the abovementioned limitations, the DR is a commonly used reference method in such validation studies (56). We also tested the reproducibility, using a repeated FFQ (sqFFQ/SI2) administered 6 weeks after the first measurement.

To evaluate the validity and reproducibility we used various approaches. The results are showing that validity varied from poor to good, and good for reproducibility. We have demonstrated that the proposed FFQ is acceptable and is therefore an appropriate tool for the effective assessment of habitual vitamin D intake on an individual level. Overall, we observed higher reproducibility than validity. However, such tools are also commonly used in population studies. Therefore, we further compared the estimated mean vitamin D intakes between the tested methods and literature data. The difference between the mean vitamin D intake according to

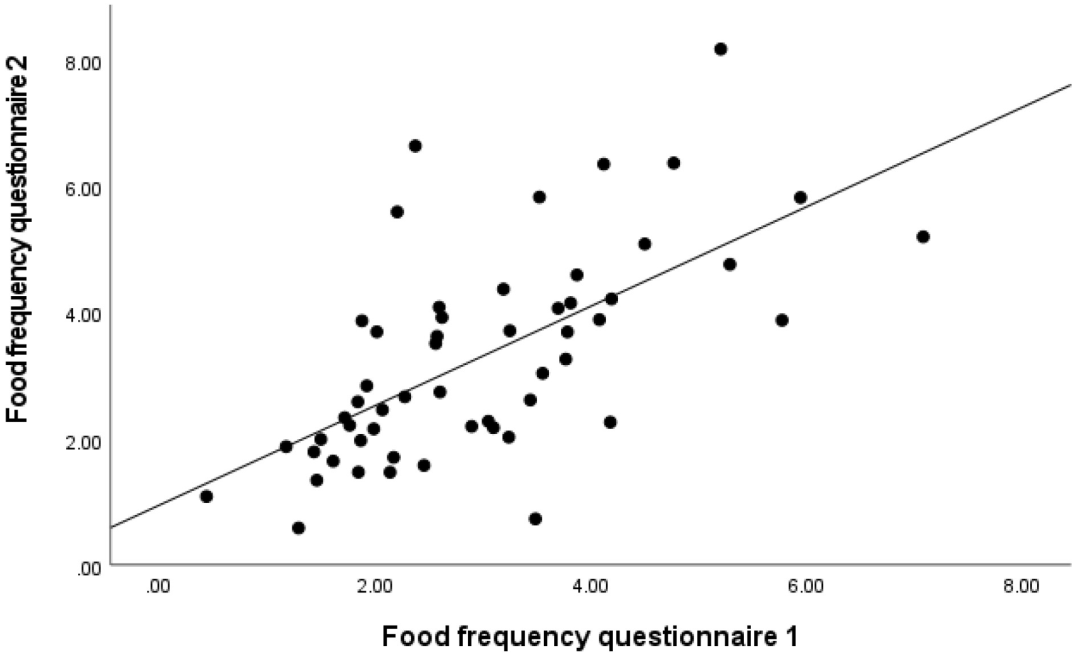


FIGURE 4
Analysis of correlation for daily vitamin D intake estimated with semi-quantitative Food Frequency Questionnaire 1 (sqFFQ/SI1) and 2 (sqFFQ/SI2) (correlation coefficient = 0.689; p < 0.001).

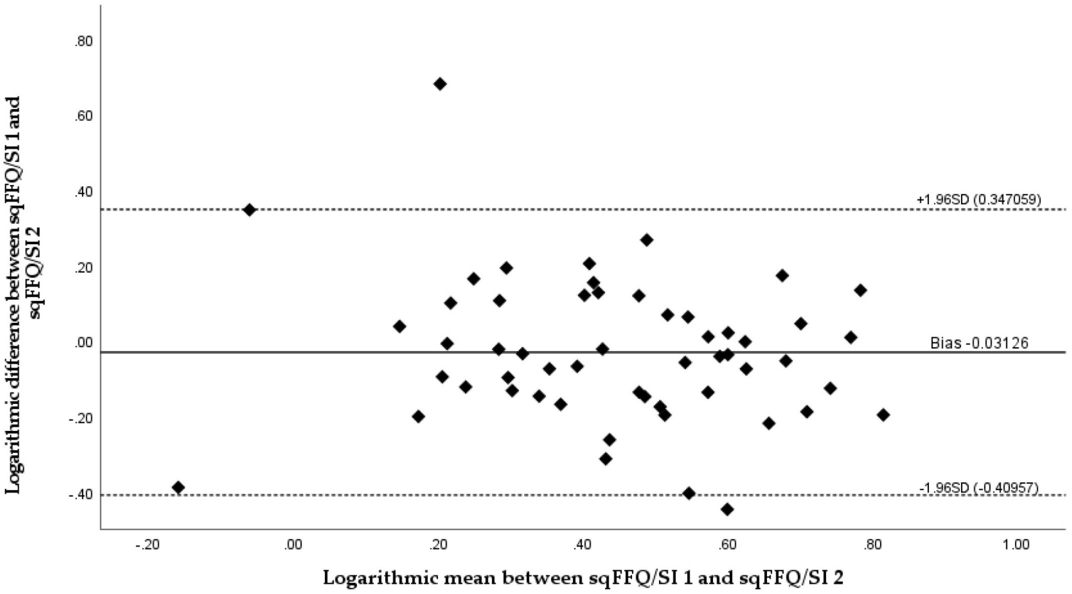


FIGURE 5
Bland–Altman plot comparing daily vitamin D intake estimated with a semi-quantitative Food Frequency Questionnaire 1 (sqFFQ/SI1) and 2 (sqFFQ/SI2) (Bland–Altman index: 3.70%).

both of the tested methods was small (0.51 μg) and statistically insignificant. With consideration of the recommended daily vitamin D intake (20 μg), the clinical importance of such a difference is minimal. Similar differences were also observed in other similar studies, for example, in the study by Kiely et al. in their comparison of the FFQ and 14-day DR results (29). Furthermore, our results are comparable with mean vitamin D intakes reported for the general Slovenian population. Vitamin

D intake was recently investigated in a nationally representative SI. Menu study (15). The weighted population mean intake was estimated with the multiple source method (MSM), using two 24 h recalls and the Food Propensity Questionnaire. The estimated mean vitamin D intake in adults (18–64 years) was 2.85 μg (15), comparable to the results in our study (sqFFQ/SI1: 2.99 μg). A recent systematic review also highlighted that vitamin D intakes in other studies in the Slovenian population were below 5 μ g (16).

Although the developed tool was shown as valid and reproducible, some limitations need to be noted. While we followed the recommendation that validity studies should be conducted on at least 50 subjects (31), a bigger sample would be beneficial to check the validity in more specific population groups. Furthermore, we did not use biological biomarkers of vitamin D status [serum 25(OH)D concentration], but we should note that this biomarker is seriously affected by UVB-induced endogenous vitamin D biosynthesis, which results in major inter-individual differences. The limited use of blood biomarkers for such validation studies in the case of vitamin D was noted also in other studies (26, 33). Moreover, we should note that while majority (72.2%) of our study participants were with BMI < 25 kg/m^2, we also had some overweight/obese subjects (18.5 and 9.3%, respectively), where food intake misreporting might be more common. We have not excluded those from the analyses, because vitamin D intake screening is also very relevant in this population group. At last, it should be said that we tested the tool on healthy, non-pregnant, no-lactating, adult, omnivore populations. We suggest that the described sqFFQ/SI is further tested on other populations of public health interest.

Conclusion

The estimation of one's usual daily vitamin D intake is a challenging task, regardless of the method used, due to its major day-to-day variability. Building on previously established methods and major contributors to vitamin D intake in the Slovenian population, we developed a simple one-page semi-quantitative FFQ (sqFFQ/SI) for the quick estimation of one's usual daily vitamin D intake. To the best of our knowledge, the described tool is the first FFQ adapted for the Slovenian population. The Bland–Altman plot analyses showed a good level of agreement between the developed sqFFQ/SI and the standard 5-day DR method, as well as a good reproducibility, with less than 5% of the outliers falling outside of the agreement limit and a significant correlation being observed. Further analyses of correlation showed acceptable and good correlation, whereas Kappa analyses of terciles showed poor and acceptable agreement tor validity and reproducibility, respectively. Considering the analyses results, this tool will be used in further population studies and for the development of a screening tool for the assessment of the risk for vitamin D deficiency in healthy non-pregnant, no-lactating, adult, and omnivore populations. Due to the high prevalence of vitamin D deficiency, such a method is important not only for researchers but also for clinical practice and policymakers. It should be noted that the developed tool is very valuable for use in other countries in the Central European region due to similar food policies and dietary patterns. However, minor modifications might be appropriate for specific populations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Nutrition Research Ethics Committee (Biotechnical Faculty, University of Ljubljana), KEP-1-2/2020. The patients/participants provided their written informed consent to participate in this study.

Author contributions

KŽm and IP: conceptualization, funding acquisition, methodology, and supervision. MH and KŽl: data curation, formal analysis, investigation, and writing—original draft. KŽm: project administration and resources. MH: validation and visualization. MH, KŽm, KŽl, and IP: writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This study was conducted within the national research program P3-0395 "Nutrition and Public Health" and project Nutri-D "Challenges in achieving adequate vitamin D status in the adult population" (L7-1849), supported by the Slovenian Research Agency and Ministry of Health of the Republic of Slovenia.

Acknowledgments

We acknowledge the support of other member of the research group, particularly Hristo Hristov (Nutrition Institute, Ljubljana, Slovenia) and Urška Blaznik (National Institute of Public Health, Ljubljana, Slovenia).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2022.950874/full#supplementary-material

References

1. Holick MF, Binkley NC, Bischoff-Ferrari HA, Gordon CM, Hanley DA, Heaney RP, et al. Evaluation, treatment, and prevention of vitamin D deficiency: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab.* (2011) 96:1911–30. doi: 10.1210/jc.2011-0385

2. Autier P, Boniol M, Pizot C, Mullie P. Vitamin D status and ill health: a systematic review. *Lancet Diabetes Endocrinol.* (2014) 2:76–89. doi: 10.1016/S2213-8587(13)70165-7

3. Molina P, Carrero JJ, Bover J, Chauveau P, Mazzaferro S, Torres PU, et al. Vitamin D, a modulator of musculoskeletal health in chronic kidney disease. *J Cachexia Sarcopenia Muscle.* (2017) 8:686–701. doi: 10.1002/jcsm.12218

4. Holick MF. Vitamin D deficiency. *New Engl J Med.* (2007) 375:266–81. doi: 10.1056/NEJMra070553

5. Calvo MS, Whiting SJ, Barton CN. Vitamin D intake: a global perspective of current status. *J Nutr.* (2005) 135:310–6. doi: 10.1093/jn/135.2.310

6. Spiro A, Buttriss JL. Vitamin D: an overview of vitamin D status and intake in Europe. *Nutr Bull.* (2014) 39:322–50. doi: 10.1111/nbu.12108

7. O'Mahony L, Stepien M, Gibney MJ, Nugent AP, Brennan L. The potential role of vitamin D enhanced foods in improving vitamin D status. *Nutrients.* (2011) 3:1023–41. doi: 10.3390/nu3121023

8. World Health Organization [WHO]. *Vitamin and Mineral Requirements in Human Nutrition.* 2nd ed. Geneva: World Health Organization (2005).

9. German Nutrition Society [GNS]. New reference values for vitamin D. *Ann Nutr Metab.* (2012) 60:241–6. doi: 10.1159/000337547

10. EFSA Panel on Dietetic Products and Allergies. Dietary reference values for vitamin D. *EFSA J.* (2016) 14:e04547. doi: 10.2903/j.efsa.2016.4547

11. European Commission [EC]. *Commission Regulation (EC) No 1170/2009 of 30 November 2009 Amending Directive 2002/46/EC of the European Parliament and of Council and Regulation (EC) No 1925/2006 of the European Parliament and of the Council as Regards the Lists of Vitamin and Minerals and Their Forms that Can Be Added to Foods, Including Food Supplements.* Brussels: European Commission (2009).

12. Mensink GB, Fletcher R, Gurinovic M, Huybrechts I, Lafay L, Serra-Majem L, et al. Mapping low intake of micronutrients across Europe. *Br J Nutr.* (2013) 110:755–73. doi: 10.1017/S000711451200565X

13. European Food Safety Authority [EFSA]. Scientific opinion on the tolerable upper intake level of vitamin D. *EFSA J.* (2012) 10:2813. doi: 10.2903/j.efsa.2012.2813

14. Roman Viñas B, Ribas Barba L, Ngo J, Gurinovic M, Novakovic R, Cavelaars A, et al. Projected prevalence of inadequate nutrient intakes in Europe. *Ann Nutr Metab.* (2011) 59:84–95. doi: 10.1159/000332762

15. Hribar M, Hristov H, Lavriša Z, Koroušić Seljak B, Matej G, Blaznik U, et al. Vitamin D intake in slovenian adolescents, adults, and the elderly population. *Nutrients.* (2021) 13:3528. doi: 10.3390/nu13103528

16. Hribar M, Benedik E, Gregorič M, Blaznik U, Kukec A, Hristov H, et al. A systematic review of vitamin D status and dietary intake in various Slovenian populations. *Zdarvstveno Varstvo.* (2022) 61:55–72. doi: 10.2478/sjph-2022-0009

17. Lichthammer A, Nagy B, Orbán C, Tóth T, Csajbók R, Molnár S, et al. A comparative study of eating habits, calcium and vitamin D intakes in the population of Central-Eastern European countries. *New Med.* (2015) 19:66–70.

18. Mithal A, Wahl DA, Bonjour JP, Burckhardt P, Dawson-Hughes B, Eisman JA, et al. Global vitamin D status and determinants of hypovitaminosis D. *Osteoporos Int.* (2009) 20:1807–20. doi: 10.1007/s00198-009-0954-6

19. LeFevre ML, U.S. Preventive Services Task Force. Screening for vitamin D deficiency in adults: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med.* (2015) 162:133–40. doi: 10.7326/M14-2450

20. Kennel KA, Drake MT, Hurley DL. Vitamin D deficiency in adults: when to test and how to treat. *Mayo Clin Proc.* (2010) 85:752–7; quiz 7-8. doi: 10.4065/mcp.2010.0138

21. Ross AC, Taylor CL, Yaktine AL, Del Valle HB. (eds). *Institute of Medicine (US) Committee to Review Dietary Reference Intakes for Vitamin D and Calcium. Dietary Reference Intakes for Calcium and Vitamin D.* Washington, DC: National Academic Press (2011).

22. Willett WC. *Nutritional Epidemiology.* New York, NY: Oxford University Press (1998). doi: 10.1093/acprof:oso/9780195122978.001.0001

23. Glabska D, Uroic V, Guzek D, Pavic E, Bival S, Jaworska K, et al. The possibility of applying the vitamin D brief food frequency questionnaire as a tool for a country with no vitamin D data in food composition tables. *Nutrients.* (2018) 10:1278. doi: 10.3390/nu10091278

24. Verbeke W, Vackier I. Individual determinants of fish consumption: application of the theory of planned behaviour. *Appetite.* (2005) 44:67–82. doi: 10.1016/j.appet.2004.08.006

25. Cade J, Thompson R, Burley V, Warm D. Development, validation and utilisation of food-frequency questionnaires - a review. *Public Health Nutr.* (2002) 5:567–87. doi: 10.1079/PHN2001318

26. Bärebring L, Amberntsson A, Winkvist A, Augustin H. Validation of dietary vitamin D intake from two food frequency questionnaires, using food records and the biomarker 25-Hydroxyvitamin D among pregnant women. *Nutrients.* (2018) 10:745. doi: 10.3390/nu10060745

27. Perez Rodrigo C, Aranceta J, Salvador G, Varela-Moreiras G. Food frequency questionnaires. *Nutr Hosp.* (2015) 31(Suppl. 3):49–56.

28. Glabska D, Guzek D, Sidor P, Wlodarek D. Vitamin D dietary intake questionnaire validation conducted among young polish women. *Nutrients.* (2016) 8:36. doi: 10.3390/nu8010036

29. Kiely M, Collins A, Lucey AJ, Andersen R, Cashman KD, Hennessy A. Development, validation and implementation of a quantitative food frequency questionnaire to assess habitual vitamin D intake. *J Hum Nutr Diet.* (2016) 29:495–504. doi: 10.1111/jhn.12348

30. Wu H, Gozdzik A, Barta JL, Wagner D, Cole DE, Vieth R, et al. The development and evaluation of a food frequency questionnaire used in assessing vitamin D intake in a sample of healthy young Canadian adults of diverse ancestry. *Nutr Res.* (2009) 29:255–61. doi: 10.1016/j.nutres.2009.03.006

31. Pritchard JM, Seechurn T, Atkinson SA. A food frequency questionnaire for the assessment of calcium, vitamin D and vitamin K: a pilot validation study. *Nutrients.* (2010) 2:805–19. doi: 10.3390/nu2080805

32. Taylor C, Lamparello B, Kruczek K, Anderson EJ, Hubbard J, Misra M. Validation of a food frequency questionnaire for determining calcium and vitamin D intake by adolescent girls with anorexia nervosa. *J Am Diet Assoc.* (2009) 109:479–85, 485.e1–3. doi: 10.1016/j.jada.2008.11.025

33. Djekic-Ivankovic M, Weiler HA, Nikolic M, Kadvan A, Gurinovic M, Mandic LM, et al. Validity of an FFQ assessing the vitamin D intake of young Serbian women living in a region without food fortification: the method of triads model. *Public Health Nutr.* (2016) 19:437–45. doi: 10.1017/S136898001500138X

34. Park Y, Kim SH, Lim YT, Ha YC, Chang JS, Kim IS, et al. Validation of a new food frequency questionnaire for assessment of calcium and vitamin d intake in korean women. *J Bone Metab.* (2013) 20:67–74. doi: 10.11005/jbm.2013.20.2.67

35. 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*. Amsterdam: Elsevier (2017). p. 5–48. doi: 10.1016/B978-0-12-802928-2.00001-1

36. Gibson RS. *Principles of Nutritional Assessment*. 2nd ed. New York, NY: Oxford University Press (2005). 908 p.

37. Krušič S, Hribar M, Hafner E, Žmitek K, Pravst I. Use of branded food composition databases for the exploitation of food fortification practices: a case study on vitamin D in the slovenian food supply. *Front Nutr.* (2022) 8:775163. doi: 10.3389/fnut.2021.775163

38. Roe M, Pinchen H, Church S, Finglas P. McCance and Widdowson's the composition of foods seventh summary edition and updated composition of foods integrated dataset. *Nutr Bull.* (2015) 40:36–9. doi: 10.1111/nbu.12124

39. Korošec M, Golob T, Bertoncelj J, Stibilj V, Seljak BK. The Slovenian food composition database. *Food Chem.* (2013) 140:495–9. doi: 10.1016/j.foodchem.2013.01.005

40. United States Department of Agriculture [USDA]. *United States Department of Agriculture: Food data central Washington: United States Department of Agriculture, Agricultural Research Service*. (2020). Available online at: https://fdc.nal.usda.gov/ (accessed April, 2021).

41. Vede T. *Slikovno Gradivo s Prikazom Velikosti Porcij*. Ljubljana: Nacionalni inštitut za javno zdravje (2016).

42. Biró L, Gee J. Development of a flexible, updatable, user-friendly electronic food frequency questionnaire. *Acta Alimentaria.* (2011) 40:117–27. doi: 10.1556/AAlim.40.2011.1.14

43. Odprta Platforma za Klinično Prehrano [OPKP]. *Odprta Platforma za Klinièno Prehrano [Spletno Orodje]*. Ljubljana: Institut Jožef Stefan (2021).

44. Fineli, Nutrition Unit of the National Institute for Health and Welfare. Fineli - the National Food Composition Database Helsinki: National Institute for Health and Welfare. (2019). Available online at: https://fineli.fi/fineli/en/index (accessed August 2021).

45. IBM Corp. *IBM SPSS Statistics for Windows. Version 22.0 ed*. New York, NY: IBM Corp. (2013).

46. Willet, W, Lenart E. Reproducibility and validity of food frequency questionnaires. 3rd ed. In: Walter Willet editor. *Nutritional Epidemiology*. Oxford, UK: Oxford University Press (2013). doi: 10.1093/acprof:oso/9780199754038.003.0006

47. Masson LF, McNeill G, Tomany JO, Simpson JA, Peace HS, Wei L, et al. Statistical approaches for assessing the relative validity of a food-frequency questionnaire: use of correlation coefficients and the kappa statistic. *Public Health Nutr.* (2003) 6:313–21. doi: 10.1079/PHN2002429

48. Lombard MJ, Steyn NP, Charlton KE, Senekal M. Application and interpretation of multiple statistical tests to evaluate validity of dietary intake assessment methods. *Nutr J.* (2015) 14:40. doi: 10.1186/s12937-015-0027-y

49. Giavarina D. Understanding bland altman analysis. *Biochem Med (Zagreb).* (2015) 25:141–51. doi: 10.11613/BM.2015.015

50. Lee Y, Park K. Reproducibility and validity of a semi-quantitative FFQ for trace elements. *Br J Nutr.* (2016) 116:864–73. doi: 10.1017/S0007114516002622

51. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England).* (1986) 1:307–10. doi: 10.1016/S0140-6736(86)90837-8

52. Perreault M, Xu VY, Hamilton S, Wright D, Foster W, Atkinson SA. Validation of a food frequency questionnaire for bone nutrients in pregnant women. *Can J Diet Pract Res.* (2016) 77:133–9. doi: 10.3148/cjdpr-2016-002

53. Doğan N. Bland-Altman analysis: a paradigm to understand correlation and agreement. *Turkish J Emerg Med.* (2018) 18:139–41. doi: 10.1016/j.tjem.2018.09.001

54. Itkonen ST, Erkkola M, Skaffari E, Saaristo P, Saarnio EM, Viljakainen HT, et al. Development and validation of an interview-administered FFQ for assessment of vitamin D and calcium intakes in Finnish women. *Br J Nutr.* (2016) 115:1100–7. doi: 10.1017/S0007114515005474

55. Hribar M, Hristov H, Gregorič M, Blaznik U, Zaletel K, Oblak A, et al. Nutrihealth study: seasonal variation in vitamin D status among the slovenian adult and elderly population. *Nutrients.* (2020) 12:1838. doi: 10.3390/nu12061838

56. Potosky AL, Block G, Hartman AM. The apparent validity of diet questionnaires is influenced by number of diet-record days used for comparison. *J Am Diet Assoc.* (1990) 90:810–3. doi: 10.1016/S0002-8223(21)01632-1

frontiers | Frontiers in Nutrition

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fnut.2022.993238

Performance and discriminatory capacity of Nutri-Score in branded foods in Greece

Antonis Vlassopoulos, Alexandra Katidi and Maria Kapsokefalou*

Laboratory of Chemistry and Food Analysis, Department of Food Science and Human Nutrition, Agricultural University of Athens, Athens, Greece

Check for updates

OPEN ACCESS

EDITED BY
Massimo Lucarini,
Council for Agricultural Research and Economics, Italy

REVIEWED BY
Mack Shelley,
Iowa State University, United States
Ivon Milder,
National Institute for Public Health and the Environment, Netherlands

*CORRESPONDENCE
Maria Kapsokefalou
kapsok@aua.gr

SPECIALTY SECTION
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

RECEIVED 13 July 2022
ACCEPTED 14 September 2022
PUBLISHED 28 September 2022

CITATION
Vlassopoulos A, Katidi A and Kapsokefalou M (2022) Performance and discriminatory capacity of Nutri-Score in branded foods in Greece.
Front. Nutr. 9:993238.
doi: 10.3389/fnut.2022.993238

Background: The harmonization of front-of-pack nutritional declaration is in the heart of food and nutrition policy discussions in Europe. The Nutri-Score system has been proposed by many countries as a potential candidate but its suitability for use across Europe is still under consideration. The current study aimed to evaluate the performance and discriminatory capacity of Nutri-Score in Greece and to test its alignment with the national food-based dietary guidelines.

Materials and methods: Data on the energy, saturated fat, total sugars, sodium, protein, and fiber content per 100°g or ml were extracted for all foods available (*n* = 4,002) in the Greek branded food composition database HelTH. Each food content in fruits, vegetables, pulses, nuts and oils was manually estimated from the ingredients list. The Nutri-Score algorithm was used both as a continuous (FSAm-NPS Score) and a categorical variable [Grades (A)–(E)].

Results: The average FSAm-NPS Score in Greece was 8.52 ± 9.4. More than half of the solid foods (53.8%) were graded from (A) to (C), while most beverages (59.2%) were graded (E). More than 50% of food categories were populated with foods in all Nutri-Score grades, indicative of a good discriminatory capacity. The system scores favorably vegetables, pulses, and low-fat dairy products and unfavourablly sweets and processed meats showing in this way good alignment with the national guidelines. Eggs and seafood scored preferably compared to meat products. Animal fats received less favorable scores and so did cereal products that were highly processed.

Discussion: Nutri-Score showed good capacity to inform consumers toward better food choices in line with the national guidelines. It showed a potential to guide consumers and manufacturers toward less energy dense and more nutrient dense options and highlighted areas of improvement in the food supply.

KEYWORDS

Nutri-Score, front-of-pack nutritional labelling, dietary guidelines, food supply, Greece

Introduction

The creation of a unified Front-of-Pack (FOP) labeling system is at the core of European discussions for the past years (1). Globally, scientists have developed numerous nutrient profiling algorithms over the years and in the past systems like the UK Traffic Lights and the Choices have been implemented in several countries (2).

Front-of-Pack (FOP) labeling has been proposed as a cost-effective tool for consumer education at the point of sale, linked to both improvements in dietary behaviors and in industry practices (3). However, the understanding and use of nutrition labeling varies greatly among European countries and population groups (4). The available evidence points toward color-coded interpretative systems that give an assessment of the healthiness of a specific food as the best option to enable consumer choices (5), but there still a need for guidance and standardization in the design and implementation of such policies (6, 7).

Since 2018, a number of European countries, namely France, Belgium, Germany, Luxembourg, Netherlands, and Spain have decided upon the adoption of Nutri-Score as the new FOP labeling system (8) and the WHO has supported the launch of this system across Europe to promote health (9). The Nutri-Score is a nutrient profiling system developed originally in France, which converts the nutritional content of foods into a five-tier score ranging from A to E (green to red) from healthier to less healthy choices within food groups (10). The system has been tested for its understanding with consumers across 12 countries (11) and there are data available about its capacity to discriminate foods based on their nutritional composition mainly in Central Europe (12). In Southern Europe and the Balkans the data are limited and often in specific food categories (13–15).

The Nutri-Score algorithm, is currently the strongest candidate for application across Europe; however there are concerns raised by some Mediterranean countries that the algorithm has not taken into account the specificities of the food system in the region (16–18). The validation of nutrient profiling systems and especially the assessment of the impact of FOP labeling policies is a data-intensive task. It requires access to granular food composition databases that cover a large number of foods currently sold in the country or region of interest and a representation of multiple food groups rather than just one food category or subcategory. Latest reports are focused on central and Western Europe exploiting regional branded food composition databases (BFCDs) available in the EuroFIR platform and in Open Food Facts (11). In Southern Europe and the Balkans, access to that type of data is still limited and hence there is a gap in the evaluation of the Nutri-Score in the region.

A new tool that could facilitate the investigation of Nurti-Score in Greece, is the Hellenic Food Thesaurus (HelTH). HelTH is a BFCD launched in 2019 by the Agricultural University of Athens, which collects and analyses all nutritional and quality data provided on labels of branded food products (19).

The aim of the present study was to assess the performance and test the discriminatory capacity of the Nutri-Score algorithm in the Greek foodscape, as well as to evaluate the alignment of Nutri-Score with the national food-based dietary guidelines (FBDGs) of Greece.

Materials and methods

The Hellenic food thesaurus database

Food composition data were extracted from HelTH, a dynamic dataset that compiles data on the nutritional composition and quality characteristics of branded foods available in Greek supermarkets.

Hellenic food thesaurus (HelTH) started as an initiative in 2018 and in its first version (11/2019), used in this analysis, contained data for n = 4,002 food products. In brief, HelTH includes information on the nutritional composition of foods, extracted from food labels available on the e-shops of large supermarket chains in Greece. Data on any health and/or nutrition claims made on pack, and information on any other quality claims written on pack (environmental claims, logos, origin, etc.) and the nutritional declaration was checked for quality by two independent researchers and curated in the database. A detailed description of the methodology and structure of HelTH has been published previously (19).

Data were selected on the basis of the availability of nutritional composition data and the availability of data to calculate Nutri-Score. Herbs and spices, alcoholic beverages, dietary supplements, and foods for special nutritional use were excluded (n = 139) as they are not included in the scope of the Nutri-Score according to the European regulation (10). All information around the nutritional composition was taken from the packaging and entered into the database. All products were classified in 13 categories and 36 subcategories following the LanguaL methodology.

All data of the HelTH BFCD were checked and cleaned. In particular, duplicates of the same product, constituting part of an offer or discount multi-package, or by human error appearing twice at the online platform, were excluded (multi-pack items were deleted where the single item was also available).

Nutri-Score calculation

The latest Nutri-Score algorithm was used in this analysis (20). In brief, the FSAm-NPS score was calculated for each food based on their nutritional composition per 100°g/ml of food/beverage (20). For each food, content of energy (kJ), total sugars (g), saturated fatty acids (SFAs) (g), and sodium (mg) were considered "negative nutrients" scored from 0 to 10 with higher scores for higher content. In the case of added fats, instead of SFA content the Ratio of SFA/Total Fat was used. Protein content (g), fiber content (g), and fruits/vegetables/pulses/nuts/specific oils content (FV%) were considered "positive nutrients" and received points from 0 to 5 with higher scores for higher content.

An overall score ranging from −15 to +40 was calculated by subtracting the "positive nutrients" score from the "negative nutrients" score. More specifically, fiber and FV scores were subtracted for all products, but the protein score was subtracted only in products with "negative nutrients" score < 11, those with an FV score > 5 or for cheeses.

The FSAm-NPS score was translated to Nutri-Score based on the following criteria (20): (A) was assigned to solid foods with a score from −15 to −1 or waters, (B) to solid foods with a score from 0 to 2 and beverages from −15 to 1, (C) to solid foods with a score 3 to 10 and beverages from 2 to 5, (D) to solid foods from 11 to 18 and beverages from 6 to 9 and (E) to solid foods from 19 to 40 and beverages from 10 to 40.

All nutrient contents were based on the labeled nutritional composition declaration. FV% was estimated based on the ingredient list in a two-step process. Firstly, all foods were screened to assess the presence of at least 40% content in fruits, vegetables, pulses, nuts and rapeseed, walnut, and olive oils, which is the minimum content required. Then for the products that met this minimum requirement a thorough quantification was carried out.

For the purpose of the study, products that did not contain any data about their energy, saturated fat, total sugar or sodium content (n = 778) were excluded, as no Nutri-Score could be calculated. Missing nutrient values could be due to lack of nutritional declaration or low-quality images obtained from the specific foods. On the contrary for "positive nutrients" missing information were imputed as zero.

Evaluation of alignment with the national food-based dietary guidelines

The latest FBDGs were developed in 2014 from a group of experts, have been endorsed by the National Nutrition Policy Committee and adopted by the Greek Ministry of Health on October 2017 as the national food-based dietary guidelines. These guidelines cover all age groups, but for this analysis only the parts on non-pregnant, healthy adults were used as a reference (21). To test the alignment between Nutri-Score and the national FBDGs (21), Langual food categories and subcategories were matched to the food categories as mentioned in the guidelines. The national FBDGs provide food-based guidance on the basis of "foods to avoid," "foods to consume in moderation" and "foods to promote." For the purpose of this analysis it was assumed that Nutri-Score grades (A) or (B) represented "foods to promote," grade (C) represented "foods to consume in moderation," and grades (D) and (E) represented "foods to avoid," following previously published methodology (22).

The national FBDGs provide an overarching guidance to avoid energy-dense foods and prefer nutrient-dense options. As the Nutri-Score algorithm follows a similar methodology in its ranking algorithm it was considered that for the algorithm to be considered aligned with the guidelines, within each food category there should be evidence of foods being ranked in multiple grades rather than all foods being clustered in a single grade.

In the same context, the national FBDGs advice toward food choices that is poorer in total fat, SFA, added sugars, salt and richer in unrefined cereals and fiber. To test the alignment of Nutri-Score with this guidance the macronutrient distribution of energy, SFA, salt and total sugars across Nutri-Score grades within each food category were tested. In the case of promoting the consumption of unrefined cereals, although the guidelines call for the promotion of wholegrain cereals, Nutri-Score does not track wholegrains and HelTH does not include wholegrain content data. As such in this analysis a food's fiber content was used as a proxy for wholegrain content. In this analysis, a decreasing fiber content with each increasing Nutri-Score grade would be considered an alignment to the FBDGs.

Statistical analysis

Statistical analysis was carried out using IBM SPSS Statistics® (version 23, Northridge, CA, USA). Nutritional composition data (content per 100 g or 100 mL of product) and the FSAm-NPS score were analyzed as continuous variables. Data were tested for normality using the Kolmogorov-Smirnov test. None of the variables followed the normal distribution. Therefore, variables were expressed as median (interquartile range). We assessed the distribution of prepacked products across different NS grades for main categories and subcategories and displayed this information in boxplots emphasizing median, 25th, and 75th percentiles. Discriminating ability was considered good when the food group comprised at least three different NS grades (12, 13). Differences were tested using the Kruskal-Wallis non-parametric test for k independent samples. Between-group differences were tested using the Mann-Whitney U test for continuous variables. Statistical significance was set at 0.01% to adjust for multiple comparisons (Bonferroni correction).

Results

Distribution of Nutri-Score

A total of 3,224 products were included in the final analysis with grain and grain products being the largest food category followed by dairy products and imitations and then non-milk beverages, sugar products and miscellaneous foods (Table 1). The median FSAm- NPS score for all categories was 10, with significant differences among the various food categories ($p < 0.001$). Vegetables had the lowest average score among all groups ($p < 0.001$, data not shown), followed by ready meals, eggs, and fruits which all received similar Nutri-Score ($p = 0.39$, data not shown). Sugar products had the highest FSAm-NPS Score compared to all food categories ($p < 0.001$ data not shown), followed by meat products, fats and oils, miscellaneous foods and non-milk beverages ($p < 0.001$ with the remaining categories, $p > 0.05$ among them, data not shown).

The distribution of FSAm-NPS Score across all categories is shown in Figure 1, separately for solids and beverages. Overall, 21.0% all of foods were rated A, 13.0% B, 16.5% C, 27.9% D, and 21.6% E. The distribution shows spikes especially around in-between Nutri-Score grades. For example, 6.7 and 5.5% of solid products were graded with score −1 (Grade A) and with score 0 (Grade B). The next highest prevalence 4.9% was seen around score 11 (start of Grade D).

Overall, 100% of egg products, 95.2% of vegetables products, 77.6% of ready meals, 67.5% of fruit products, and 48.6% of milk and milk products categories were graded as (A) or (B). On the contrary, 85.8% of meat products, 90.3% of sugar products, 65.1% of fats and oils, and 56.8% of miscellaneous foods were graded as (D) or (E). The same was true for beverages with 74.9% of all beverages being graded as (E) (Figure 2, Table 2).

In 8 out of the 13 food categories there was at least one product in every Nutri-Score grade. The categories with the lowest variability were egg products and fats and oils, in that order.

In terms of the distribution in subcategories within the milk products, milk and yogurt had the highest number of products rated (A) or (B), 93.5 and 86.8% respectively. On the other hand, cheeses were graded mostly (D) (78.6%), however a small proportion of cheeses (< 2%) were graded (A) and (B) (Table 2). Imitation milk products received an overall positive Nutri-Score with 79.2% being graded (A) or (B).

TABLE 1 Mean Nutri-Score per food category in the n = 3,224 branded food products of the Hellenic food thesaurus (HelTH) branded food composition databases (BFCD) analyzed.

Food category	Nutri-Score median (Q1, Q3)
Milk, milk products, and substitutes (n = 574)	3 (0, 15)
Eggs or egg products (n = 30)	0 (−1, 0)
Meat or meat products (n = 103)	15,5 (11, 19)
Fish and seafood (n = 58)	5 (2, 14)
Fats and oils (n = 63)	13 (9, 19)
Grains or grain products (n = 935)	9 (−1, 15)
Nuts and seeds (n = 114)	8 (2,13)
Vegetables or vegetable products (n = 210)	−6 (−10, −5)
Fruits or fruit products (n = 37)	1 (−2, 4)
Sugar or sugar products (n = 288)	22 (14, 26)
Non-milk beverages (n = 370)	11 (5, 15)
Ready meals (n = 76)	−0.5 (−4, 2)
Miscellaneous (n = 278)	11 (5,15)
Total (n = 3,085)	10 (0, 16)

For meat products, the most common Nutri-Score was (D) for all subcategories. Preserved meats and meat dishes showed some variability in Nutri-Score but the absolute numbers are very small ($n < 10$) (Table 2). In the case of fats and oils, animal sources were at large (96%) graded as (D) or (E), while plant-based margarines were graded either (C) or (D). At this point is worth mentioning that no vegetable fats were available in the version of HelTH used for the current analysis.

Grains and grain products, the largest food category, showed great variability in Nutri-Score. As the largest and most versatile food category, this variability was also seen among its subcategories with pasta, breads, rice, and cereal milling products receiving Nutri-Scores across the spectrum with larger numbers at the lower scores (Table 2). On the other hand, fine bakery ware and savory cereal dishes had a Nutri-Score distribution that technically started from grade (C) onward. For vegetables, the distributions were all skewed toward grades (A) and (B) for all vegetables, starchy or not, and for pulses alike. Some vegetable products existed with Nutri-Score grades above (C) but those represented less than 5% of the subcategory (Table 2). In contrary to vegetables, processed fruit products graded as (C) or (D) were 32.4% of all foods in the category.

Jams, non-chocolate confectionary, condiments and prepared food products were subcategories where Nutri-Score (D) was the dominant grade and that was more pronounced in the sweet options rather than the savory options. In general, those subcategories represent mainly sweet and savory snack foods. Sweet snacks are predominantly graded as (D) with the exception of chocolates with 89.6% of all products graded as (E). For savory snacks the main volume of products is split between grades (C) and (D).

More complex food products as they are represented by the composite dishes in the ready-to-eat and frozen foods subcategories receive overall positive grades, with > 70% of all products in (A) and (B). Semi-prepared dishes performed better than ready-to-eat products but even then, ready-to-eat foods were equally split between (A) and (B).

Finally, almost all juices and nectars (98%) were graded as (E), while for the remaining non-alcoholic beverages a quarter of the products were graded as (B) which is the lowest Nutri-Score for beverages other than water.

Alignment with food-based dietary guidelines

In terms of agreement with the national food-based guidelines, Nutri-Score shows a preliminary good agreement

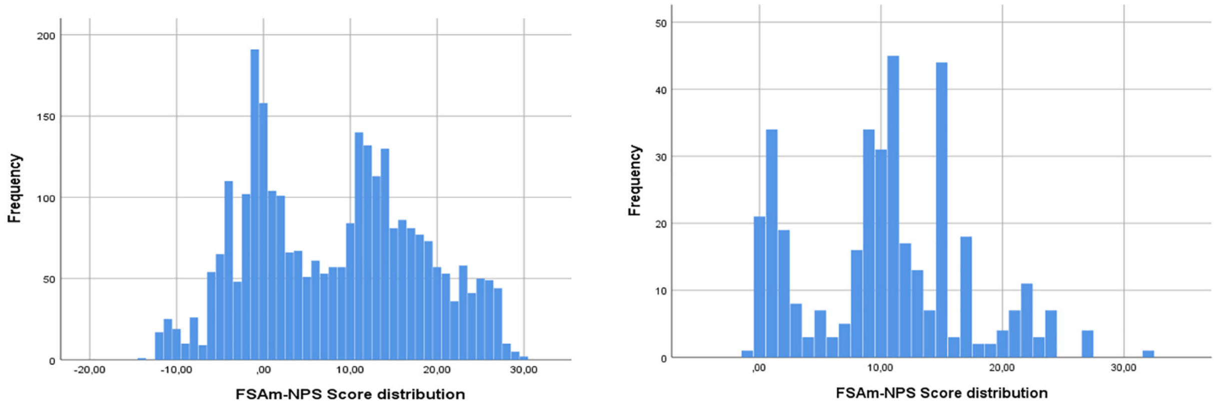


FIGURE 1
Distribution of FSAm-NPS Score among branded solid foods and beverages in the He1TH FCDB (*n* = 3,224).

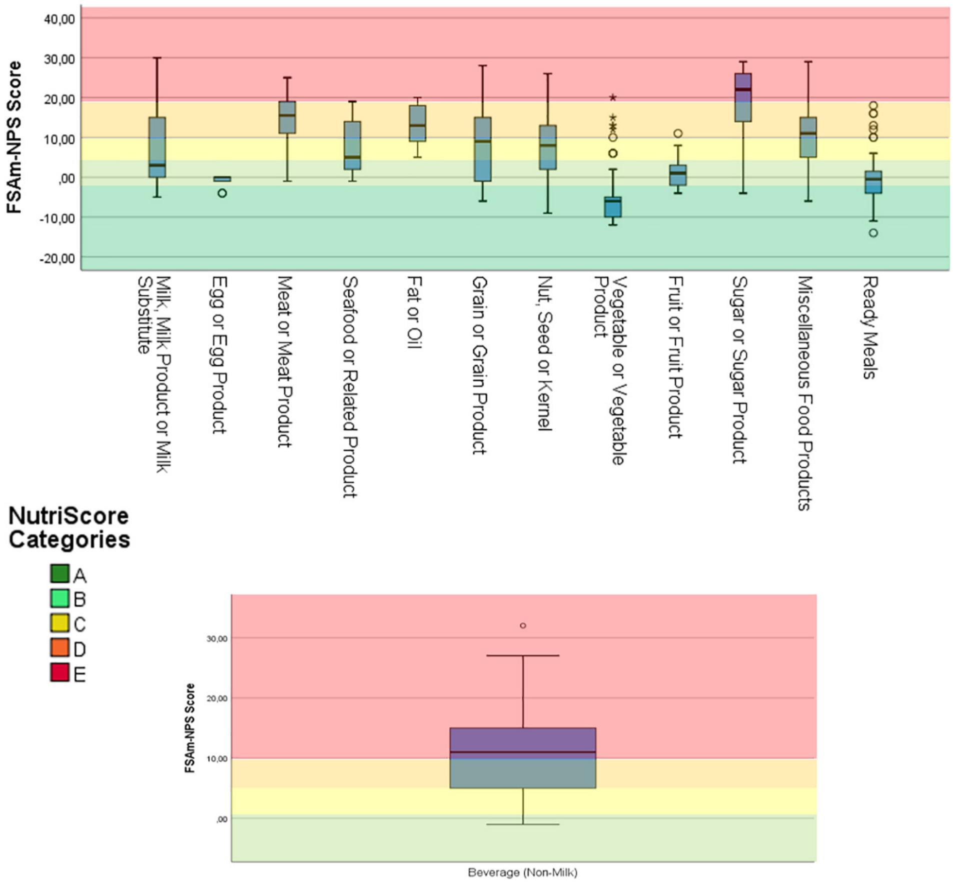


FIGURE 2
Overall distribution of products within the main food categories. Dark green: Nutri-Score "A", light green: Nutri-Score "B", yellow: Nutri-Score "C", light orange: Nutri-Score "D", and dark orange: Nutri-Score "E". No Nutri-Score "A" was represented on the graphic of beverages, given that only waters can be classified as "A" and were thus excluded from the graphic (*n* = 3,224).

as shown in Table 3. Overall, food groups like vegetables, fruits, and pulses that are mentioned in a positive manner in the guidelines are also scored preferably by Nutri-Score. On the other hand, animal sources of protein are more strictly judged by the system. From animal protein sources, Nutri-Score shows a tendency to favor eggs and seafood and to unfavour

TABLE 2 Nutri-Score distribution in food subcategories (n = 3,224) of the Hellenic food thesaurus (HelTH) branded food composition databases (BFCD).

Food category	Food subcategory	A n (%)	B n (%)	C n (%)	D n (%)	E n (%)
Milk, milk products, and substitutes (*n* = 574)	Milk (*n* = 147)	51 (34.7)	86 (58.8)	6 (4.1)	2 (1.4)	2 (1.4)
	Yogurt (*n* = 152)	78 (51.3)	54 (35.5)	20 (13.2)	–	–
	Cheese (*n* = 159)	2 (1.3)	3 (1.9)	15 (9.4)	125 (78.6)	14 (8.8)
	Cream (*n* = 30)	–	2 (6.7)	7 (23.3)	21 (70.0)	–
	Frozen dairy desserts (*n* = 38)	1 (2.6)	1 (2.6)	2 (5.3)	23 (60.5)	11 (28.9)
	Imitation milk products (*n* = 48)	–	1 (2.1)	23 (47.9)	15 (31.3)	9 (18.8)
Eggs or egg products (*n* = 30)		11 (36.7)	19 (63.3)	–	–	–
Meat or meat products (*n* = 105)	Preserved meat (*n* = 68)	–	1 (1.5)	10 (14.7)	37 (41.6)	20 (29.4)
	Sausages (*n* = 21)	–	–	–	12 (57.1)	9 (42.9)
	Meat dish (*n* = 16)	1 (6.3)	2 (12.5)	1 (6.3)	12 (61.5)	–
Fish and seafood (*n* = 58)		4 (6.9)	15 (25.9)	14 (24.1)	22 (37.9)	3 (5.2)
Fats and oils (*n* = 63)	Margarine or mixed fats/oils (*n* = 38)	–	–	21 (55.3)	17 (44.7)	–
	Animal fat/butter (*n* = 25)	–	–	1 (4.0)	8 (32.0)	16 (64.0)
Grains or grain products (*n* = 935)	Cereal or cereal-like milling products (*n* = 36)	4 (11.1)	5 (13.9)	7 (19.4)	13 (36.1)	7 (19.4)
	Rice (*n* = 63)	26 (41.3)	18 (28.6)	10 (10.3)	5 (7.9)	4 (6.3)
	Pasta (*n* = 201)	172 (85.6)	6 (3.0)	10 (5.0)	13 (6.5)	–
	Breakfast cereal and bars (*n* = 152)	18 (11.6)	13 (8.6)	62 (40.8)	59 (38.8)	–
	Bread or similar products (*n* = 180)	33 (18.3)	45 (25.0)	54 (30.0)	35 (19.4)	13 (7.2)
	Fine bakery ware (*n* = 227)	–	1 (0.4)	18 (7.9)	87 (38.3)	121 (53.3)
	Savory cereal dish (n = 76)	–	1 (1.3)	19 (25.0)	51 (67.1)	5 (6.6)
Nuts and seeds (*n* = 114)	Nuts (*n* = 54)	17 (31.5)	14 (25.9)	18 (33.3)	3 (5.6)	2 (3.7)
	Seeds (*n* = 34)	–	1 (2.9)	10 (29.4)	22 (64.7)	1 (2.9)
	Nuts or seeds products (*n* = 26)	–	–	5 (18.5)	19 (73.1)	2 (7.7)
Vegetables or vegetable products (*n* = 210)	Vegetables (*n* = 149)	135 (90.6)	4 (2.7)	5 (3.4)	4 (2.7)	1 (0.7)
	Starchy roots (*n* = 17)	10 (58.8)	7 (41.2)	–	–	–
	Pulses and products (*n* = 44)	43 (97.7)	1 (2.3)	–	–	–
Fruits or fruit products (*n* = 37)		15 (40.5)	10 (27.0)	11 (29.7)	1 (2.7)	–
Sugar or sugar products (*n* = 288)	Jams (*n* = 56)	2 (3.6)	–	21 (37.5)	32 (57.1)	1 (1.8)
	Non-chocolate confectionary (*n* = 40)	1 (2.5)	2 (5.0)	1 (2.5)	34 (85.0)	2 (5.0)
	Chocolate (*n* = 192)	–	–	1 (0.6)	19 (9.9)	172 (89.6)
Ready meals (*n* = 76)	Ready-to-eat (*n* = 43)	11 (30.6)	14 (38.9)	8 (22.2)	3 (8.3)	–
	Frozen, semi-ready (*n* = 41)	27 (67.5)	7 (17.5)	4 (10.0)	2 (5.0)	–
Miscellaneous (*n* = 278)	Spice, condiment (*n* = 144)	10 (4.4)	20 (8.7)	78 (34.1)	74 (32.3)	47 (20.5)
	Prepared food product (*n* = 135)	5 (3.7)	10 (7.5)	34 (25.4)	70 (52.2)	15 (11.2)
Non-milk beverages (*n* = 370)	Juice/nectar (*n* = 157)	–	1 (0.6)	10 (6.4)	48 (30.6)	98 (62.4)
	Non-alcoholic beverages (*n* = 213)	–	55 (25.8)	27 (12.7)	10 (4.7)	121 (56.8)

processed and cured meat products, in line with the national FBDGs. Increasing Nutri-Score in meat products and seafood was associated with higher sodium and SFA content ($p < 0.001$, data not shown). Sweets are overall unflavored and graded as (D) and (E), as are juices.

Fats, oils and nuts are mentioned as food groups to be consumed in moderation and with close consideration in their nutritional composition, in the case of Nutri-Score grading all food groups that contained the statement in moderation did not receive any grade below (C), which could be considered

TABLE 3 Presentation of Greek food-based dietary guidelines for adults (21) per food group/subgroup and the relevant distribution of Nutri-Score calculated for the branded food products of the Hellenic food thesaurus (HelTH) branded food composition databases (BFCD) (*n* = 3,224).

Food group/Subgroup	Guideline	A n (%)	B n (%)	C n (%)	D n (%)	E n (%)
Vegetables	Consume 4 portions a day Prefer fresh and uncooked vegetables Consume vegetable based main dishes 1–2 times/week	135 (90.6)	4 (2.7)	5 (3.4)	4 (2.7)	1 (0.7)
Fruits	Prefer fresh fruits Consume dried fruits in moderation Avoid canned fruit especially if preserved in syrup	15 (40.5)	10 (27.0)	11 (29.7)	1 (2.7)	–
Juices	Prefer fresh fruits to juices and consume up to 1/2 cup a day Avoid prepacked juices	–	1 (0.6)	10 (6.4)	48 (30.6)	98 (62.4)
Cereals[1]	Prefer wholegrain cereals, pasta and rice When choosing bread and breakfast cereals read the labels carefully as they can be hidden sources of salt and/or sugars	253 (39.7)	88 (13.8)	92 (14.4)	176 (27.6)	29 (4.5)
Potatoes	Consume 3 times a week Avoid French fries	10 (58.8)	7 (41.2)	–	–	–
Dairy products	Consume 2 portion/day with preference toward low fat milk, low fat yogurt and low-fat cheese Prefer foods naturally lower in fat and sodium	146 (25.4)	170 (29.6)	54 (9.4)	171 (29.8)	33 (5.7)
Milk	Prefer low fat milk Avoid sugar sweetened milk	51 (34.7)	86 (58.8)	6 (4.1)	2 (1.4)	2 (1.4)
Cheese	Prefer low fat and low sodium cheese	2 (1.3)	3 (1.9)	15 (9.4)	125 (78.6)	14 (8.8)
Yogurt	Prefer low fat yogurt	78 (51.3)	54 (35.5)	20 (13.2)	–	–
Cream	Avoid cream and replace it with yogurt when possible	–	2 (6.7)	7 (23.3)	21 (70.0)	–
Pulses	Consume 3 times/week Source of plant protein, fiber and micronutrients	40 (90.9)	2 (4.5)	2 (4.5)	–	–
Eggs	Consume up to 4 eggs/week Source of high-quality protein	11 (36.7)	19 (63.3)	–	–	–
Meat	Consume up to 1 portion/week red meat Consume 1–2 portions/week white meat Avoid processed or cured meats	1 (0.9)	3 (2.8)	11 (10.4)	61 (57.5)	30 (28.3)
Fish and seafood	Consume 2–3 portions/week Prefer fresh fish to seafood Avoid any processed fish/seafood	4 (6.9)	15 (25.9)	14 (24.1)	22 (37.9)	3 (5.2)
Fats and oils	Consume all fats and oils in moderation (total 4–5 portions/day) Prefer olive oil as the main oil followed by other vegetable oils (except palm oil) Avoid animal fats and hard margarines	–	–	22 (34.9)	25 (39.7)	16 (25.4)
Nuts and products	Consume in moderation Count toward the 4–5 portions/day of fats and oils Prefer unsalted nuts Use nut spreads as a snack	–	–	15 (25.0)	41 (68.3)	4 (6.7)
Sweets[2]	Reduce all sweets to 1 portion/week	3 (0.6)	3 (0.6)	34 (6.6)	179 (34.8)	296 (57.5)
Spices and condiments	Avoid commercial spices and condiments as they are sources of sodium and sugar	1 (0.7)	6 (4.2)	59 (41.0)	63 (43.8)	15 (10.4)
Beverages	Prefer water and unsweetened beverages Avoid sugar-sweetened beverages	–	55 (25.8)	27 (12.7)	10 (4.7)	121 (56.8)

[1] Not including fine bakery ware.
[2] Including fine bakery ware.

in agreement with the guideline. In the case of grains and cereals, Nutri-Score showed a wide variability but an analysis of the fiber content showed that foods graded as (D) and (E) had significantly lower fiber content compared to all other Nutri-Score grades ($p < 0.01$). More specifically cereal products in Nutri-Score grades (D) and (E) had an average fiber content of 3.91 ± 2.7 and 2.50 ± 1.2 g/100°g respectively, as opposed to products graded (A) to (C) which had an average content of 5.24 ± 4.6 g/100°g. The majority of wholegrain/non-refined cereals (76%) were graded either (A) or (B) which indicates a greater capacity to highlight the differences in the nutritional composition of this subcategory.

The Greek food-based guidelines mention a preference toward dairy foods that are low in fat naturally, meaning a prioritization of milks, yogurts which is documented in the Nutri-Score performance in the dairy subcategories. Only a small number of cheeses were graded as (A) or (B) but there was good discriminatory capacity among cheeses as all Nutri-Score grades were populated. In dairy products, increasing Nutri-Score was associated with increasing energy, SFA, sodium, and total sugars content ($p < 0.001$, data not shown). On the other hand, although the National FBGs include a mention on avoiding sweetened dairy products, only a few products (namely sweetened condensed milk) received a Nutri-Score grade above (C). Sweetened yogurts (either kid's yogurts or yogurt desserts) were graded as (B) or (C) even when sweetened with fruit juices/jams.

Discussion

This study is the first to apply the Nutri-Score algorithm in a large sample of branded food products currently available in Greece. In that context this study also expands previous work on the application of Nutri-Score in countries of the European south and to test its alignment with the national food-based dietary guidelines (23).

Distribution and discriminatory capacity of Nutri-Score

The overall aim of Nutri-Score is to facilitate consumers' understanding of the nutritional information and thus to help them in making informed choices (20). For this to be achieved Nutri-Score needs to be able to identify alternatives within the same food group. In the current analysis ∼50% of all food groups were populated with products that were graded across the whole Nutri-Score spectrum (A)–(E). In fact, only three food groups, eggs, juices, and fats and oils showed narrow distributions. The discriminatory capacity of Nutri-Score was less apparent in subcategories, ∼44% of the subcategories covered all the Nutri-Score range. Larger categories and subcategories showed better discriminatory capacity and on the opposite side very homogeneous categories showed limited discriminatory capacity. These results are in agreement with previous reports from various European countries (12, 13, 22) but also from Mediterranean countries like Italy and Spain (23).

When the FSAm-NPS score variability is studied it becomes apparent that there is a clustering of products around cut-off values, indicating that Nutri-Score once rolled out could be used as a stimulus for food reformulation. In fact, the highest clustering of food products is seen in the cut-off value between grades (B) and (C) (FSAm-NPS Score = 1) with a second peak at FSAm-NPS Score = 11, the cut-off point between grades (C) and (D). That shows that although currently 27.9% of all products are graded (D) and 16.5% are graded (C), it is possible for a substantial proportion of those foods to improve without extensive reformulation. In fact, 9.5% of all foods have an FSAm-NPS score = 11–12 and 7.1% of all foods have scores at FSAm-NPS = 1–2. Similar results were seen in the Netherlands (22) in France (24) were the potential of Nutri-Score to guide reformulation was deemed high. The phenomenon of clustering around cut-off points is documented in multiple countries across Europe (13, 23) but most importantly it is more apparent in countries with higher average Nutri-Score. Overall, the FSAm-NPS Score in Europe ranges for 7.6–9.9, with Slovenia, France reporting the highest scores (13, 23). However, there is a positive association between the number of foods analyzed and the average Nutri-Score for the country (23). This could be explained by the type of data included in each analysis, the same analysis when performed in branded food composition databases only leads to greater average FSAm-NPS scores as compared to analyses carried out using a combination of branded and generic food composition databases (12, 23). In that context as Nutri-Score is designed to be implemented on packed foods, one could argue that branded food composition databases are more appropriate to test the algorithm's performance in conditions that mimic the foodscape. In fact, an analysis in the Slovenian foodscape highlighted that branded food composition data combining with market share data are even more appropriate to describe the performance of Nutri-Score as often the products with the less desirable nutritional compositions are the ones that are preferred from the consumers (13).

In the case of Greece, Nutri-Score managed to successfully identify "healthier" options for consumers in all food categories and subcategories allowing for product substitutions up to two Nutri-Score grades below. The agreement of our findings with previous analyses in other countries also adds to the discussion of the potential for extrapolation of the findings across Europe and even in other regions, suggesting that Nutri-Score performance is rather homogenous in multiple settings.

Alignment with the national food-based dietary guidelines

As FOP labeling's main purpose is consumer information on healthier food choices, a key stage in its validation is testing its alignment with national and international guidelines (7, 25, 26). In the past, Nutri-Score has been validated against the dietary guidelines of various countries (10, 12, 13, 23), while some controversies were raised in others, like in the Netherlands (22, 27). In our analysis of the alignment of

Nutri-Score with the food-based dietary guidelines for Greece we found good agreement between the two both in principle (nutrients to be reduced, nutrients to be promoted) but also among specific subgroups. Overall, all food groups that were mentioned in the guidelines as foods to be promoted like vegetables, fruits and pulses received the lowest Nutri-Score. Although the guidelines mention fruits as foods to be promoted in our analysis approximately 30% of all foods were graded (C). This can be explained from the nature of the foods available in HelTH, which in the case of fruits would include mainly dried and canned fruit (19). In this context, the Nutri-Score outcome in this analysis reflects quite closely the spirit of guidelines, that call for an increased intake of fresh fruit, the consumption of dried fruit in moderation and avoidance of canned fruit and fruit juices (21).

A similar explanation could be offered for the unfavorable grading of the meat and meat products group, which in the case of HelTH is mainly populated by sausages, cured or dried meats which are discouraged both as potential carcinogens and for their high fat and sodium content (21, 28). When studied collectively in our analysis, animal protein sources like eggs and seafood were prioritized by the Nutri-Score algorithm over processed meat. Plant based protein from pulses were even further promoted. In the case of ready meals, that was also true as meals higher in protein but poorer SFA and sodium received better FSAm-NPS score, directing consumers toward white meat and fish/seafood options. Although not covered by the national guidelines, even among dairy products, plant based dairy imitations also received better FSAm-NPS Scores. For dairy products, the Nutri-Score algorithm showed good alignment with the guidelines asking for a prioritization over lower fat and sodium dairy options such as milk and yogurt and then the consumption of cheeses that are naturally low in fat. In our analysis we were able to identify a small number of such products, both traditional and low-fat versions of traditional foods. Previous work target in the most commonly consumed traditional Greek cheeses, confirmed epidemiological data suggesting that traditional cheeses are generally discouraged by Nutri-Score (14, 15) but there might be a need for a targeted expansion of such databases to include less popular traditional cheeses that are naturally low in fat and/or sodium (29). In the case of sweetened dairy, Nutri-Score graded sweetened yogurts as (B) or (C), as opposed to (A) for the low fat, unsweetened alternatives. As far as within category comparisons are concerned the algorithm shows a fair discriminatory capacity between the sweetened and unsweetened variant. The discriminatory capacity is stronger across categories when comparisons between sweetened dairy products and sweets and confectionaries are concerned. The Nutri-Score algorithm also shows a good capacity to differentiate refined and non-refined cereal as non-refined cereals were in their majority graded as (A) or (B) and were all concentrated in the lower part of the FSAm-NPS distribution.

The lowest discriminatory capacity was seen among sweets and more so among chocolates. Although discriminatory capacity is always better to help identify "healthier" options in the case of those food subcategories, the lack of discriminatory capacity is in line with national and international guidelines that call for a reduction in sugar intake and the avoidance of sweets to a maximum of one portion per week (21, 30).

Finally, a key consideration for Nutri-Score in Greece is its performance vis-à-vis fats and oils. In the case of our analysis, the dataset used did not include any data on vegetable oils (19), as such the results presented herein do not include any data on vegetable oils including olive oil. On the contrary the dataset includes data on vegetable and animal fats. As per the latest Nutri-Score algorithm (20), olive oil is automatically graded as (C). With this in mind, there are two important considerations in the topic, with the first being the discriminatory capacity of the algorithm. Based on our data, no fats available in our dataset received a Nutri-Score grade lower than (C), margarines received either (C) or (D), while animal fats were primarily graded as (E). That indicates an agreement with the FBDGs that propose the avoidance of animal fats and the larger uptake of vegetable fats and oils. As far as, olive oil is concerned it is true that it is not graded more favorably than all fats available in the Greek marketplace and in fact a large proportion of margarines would receive a similar Nutri-Score to olive oil. It is important that future research performs more targeted analysis in fats and oils, in order to understand whether there is still a need for further finetuning of the algorithm although preliminary data suggest against it (17). The second issue is linked with the nutritional composition of the fats that are graded similarly to olive oil. A preliminary analysis of our data, indicates that the majority of margarines with a Nutri-Score (C), are reformulated products with higher olive oil content, or fortified with plant-sterols, or even products with higher protein content (yogurt fortified margarines). Overall, the caping of Nutri-Score at grade (C) as the lowest possible grade for fats and oils could be considered in line with the FBDGs asking for moderate consumption of such products.

Limitations of the current study linked to the nature of the HelTH dataset have already been mentioned in the relevant sections. Although HelTh is the only available branded food composition database for Greece and covers an important part of the market, it is still in need of targeted expansions as for the case of oils and potentially novel foods like plant-based meat alternatives etc. Despite, its gaps the use of branded food composition databases is linked with substantial improvement in the relevance of the results for the consumer and the food industry as it is a direct reflection of the marketplace as compared to analyses performed on generic food composition data (31, 32).

This study also faced issues with missing data, especially for positive nutrients like protein and fiber. Although data completeness is relatively high for protein, fiber is only declared in food categories that are relevant or in foods that carry a nutrition claim for the specific nutrient (19). In the case of missing nutrient data, those were common among traditional artisanal foods that are not required by the regulation to carry a full nutritional declaration or due to the inability of the researchers to obtain access to the physical packaging of the foods. As described earlier, HelTH obtained data from products sold on e-shops of large supermarket chains. Often foods were missing clear images or images altogether from the nutritional declaration, in some cases those data have been added to the database through sampling in the physical supermarket but this process is still ongoing. The choice to impute positive nutrients with zero was merely of a mathematical nature. Imputation with zero for the positive nutrients was only likely to underestimate a food's performance and that was decided to be the safest and prudent approach. However, the wider implementation of Nutri-Score as a FOP scheme is likely to resolve the data completeness issue as more manufactures would be displaying positive nutrients included in the Nutri-Score algorithm.

The hardest part of the Nutri-Score calculation is the calculation of the Fruits, Vegetables, Nuts, Pulses, and Oils component. This calculation has to be done manually and it is always linked to underestimation. Especially, in the context of the Mediterranean foodscape the importance of this component is vital as both national and Mediterranean Diet guidelines suggest that foods that contain vegetables or pulses and use olive oil as their main fat should be preferred (21, 33). Another area of importance is the use of dietary fiber as a proxy for wholegrain cereal content. As fiber and wholegrain content do not always correlate, the addition of wholegrain content as part of the Nutri-Score algorithm could be considered (34).

In the case of testing the alignment with existing FBDGs there are additional limitations to be considered. These include issues like the misalignment of the food categories as mentioned in the guidelines as opposed to the food categories proposed by systems like Langual or FoodEx2. For example, although the guidelines considered fine bakery ware to be considered sweets from a food technology point of view these foods are more likely to be classified as cereal-based foods. Similarly, the guidelines often refer to decreased intake of trans-fatty acids and increased intake of fiber and wholegrains, however this information is not mandatory as part of the nutritional declaration in EU and is often missing or it needs to be manually estimated from the ingredients list.

Although, this work offers evidence on the alignment of Nutri-Score with national FBDGs, it is important for Nutri-Score to be tested against dietary patterns with a documented beneficial effect on health. The Mediterranean Diet Pyramid is such a pattern and its principles expand beyond the nutritional composition of foods to cover elements of locality, tradition, seasonality, culinary, and cultural elements. Testing the alignment of Nutri-Score with dietary patterns like the Mediterranean Diet would require targeted analysis and testing. Future work should aim to directly test the alignment of Nutri-Score with these guidelines as it will allow to answer questions around the type of reformulation that Nutri-Score will promote in the Mediterranean and whether the traditional cooking techniques will be favored and what the impact of Nutri-Score would be on the Mediterranean agri-food value chain.

Conclusion

Overall, this study is the first to report the performance and discriminatory capacity of Nutri-Score in the Greek foodscape using a branded food composition database. It highlights an overall good discriminatory capacity and satisfactory agreement with the national FBDGs. However, the evaluation of an upcoming food policy requires further data on the consumer perception and likelihood for adoption, as well as an analysis of the alignment with the existing agricultural policies and agroeconomic strategies. After this complete description of the risks and benefits a roadmap of implementation could be developed.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

AV: formal analysis and writing—original draft preparation. AV and AK: investigation and data curation. MK: resources and supervision. All authors have conceptualization, methodology, writing—review and editing, read, and agreed to the published version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. European Commission. *Proposal Revision Regulation of Food Information to Consumers*. Luxembourg: European Commission (2020).

2. Labonté MÈ, Poon T, Gladanac B, Ahmed M, Franco-Arellano B, Rayner M, et al. Nutrient profile models with applications in government-led nutrition policies aimed at health promotion and noncommunicable disease prevention: a systematic review. *Adv Nutr.* (2018) 9:741–88. doi: 10.1093/advances/nmy045

3. Shangguan S, Afshin A, Shulkin M, Ma W, Marsden D, Smith J, et al. A Meta-Analysis Of Food Labeling Effects On Consumer Diet behaviors and industry practices. *Am J Prev Med.* (2019) 56:300–14. doi: 10.1016/j.amepre.2018.09.024

4. Grunert KG, Fernández-Celemín L, Wills JM, Storcksdieck Genannt Bonsmann S, Nureeva L. Use and understanding of nutrition information on food labels in six European countries. *Z Gesundh Wiss.* (2010) 18:261–77. doi: 10.1007/s10389-009-0307-0

5. Muller L, Ruffieux B. What makes a front-of-pack nutritional labelling system effective: the impact of key design components on food purchases. *Nutrients.* (2020) 12:2870. doi: 10.3390/nu12092870

6. Goiana-da-Silva F, Cruz-e-Silva D, Miraldo M, Calhau C, Bento A, Cruz D, et al. Front-of-pack labelling policies and the need for guidance. *Lancet Public Health.* (2019) 4:e15. doi: 10.1016/S2468-2667(18)30256-1

7. Kelly B, Jewell J. *What is the Evidence on the Policy Specifications, Development Processes and Effectiveness of Existing front-of-Pack Food Labelling Policies in the WHO European Region?*. Copenhagen: WHO Regional Office for Europe (2018).

8. Southey F. *7 European Countries team up to Propel Nutri-Score Rollout*. Crawley: Foodnavigator (2021).

9. IARC. *Nutri-Score: Harmonized and Mandatory Front-of-pack Nutrition Label Urgently Needed at the European Union Level and Beyond*. Lyon: IARC (2021).

10. Chantal J, Hercberg S, World Health Organization. Development of a new front-of-pack nutrition label in France: the five-colour nutri-score. *Public Health Panorama.* (2017) 3:712–25.

11. Egnell M, Talati Z, Hercberg S, Pettigrew S, Julia C. Objective understanding of front-of-package nutrition labels: an international comparative experimental study across 12 countries. *Nutrients.* (2018) 10:1542. doi: 10.3390/nu10101542

12. Dréano-Trécant L, Egnell M, Hercberg S, Galan P, Soudon J, Fialon M, et al. Performance of the front-of-pack nutrition label nutri-score to discriminate the nutritional quality of foods products: a comparative study across 8 European countries. *Nutrients.* (2020) 12:1303. doi: 10.3390/nu12051303

13. Hafner E, Pravst I. Evaluation of the ability of nutri-score to discriminate the nutritional quality of prepacked foods using a sale-weighting approach. *Foods.* (2021) 10:1689. doi: 10.3390/foods10081689

14. Katsouri E, Magriplis E, Zampelas A, Nychas GJ, Drosinos EH. Nutritional characteristics of prepacked feta PDO cheese products in greece: assessment of dietary intakes and nutritional profiles. *Foods.* (2020) 9:E253. doi: 10.3390/foods9030253

15. Katsouri E, Magriplis E, Zampelas A, Drosinos EH, Nychas GJ. Dietary intake assessment of pre-packed graviera cheese in Greece and nutritional characterization using the nutri-score front of pack label scheme. *Nutrients.* (2021) 13:295. doi: 10.3390/nu13020295

16. Delhomme V. Front-of-pack nutrition labelling in the European Union: a behavioural, legal and political analysis. *Eur J Risk Regulat.* (2021) 12:825–48. doi: 10.1017/err.2021.5

17. Fialon M, Salas-Salvadó J, Babio N, Touvier M, Hercberg S, Galan P. Is FOP nutrition label nutri-score well understood by consumers when comparing the nutritional quality of added fats, and does it negatively impact the image of olive oil? *Foods.* (2021) 10:2209. doi: 10.3390/foods10092209

18. Fialon, M, Nabec L, Julia C. Legitimacy of front-of-pack nutrition labels: controversy over the deployment of the nutri-score in Italy. *Int J Health Policy Manag.* (2022) doi: 10.34172/ijhpm.2022.6127

19. Katidi A, Vlassopoulos A, Kapsokefalou M. Development of the Hellenic food thesaurus (HelTH), a branded food composition database: aims, design and preliminary findings. *Food Chem.* (2021) 347:129010. doi: 10.1016/j.foodchem.2021.129010

20. Sante Publique France. *Nutri-Score*. (2022).

21. Linou A. *Institute of Preventive, Environmental & Occupational Medicine. National Food Based Dietary Guidelines for Adults*. Athens: Lampakis SA (2014).

22. ter Borg S, Steenbergen E, Milder IEJ, Temme EHM. Evaluation of nutri-score in relation to dietary guidelines and food reformulation in the Netherlands. *Nutrients.* (2021) 13:4536. doi: 10.3390/nu13124536

23. de Edelenyi FS, Egnell M, Galan P, Druesne-Pecollo N, Hercberg S, Julia C. Ability of the Nutri-Score front-of-pack nutrition label to discriminate the nutritional quality of foods in the German food market and consistency with nutritional recommendations. *Arch Public Health.* (2019) 77:28. doi: 10.1186/s13690-019-0357-x

24. Julia C, Kesse-Guyot E, Touvier M, Méjean C, Fezeu L, Hercberg S. Application of the British food standards agency nutrient profiling system in a French food composition database. *Br J Nutr.* (2014) 112:1699–705. doi: 10.1017/S0007114514002761

25. Combris P, Goglia R, Henini M, Soler LG, Spiteri M. Improvement of the nutritional quality of foods as a public health tool. *Public Health.* (2011) 125:717–24. doi: 10.1016/j.puhe.2011.07.004

26. Jones A, Neal B, Reeve B, Mhurchu CN, Thow AM. Front-of-pack nutrition labelling to promote healthier diets: current practice and opportunities to strengthen regulation worldwide. *BMJ Global Health.* (2019) 4:e001882. doi: 10.1136/bmjgh-2019-001882

27. van Tongeren C, Jansen L. Adjustments needed for the use of nutri-score in the Netherlands: lack of selectivity and conformity with dutch dietary guidelines in four product groups. *Int J Nutr Food Sci.* (2020) 9:33. doi: 10.11648/j.ijnfs.20200902.11

28. Iarc Working Group on the Evaluation of Carcinogenic Risks to Humans. *Red Meat and Processed Meat [Internet]*. Lyon:International Agency for Research on Cancer. (2018).

29. Andrikopoulos NK, Kalogeropoulos N, Zerva A, Zerva U, Hassapidou M, Kapoulas VM. Evaluation of cholesterol and other nutrient parameters of Greek cheese varieties. *J Food Composit Analy.* (2003) 16:155–67. doi: 10.1016/S0889-1575(02)00164-3

30. World Health Organization Guidelines Approved by the Guidelines Review Committee. *Guideline: Sugars Intake for Adults and Children*. Geneva: World Health Organization (2015)

31. Pravst I, Hribar M, Žmitek K, Blažica B, Koroušić Seljak B, Kušar A. Branded foods databases as a tool to support nutrition research and monitoring of the food supply: insights from the slovenian composition and labeling information system. *Front Nutr.* (2022) 8:798576. doi: 10.3389/fnut.2021.798576

32. Kapsokefalou M, Roe M, Turrini A, Costa HS, Martinez-Victoria E, Marletta L, et al. Food composition at present: new challenges. *Nutrients.* (2019) 11:E1714. doi: 10.3390/nu11081714

33. Serra-Majem L, Tomaino L, Dernini S, Berry EM, Lairon D, Ngo de la Cruz J, et al. Updating the mediterranean diet pyramid towards sustainability: focus on environmental concerns. *Int J Environ Res Public Health.* (2020) 17:E8758. doi: 10.3390/ijerph17238758

34. Drewnowski A, McKeown N, Kissock K, Beck E, Mejborn H, Vieux F, et al. Perspective: why whole grains should be incorporated into nutrient-profile models to better capture nutrient density. *Adv Nutr.* (2021) 12:600–8. doi: 10.1093/advances/nmaa172

TYPE Original Research
PUBLISHED 14 December 2022
DOI 10.3389/fnut.2022.1025929

Check for updates

OPEN ACCESS

EDITED BY
Alessandra Durazzo,
Council for Agricultural Research
and Economics, Italy

REVIEWED BY
Şenol Çelik,
Bingöl University, Turkey
Nasir Abbas,
Government Graduate College,
Adhiwal Chowk, Pakistan

*CORRESPONDENCE
Nadia Shabnam
nadia.shabnam@numspak.edu.pk

SPECIALTY SECTION
This article was submitted to
Nutrition Methodology,
a section of the journal
Frontiers in Nutrition

RECEIVED 23 August 2022
ACCEPTED 04 November 2022
PUBLISHED 14 December 2022

CITATION
Shabnam N (2022) Assessment
of non-linearity in calorie–income
relationship in Pakistan.
Front. Nutr. 9:1025929.
doi: 10.3389/fnut.2022.1025929

Assessment of non-linearity in calorie–income relationship in Pakistan

Nadia Shabnam*

Department of Health Professions Education, National University of Medical Sciences, Rawalpindi, Pakistan

This article considers the issue of assessing non-linearity in the relationship between calorie consumption and income using non-parametric and semi-parametric approaches. These methodologies are implemented on the cross-sectional household survey data conducted in Pakistan in 2010–2011. This framework takes account of the heterogeneity among families and potential non-linearity in the relationship. The findings show that the calorie–income elasticity is considerable and statistically significant across estimating methodologies. The results also demonstrate that the elasticity is larger for the substantially poorer households of the sample. By incorporating the explanatory variables in a manageable way in the parametric section of regression procedures, the semi-parametric analysis also reveals a slight increase in calorie response to increases in income at various income levels.

KEYWORDS

calorie–income, non-parametric regression, semi-parametric regression, single-index model, non-linear elasticities, Pakistan

Introduction

One of the most significant issues affecting the impoverished, in both developed and developing countries alike, is possibly inadequate nutrition. Malnutrition would make people less productive and make them more susceptible to illness, both of which would contribute to the continued poverty and further problems for the poor. Calorie consumption has been demonstrated to have a substantial correlation with both productivity and human health, making it one of the most significant aspects from the perspective of policymakers (1). On the one hand, the human body needs calories to preserve its natural metabolism. On the other hand, calorie consumption is the top priority for policymakers when creating programs helpful for the underprivileged parts of society. These policies, which are being implemented in various countries, can be categorized as (i) basic food subsidies, (ii) cash transfers, (iii) food vouchers, and (iv) conditional finance. The success of these policies is based on the strategy used in designing the program (2) or the sensitivity of food demand to changes in income (3). As a result, we decided to use calorie consumption as the subject of our research in this work. The role of income in calorie consumption continues to generate serious investigations, with contrasting results appearing throughout the literature. The debate regarding the size of the calorie–income relationship is well-documented in the literature [details are given in (4)]. Recently, Santeramo and Shabnam (5) well-summarized this debate by providing a meta-analysis of articles published on this issue in several countries of the world. Most of the studies in the literature used the parametric approach, while non-linear specifications were also in many studies. Following Gibson and Rozelle (6), only few studies used semi-parametric specifications to deal with the non-linearity of the calorie–income relationship (3, 7, 8).

Previous studies focusing on the parametric approach have revealed that the relationship between income and calorie is linear. While poleman (9) and Lipton (10) have argued that the calorie–income curve may be elbow-shaped for samples from the very poor category, indicating that share of food budget initially increases with the increase in income for the poor households. Similarly, Strauss and Thomas (11) reported that elasticity for the lowest decile increased up to 0.26 and then decreased to 0.03 for the highest decile. Thus, following Ravallion (12), the literature generally agrees that the calorie–income relationship is non-linear. It shows that with the increase in income, per capita calorie consumption increases and then tends to decrease with a further increase in income. However, non-linear specifications such as the quadratic term of income and expenditure may not always be appropriate to capture the non-linearity or shape of the calorie–income relationship.

Another way to capture this non-linearity existing in the calorie–income relationship is by using non-parametric procedures. Non-parametric smoothing techniques represent a set of flexible tools for analyzing unknown regression relationships. These techniques can search for appropriate non-linear forms that can best describe the available data and also provide useful tools for parametric non-linear modeling and helpful diagnostics.[1] Gibson and Rozelle (6), Abdulai and Aubert (13, 14), Skoufias et al. (15), Babatunde et al. (1), Skoufias et al. (16), among others, used the non-parametric approach to capture the potential non-linearity in the calorie–income relationship. Although non-linear items in the calorie–income relationship can be investigated using a non-parametric technique in general, this approach is limited to bivariate relationships. When we take into account the impact of additional potential variables, the situation gets worse. The "curse of dimensionality" refers to the issues related to this non-parametric method. The precision of the non-parametric estimator diminishes as the component of X grows. Thus, some authors emphasize on this point and favor the use of parametric estimates to examine the impact of additional factors other than expenditure on the consumption of calories and nutrients. But in this study, we prefer to use semi-parametric regression methods to deal with the curse of dimensionality. This article aims at contributing to the body of knowledge regarding calorie–income estimation using current advancements in semi-parametric estimation methods and model selection as well (17) in order to address the non-linearity problem mentioned beforehand.

In general, semi-parametric methods combine parametric and fully non-parametric models in a specific mode. Semi-parametric methods are supposed to impose assumptions that are stronger than the fully non-parametric method but less restrictive than the parametric method of estimation. This allows the semi-parametric methods to trim down the effective dimension of the estimation problem, thus increasing the precision of estimation relative to that obtained by the non-parametric estimation, while allowing greater flexibility and lowering the risk of specification errors that are possible with the parametric model. Semi-parametric methods represent some widely accepted methods that provide a flexible estimation. However, the use of the semi-parametric approach is still very limited in the literature.

As a result, our goal in this article is to explore the calorie–income link by employing non-parametric and semi-parametric techniques for analyzing household survey data (2010–2011). In a fully non-parametric regression framework, we used the logarithm of per capita calorie intake conditional on the logarithm of per capita expenditure, while in a semi-parametric framework, some other control variables can be added. Here, we consider the partially linear regression approach and the semi-parametric single-index model from the family of semi-parametric specifications. Several potential options such as GAM specifications and parametric double-log specifications are available, and we must choose among them. We used a procedure proposed by Hasio et al. (18) to choose among these various competing parametric, non-parametric, and semi-parametric specifications.

Following the Introduction, in section "Methodology" of the article, we give an overview of both the non-parametric regression method and the semi-parametric regression approach. Data, models, and descriptive statistics are presented in section "Data." Finally, in section "Results," we present the estimated results and contrast them with the parametric results to draw conclusions about the study. Section "Discussion" concludes the study.

Methodology

In this section, we provide an overview of the estimation techniques used to explore the issue of the calorie–income relationship.

Non-parametric estimation method

In the non-parametric method, no assumption is made regarding the functional form of conditional mean function and

[1] Details regarding non-parametric regression methods are given in Li and Racine (19).

assumed that $r(x)$ satisfies the smoothness condition such as differentiability. The technical detail is given in Li and Racine (19). We use the local linear kernel regression to estimate $r(x)$, and the procedure for this technique is given as follows. At any given point x, we run a weighted linear regression of Y on X. The weights are chosen for the observations of *Yi* are higher for which *Xi* is close to x than the observations which are far from x. The estimate of $r(x)$ is the predicted value from the local regression at x, and the estimated slope coefficient of local regression "$\hat{\beta}(x)$" is considered an estimate of the slope $\hat{r}(x)$. Let (h) be a sequence of positive numbers, known as the bandwidth that converges to 0 as $n \to \infty$.

Semi-parametric estimation methods

Previous studies used two methods to incorporate other control variables in calorie demand models. For example, Subramanian and Deaton (20) split the sample according to household size and then estimated the non-parametric regression for the calorie–income relationship within each subsample. Strauss and Thomas (21) first used the non-parametric locally weighted smoothing scatter plot technique to capture the non-linear items in the calorie–income relationship and then used the log-inverse of the quadratic term (parametric functional form) to approximate the shape they observed in their non-parametric framework. The major advantage of using the parametric approach is that other potential control variables can be added to the model. In this article, we implemented new methods to incorporate the covariates into the non-parametric model that are semi-parametric methods, as follows: semi-parametric partially linear regression method (two or three studies implemented this methodology, as mentioned before) and semi-parametric single-index method (none of the studies in literature implemented this approach).

Partially linear model

The semi-parametric partially linear regression model combines both non-parametric and parametric components and is given as follows:

$$Y_j = X_j'\beta + G(R_j) + u_j, \qquad j = 1, ..., n \tag{2.1}$$

where X_j is $q \times 1$ vector, β is $q \times 1$ vector of unknown parameters, G is an unknown function, and $R_j \in \mathbb{Z}^p$. The finite-dimensional parameter β represents the parametric part, and the unknown function G (.) represents the non-parametric part of the model. The data are supposed to be independent and identically distributed random variables (i.i.d) that are given as follows:

$$E\left(u_j|\mathrm{X}_j, R_j\right) = 0 \tag{2.2}$$

$$E\left(u_j^2|\mathrm{X}_j = x, R_j = r\right) = \sigma^2\,(x, r) \tag{2.3}$$

In the partially linear model, the foremost issue is the identification of β ; once this is carried out, an estimator of G (.) can be easily obtained. The partially linear model was first proposed by Robinson (22), and then Li and Racine (19) extended this work to handle the presence of qualitative variables in this model.

Single-index model

A semi-parametric single-index model has a form of a conditional mean function given as follows:

$$Y = G(X'\beta) \tag{2.4}$$

where Y is the dependent variable, $X \in \mathbb{Z}^p$ is the vector of covariates, β is an unknown parameter vector of order $p \times 1$, and G is an unknown function. The quantity $X'\beta$ is known as *single index* as it is scalar, even though x is vector. From Equation (2), we can see that our model is only a function of $X'\beta$ because when the functional form of G (.) is unspecified, then the location parameter α cannot be identified. This implies that Y depends on x only by the way of linear combination of $X'\beta$, and the relationship is characterized by the link function G (.). Thus, the main statistical issue is to estimate G and β from the data (Y,X). Model (2) involves many widely used parametric models as special cases. Such as, if G is the identity function, then (2) is the linear model. If G is observed to be cumulative normal or logistic distribution, then Equation (2) is a discrete-choice logit or probit model. In a case where $G$is unspecified, Equation (2) gives a specification that is more flexible than a parametric model. Thus, the semi-parametric single-index model just like the partially linear model is designed to lessen the effects emerging due to the curse of dimensionality.

Identification condition

For the estimation of β and G, some restrictions are required for their identification. That is, β and G must be obtained through the population distribution of (Y, X), as follows:

$$\mathrm{E}\,[Y|x] = G\left(x'\beta\right) \tag{2.5}$$

The identification conditions for the single-index model were first investigated by Ichimura (23), and then in the case of the binary response model, Manski (24) and Horowitz (25) presented identifiability conditions for the single-index model. The identification of β and G in a semi-parametric single-index model requires that

(a) $G(.)$ cannot be a constant function; otherwise, β is not identified.

(b) Perfect multicollinearity is not permissible among components of x.

(c) x should include at least one continuous random variable. The intuition behind this can be explained by the following reason. Suppose x has only a binary (0–1 dummy) variable, then the range of x is finite as well as the range of $X'\beta$ for any vector β. Of course, there exists an infinite number of functions $G(.)$ and β vectors that satisfy the finite number of restrictions imposed by$\mathrm{E}\left[Y|x\right] = G\left(x'\beta\right)$. For more details, refer to the study by Horowitz (25), who explained this condition for a specific example.

(d) x should not include a constant term (intercept) as long as β does not include the location parameter. It should only be identifiable up to a scale. For example, $\mathrm{E}\left[Y|x\right] = G\left(x'\beta\right)$ and $\mathrm{E}\left[Y|x\right] = G^{*}\left(\lambda + \theta x'\beta\right)$ are observationally equivalent models, where λ and θ are both arbitrary and not equal to zero and G^{*} is defined by the relation $G^{*}\left(\lambda + \theta\omega\right) = G\omega$ for all ω in the range of $X'\beta$. So, β and G cannot be identified, unless we imposed the restrictions that uniquely identify λ and θ. The restriction imposed on λ is called location normalization and involves that X should not include the intercept term. The restriction on θ is called scale normalization, and this can be attained by assuming the first component of X is equal to 1, that is, β has unit length ($||\beta|| = 1$), and this component is assumed to be continuous.

Ichimura's method

Several estimation methods are available to estimate β, but we describe the estimation method proposed by Ichimura (23) and used this method for analysis. If the function G were specified, then Equation (2.5) would be a standard non-linear regression model and β could be estimated through a non-linear least square (NLS) method with possible weights by minimizing $\sum_{i=1}^{n}\left[Y_i - G\left(X_i'\beta\right)\right]^2$ with respect to β. Then, the estimator would be as follows:

$$\hat{\beta} = \underset{\hat{\beta}\in Z^{a}}{\arg\min} \sum_{i=1}^{n} \eta\left(X_i\right)\left[Y_i - G\left(X_i'\beta\right)\right]^2 \qquad (2.6)$$

However, if the function G is unknown, then we first need to estimate $G(.)$. In this situation, the kernel method cannot be directly applied to estimate $G\left(X'\beta\right)$ because both β and β are unknown. In this situation, we can estimate $Y_i = G\left(X'\beta\right) + \varepsilon_i$ and $\mathrm{E}\left(\varepsilon_i|X_i\right) = 0$ for a given value of β by using the kernel method, which is given as follows:

$$G\left(X_i'\beta\right) \equiv \mathrm{E}\left[Y_i|X_i'\beta\right] = \mathrm{E}\left[g\left(X_i'\beta\right)|X_i'\beta\right] \qquad (2.7)$$

If $\beta = \hat{\beta}G\left(X_i'\beta\right) = g\left(X_i'\beta\right)$, then $G\left(X_i'\beta\right) \neq g\left(X_i'\beta\right)$ if $\beta \neq \hat{\beta}$ in general. A leave-one-out non-parametric kernel estimator of $G\left(X_i'\beta\right)$is given as follows:

$$\hat{G}_{-i}\left(X_i'\beta\right) \equiv \hat{\mathrm{E}}_{-i}\left(Y_i|X_i'\beta\right)$$

$$= \frac{(nh)^{-1}\sum_{j=1,j\neq i}^{n} Y_j\left(\frac{X_j - X_i'\beta}{h}\right)}{\hat{s}_{-1}\left(X_i'\beta\right)} \qquad (2.8)$$

$\hat{s}_{-i}\left(X_i'\beta\right) = (nh)^{-1}\sum_{j=1,j\neq 1}^{n} k\left(\frac{X_j' - X_i'\beta}{h}\right)$. Thus, Ichimura (23) suggested the estimation of $G\left(X_i'\beta\right)$ by replacing with the leave-one-out estimator $\hat{G}_{-i}\left(X_i'\beta\right)$ and choosing β using the semi-parametric NLS method. In this method, Ichimura also used a trimming function to trim out the small values of $\hat{s}_{-i}\left(X_i'\beta\right)$. Consider the following:

$$A_{\upsilon} = \left\{s(x'\beta) \geq \upsilon, \forall\beta B\right\} \qquad (2.9)$$

$$A_m = \left\{x : \left|x - x^{*}\right| \leq 2h \text{ for some } x^{*} \in A_m\right\} \qquad (2.10)$$

$\upsilon > 0$ is a constant, B is a compact subset in $\mathbb{Z}^{p}$, $A_{\vartheta} \subset A_m$ as $n \to \infty$, $h \to 0$ than A_m get smaller too A_{ϑ} . Thus, Ichimura (23) estimator is as follows:

$$\hat{\beta}_I = \arg \underset{\beta}{\min} \sum_{i=1}^{n}\left[Y_i - \hat{G}_{-i}\left(X'\beta\right)\right]^2 \eta\left(x_i\right) 1\left\{X_i \in A_{\vartheta}\right\} \qquad (2.11)$$

$\eta(X_i)$ is a non-negative weight function that is bounded in A_{ϑ}, I(.) is an indicator function, $1\left\{X_i \in A_{\vartheta}\right\}$is a trimming function that equals 1 if $X_i \in A_{\vartheta}$, or zero otherwise. The trimming function provides guarantee that the random denominator in the kernel estimator is non-negative, with high probability so as to simplify the asymptotic analysis.

Model specification test

The Hsiao test is based on the moments that hold value zero if a parametric specification (H$_0$) is correct, or greater than zero otherwise. In this case, the null hypothesis is given as follows:

$$H_0^a = \mathrm{E}(Y|x) = \theta(x, \beta_0) = 1 \textit{ for some } \beta_0 \in \mathrm{B} \subset \mathbb{Z}^{p}$$

where $\theta(x, \beta_0)$ is a known function with β_0 as a vector of unknown parameters of order $p \times 1$. Under the alternative hypothesis, we have a function that is negation of H_0^a :

$$H_1^a = \mathrm{E}(Y|x) = m(x) \neq \theta(x, \beta_0) < 1 \textit{ for all } \beta_0 \in \mathrm{B}$$

The test statistics are based on $I = \mathrm{E}\{U\mathrm{E}(U|X)f(X)\}$, where $U = Y - \theta(x, \beta_0)$ is independently proposed by Fan and Gijbels (26) and Zheng (27). Consider that $I = \mathrm{E}\{[\mathrm{E}(u_i|x_i)]^2 f(x_i)\} \geq 0$ and $I = 0$ if the null hypothesis is true. Thereby, I is a valid candidate for testing H_0^a . The sample analog of I is given as follows:

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \hat{\mathrm{E}}_{-i}(u_i|x_i)\hat{f}_{-i}(x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i\{n^{-1}\sum_{j=1,j\neq i}^{n} \hat{u}_j K_{\omega,ij}\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \hat{u}_i\hat{u}_j K_{\omega,ij} \qquad (2.12)$$

where $\hat{u}_i = y_i - \theta(x_i, \hat{\beta})$ is the residual of the parametric null model, $\hat{\beta}$ is $\sqrt{n}$ consistent estimator of β, $K_{\omega,ij} = M_{h,ij}W_{\gamma,ij}(\omega = (h, \gamma))$, and $\hat{E}_{-i}(u_i|x_i)\hat{f}_{-i}(x_i)$ is the leave-one-out kernel estimator of $E(y_i|x_i)f(x_i)$. A CV method is used for the selection of h and γ, and $\hat{I}_n$ is used to denote the CV-based test and can be defined the same way as I_n in previous equation but replacing them $(h_1 ... h_q, \gamma_1 ... \gamma_r)$with CV smoothing parameters$(\hat{h}_1 ... \hat{h}_q, \hat{\gamma}_1 ... \hat{\gamma}_r)$. The rejection region for the test at the α level of significance is $J_n > c_\alpha$, and the critical value c_α can be obtained by the wild bootstrap method. For detail about the wild bootstrap method, see the monograph of Li and Wang (28), Li and Racine (19, pp. 357), and Hsiao et al. (29).

Data

This study uses data of Household Integrated and Economic Survey (HIES) 2010–2011 (30), carried out from July 2010 to June 2011. The sample comprises 16,341 households and is a nationally representative survey covering 14 large cities and 81 districts, as well as urban and rural areas. The HIES also reports information on a variety of social issues, including education, health, employment and income, immunization, use and satisfaction with facilities and services, and household consumption and details. In the consumption module, the survey collects information on the quantities and values of 69 food items, and along with this, the survey also records information on 79 non-food items. The food consumption module of the HIES provides the main data for our analysis.

To compute the calorie consumption amount from the reported food quantities, we applied the conversion factors from Food Composition Table for Pakistan (31), which contains data on nutrient contents for various food items [for details of data extraction, refer to the study by (4)]. For example, the nutrient consumed of a particular type like calorie is as follows:

$$N \quad \Sigma\theta_i Q_i$$

where N is the quantity of the particular nutrient consumed, θi is the average nutrient content of a unit of food i, and Qi is the number of units consumed of food i (4).

For the purpose of analysis, we have Y (the response variable) as logarithm of per capita daily calorie consumption, and control variables on the right-hand side are logarithm of per capita expenditure (ln_PCME), household size (HHsize), gender of household head (F_HHH), age of the household head (Age_HHH), and employment status of the household head (E_Status). There are also other potential variables that can be used, but we used the dimension reduction method named least absolute shrinkage and selection operator (LASSO)[2] method to select the variables to better explain these estimation procedures. The flaw of the non-parametric method of considering only the bivariate relationship can be handled by using semi-parametric methods by including other covariates in the model in a tractable manner, but, in our study, explanatory variables were around 26, and data size was also large enough, so it is not feasible to include the entire set of variables in the semi-parametric method and obtain the results. Thus, we used LASSO as a variable selection procedure in order to ease the computational burden and increase the prediction accuracy.

Stepwise regression normally chooses models that include just a subset of the variables, while ridge regression includes all variables in the final model, and the penalty factor (θ) in ridge regression shrinks the coefficients toward zero but does not set any of the coefficients exactly to zero and does not exclude any of those from the final model (32). The LASSO overcomes this disadvantage of ridge regression. The LASSO minimizes the quantity as follows:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{q}\beta_j x_{ij}\right)^2 + \theta\sum_{j=1}^{q}|\beta_j|$$

$$RSS + \theta\sum_{j=1}^{q}|\beta_j| \tag{3.1}$$

Statistically, the LASSO uses an ℓ_1 norm and a coefficient vector associated with this norm is as $||\beta||_1 = \sum|\beta_j|$. The LASSO not only shrinks the coefficients toward zero but also forces to be exactly equal to zero when parameter θ is sufficiently large. Consequently, the model generated can be easily interpreted and produced by ridge regression [for more details, see (32)]. We first used LASSO for the variable selection in the linear model in this study, then we used the same set of explanatory variables (excluding expenditure variable) for the parametric part in semi-parametric methods.

Results

A sample of 16,290 households were used for the analysis. Descriptive statistics for the variable used in the study are given in **Table 1**. For non-parametric estimation, we restricted our analysis to only per capita expenditure and used it as the independent variable to avoid the curse of dimensionality. The results obtained from LASSO show that per capita expenditure first enters the model. Then, E_Status, shortly followed by F_HHH, HHsize, and Age_HHH, simultaneously enters in the model.

The models that we estimated are as follows:

[2] Codes and results from LASSO can be provided on request.

TABLE 1 Summary statistics of the variables used in semi-parametric models.

Variables	Mean	Standard deviation	Minimum	Maximum
ln_PCDCC	7.86	0.28	6.43	8.98
ln_PCME	7.94	0.53	6.21	11.64
HHsize	6.68	3.04	1.00	25.00
F_HHH	0.08	0.27	0.00	1.00
Age_HHH	46.20	13.23	45.00	75.00
E_Status	0.20	0.40	0.00	1.00

$$E\left(Y|\ln\_PCME, regressors\right) = \beta_0 + \beta_1.\ln\_PCME + \beta_2.HHsize + \beta_3.\text{F\_HHH}+ \beta_4.\text{Age\_HHH}+ \beta_5.\text{E\_Status} \quad (4.1)$$

$$E\left(Y|\ln\_PCME, regressors\right) = G(\beta_1.\ln\_PCME + \beta_2.HHsize + \beta_3.\text{F\_HHH}+ \beta_4.\text{Age\_HHH}+ \beta_5.\text{E\_Status}) \quad (4.2)$$

$$E\left(Y|\ln\_PCME, regressors\right) = \beta_1.HHsize + \beta_2.\text{F\_HHH}+\beta_3.\text{Age\_HHH}+ \beta_4.\text{E\_Status} + m(\ln\_PCME) \quad (4.3)$$

$$E\left(Y|\ln\_PCME\right) = r(\ln\_PCME) \quad (4.4)$$

where G, m, and r are unknown functions, and β's are unknown model parameters that may have different values in different models. Model (4.1) is a parametric model; model (4.2) and (4.3) are semi-parametric in nature, particularly model (4.2) is a single-index model and (4.3) is a partially linear model; and model (4.4) is fully non-parametric. For model (4.1), the parameters were estimated by using the standard OLS method. In the semi-parametric single index model, scale normalization was attained by setting $\beta_1 = 1$ using the non-linear least square method of Ichimura (23). This method uses a kernel estimator to estimate the unspecified function G. In the case of partially linear regression model, $E\left(Y|\ln\_PCME\right)$ and $E(regressors|\ln_PCME)$were estimated by local linear regression using the second-order Gaussian kernel.

$$K(w) = \exp\left(-Z^2/2\right)/2\sqrt{\pi}$$

where $z = (x_i - x)/h$ and $h > 0$. The fully non-parametric model (4.4) was estimated by using the local linear kernel method, and the method of least square cross-validation was used for the bandwidth selection in all estimation methods.

Non-parametric results

The calorie–expenditure curve in Figure 1A is positively sloped, then slightly flattens, and last demonstrates a sudden dip for very high-income group of expenditure distribution, but not flattens out at the tail. The possible reason for the dip could be the presence of outliers at the tail or the fact that the tail behavior in the non-parametric regression is not always good because of having few observations in the tails (33). Only 10% of the sample belongs to a very high-income group, and this may be the reason that the curve is not flattening out and showing a decreasing trend at the tail.

The gradients related to the non-parametric regression model provide a noticeable picture of the relationship. The gradients are shown in Figure 1B, which also shows 95% confidence bands for the gradients of the local linear non-parametric regression. Its shows the local linear fit by using a second-order Gaussian kernel method, with a CV bandwidth of 0.378 and bootstrapped standard error to construct a 95% confidence interval. The bootstrap procedure does not consider the cluster effect, thereby correcting the possible heteroscedasticity in errors. The procedure of bootstrapping was performed with 50, 100, and 200 replications, but in all procedures, the confidence bands obtained from standard errors were identical. Efron and Tibshirani (34) suggested that 200 replications are enough for the estimation of standard errors. In our case, the bands were fairly tight around the lower and middle of the regression and wide at the upper tail.

The income elasticity of calories with bootstrap standard error in Figure 1B shows that the curve slopes downward, which means calorie consumption falls less rapidly for poorer households because their income constraints either the quantity or quality of their food budget. The overall representation of this simple bivariate relation by using the non-parametric estimation method implies that calorie–income elasticity is statistically different from zero for almost all income levels, except for the very high-income level, where income elasticity is negative and insignificant, and it shows that local linear regression estimates the relationship with relative precision. We also ran a parametric regression of the log per capita daily calorie consumption on log of per capita expenditure to determine how well it demonstrates the true relationship by using the non-parametric model.

Semi-parametric results

This section describes the results of semi-parametric regression methods. Table 2 shows the β parameter estimates of models (4.1–4.3). To get a clear picture as compared to the point estimate, we semi-parametrically modeled the relation between calorie consumption and expenditure for a given parametric specification of the effect of household characteristics on consumption of calories. The basic aim, throughout the analysis,

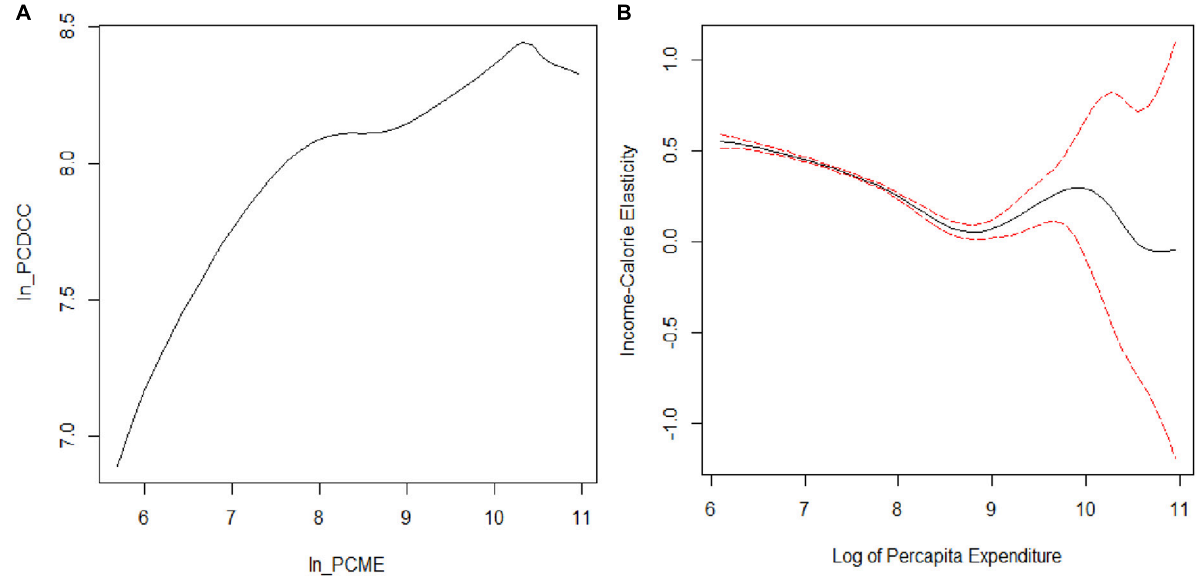


FIGURE 1
Non-parametric estimation of calorie–income relationship. **(A)** Calorie–expenditure curve and **(B)** income–calorie elasticity.

TABLE 2 Parameter estimates of parametric and semi-parametric methods.

Independent variables	Parametric model	Partially linear model	Single-index model
Constant	5.658*** (0.359)	–	–
ln_PCME	0.307*** (0.005)	–	1
HHsize	−0.007*** (0.001)	−0.005*** (0.001)	−0.010*** (0.002)
F_HHH	0.027*** (0.007)	0.033*** (0.006)	0.082*** (0.028)
Age_HHH	−0.001*** (0.000)	−0.002*** (0.000)	−0.002*** (0.001)
E_Stauts	0.175*** (0.004)	0.162** (0.004)	0.376*** (0.010)
R^2	0.371	0.41	0.42

Standard errors are within parentheses. *, **, and *** indicate statistical significance at 10, 5, and 1%, respectively.

is to explore the response of calorie consumption over a range of income distribution to income changes, rather than at a single point.

The income elasticity of calorie consumption is lower in multivariate parametric regression than in the bivariate regression model with the per capita expenditure as the only regressor (0.32). Indeed, there is a small difference between the parametric and partially linear estimates, but there is a relatively higher difference between parametric and single-index estimates.

Figure 2 shows the elasticity for different levels of income distribution and also demonstrates a higher and statistically significant estimate for the lower income group. The figure shows that expenditure elasticities are less than unity and remain fairly constant between 0.7 and 0.8 over a range of the low-income group. It is only at levels of Y above the sample mean of monthly per capita expenditure of Rs. 2,596 (in terms of log as 7.2). This elasticity decreases with the increase in income, and beyond the mean income, it begins to decrease and then becomes insignificant (as zero line is included in the confidence band). One possible reason for income elasticity being not statistically different from zero for the higher income group could be their interest in non-nutritive attributes of food items (33). Overall, the picture illustrates the fact that calorie consumption will improve with the change in income for poorer households as compared with their rich households. This result is consistent with the study of Subramanian and Deaton (20), Gibson and Rozelle (6), Tian and Yu (7), Nie and Sousa-Poza (8), and Trinh et al. (3). Trinh et al. (3) used semi-parametric specifications belonging to the family of generalized additive models to estimate the relationship for China and Vietnam. We also observed that the plot of non-parametric and semi-parametric regressions is almost the same in scale and shape, and this is consistent with the study of Bhalotra and Attfield (35) and Roy (33).

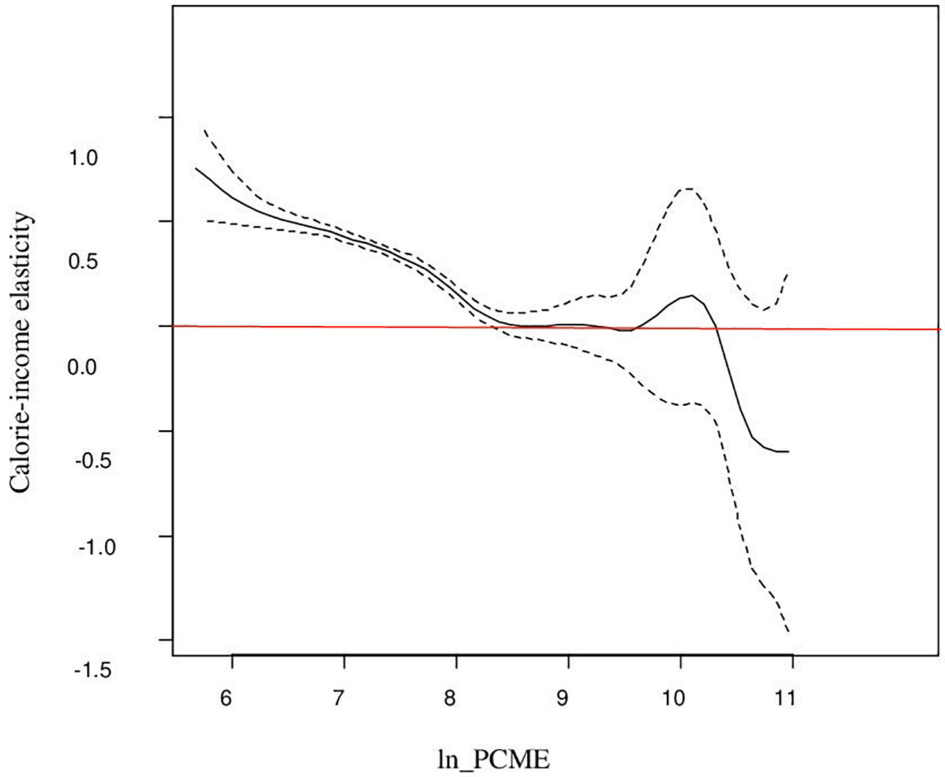


FIGURE 2
Semi-parametric partially linear estimation of calorie–income relationship.

However, in terms of point estimate, the average elasticity is slightly higher in the semi-parametric partially linear regression model than in the fully non-parametric regression model. It shows that by adding covariates to the model causes only a small increase in the elasticity of calories. Moreover, Gibson and Rozelle (6) showed a slightly downward shift in elasticity estimates by adding covariates in the semi-parametric model. The coefficient of per capita expenditure in the single-index model is set by normalization. Thus, the eminent feature of the single-index model is that $E(Y|regressors)$ is constant along curves such as $\ln\_PCME + \beta_2.x_2 + \beta_3.x_3 + \beta_4.x_4 + \beta_5.x_5$ is constant for the parameter β . The curve in **Figure 3** shows that the index is increasing and has a similar trend as in the non-parametric model but has a fluctuating behavior at the upper end of the tail. However, providing an average value for non-parametric and semi-parametric models really wipes out the significant contribution of this kind of analysis.

The household size has a negative magnitude in all models (**Table 2**). It shows that economies of size decrease the calorie consumption by 0.5–1 in the percentage point. Similarly, age of the household head has a negative effect on household calorie consumption. Gender of the household head also has a positive and significant effect on calorie consumption. Results reported in **Table 2** show that a female household head increases the calorie consumption by 3–8% compared with a male household head. In addition to this, if a household head is employed, then the head will perform better care of welfare of the members of household in terms of increasing calorie consumption. Finally, the last row in **Table 2** provides the goodness of fit of the parametric and semi-parametric models. The value of R^2 shows that the parametric fit is poor compared with the fit of semi-parametric models, and the single-index model has a better fit than the partially linear model. Thus, the single-index model emerges out to be a better specified semi-parametric model on the basis of goodness of fit.

We have also used some formal specification model tests (6.37–6.40) based on residual analysis for the purpose of comparison among models. Many procedures are available for testing a parametric model against its non-parametric alternative, but here, we used the test proposed by Hsiao et al. (29) due to its number of desirable properties in comparison to others. Hsiao et al. (29) proposed a non-parametric kernel-based model specification test and used a cross-validation (CV) method of bandwidth selection. This test used a residual-based wild bootstrap method to approximate the null distribution of the test statistics.

In our case, the implication of Hsiao's test rejects the parametric model against the non-parametric model at the 1%

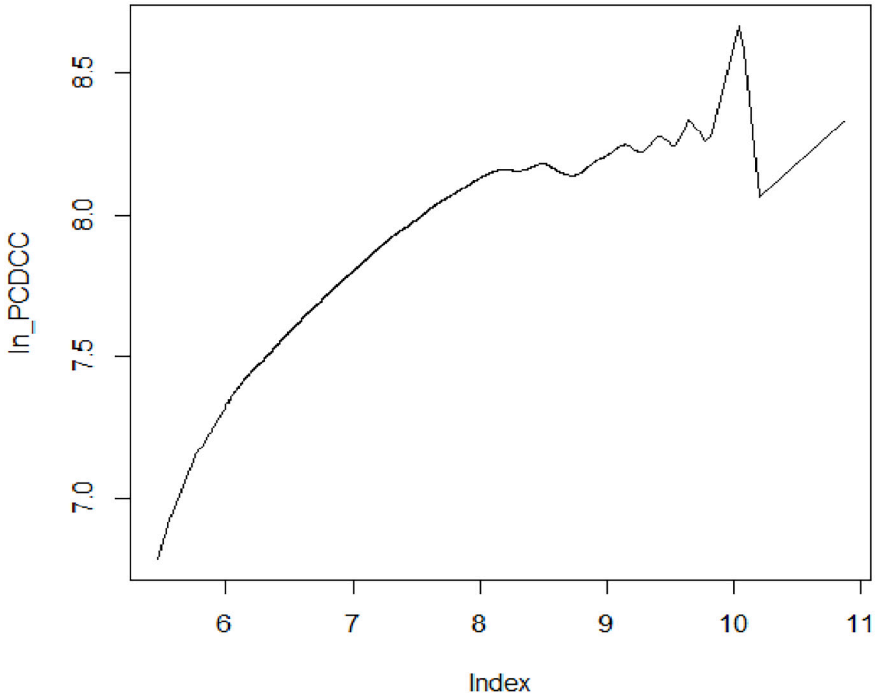


FIGURE 3
Semi-parametric single-index estimation of calorie–income relationship.

level of significance (J_n: 78.596, $p < 0.001$) and turns out to be significant for the non-parametric regression model. This formal specification test shows that the non-parametric model outperforms the usual parametric model, and this result is consistent with the results of the informal graphical analysis, as shown in Figure 1.

We also used Hsiao's test to test the significance of semi-parametric methods with the parametric model in the null hypothesis, and again, this specification test rejects the parametric model (J_n: 12.032, $p < 0.001$) and supports the implication of semi-parametric methods. Thus, the formal test is consistent with the informal graphical analysis, as shown in Figures 2, 3. This result is consistent with the findings of Trinh et al. (3), although they have used a preference test for model selection.

Discussion

Non-parametric and semi-parametric estimation methods have attracted a great deal of attention from statisticians in the last decade. Horowitz and Lee (36) reported that the expediency of semi-parametric models in applied statistics is not well-understood in the literature yet, and any new application of semi-parametric models will generate valuable additional piece of information about these models. This article sheds light on the non-parametric and various semi-parametric estimators and demonstrated them with an application of consumption survey data (2010–2011) to identify the calorie–income relationship.

The analysis reported in this article shows that non-parametric and semi-parametric estimation methods achieved the proposed goal to capture the non-linearity in the calorie–income relationship. The fully non-parametric estimate embodies the true conditional mean function up to random sampling errors. Figure 1B shows a downward trend from the lower tail to the upper tail and demonstrates that calorie consumption decreases less rapidly for poorer households. Of course, the slopes at the extreme of the distribution are quite imprecisely estimated, but at the median level of expenditure, the slope is around 0.40 and is precisely estimated. However, it shows that local linear kernel regression estimates the relationship with relative precision. In addition to this, the article demonstrates the implication of two classes of semi-parametric regression: One is the partially linear model and the second is the single-index model. The plot of partial linear regression (Figure 2) shows that calorie consumption improves with the change in income for poorer households as compared with their rich counterparts. This result is consistent with the findings of Subramanian and Deaton (20) and Gibson and Rozelle (6), Tian and Yu (7), Nie and Sousa-Poza (8), and Trinh et al. (3). While the curve in Figure 3 shows that the index is increasing and has a similar trend as the non-parametric model but has a fluctuating behavior at the upper end of tail. Last, the comparison of non-parametric and semi-parametric estimation methods with the parametric method shows that the parametric fit is poor compared with the fit of the semi-parametric models, and the single-index model has a better fit than the partially linear model. Thus, the single-index model emerges to be a better specified semi-parametric model on the basis of goodness of fit. Moreover, the study revealed that fully non-parametric and semi-parametric models highlight the significant feature of the calorie–income relationship, which was not accounted for by using the parametric model.

Strength and limitations

The calorie–income elasticity was calculated using information from a household survey. Otherwise, it would not have been possible for us to obtain comprehensive nutritional data from a large sample from different locations throughout Pakistan. In addition, this study concentrated on households in which the daily caloric intake ranged from 600 to 8,000 kcal. However, because of the sizeable sample size and thorough measurement of the overall calorie intake, this study was able to generate accurate estimations and significant insights into the general nutritional condition of the Pakistani population. From the methodological point of view, this study contributes to the literature the applying the single-index model and providing a test for model specification. The data used in the study are cross-sectional for a single year, but these methods can also be used for multiple waves of data from 2010 onward to get complete insights into the calorie–income relationship. We restricted the analysis to the calorie–income relationship, but the same methodology can also be applied to explore this relationship across different food groups, region-wise as well as gender-wise.

Policy recommendations

The findings of this study suggest several significant policy changes that could be made to enhance the nutrition intake of the Pakistani population. The key concern is giving complete knowledge to eliminate nutritional gaps between average consumption and the ideal daily intake of calories in low-income households. This could be accomplished by increasing food subsidies, such as through networks of discounted grocery stores, direct nutrient supplementation plans, or in-kind transfers of food items, pricing interventions, cash transfer plans, and social safety net initiatives. Finally, an increase in money might not be enough to combat hunger; other socioeconomic and environmental issues, such as access to clean water, improved healthcare, and quality education, should also be taken into consideration. These elements might encourage better food consumption.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found here: https://www.pbs_gov_pk/content/pakistan-social-and-living-standards measurement.

Author contributions

NS is the solo author of this manuscript, and conceptualized, analyzed, and interpreted the data and all the write-up.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Babatunde R, Adejobi A, Fakayode S. Income and calorie intake among farming households in nigeria: Results of parametric and nonparametric analysis. *J Agric Sci.* (2010) 2:135–46. doi: 10.5539/jas.v2n2p135

2. Banerjee AV. Policies for a better-fed world. *Rev World Econ.* (2016) 152:3–17. doi: 10.1007/s10290-015-0236-7

3. Trinh HT, Simioni M, Thomas-Agnan C. Assessing the nonlinearity of the calorie-income relationship: an estimation strategy – with new insights on nutritional transition in Vietnam. *World Dev.* (2018) 110:192–204. doi: 10.1016/j.worlddev.2018.05.030

4. Shabnam N, Ashraf MA, Laar RA, Ashraf R. Increased household income improves nutrient consumption in pakistan: A cross-sectional study. *Front Nutr.* (2021) 8:672754. doi: 10.3389/fnut.2021.672754

5. Santeramo FG, Shabnam N. The income-elasticity of calories, macro-and micro-nutrients: What is the literature telling us? *Food Res Int.* (2015) 76:932–7. doi: 10.1016/j.foodres.2015.04.014

6. Gibson J, Rozelle S. How elastic is calorie demand? Parametric, nonparametric, and semiparametric results for papua new guina. *J Dev Stud.* (2002) 38:23–46. doi: 10.1080/00220380412331322571

7. Tian X, Yu X. Using semiparametric models to study nutrition improvement and dietary change with different indices: The case of China. *Food Policy.* (2015) 53:67–81. doi: 10.1016/j.foodpol.2015.04.006

8. Nie P, Sousa-Poza A. A fresh look at calorie-income elasticities in China. *China Agric Econ Rev.* (2016) 8:55–80. doi: 10.1108/CAER-09-2014-0095

9. Poleman TT. Quantifying the nutrition situation in developing countries. *Food Res Instit Stud.* (1981) 18:1–58.

10. Lipton M. *Poverty, undernutrition and hunger. World Bank Staff Working Papers no. 597*. Washington, D.C: World Bank (1983).

11. Strauss J, Thomas D. *The Shape of the CalorieExpenditure Curve, Center Discussion Paper, No. 595*. New Haven, CT: Yale University, Economic Growth Center (1990).

12. Ravallion M. Income effects on undernutrition. *Econ Dev Cult Change.* (1990) 38:323–37. doi: 10.1086/451812

13. Abdulai A, Aubert D. A cross-section analysis of household demand for food and Nutrients in Tanzania. *Agric Econ.* (2004) 31:67–9. doi: 10.1111/j.1574-0862.2004.tb00222.x

14. Abdulai A, Aubert D. Nonparametric and parametric analysis of calorie consumption in tanzania. *Food Policy.* (2004) 29:113–29. doi: 10.1016/j.foodpol.2004.02.002

15. Skoufias E, Di Maro V, Gonzalez-Cassio T, Sonia RR, Emmanuel S. Nutrient consumption and household income in rural mexico. *Agric Econ.* (2009) 40:657–75. doi: 10.1111/j.1574-0862.2009.00406.x

16. Skoufias E, Tiwari S, Zaman H. Crisis, food prices and the income elasticity of micronutrients: Estimates from Indonesia. *World Bank Econ Rev.* (2011) 26:415–42. doi: 10.1093/wber/lhr054

17. Wood SN. *Generalized additive models: An introduction with R.* 2nd ed. London: Chapman and Hall/CRC (2017). doi: 10.1201/9781315370279

18. Hsiao C, Li Q, Racine JS. A consistent model specification test with mixed discrete and continuous data. *J Econom.* (2007) 140:802–26.

19. Li Q, Racine JS. *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press (2007).

20. Subramanian S, Deaton A. The demand for food and calories. *J Polit Econ.* (1996) 104:133–62. doi: 10.1086/262020

21. Strauss J, Thomas D. Human resources: empirical modeling of household and familiy decisions. In: Chenery H, Srinivasan TN, Behrman JR editors. *Handbook of Develoment Economics*. Amsterdam: Elsevier (1995). p. 1883–2023. doi: 10.1016/S1573-4471(05)80006-3

22. Robinson PM. Root-N consistent semiparametric regression. *Econometrica.* (1988) 56:931–54. doi: 10.2307/1912705

23. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *J Econom.* (1993) 58:71–120. doi: 10.1016/0304-4076(93)90114-K

24. Manski CF. Identification of binary response models. *J Am Stat Assoc.* (1988) 83:729–38. doi: 10.1080/01621459.1988.10478655

25. Horowitz JL. *Semiparametric methods in econometrics*. New York, NY: Springer-Verlag (1998). doi: 10.1007/978-1-4612-0621-7

26. Fan J, Gijbels I. *Local polynomial modelling and its application*. London: Chapman and Hall (1996).

27. Zheng JX. A consistent test of functional form via nonparametric estimation techniques. *J Econom.* (1996) 75:263–89.

28. Li Q, Wang S. A simple consistent bootstrap test for a parametric regression function. *J Econom.* (1998) 87:145–65.

29. Hsiao C, Li Q, Racine JS. A consistent model specification test with mixed discrete and continuous data. *J Econom.* (2007) 140:802–26. doi: 10.1016/j.jeconom.2006.07.015

30. Government of Pakistan. *Household Integerated Economic Survey*. Islamabad: Pakistan Bureau of Statistics (2010).

31. UNICEF. *Food Composition Table for Pakistan (Revised 2001)*. Islamabad: UNICEF (2001).

32. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc.* (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x

33. Roy N. A semiparametric Analysis of Calorie Response to Income change across income groups and gender. *J Int Trade Econ Dev.* (2001) 10:93–109. doi: 10.1080/09638190010015287

34. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. London: Chapman and Hall (1993). doi: 10.1007/978-1-4899-4541-9

35. Bhalotra SR, Attfield C. *Intrahousehold resource allocation in rural pakistan: A semiparametric analysis*. London: The Suntory Centre, London School of Economics and Political Science (1998). 11 p. doi: 10.1002/(SICI)1099-1255(1998090)13:5<463::AID-JAE510>3.0.CO;2-3

36. Horowitz JL, Lee S. Semiparametric methods in applied econometrics: do the models fit the data. *Stat Model Issue.* (2002) 2:3–22. doi: 10.1191/1471082x02st024oa

frontiers | Frontiers in Nutrition

TYPE Review
PUBLISHED 05 January 2023
DOI 10.3389/fnut.2022.1052934

Check for updates

OPEN ACCESS

EDITED BY
Massimo Lucarini,
Council for Agricultural Research and Economics, Italy

REVIEWED BY
Ali Ben Mrad,
University of Sfax, Tunisia
Barbara Burlingame,
Massey University, New Zealand
Susanne Westenbrink,
National Institute for Public Health and the Environment, Netherlands
Sian Astley,
European Food Information Resource (EuroFIR AISBL), Belgium

*CORRESPONDENCE
Henrique Ferraz de Arruda
✉ h.f.arruda@gmail.com

SPECIALTY SECTION
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

RECEIVED 24 September 2022
ACCEPTED 07 December 2022
PUBLISHED 05 January 2023

CITATION
Ferraz de Arruda H, Aleta A and Moreno Y (2023) Food composition databases in the era of Big Data: Vegetable oils as a case study. *Front. Nutr.* 9:1052934. doi: 10.3389/fnut.2022.1052934

Food composition databases in the era of Big Data: Vegetable oils as a case study

Henrique Ferraz de Arruda[1,2]*, Alberto Aleta[1,3,4] and Yamir Moreno[1,2,3,4]

[1]ISI Foundation, Turin, Italy, [2]CENTAI Institute, Turin, Italy, [3]Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain, [4]Department of Theoretical Physics, Faculty of Sciences, University of Zaragoza, Zaragoza, Spain

Understanding the population's dietary patterns and their impacts on health requires many different sources of information. The development of reliable food composition databases is a key step in this pursuit. With them, nutrition and health care professionals can provide better public health advice and guide society toward achieving a better and healthier life. Unfortunately, these databases are full of caveats. Focusing on the specific case of vegetable oils, we analyzed the possible obsolescence of the information and the differences or inconsistencies among databases. We show that in many cases, the information is limited, incompletely documented, old or unreliable. More importantly, despite the many efforts carried out in the last decades, there is still much work to be done. As such, institutions should develop long-standing programs that can ensure the quality of the information on what we eat in the long term. In the face of climate change and complex societal challenges in an interconnected world, the full diversity of the food system needs to be recognized and more efforts should be put toward achieving a data-driven food system.

KEYWORDS

food composition database, food composition, food nutrient, vegetable oils, healthy nutrition, food guidelines

1. Introduction

In their seminal book published in 1940, McCance and Widdowson started by stating that: *"The nutritional and dietetic treatment of disease, as well as research into problems of human nutrition, demands an exact knowledge of the chemical composition of food"* (1). In the years to follow, researchers from all over the world—together with private companies and non-profit organizations—delved into the basic constituents of food in an effort to understand what we eat and how it affects us. However, despite the many advances, this task may be far from being complete.

The information on the composition of food items used to be compiled in Food Composition Tables (FCTs), although nowadays, many countries have updated them into Food Composition Data Bases/Data Banks (FCDBs). For instance, the EuroFIR project hosts the databases from 30 different countries. This information may come from four different sources: primary publications, secondary publications, unpublished reports, and analytical reports (2). In the first, the composition is extracted directly from journal articles. In the second, food information is compiled from other sources, which include reviews, books, reports, or other FCDBs/FCTs. The third source includes reports that are not publicly available, which include internal use reports. Lastly, analytical reports can be divided into two subcategories, namely specific and non-specific. The specific analytical reports are created to feed a particular FCDB/FCT, and the non-specific analytical reports contain data not obtained for this aim. The diversity of sources, analytical methods, variations of the same food, terminologies, and economic cost makes the procedure of collecting and integrating all this information a huge challenge.

The problem with these data is further exacerbated by the different criteria used in each country to create them, which can be partially explained by the diverse object pursued by each project. At least since 1982, there have been several initatives to harmonize procedures for better data comparability and interchange, such as INFOODS from FAO or EUROFOODS from COST (3). Even so, studies from the late 90s showed that mean intakes of individual nutrients for the same diet could vary up to 20–45% depending on the database used to estimate food composition (4). These differences were associated with systematic and random errors, variations in naming, terminology, or calculation procedures across databases, and to the intrinsic differences of food items in different countries. Many initiatives aimed at solving these problems either failed or succeeded only for a brief period but then aged badly due to lack of funding or a driving force (5).

In 1989, the United Nations published the INFOODS data interchange handbook with guidelines to improve data on the nutrient composition of foods, as they observed that such "data do not exist or are incomplete, incompatible, and inaccessible" (6, 7). In the European context, a clear indicator that this problem still needs to be solved is the number of initiatives that periodically appear to address these issues. From 1985 the NORFOODS group practiced data interchange among Nordic countries. Even though pioneering, it had some limitations. For instance, the data exchange was restricted to the data contained in the databases, but not the metadata (8). From 1995 to 1999, the EUROFOODS project created a working group to address the issues of food composition data management and interchange. The project led to a set of recommendations to make national databases compatible and facilitate data interchange (9). Independently, in 1990 the European Prospective Investigation into Cancer and Nutrition (EPIC) started. Its objective was to investigate the relationships between diet, nutritional status, lifestyle, and environmental factors and the incidence of cancer and other chronic diseases. Between 1992 and 1999, their participants collected data on the food intake of a large prospective cohort of over half a million individuals. The next step was to use FCDBs to estimate the nutritional intake. By that time, it was already known that FCDBs could be a significant source of imprecisions in this estimation (10). For this reason, in 1999, researchers reviewed the FCDBs available in the participant countries. They determined that: (i) the FCTs of different countries were not comparable due to the lack of reporting standards; (ii) comparisons within tables were problematic due to the use of very different sources; (iii) some tables were compiled with outdated information; (iv) there were inconsistent values of several nutrients across tables. As such, they established the necessity of creating standardized food composition tables for the countries involved in EPIC (11). This led to the creation of the Epic Nutrient Database (ENDB) in 2007, which had information for 26 components of 550–1,500 foods in 10 countries (12).

Concurrently, from 2005 to 2009, the European Commission sponsored the EuroFIR project, whose aim was to develop a pan-European system on food information (13). In 2010, the project was extended for 2 years with funding from the European Community's Seventh Framework Program under the name EuroFIR Nexus. It later evolved into a financially independent Association based in Brussels with the mission of maintaining high quality, validated national food composition data. An analysis carried out in 2021 showed that the documentation of the 26 European datasets included in EuroFIR was successfully standardized. Yet, full comparability of the datasets was not guaranteed as there were still many differences and inconsistencies. For instance, 15 out of the 26 datasets reported energy values calculated by factored summation with up to five different methods, while the others reported that the method was unknown or measured through analytical procedures (14).

In 2009, the European Food Safety Authority launched a pan-European food consumption survey - the EU Menu project (15). To create the methodological guidelines, it was necessary to obtain certain information on the nutritional composition of food. To do so, they funded the "Updated food composition database for nutrient intake project." This project, finished in 2013, compiled information from 14 national food composition databases for almost 1750 food products. National datasets that did not contain all the information borrowed it from the datasets of other countries that, in the opinion of each compiler, consumed similar types of food. The percentage of values borrowed for each dataset ranged from 5 to 90%. Even extensive datasets borrowed 40% of the values as they were required to provide data for all elements included in the EFSA food list. It was presumed that most of the borrowed values belonged to seldom consumed foods in the country, and thus this should not have an important impact on the final intake estimation (16).

In 2018, the European Commission funded the Stance4Health project, in which one of the main tasks was to build a nutritional database to complete the national FCDBs from the countries involved in the project (Germany, Spain, and Greece) with as many foods and nutrients as possible. To add information for nutrients and biocompounds not present in those databases, they complemented the information with the FCDBs of Italy, the Netherlands and the United Kingdom, together with the FAO, USDA, and Phenol-Explorer databases. This way, the total contribution to the unified database of each of the three original sources of data was 15 9 and 2% for Germany, Spain, and Greece. Excluding polyphenols and focusing only on the 40 nutrients most commonly used in epidemiology, Spain and Germany had information about 88% of the nutrients and Greece 40% (17).

It is thus clear that, despite the huge advances produced in the last 40 years, the creation of a unified, reliable, and comprehensive database on food composition at the European level is far from complete. Globally, the situation is not better. Even though it is crucial to have regional and country-specific FCDBs, many countries do not have a national FCDB, and many of the ones that do are outdated, do not follow international standards in terms of quality, coverage, accessibility, and documentation. Furthermore, most of them, including European ones, contain $<$ 25% of analytical data, and these data are usually old and not generated specifically for the FCDB (18). The situation is particularly challenging in Africa, where micronutrient deficiencies are one of the major public health challenges, and the data gaps on what people are eating (and their contents) make it very difficult to devise optimal strategies to improve diets and malnutrition (19).

The main contribution of this paper is to depict and exemplify the main limitations of the FCDBs. In particular, we quantify the differences in terms of the number of nutrients and compounds found in these databases. Further, we explore the age of the information contained in them, which is a problem often overlooked. Next, we examine the differences in terms of the compositions of various vegetable oils in a given FCDB and the discrepancies between FCDBs for the same oil. Although differences might be expected, we highlight inconsistencies that can be found even within the same database. Finally, we discuss how the development of current machine learning methods could help solve the problems identified in FCDBs.

2. The problems of current FCDBs

Currently available datasets have important limitations, including incomplete and outdated data, and not enough documentation on their sources and assumptions (18, 20). The nutrient content of food changes over time as a function of very complex processes involving many different aspects, from agricultural practices to policy and consumer pressure, and this is seldom reflected in them (21). Besides, it is common practice to complete missing elements with values obtained from other FCDBs or from the general literature. This might be reasonable for food items produced in few countries and traded globally, but it is problematic for locally produced food (22–24).

These errors may then propagate to other datasets and studies that might not be aware of those borrowings. Furthermore, research data, even if they are of sound analytical quality, may be biased in the selection of foods. Many FCDBs also do not include fortified foods, branded foods or a proper representation of the biodiversity of the food chain, which may lead to systematic errors in intake estimation (18). Besides the problems directly related to human nutrition and epidemiology, data gaps also have significant consequences for other surrounding areas, such as the sustainability of food systems and the pursuit of Sustainable Development Goals (SDGs).

In the following, we describe some of the most important problems of FCDBs, while in Section 3 we provide the analysis for the particular case of vegetable oils:

2.1. Missing values

While many countries possess their own FCDBs, the majority of them contain outdated and incomplete information. For instance, in 2021, it was determined that in the Dutch database (NEVO) about 50% of the items were missing information on the amount of vitamin K, hindering the assessment of the portion of the population with an adequate intake (25). This could be explained by the relatively recent discovery of the precise function of vitamin K in the 1970s (26), but at the same time shows the complexity of interpreting missing values. In fact, it is not always clear if there is a distinction between missing data and a value of zero for certain nutrients (17, 25). Thus, it is not possible to be sure if the compound is present or not in the food item, which may lead to important underestimations of nutritional intake.

2.2. Lack of sources

The lack of proper documentation of these databases may also hide important issues. Despite recent efforts, many databases are still missing their source of information, or the references might be incomplete (27). In the ones that report their sources, one can see that the information is extracted from the literature without taking into account important regional differences. To exemplify the problem, a literature review constrained to food produced and marketed in Brazil showed that the reported values were compatible with the ones contained in their national database for 81% of the products, decaying to 37.5% when comparing it with the database from the United States (USDA) (28).

Similarly, a comparison between the ENDB database (used in the European project EPIC) and the USDA found a strong agreement for macronutrients, but a weak agreement for starch, vitamin D and E, and thiamine-14% of the 28 compounds common in both datasets (24). This could be read as a sign of the small differences between the food composition of these two regions, at least in terms of macronutrients but, as we will see in Section 3.2 for the particular case of vegetable oils, many national databases extract their information from a common source, rather than by direct analysis. As such, when two databases report similar values, unless their source is stated, it is not possible to determine if that is because the food is similar in both regions or if they simply share the source.

2.3. Food fortification

In 2021, the United Kingdom joined the group of over 80 countries in which folic acid fortification of staples is mandatory in an effort to reduce the risk of neural tube defects in babies (29). Food fortification is becoming especially important in low- to middle-income countries, where micronutrient deficiencies are a widespread problem (30). In 2022, a study carried out in the Netherlands showed that up to 75% of the population consumed voluntarily fortified foods, resulting in a 64% higher intake of habitual micronutrients compared to non-users (31). Importantly, the study used the values reported on the labels of the products as an indicator of their composition. However, a Dutch study from 2017 showed that the vitamin D of some selected products ranged from 50 to 153% compared to the declared values (32). Similar results have been reported in the US (33), which could be related to the overages of vitamins added by producers to account for shelf life. Thus, using the reported values might produce under- or over-estimation of micronutrient intake in the population.

If, instead, one uses the information contained in FCDBs the problem might be even worse. The lack of information on the source, the outdated values and the important regional differences in terms of fortification policies may severely impact any estimations. For instance, Nordic countries have mild iodine deficiency and their fortification practices vary. A study from 2016 carried out by the NORFOODS project found out that the national animal feeding practices could produce two-fold differences in the iodine content of milk and eggs (34). As such, even for neighboring countries, the use of borrowed values for certain products might severely impact the estimations of nutrient intake and, hence, give rise to misguided policies. Thus, if it is necessary, the insertion of borrowed data in FCDBs shall be done only by experts adequately trained to understand the local nuances of the food and region.

In terms of coverage, even in countries with mandatory food fortification programs, data is not routinely collected (35, 36). For voluntarily fortified foods, given that they depend mainly on their producer, more information can be obtained in databases of branded foods. However, even though that, in Europe, declaring nutrients added for fortification purposes is mandatory, the information is not always clear. For instance, the authors of the Dutch branded food database (LEDA) could not determine the coverage of data on fortified nutrients due to unclear food name, ingredient descriptions and missing nutrient values (37).

2.4. Nutritional dark matter

Borrowing the term from genetics, nutritional "dark matter" refers to all those dietary factors that can influence our health but that remain largely invisible (38). For instance, it was recently shown that microRNAs present in plant foods can influence the genetic expression of enteric bacteria (39). There are thousands of biochemical compounds present in our food, but FCDBs were built to study only the nutrients that are essential for life. Due to the lack of data, nutritional epidemiology has focused on these few dozens of nutrients, disregarding elements such as amino acids and biogenic amines (40). While the USDA reports information on about 150 nutritional components present in food, FooDB, a large database on the chemical composition of food, contains more than 70,000 distinct biochemical compounds as of June 2022 (41). Yet, only 5% of them have been quantified. All this chemical diversity that remains invisible in common epidemiology may have an important effect on our health (42). Numerous initiatives are trying to compile this information from validated peer-review sources, such as Phenol-Explorer (43), but the current lack of harmonization introduces important challenges (44).

2.5. Branded foods

National FCDBs usually only document generic, non-branded foods. There are commercial databases that may provide this information, but they tend to be expensive and contain only details on macronutrients. In the European Union, pre-packaged foods must display their amount of some selected nutrients, but it is hard to validate their accuracy (45). Reformulation of processed foods is frequent as manufacturers try to keep their market share, increase their profits, make the food healthier, or are even forced to change due to government policies or consumer pressure. A study on the pizzas offered on the website of six supermarkets in the United Kingdom showed that, out of 903 pizzas, 10.8% of them changed their composition over 6 months and that 29.9% of them were either discontinued or new market entries (46). This information is hardly captured in most studies and, if it is, it might be restricted to the few nutrients reported in the labels of the product. Furthermore, many companies may rely on national FCDBs to estimate the nutritional value of their products rather than using direct measurements. If the limitations of the data are not clear, errors may propagate throughout the whole chain (21).

2.6. Outdated and misdated information

Even for raw products, the nutritional composition changes over time as a consequence of genetic selection, changes in agricultural practices or feed ingredients for farmed animals (47, 48). If FCDBs are not routinely updated, they may easily become obsolete (49). And, if they are, they should properly document all the changes so that one is aware if they were produced because the composition of food has changed or due to the improvement of the analytical techniques. Otherwise, for research studies over extended periods, variations in nutrient intake may reflect changes in the data rather than in the dietary patterns of the population (50). Similarly, dietary surveys must be analyzed with FCDBs compiled in the same period, or one risks finding spurious patterns due to the expected composition changes. Many institutions invest significant efforts in keeping the information updated, but this is not homogeneous. For instance, as we will see in Section 3.2, while the Spanish database has not been updated since 2010, the Danish database updated its information on coconut oil in 2022. More broadly, in a survey performed in 2019, researchers found that only 30 out of 107 available FCDBs had been updated in the previous 5 years (51).

2.7. Biodiversity

The differences in nutrient composition among varieties of the same product can be as important as between different species. For instance, an orange-fleshed banana from Micronesia can have 50 times more vitamin A than the common white-fleshed bananas, representing the border between nutrient deficiencies and nutrient adequacy (52). This biodiversity is seldom acknowledged, and general FCDBs usually report the information of a single sample or a naive average over different varieties of the same product. Over 15 years ago, FAO recognized the importance of biodiversity in nutrition and launched an initiative to create a database on biodiversity which could mitigate this lack of information (53, 54). Yet, despite the great advances produced by this initiative, and the relatively large size of the database, many common food items are not well-characterized yet (55). For instance, in the latest version of the food composition database for biodiversity, published in 2017, there is no information about olives, coconuts, palm or soybeans (56).

2.8. Climate change

Even though it is still early, research so far depicts a very complex picture in which some crops might benefit from higher temperatures—thanks to warmer temperatures—while others, specially those that require vernalization, will suffer (57). At the same time, faster growth might result in lower quality products both in terms of external appearance and internal composition (58). Changes in CO_2 concentration may also have an impact on nutritional composition (59, 60). Furthermore, besides the changes directly produced by climate change, it may also be necessary to select and adapt crops to the new environmental conditions (51). Maintaining updated FCDBs will be a key element in devising the sustainable food supply of the future. And, at the same time, FCDBs can be a great resource for monitoring biodiversity and climate impacts in food systems.

3. Composition of vegetable oils in selected FCDBs

Although everyday there are more FCDBs in electronic format, accessing certain FCDBs can be complicated since some of them are not free, others are in analogical formats, and they may even lack an English translation (61). For those in digital format, in comparing several foods and nutrients, it can also be challenging to query automatically, and one needs to resort to manual exploration. Since this is a quite demanding process, and given their importance in the total caloric intake of the population, for this analysis, we focus on the particular case of vegetable oils as a case study.

To provide an overview of the current state of FCDBs, we have selected six databases covering several regions of the world:

- BEDCA: Is the Spanish FCDB, developed in 2010 as part of the EuroFIR project and has not been updated since then (62). It was compiled using the indirect method, that is, collecting all the information from different sources. Thus, it may not reflect the regional variability of certain products. It reports the source of information, but many references are empty. It does not contain information on fortified foods, which may impact the estimation of micronutrient intakes (63). Due to its weaknesses, Spanish commercial nutritional programs use a variety of other FCDBs (20). A study from San Mauro Martín and Hernández Rodríguez (64) studied the nutritional composition of the same diet estimated using different Spanish commercial nutritional programs. They showed that the estimated intake for each nutrient was highly heterogeneous, with differences in the range of 8–84% depending on the program.

- FRIDA: Is the Danish FCDB maintained by the National Food Institute (Technical University of Denmark) (65). It is easily accessible, updated frequently, and well-documented. It is composed by a mixture of direct analysis, information provided by several danish stakeholders and indirect information extracted from the scientific literature (66).
- USDA (Food Data Central): Is the FCDB from the United States Department of Agriculture (67). It is composed of five different databases, of which the Foundation Foods is the newest and most advanced. Until 2018 the main database was the SR Legacy, and it has been regarded for many years as a gold standard in the field, up to the point that many FCDBs in the world extracted their information from it. It is composed of data obtained from direct analysis, calculations as well as extracted from published literature.
- TBCA: Is the Brazilian FCDB (68). It is easily accessible and well-documented, although it lacks English translation. It contains an extensive selection of local products and their biodiversity, including many varieties for the same product. These products are mostly directly analyzed in Brazilian institutions, while for common foods in the world it comes from international databases such as the USDA.
- NIGERIA: The FCDB from Nigeria is small in terms of products but has an extensive selection of the most commonly used in the country (69). The documentation is scarce, although it reports the source of information for each product. Yet, they are not linked to each individual nutrient, and thus where each value comes from is unclear. The information is mostly extracted from published literature from Nigerian institutes—specially for local products—but also contains information from global sources such as the USDA.
- SMILING: The SMILING project aims at reducing micronutrient deficiency among children and women in South East Asia (Indonesia, Thailand, Cambodia, and Vietnam) (70). To create optimal diets for those countries, the first step was to compile regional FCDBs with information about the most commonly consumed products in the area, which they did in 2018. Due to the limited resources, they had to resort to indirect compilation. For several micronutrients they were not able to obtain local information and had to use international FCDBs. Besides, they realized that some of the sources they used were quite old and might have copied their values from non-regional sources. Thus, they claim, there is an urgent need to produce high quality data for local foods in the region (71). Note that many of these limitations are also present to some extent in databases of highly developed countries.

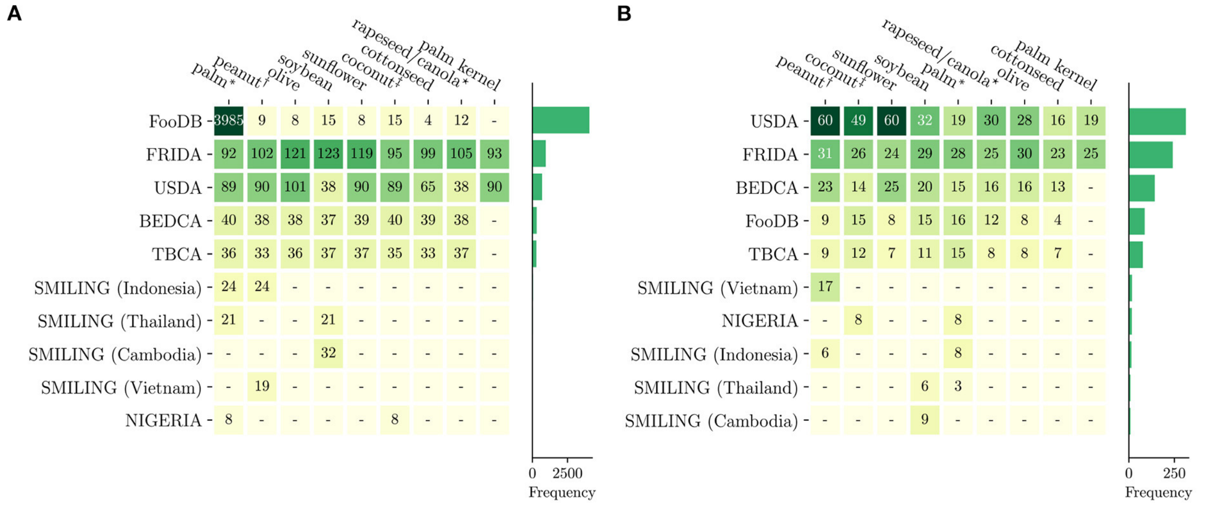
It must be noted that these databases were created through very different projects and budgets. For instance, BEDCA is the result of a project to build the first Spanish database using the EuroFIR standards. The project started in 2004 and finished in 2010, and thus it has not been updated ever since (20). In contrast, Food Data Central is a platform hosted by the USDA, a federal agency from the United States that has been analyzing foods and conducting human nutrition research for over 100 years (72).

We also complement the previous set with two other databases:

- FooDB: It is an online database that aims to be the largest resource on food constituents. It is easily accessible and well-documented, and reports thousands of chemical compounds with each food item. Unfortunately, most of them are not quantified, so the actual amount of reported nutrients is similar to national FCDBs. The information is extracted from other FCDBs as well as from public databases on phenols or pathways. Besides, the main source of information on nutrients are the USDA and FRIDA databases, and thus lacks information on regional biodiversity (41).
- EuroFIR: As previously described, the EuroFIR database is the result of the original EuroFIR project, which intended to create a homogeneous database for Europe. In contrast to the other databases, to access EuroFIR data, it is necessary to purchase a membership, which was imposed to assure the long-term sustainability of the initiative (73). The EuroFIR guidelines are one of the standards used in the field, and thus the scheme of the database is detailed and well-documented. However, since the information populating the database is provided by third parties, its quality varies greatly (14).

3.1. Data description

For this analysis, we focus on the major vegetable oils in terms of world supply and distribution: coconut, cottonseed, olive, palm, palm kernel, peanut, rapeseed/canola, soybean, and sunflower oils (74). Figure 1 illustrates the amount of information on these vegetable oils contained in the selected databases. In Figure 1A we report the total number of compounds present in each database. According to this, FRIDA is the database with the largest amount of information, superseding the USDA database, except for palm oil, for which FooDB provides an enormous amount of compounds. In terms of the overall coverage of each oil, palm oil is the most studied one, followed by peanut oil and olive oil. However, many of the entries in these databases are 0 (the distinction between measured 0 and logical 0 is seldom made). If those nutrients are removed, the depicted scenario changes completely Figure 1B.



* Since there is no Palm oil in FooDB, we consider the entry "Oil palm". In NIGERIA, the data refers to fresh palm oil;
† In SMILING (Vietnam) and NIGERIA, peanut oil consists of fried and fresh, respectively;
‡ Coconut oil is fresh in NIGERIA;
⋆ Canola/rapeseed oils are classified as different varieties for each dataset. The varieties considered in FRIDA, USDA, BEDCA, and TBCA are "rape seed (no eruca acid)", "canola", "rape oil", and "canola, Arachis hypogaea L.", respectively. In FooDB, canola and rapeseed can be found. We choose canola because it contains more compounds than rapeseed.

FIGURE 1
Number of compounds and nutrients reported for each vegetable oil in the selected databases (EuroFIR not included). **(A)** The total number of nutrients for which the database provides some information, while **(B)** contains only those nutrients with a quantity larger than 0. In both panels, the bar plots represent the sum of the values of the same line, and the rows and columns of the matrix are ordered in decreasing order from left to right and from top to bottom.

Indeed, if we focus only on those nutrients with a reported presence larger than 0, the USDA turns out to be the database with the largest amount of information. Besides, the most studied oil is peanut oil, followed by coconut and sunflower oils, while palm and olive oil move to the 5th and 7th positions, respectively. In FRIDA, the number of nutrients with a quantity larger than 0 is one-third of the total amount of nutrients studied, contrasting with the USDA database in which only half of the nutrients are quantified as 0. This depicts a very different scenario in terms of micronutrients present in vegetable oils depending on the database analyzed.

The rest of the databases contain much less information than the first two. The smallest ones (NIGERIA and SMILING) focus specifically on regional foods, and thus it is expected that these datasets do not report much information on vegetable oils that are not common in these countries. It is also interesting to note that while FooDB contains information about thousands of chemical compounds, the quality of said information is relatively low since the actual amount of quantified compounds per vegetable oil is even lower than in the FCDBs that it uses as source. The low quality of the metadata—if present—is also a major problem, as it is usually impossible to know the analytical procedure used, the cultivar, variety or simply the species of the element.

Lastly, we must note that we have not included EuroFIR due to the heterogeneity of its data. Currently, the database has a set of guidelines that contributors have to follow when uploading information to the system, but that does not guarantee that they will follow them, nor that the original information has sufficient quality (27). For instance, the number of countries reporting information is quite variable: 36 for olive oil; 30 for sunflower oil; 22 for palm oil; 22 for coconut oil; 21 for peanut oil; 13 for soybean oil; and 9 for cottonseed oil (note that EuroFIR now includes some non-European countries). Regarding the quality of the documentation, even though EuroFIR requires information on the analytical method used to measure the composition, in most cases, it is reported as "unknown." Similarly, it is mandatory to provide the source of the data, but in many cases, it is either not reported or not well-described (e.g., "No change from USDA"). Even though the platform is a huge step forward in the right direction, there are still many values that are not fully comparable (14). Solving these issues is beyond the scope of our paper, and thus we have not included it in the subsequent analysis.

3.2. Qualitative comparison

Next, we look at the age of the information to evaluate the validity of the data. From a broader perspective, the problem of outdated information can be related to the issue of data obsolescence. Obsolescence refers to the appearance of a new piece of information that supersedes an existing one that is still available (75). Some authors propose the use of machine-learning techniques to detect when data becomes obsolete or contradicts previous knowledge (76). In our context, as previously discussed, the composition of food is continuously changing as a consequence of both natural and human interventions. Besides, analytical techniques keep improving, giving more detailed and precise estimations. As such, it is important to both keep the databases updated and, at the same time, store the old information so that dietary studies carried out in the past can use the proper composition.

From the databases analyzed, only USDA, BEDCA, FRIDA, and TBCA provide detailed information on the year when the content was measured. In the subsequent analysis, for USDA, we considered only SR Legacy, when available, so as to be able to analyze the dates of all compounds separately. Yet, note that when the data are extracted from a scientific publication, the date that is associated is the one when it was published, not when the product was actually analyzed. Thus, unless it comes from direct estimation, any value might have been measured at the depicted date or before. In fact, many compounds share the same date, but that is because they were extracted from a compilation or a database published in that year and not because they were measured in that year.

Figure 2 shows the number of compounds classified by the decade corresponding to the listed year in their source. As we can see, the information tends to be decades old, questioning its validity. The selected USDA database does not contain any information collected after 2010, except for coconut oil which was substantially updated in 2015. Similarly, all the information contained in BEDCA comes from the decade of 2000. A closer inspection reveals that most information comes from either a book published in 2004 or from the USDA database that was available back then. Yet, as we can see, even though they used the version that was available at the time, the information contained there could already be decades old. For the FRIDA database, we observe that most of the information comes from three different dates separated by a decade, signaling that the speed of the updates is relatively low. Lastly, TBCA is the most updated one, which is to be expected since it started in the decade of 2010 and most of the information comes from direct analyses.

Yet, a closer inspection reveals more weaknesses. For instance, focusing on the case of palm oil, we observe that in the USDA database most of the information comes either from 1979 or from the early 2000s, with the last update in 2009 (folate). It is important to note that the values which are assumed to be 0 are usually not updated, explaining why there are so many compounds that have not been updated since 1979 in Figure 2A. Note also that, as previously discussed, a value of 0 might mean different things in each database: below detection limits, not analyzed, assumed to be 0, etc. In Figure 2B, when we remove those elements whose concentration is reported as 0, we observe an important reduction of compounds with information from that period. A similar result was found for FRIDA, with the majority of the data also included in two dates. In the case of TBCA, even though the source is supposed to be recent, there are several compounds whose information was extracted from the USDA in 2017. Given that the USDA has not updated the information on palm oil since 2009, the information contained in TBCA is actually a decade older than reported. All in all, if we consider USDA, FRIDA and BEDCA, 57.9% of the information was collected before 1990 and the remaining before 2010.

3.3. Quantitative comparison

In Figure 3 we show the fatty acid composition of the oils contained in the USDA. Specifically, canola, coconut, peanut, soybean, and sunflower oils are from Foundation foods, palm, palm kernel, and olive oils are from SR Legacy, and cottonseed oil is from Survey Foods (FNDDS). As expected, each vegetable oil has a very different composition, which highlights why it is so important to have precise information about as many foods as possible. If a product is substituted simply with one that seems similar, one may incur in important errors when estimating the actual intake. Having extensive documentation is also very important to understand the information. For instance, common sunflower oil usually has a concentration of 20% monounsaturated fatty acids, which contrasts with the 60% reported in the USDA. A closer inspection of this database shows that the value is the average of eight samples, two of which have about 20% of MUFA and six with around 75–80%. In other words, they are averaging the composition of two samples of common sunflower oil and 6 samples of high oleic sunflower oil. If the information on individual samples is not available, it is impossible to understand the origin of certain discrepancies, and any estimation done with these values might be biased.

Following the previous example of palm oil, we now look at the composition in terms of fatty acids in the databases explored in the previous section (Figure 4). It is worth noting that for BEDCA the sum of all fatty acids is 100.94 g per 100 g, a common inconsistency found in FCDBs that extract their information from a combination of scientific publications. Furthermore, the only components that are not 0 are fatty acids and alpha-tocopherol. In contrast, the sum of all fatty acids in the USDA and Frida databases is 95.4 g per 100 g, and they also report the presence of vitamin K.

At first glance, all databases share similar values. However, upon closer inspection, one can see that the amount of polyunsaturated fatty acids reported by TBCA is 78% higher than the one in USDA and FRIDA. As previously discussed, there are many factors that can alter the composition of a product. Regarding the source material, different species will have different nutrient contents, and even within the

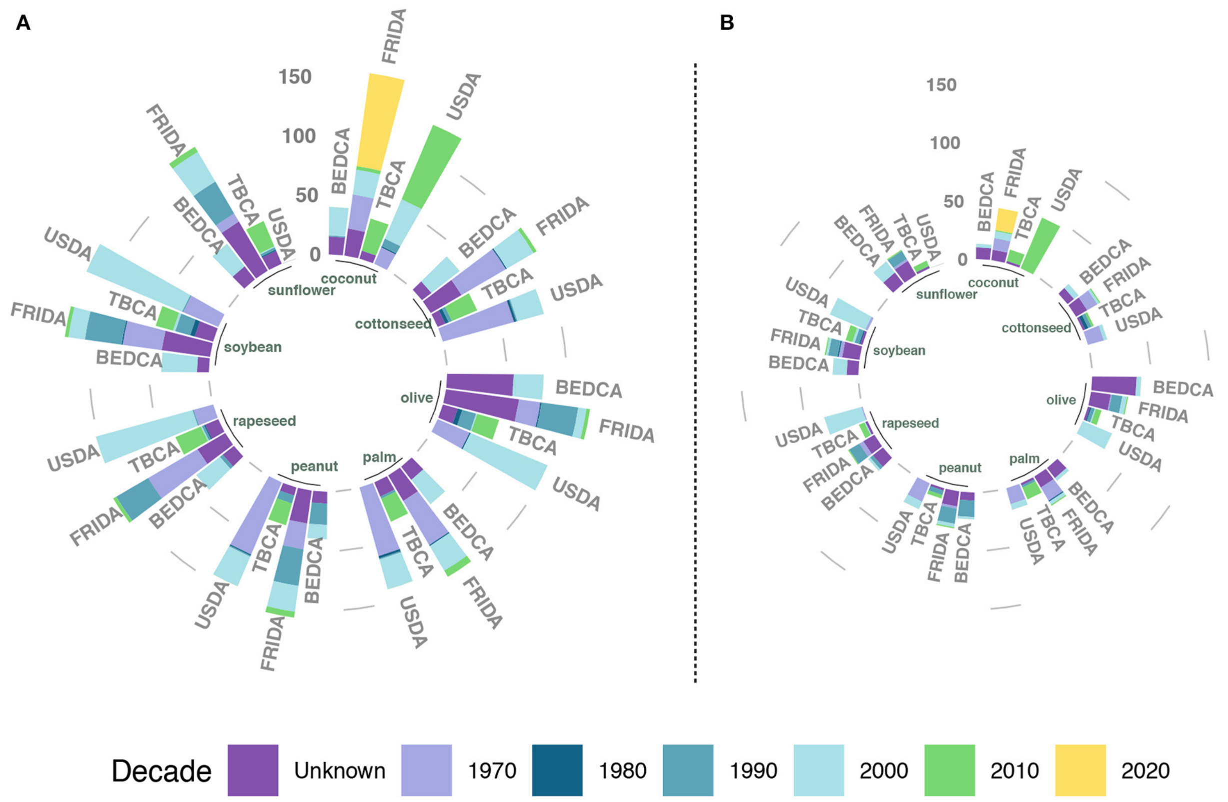


FIGURE 2
Date associated with each compound or nutrient in the selected databases that provide said information. For each vegetable oil, the length of the bar indicates the number of compounds, while the color represents the amount of them that have an associated date within a decade. Palm kernel is not included as it is only reported in FRIDA and USDA. **(A)** The information for all elements present in the database, regardless of their actual value, while **(B)** shows the information only for those with a reported quantity larger than 0.

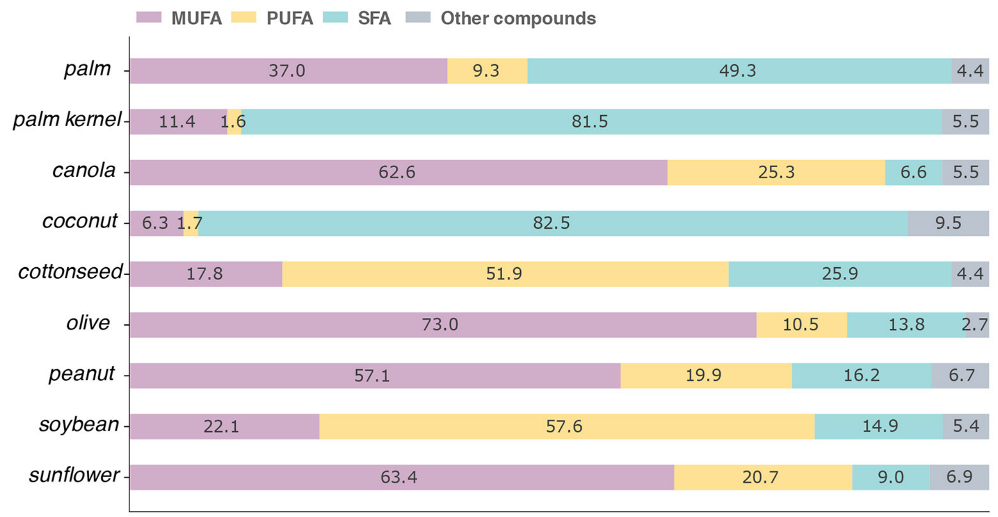


FIGURE 3
Comparison of fatty acids among different oils in the USDA database. The values are shown as the percentage out of 100 g of the product. MUFA, PUFA, and SFA represent monounsaturated, polyunsaturated, and saturated fatty acids, respectively. Other compounds denote elements that were not described in the database and that would be necessary to reach 100%.

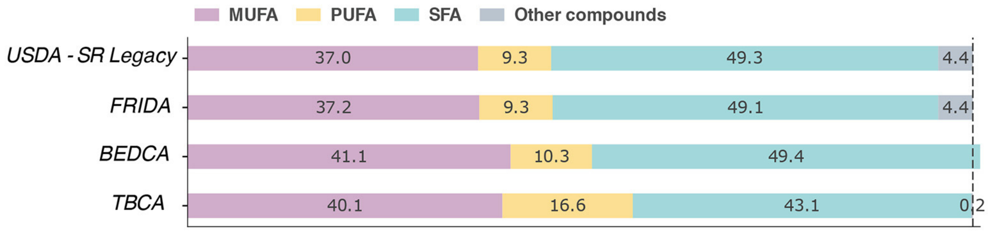


FIGURE 4
Comparison of fatty acids in palm oil among different FCDBs. The values are shown as the percentage out of 100 g of the product. MUFA, PUFA, and SFA represent monounsaturated, polyunsaturated, and saturated fatty acids, respectively. Other compounds denote elements that were not described in the database and that would be necessary to reach 100%.

same species the composition may change substantially from cultivar to cultivar (2). Besides, the particular season when it was harvested, or the production processes can also alter the composition. Thus, the main concern is not that the values are different but that there is no information in the databases that allows one to determine what could be their cause. One solution to this problem would be to include additional metadata with information on species, variety, cultivar, etc.

Another example is the reported concentration of palmitic acid. A study from 1973 showed that samples from Zaire, Indonesia, and Malaysia contained, on average, 42, 48.6, and 49.2 g, respectively (77). In contrast, in the FCDBs considered the concentrations are much closer to one another: 43.50 g for USDA; 43.68 g for FRIDA; and 43.04 g for BEDCA (TBCA does not report the quantity of this fatty acid). This may be caused by the importation of palm oil from the same area, which would explain the similarities. However, it signals that the values might be ill-suited for countries that may not obtain it from the same source. Lastly, even though it was not included directly in the analysis, if we look at the value reported by FooDB we get that the median concentration is 25.8 g, wildly different than in any other database. Fortunately, it is possible to download the raw information, which reveals that the website is averaging the values of both palm oil and palm kernel oil, even though the latter has a completely different composition.

4. The challenges for a Big Data approach

As we have seen, FCDBs collect a lot of information from scientific publications, and they may lose very valuable information in the process. Besides, they also tend to neglect biodiversity and the temporal and spatial dimensions of food composition, weakening the conclusions that can be reached using that data. One possibility to update and enrich the quality of FCDBs would be to systematically review the literature and extract as much information as possible, which can then be studied using Big Data techniques—a task full of challenges.

Continuing with the example of palm oil, we can estimate the number of scientific records that are relevant for this purpose using the information of scientific records from Microsoft Academic Graph (MAG) (78). In particular, we used the version that contains publications up to 25th of June 2020, provided by the CADRE project from Indiana University (79). Considering the abstracts and titles, we recovered all entries with the words "palm" or "elaeis" and "oil." As a result, we obtained 79,210 documents. Taking this information, we created a network of citations between these documents. Specifically, each document (e.g., a paper or a book) represents a node in the network, and two documents are linked if they reference each other. This allows us to classify documents according to their content since papers that belong to the same subfield tend to cite each other more. After removing the nodes that are disconnected from the rest of the network (they have no citations with any other documents of this subset of scientific records), we end up with a network of 29,912 nodes. Next, we extract the communities from this network (80, 81). In network science, communities are groups of nodes which have much more connections between each other than expected. Furthermore, we automatically assign them descriptive labels using a topic modeling technique (82, 83).

In Figure 5, we depict the communities obtained, ordered in decreasing order according to their sizes. The keywords within each community are ordered in decreasing order for each community according to their importance. We define importance as the difference between the normalized frequencies of *n*-grams of a given community, and the normalized frequencies of *n*-grams excluding it (an *n*-gram is a set of *n* consecutive words in a text) (83). We can see clearly that the keywords of community A—the largest one—are related to food composition. In the case of community B, it seems to be related to the processing and resulting waste. Community C keywords can be related to the regions and plantations. In D, the keywords are related to applications for palm oil, such as its use

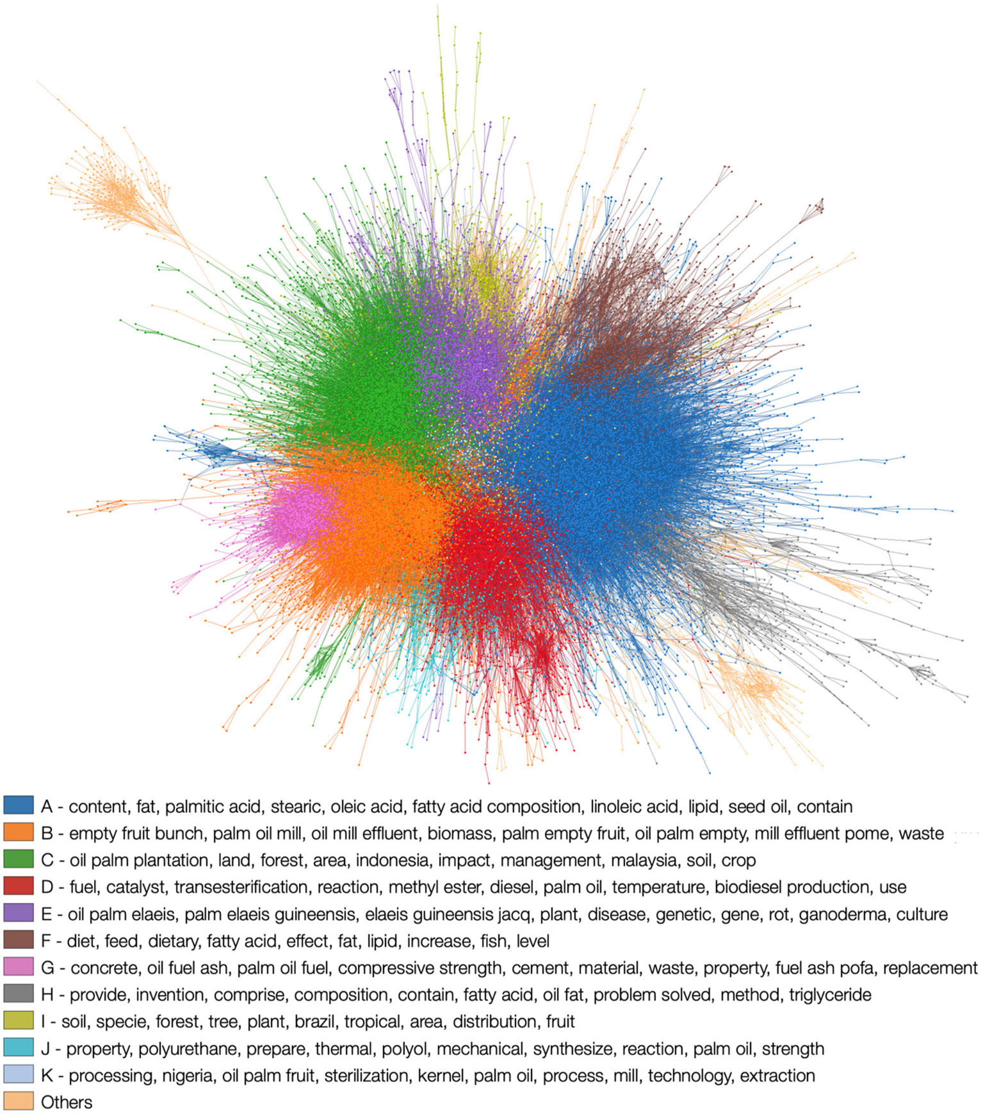

🟦 A - content, fat, palmitic acid, stearic, oleic acid, fatty acid composition, linoleic acid, lipid, seed oil, contain
🟧 B - empty fruit bunch, palm oil mill, oil mill effluent, biomass, palm empty fruit, oil palm empty, mill effluent pome, waste
🟩 C - oil palm plantation, land, forest, area, indonesia, impact, management, malaysia, soil, crop
🟥 D - fuel, catalyst, transesterification, reaction, methyl ester, diesel, palm oil, temperature, biodiesel production, use
🟪 E - oil palm elaeis, palm elaeis guineensis, elaeis guineensis jacq, plant, disease, genetic, gene, rot, ganoderma, culture
🟫 F - diet, feed, dietary, fatty acid, effect, fat, lipid, increase, fish, level
🟪 G - concrete, oil fuel ash, palm oil fuel, compressive strength, cement, material, waste, property, fuel ash pofa, replacement
⬛ H - provide, invention, comprise, composition, contain, fatty acid, oil fat, problem solved, method, triglyceride
🟨 I - soil, specie, forest, tree, plant, brazil, tropical, area, distribution, fruit
🟦 J - property, polyurethane, prepare, thermal, polyol, mechanical, synthesize, reaction, palm oil, strength
🟦 K - processing, nigeria, oil palm fruit, sterilization, kernel, palm oil, process, mill, technology, extraction
🟧 Others

FIGURE 5
Visualization of the palm oil citation network. Scientific records that cite each other can form communities, signaling that they contain similar information. Each color represents a community detected in the network and the labels are the keywords that determine the contents present in the community. Communities are ordered according to their size, with A being the largest. This network was plotted using the software implemented in Silva et al. (82).

for producing biodiesel, etc. Thus, one could focus on studying the 2,293 publications belonging to community A.

The next step would require the application of advanced text mining techniques that could extract the information contained in the papers (84, 85). However, the unstructured nature of these publications makes this a very complex task (86). Furthermore, it is not clear if the results would be valuable enough. One of the problems of Big Data is the high dimensionality of the information, which brings noise accumulation and may introduce spurious correlations. If the information is of low quality, increasing the amount of papers will only exacerbate these issues. Besides, aggregating information from so many different sources will inevitably mix results obtained in different locations, times and with different technologies, which introduces further systematic biases and quality issues (87). As such, simply extracting the pair nutrient - quantity is not enough. Instead, it is necessary also to determine exactly how the sample was analyzed, its specific variety, when and where it was harvested/produced, etc. Not only this represents a much harder task, but it also may not be achievable since much of this metadata might not be contained in the own publication in the first place (88).

Among the AI approaches, Natural Language Processing (NLP) techniques (89) are particularly useful because they can help extract information from the scientific literature. Additionally, recommendation systems have been explicitly created to retrieve and filter scientific papers, which can combine information of different natures (e.g., citation network and paper content) (90). With a set of documents adequately selected, it may be easier for specialists to extract and validate data from the literature. However, one can also automatically look into the content of the papers using NLP. Many techniques have been used to extract and represent the semantics of the texts. Some successfully used methods are the embeddings (91–93), such as *word2vec* (92) and *doc2vec* (93), which represent words and documents, respectively. More recently, transformers were proposed (94). Among the most successful ones are the Bidirectional Encoder Representations from Transformers (BERT) (95) and the Generative Pre-trained Transformer 3 (GPT-3) (96), which can be used as part of systems devoted to retrieving information from documents of different domains (97–101). As such, developing and extending these approaches to assist in constructing better FCDBs is a promising area of research that could help to improve our knowledge about food, nutrition and health (42).

5. Conclusion

The decade of 1980 kicked off a global effort to homogenize and standardize the way in which nutritional composition is collected in order to make meaningful comparisons between countries. Since then, initiatives like INFOODS or EuroFIR have established very clear guidelines and best practices that should be followed to properly obtain, document, store and share this type of data. However, many times, probably due to a lack of funding, these guidelines are not fully adhered to. Furthermore, the end users of these data are usually not fully aware of its limitations and many times complement it with information extracted from sources that are not totally compatible. This may lead to wrong conclusions and misguided policies, with impacts that can take years to fix.

In the age of data, it is more important than ever to ensure that it is correctly captured and displayed. In this contribution, we have discussed that most FCDBs already have many problems with the little information they report. In the particular case of vegetable oils, we have demonstrated that missing information is not always handled properly, that many sources commonly used are old (see Supplementary Table 1) or have mistakes in them (assuming that the source is provided, which is not always the case), and that the quantitative composition can either vary a lot or not at all, without knowing the reasons behind that. This problem is also present in the global scientific literature, not only in FCDBs, which hinders the possibility of reaching precision nutrition. Initiatives such as Foundation foods from USDA are heading in the right direction, but the effort should be much more global and, importantly, sustained in time.

In terms of FCDBs and artificial intelligence (AI), there are two crucial points. The first issue is the urge for good quality data to train AI models properly, and the second is how AI can help feed these databases. Both aspects are related and interdependent because without having data, it is challenging to train models, and without good models, it is much more demanding to enhance the databases. In this paper, we have shown a perspective on the amount of scientific data that has to be processed to extract information regarding a single food item, palm oil, if one wants to scan the information already present in the literature. If this is to be done for many food items, the volume and challenges will increase even further. Nonetheless, we expect that the development of AI in food-related research can positively impact the overall quality of FCDBs, as it has done in other areas of nutrition (102, 103).

There will be many new challenges in this process. This type of analysis will require the collaboration of researchers from different knowledge areas, including network science, neural language processing, food chemistry or nutrition. For the development of new machine learning approaches, it will be essential to include experts in food composition data to evaluate the quality of the information and guarantee the overall quality of the database. As noted, this is a complex task, and the problems related to FCDBs can only be mitigated if experts in many areas put their efforts together. A related problem is the necessity of new funding opportunities for interdisciplinary research projects. Even though large funding agencies actively encourage proposals that cross disciplinary boundaries, in practice most funded projects remain firmly in a disciplinary framework (104).

To conclude, nowadays, the sustainability of the food system is being questioned in the pursuit of the SDGs. Food production is closely related to public health and the environment, and proper knowledge of what we eat is key to improve both. The lack of information on many aspects, such as food fortification or biodiversity is inevitably hindering the progress toward a better food system. Besides, climate change is already having a measurable effect on crops, and not only it will increase in the future, but as we adapt to it, our consumption patterns might change. To mitigate possible further nutritional problems and to solve the ones that we already have, gathering and curating much more and better data is imperative. As the food sector digitizes, it is essential to acknowledge the importance of pursuing a holistic view of nutrition and to move toward a data-driven food system. Only then relevant players will be able to issue evidence-based and timely policy recommendations.

Author contributions

HF, AA, and YM: conceptualization, methodology, investigation, writing–original draft preparation, and writing–review and editing. HF: software, validation, and data curation. HF and AA: visualization. All authors have read and agreed to the published version of the manuscript.

Funding

This paper was developed with the financial support of Soremartec SA and Soremartec Italia, Ferrero Group, in the frame of the Sustainable Nutrition Scientific Board. We also acknowledge partial support from the Government of Aragon, Spain and ERDF A way of making Europe through grant E36–20R (FENOL) to AA and YM, from Ministerio de Ciencia e Innovacin, Agencia Espaola de Investigacin (MCIN/AEI/10.13039/501100011033) Grant No. PID2020115800GBI00 to AA and YM, from Banco Santander (SantanderUZ 2020/0274) to AA and YM, and from Grant RYC2021033226I funded by MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR to AA. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2022.1052934/full#supplementary-material

References

1. McCance RA, Widdowson EM. *The Chemical Composition of Foods*. H.M. Stationery Office (1940).

2. Greenfield H, Southgate DA. *Food Composition Data: Production, Management, and Use*. Rome: Food and Agriculture Organisation (2003).

3. Murphy SP, Charrondiere UR, Burlingame B. Thirty years of progress in harmonizing and compiling food data as a result of the establishment of INFOODS. *Food Chem*. (2016) 193:2–5. doi: 10.1016/j.foodchem.2014.11.097

4. Charrondiere UR, Vignat J, Riboli E. Comparable nutrient intake across countries is only possible through standardization of existing food composition tables (FCT). In: Riboli E, Lambert R, editors. *Nutrition and Lifestyle: Opportunities for Cancer Prevention*. Lyon: IARC Press (2002). p. 45–9.

5. Charrondiere UR, Vignat J, Moller A, Ireland J, Becker W, Church S, et al. The European nutrient database (ENDB) for nutritional epidemiology. *J Food Compos Anal*. (2002) 15:435–51. doi: 10.1006/jfca.2002.1089

6. Klensin JC, Feskanich D, Lin V, Truswell AS, Southgate DAT. *INFOODS Food Composition Data Interchange Handbook*. Tokyo: United Nations University Press (1989).

7. Klensin JC. *INFOODS Food Composition Data Interchange Handbook*. Tokyo: United Nations University Press (1992)

8. Moller A. NORFOODS computer group. Food composition data interchange among the nordic countries: a report. *World Rev Nutr Diet*. (1992) 68:104–20.

9. Schlotke F, Becker W, Ireland J, Moller A, Ovaskainen ML, Monspart J. *Cost Action 99 - EuroFOODS recommendations for food composition database management and data interchange (Report No. EUR 19538)*. European Commission (2000). doi: 10.1006/jfca.2000.0891

10. Slimani N, Charrondiere UR, van Staveren W, Riboli E. Standardization of food composition databases for the European prospective investigation into cancer and nutrition (EPIC): general theoretical concept. *J Food Compos Anal*. (2000) 13:567–84. doi: 10.1006/jfca.2000.0910

11. Deharveng G, Charrondiere UR, Slimani N, Southgate DA, Riboli E. Comparison of nutrients in the food composition tables available in the nine European countries participating in EPIC. *Eur J Clin Nutr*. (1999) 53:60–79. doi: 10.1038/sj.ejcn.1600677

12. Slimani N, Deharveng G, Unwin I, Southgate DAT, Vignat J, Skeie G, et al. The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study. *Eur J Clin Nutr*. (2007) 61:1037–56. doi: 10.1038/sj.ejcn.1602679

13. Castanheira I, André C, Oseredczuk M, Ireland J, Owen L, Robb P, et al. Improving data quality in food composition databanks: a EuroFIR contribution. *Accredit Qual Assur*. (2007) 12:117–125. doi: 10.1007/s00769-006-0225-6

14. Westenbrink S, Presser K, Roe M, Ireland J, Finglas P. Documentation of aggregated/compiled values in food composition databases; EuroFIR default to improve harmonization. *J Food Compos Anal*. (2021) 101:103968. doi: 10.1016/j.jfca.2021.103968

15. EFSA. Guidance on the EU Menu methodology. *EFSA J*. (2014) 12:3944. doi: 10.2903/j.efsa.2014.3944

16. Roe MA, Bell S, Oseredczuk M, Christensen T, Westenbrink S, Pakkala H, et al. Updated food composition database for nutrient intake. *EFSA Support Public*. (2013) 10:355E. doi: 10.2903/sp.efsa.2013.EN-355

17. Hinojosa-Nogueira D, Pérez-Burillo S, Navajas-Porras B, Ortiz-Viso B, de la Cueva SP, Lauria F, et al. Development of an unified food composition database for the European project "Stance4Health". *Nutrients*. (2021) 13:4206. doi: 10.3390/nu13124206

18. Micha R, Coates J, Leclercq C, Charrondiere UR, Mozaffarian D. Global dietary surveillance: data gaps and challenges. *Food Nutr Bull*. (2018) 39:175–205. doi: 10.1177/0379572117752986

19. Ene-Obong H, Schönfeldt HC, Campaore E, Kimani A, Mwaisaka R, Vincent A, et al. Importance and use of reliable food composition data generation by nutrition/dietetic professionals towards solving Africa's nutrition problem: constraints and the role of FAO/INFOODS/AFROFOODS and other stakeholders in future initiatives. *Proc Nutr Soc*. (2019) 78:496–505. doi: 10.1017/S0029665118002926

20. Lupiañez-Barbero A, Blanco CG, de Leiva Hidalgo A. Spanish food composition tables and databases: need for a gold standard for healthcare professionals (review). *Endocrinol Diabetes Nutr*. (2018) 65:361–73. doi: 10.1016/j.endien.2018.05.011

21. Kapsokefalou M, Roe M, Turrini A, Costa HS, Martinez-Victoria E, Marletta L, et al. Food composition at present: new challenges. *Nutrients*. (2019) 11:1714. doi: 10.3390/nu11081714

22. Ispirova G, Eftimov T, Seljak BK. Evaluating missing value imputation methods for food composition databases. *Food Chem Toxicol*. (2020) 141:111368. doi: 10.1016/j.fct.2020.111368

23. Ispirova G, Eftimov T, Korošec P, Koroušić Seljak B. MIGHT: statistical methodology for missing-data imputation in food composition databases. *Appl Sci*. (2019) 9:4111. doi: 10.3390/app9194111

24. Van Puyvelde H, Perez-Cornago A, Casagrande C, Nicolas G, Versele V, Skeie G, et al. Comparing calculated nutrient intakes using different food composition databases: results from the European prospective investigation into cancer and nutrition (EPIC) cohort. *Nutrients*. (2020) 12:2906. doi: 10.3390/nu12102906

25. Ocké MC, Westenbrink S, van Rossum CT, Temme EH, van der Vossen-Wijmenga W, Verkaik-Kloosterman J. The essential role of food composition databases for public health nutrition – experiences from the Netherlands. *J Food Compos Anal*. (2021) 101:103967. doi: 10.1016/j.jfca.2021.103967

26. Ferland G. The discovery of vitamin K and its clinical applications. *Ann Nutr Metab*. (2012) 61:213–8. doi: 10.1159/000343108

27. Westenbrink S, Kadvan A, Roe M, Seljak BK, Mantur-Vierendeel A, Finglas P. 12th IFDC 2017 special issue-evaluation of harmonized EuroFIR documentation for macronutrient values in 26 European food composition databases. *J Food Compos Anal*. (2019) 80:40–50. doi: 10.1016/j.jfca.2019.03.006

28. Grande F, Giuntini EB, Lajolo FM, de Menezes EW. How do calculation method and food data source affect estimates of vitamin A content in foods and dietary intake? *J Food Compos Anal*. (2016) 46:60–9. doi: 10.1016/j.jfca.2015.11.006

29. Haggarty P. UK introduces folic acid fortification of flour to prevent neural tube defects. *Lancet*. (2021) 398:1199–201. doi: 10.1016/S0140-6736(21)02134-6

30. Olson R, Gavin-Smith B, Ferraboschi C, Kraemer K. Food fortification: the advantages, disadvantages and lessons from sight and life programs. *Nutrients*. (2021) 13:1118. doi: 10.3390/nu13041118

31. de Jong MH, Nawijn EL, Verkaik-Kloosterman J. Contribution of voluntary fortified foods to micronutrient intake in The Netherlands. *Eur J Nutr*. (2022) 61:1649–63. doi: 10.1007/s00394-021-02728-4

32. Verkaik-Kloosterman J, Seves SM, Ocké MC. Vitamin D concentrations in fortified foods and dietary supplements intended for infants: implications for vitamin D intake. *Food Chem*. (2017) 221:629–35. doi: 10.1016/j.foodchem.2016.11.128

33. Patterson KY, Phillips KM, Horst RL, Byrdwell WC, Exler J, Lemar LE, et al. Vitamin D content and variability in fluid milks from a US department of agriculture nationwide sampling to update values in the national nutrient database for standard reference. *J Dairy Sci*. (2010) 93:5082–90. doi: 10.3168/jds.2010-3359

34. Christensen T, Saxholt E, Pilegaard K, Trolle E, Knuthsen P, Virtanen S, et al. *Nordic co-operation on Food information. Activities of the Nordic Food Analysis Network 2013-2016*. Nordic Council of Ministers, TemaNor (2017). Available online at: https://www.duo.uio.no/handle/10852/60710 (accessed June 01, 2022).

35. GFDx. *Gloal Status of Food Fortification Compliance or Quality*. Global Fortification Data Exchange (2021). Available online at: https://fortificationdata.org (accessed June 01, 2022).

36. Mkambula P, Mbuya MNN, Rowe LA, Sablah M, Friesen VM, Chadha M, et al. The unfinished agenda for food fortification in low- and middle-income countries: quantifying progress, gaps and potential opportunities. *Nutrients*. (2020) 12:354. doi: 10.3390/nu12020354

37. Westenbrink S, van der Vossen-Wijmenga W, Toxopeus I, Milder I, Ocké M. LEDA, the branded food database in the Netherlands: data challenges and opportunities. *J Food Compos Anal*. (2021) 102:104044. doi: 10.1016/j.jfca.2021.104044

38. Bland JS. The dark matter of nutrition: dietary signals beyond traditional nutrients. *Integrat Med Clin J*. (2019) 18:12. Available online at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6601448/# (accessed June 01, 2022).

39. Teng Y, Ren Y, Sayed M, Hu X, Lei C, Kumar A, et al. Plant-Derived exosomal MicroRNAs shape the gut microbiota. *Cell Host Microbe*. (2018) 24:637–52.e8. doi: 10.1016/j.chom.2018.10.001

40. Sarkadi LS. Amino acids and biogenic amines as food quality factors. *Pure Appl Chem*. (2019) 91:289–300. doi: 10.1515/pac-2018-0709

41. FooDB. *Listing Compounds - FooDB* (2022). Available online at: https://foodb.ca/compounds (accessed June 1, 2022).

42. Barabsi AL, Menichetti G, Loscalzo J. The unmapped chemical complexity of our diet. *Nat Food*. (2020) 1:33–7. doi: 10.1038/s43016-019-0005-1

43. Phenol-Explorer. *Database on Polyphenol Content in Foods - Phenol-Explorer*. (2022). Available online at: http://phenol-explorer.eu (accessed November 25, 2022).

44. Durazzo A, Astley S, Kapsokefalou M, Costa HS, Mantur-Vierendeel A, Pijls L, et al. Food composition data and tools online and their use in research and policy: EuroFIR AISBL contribution in 2022. *Nutrients*. (2022) 14:4788. doi: 10.3390/nu14224788

45. World Health Organization. *Using Third-Party Food Sales and Composition Databases to Monitor Nutrition Policies*. World Health Organization Regional Office for Europe (2021). Available online at: https://apps.who.int/iris/handle/10665/339075 (accessed June 01, 2022).

46. Harrington RA, Adhikari V, Rayner M, Scarborough P. Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. *BMJ Open*. (2019) 9:e026652. doi: 10.1136/bmjopen-2018-026652

47. Sissener NH, Suarez RK, Hoppeler HH. Are we what we eat? Changes to the feed fatty acid composition of farmed salmon and its effects through the food chain. *J Exp Biol*. (2018) 221 (Suppl_1):jeb161521. doi: 10.1242/jeb.161521

48. Carnovale E, Nicoli S. Changes in fatty acid composition in beef in Italy. *J Food Compos Anal*. (2000) 13:505–10. doi: 10.1006/jfca.2000.0908

49. Gnagnarella P, Parpinel M, Salvini S, Franceschi S, Palli D, Boyle P. The update of the Italian food composition database. *J Food Compos Anal*. (2004) 17:509–22. doi: 10.1016/j.jfca.2004.02.009

50. Hulshof KF, Beemster CJ, Westenbrink S, Lwik MR. Reduction in fat intake in The Netherlands: the influence of food composition data. *Food Chem*. (1996) 57:67–70. doi: 10.1016/0308-8146(96)00076-3

51. Grande F, Vincent A. The importance of food composition data for estimating micronutrient intake: What do we know now and into the future? In: *Global Landscape of Nutrition Challenges in Infants and Children*. Vol. 93. Basel: Karger Publishers (2020). p. 39–50.

52. Englberger L. Revisiting the vitamin A fiasco: going local in Micronesia. In: Burlingame B, Dernini S, editors. *Sustainable Diets and Biodiversity: Directions and Solutions for Policy, Research and Action*. Rome: FAO (2012). p. 126–33.

53. Toledo A, Burlingame B. Biodiversity and nutrition: a common path toward global food security and sustainable development. *J Food Compos Anal*. (2006) 19:477–83. doi: 10.1016/j.jfca.2006.05.001

54. Burlingame B, Charrondiere R, Mouille B. Food composition is fundamental to the cross-cutting initiative on biodiversity for food and nutrition. *J Food Compos Anal*. (2009) 22:361–5. doi: 10.1016/j.jfca.2009.05.003

55. Charrondire UR, Stadlmayr B, Rittenschober D, Mouille B, Nilsson E, Medhammar E, et al. FAO/INFOODS food composition database for biodiversity. *Food Chem*. (2013) 140:408–12. doi: 10.1016/j.foodchem.2012.08.049

56. FAO. *FAO/INFOODS Food Composition Database for Biodiversity Version 4.0 - BioFoodComp 4.0)*. FAO (2017).

57. Scheelbeek Pauline FD, Bird Frances A, Tuomisto Hanna L, Rosemary G, Harris Francesca B, Joy Edward JM, et al. Effect of environmental changes on vegetable and legume yields and nutritional quality. *Proc Natl Acad Sci USA*. (2018) 115:6804–9. doi: 10.1073/pnas.1800442115

58. Bisbis MB, Gruda N, Blanke M. Potential impacts of climate change on vegetable production and product quality - a review. *J Cleaner Prod*. (2018) 170:1602–20. doi: 10.1016/j.jclepro.2017.09.224

59. Broberg MC, Hagy P, Pleijel H. CO2-Induced changes in wheat grain composition: meta-analysis and response functions. *Agronomy*. (2017) 7:32. doi: 10.3390/agronomy7020032

60. Leisner CP. Review: climate change impacts on food security- focus on perennial cropping systems and nutritional value. *Plant Sci*. (2020) 293:110412. doi: 10.1016/j.plantsci.2020.110412

61. INFOODS. *INFOODS: Tables and Databases*. (2022). Available online at: https://www.fao.org/infoods/infoods/tables-and-databases/en (accessed November 23, 2022).

62. BEDCA. *Base de Datos Espaola de Composicin de Alimentos*. (2021). Available online at: https://www.bedca.net (accessed November 21, 2021).

63. Samaniego-Vaesken ML, Alonso-Aperte E, Varela-Moreiras G. Voluntary fortification with folic acid in Spain: an updated food composition database. *Food Chem*. (2016) 193:148–53. doi: 10.1016/j.foodchem.2014.06.046

64. San Mauro Martín I, Hernández Rodríguez B. Herramientas para la calibración de menús y cálculo de la composición nutricional de los alimentos: validez y variabilidad. *Nutr Hosp*. (2014) 29:929–34. doi: 10.3305/nh.2014.29.4.7096

65. FRIDA. *FRIDA Food Data, Version 4, 2019*. National Food Institute, Techincal University of Denmark (2021). Available online at: https://frida.fooddata.dk (accessed June 26, 2022).

66. Delgado A, Issaoui M, Vieira MC, Saraiva de Carvalho I, Fardet A. Food composition databases: does it matter to human health? *Nutrients*. (2021) 13:2816. doi: 10.3390/nu13082816

67. USDA. *Food Data Central*. (2021). Available online at: https://fdc.nal.usda.gov (accessed November 21, 2021).

68. TBCA. *Tabela Brasileira de Composição de Alimentos (TBCA). Universidade de São Paulo (USP). Version 7.1*. São Paulo: Food Research Center (2021). Available online at: http://www.fcf.usp.br/tbca (accessed June 26, 2022).

69. NIGERIA. *Nigeria Food Database*. (2021). Available online at: http://nigeriafooddata.ui.edu.ng (accessed November 21, 2021).

70. SMILING. *SMILING/IRDSMILING*. (2021). Available online at: http://www.nutrition-smiling.eu (accessed November 21, 2021).

71. Hulshof P, Doets E, Seyha S, Bunthang T, Vonglokham M, Kounnavong S, et al. Food composition tables in southeast Asia: the contribution of the SMILING project. *Matern Child Health J*. (2019) 23:46–54. doi: 10.1007/s10995-018-2528-8

72. Fukagawa NK, McKillop K, Pehrsson PR, Moshfegh A, Harnly J, Finley J. USDA's FoodData central: what is it and why is it needed today? *Am J Clin Nutr*. (2022) 115:619–24. doi: 10.1093/ajcn/nqab397

73. EuroFIR. *EuroFIR Association International Sans But-Lucratif. Annual Report 2015*. Parma: EuroFIR (2016).

74. USDA. *Oilseeds: World Markets and Trade*. United States Department of Agriculture (2022). Available online: https://apps.fas.usda.gov/psdonline/circulars/oilseeds.pdf (accessed June 1, 2022).

75. Mellal MA. Obsolescence-A review of the literature. *Technol Soc*. (2020) 63:101347. doi: 10.1016/j.techsoc.2020.101347

76. Grichi Y, Beauregard Y, Dao TM. An approach to obsolescence forecasting based on hidden Markov model and compound poisson process. *Int J Indust Eng*. (2019) 1:111–24. doi: 10.46254/j.ieom.20190202

77. Clegg AJ. Composition and related nutritional and organoleptic aspects of palm oil. *J Am Oil Chem Soc*. (1973) 50:321–4. doi: 10.1007/BF02641365

78. Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu BJP, et al. An overview of microsoft academic service (MAS) and applications. In: *WWW 15 Companion: Proceedings of the 24th International Conference on World Wide Web*. New York, NY: Association for Computing Machinery (2015). p. 243–6. doi: 10.1145/2740908.2742839

79. Mabry PL, Yan X, Pentchev V, Van Rennes R, McGavin SH, Wittenberg JV. CADRE: a collaborative, cloud-based solution for big bibliographic data research in academic libraries. *Front Big Data*. (2020) 3:556282. doi: 10.3389/fdata.2020.556282

80. Rosvall M, Axelsson D, Bergstrom CT. The map equation. *Eur Phys J Spec Top*. (2009) 178:13–23. doi: 10.1140/epjst/e2010-01179-1

81. Martin R, Bergstrom Carl T. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci USA*. (2008) 105:1118–23. doi: 10.1073/pnas.0706851105

82. Silva FN, Amancio DR, Bardosova M, da Fontoura Costa L, Novais de Oliveira O Jr. Using network science and text analytics to produce surveys in a scientific topic. *J Informetr*. (2016) 10:487–502. doi: 10.1016/j.joi.2016.03.008

83. Ceribeli C, Ferraz de Arruda H, da Fontoura Costa L. How coupled are capillary electrophoresis and mass spectrometry? *Scientometrics*. (2021) 126:3841–51. doi: 10.1007/s11192-021-03923-0

84. Westergaard D, Strfeldt HH, TAnsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput Biol*. (2018) 14:e1005962. doi: 10.1371/journal.pcbi.1005962

85. Pulla P. The plan to mine the world's research papers. *Nature*. (2019) 571:316–318. doi: 10.1038/d41586-019-02142-1

86. Hooton F, Menichetti G, Barabsi AL. Exploring food contents in scientific literature with FoodMine. *Sci Rep*. (2020) 10:16191. doi: 10.1038/s41598-020-73105-0

87. Fan J, Han F, Liu H. Challenges of big data analysis. *Natl Sci Rev*. (2014) 1:293. doi: 10.1093/nsr/nwt032

88. Gossner CME, Schlundt J, Embarek PB, Hird S, Lo-Fo-Wong D, Beltran JJO, et al. The melamine incident: implications for international food and feed safety. *Environ Health Perspect*. (2009) 117:1803. doi: 10.1289/ehp.0900949

89. Eisenstein J. *Introduction to natural language processing*. Cambridge, MA: MIT Press (2019)

90. Bai X, Wang M, Lee I, Yang Z, Kong X, Xia F. Scientific paper recommendation: a survey. *IEEE Access*. (2019) 7:9324–39. doi: 10.1109/ACCESS.2018.2890388

91. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: *Proceedingsă of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha (2014). p. 1532–43. doi: 10.3115/v1/D14-1162

92. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representationsă of words and phrases and their compositionality. In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. *Advances in Neural Information Processing Systems*. Vol. 26. Lake Tahoe, NV: Curran Associates, Inc. (2013). p. 3111–9.

93. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. Bejing: PMLR (2014). p. 1188–96.

94. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inform Process Syst*. (2017) 30:5998–6008.

95. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics (2019). p. 4171–86. Available online at: https://aclanthology.org/N19-1423 (accessed June 01, 2022).

96. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inform Process Syst*. (2020) 33:1877–901.

97. Nguyen MT, Le DT, Son NH, Minh BC, Shojiguchi A, et al. Information extraction of domain-specific business documents with limited data. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. Shenzhen: IEEE (2021). p. 1–8. doi: 10.1109/IJCNN52387.2021.9534328

98. Friedrich A, Adel H, Tomazic F, Hingerl J, Benteau R, Marusczyk A, et al. The SOFC-Exp corpus and neural approaches to information extraction in the materials science domain. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (2020). p. 1255–68. doi: 10.18653/v1/2020.acl-main.116

99. Gutirrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about GPT-3 in-context learning for biomedical IE? Think again. *arXiv Preprint*. (2022) arXiv:220308410.

100. Stammbach D, Antoniak M, Ash E. Heroes, villains, and victims, and GPT-3-automated extraction of character roles without training data. *arXiv Preprint*. (2022) arXiv:220507557. doi: 10.18653/v1/2022.wnu-1.6

101. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are zero-shot clinical information extractors. *arXiv Preprint*. (2022) arXiv:220512689.

102. Morgenstern JD, Rosella LC, Costa AP, de Souza RJ, Anderson LN. Perspective: big data and machine learning could help advance nutritional epidemiology. *Adv Nutr*. (2021) 12:621–31. doi: 10.1093/advances/nmaa183

103. Côté M, Lamarche B. Artificial intelligence in nutrition research: perspectives on current and future applications. *Appl Physiol Nutr Metab*. (2021) 47:1–8.

104. König T, Gorman ME. The Challenge of Funding Interdisciplinary Research: A Look inside Public Research Funding Agencies. Oxford: OUP Academic (2017).

TYPE Data Report
PUBLISHED 02 February 2023
DOI 10.3389/fnut.2023.1066737

Check for updates

OPEN ACCESS

EDITED BY
Alessandra Durazzo,
Council for Agricultural Research and Economics, Italy

REVIEWED BY
Yu Zhang,
Chinese Center for Disease Control and Prevention, China
Christian Ritzel,
Conthey Research Center, Agroscope, Switzerland

*CORRESPONDENCE
Elise Mognard
✉ eliseline.mognard@taylors.edu.my

SPECIALTY SECTION
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

RECEIVED 11 October 2022
ACCEPTED 05 January 2023
PUBLISHED 02 February 2023

CITATION
Mognard E, Naidoo K, Laporte C, Tibère L, Alem Y, Khusun H, Februhartanty J, Niiyama Y, Ueda H, Dasgupta A, Dupuy A, Rochedy A, Yuen JL, Ismail MN, Nair PK, Ragavan NA and Poulain J-P (2023) "Eating Out", spatiality, temporality and sociality. A database for China, Indonesia, Japan, Malaysia, Singapore and France. *Front. Nutr.* 10:1066737.
doi: 10.3389/fnut.2023.1066737

"Eating Out", spatiality, temporality and sociality. A database for China, Indonesia, Japan, Malaysia, Singapore and France

Elise Mognard[1,2,3]*, Kremlasen Naidoo[1,2,3], Cyrille Laporte[1,2,3,4], Laurence Tibère[1,2,3,4], Yasmine Alem[1,2,3], Helda Khusun[5], Judhiastuty Februhartanty[5], Yoko Niiyama[6], Haruka Ueda[7,8], Anindita Dasgupta[1,3], Anne Dupuy[1,2,3,4], Amandine Rochedy[1,2,3,4], Jan Li Yuen[1,2,3], Mohd Noor Ismail[3,9], Pradeep Kumar Nair[1], Neethianhantan Ari Ragavan[1,3] and Jean-Pierre Poulain[1,2,3,4]

[1]Faculty of Social Sciences and Leisure Management, Taylor's University, Subang Jaya, Malaysia, [2]"Food Studies: Food, Cultures and Health", Université de Toulouse, Toulouse, France, [3]Center for Asian Modernisation Studies (CAMS), Taylor's University, Subang Jaya, Malaysia, [4]Centre d'Études et de Recherche: Travail, Organisation, Pouvoir (CERTOP) Unité Mixte de Recherche 5044 (UMR), Centre National de la Recherche Scientifique (CNRS), Université de Toulouse, Toulouse, France, [5]SEAMEO Regional Center for Food and Nutrition (RECFON) – Pusat Kajian Gizi Regional Universitas Indonesia, Jakarta, Indonesia, [6]College of Gastronomy and Management and Graduate School of Economics, Ritsumeikan University, Kyoto, Japan, [7]Japan Society for the Promotion of Sciences, Tokyo, Japan, [8]Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan, [9]Centre for Community Health Studies (ReaCH), Faculty of Health Sciences, Universiti Kebangsaan Malaysia, Bandar Baru Bangi, Malaysia

KEYWORDS

sociology of food, health, nutrition, eating out, food practices, food transition, compacted modernization, open science

1. Introduction

Eating out is a central dimension of the food and nutrition transitions (1–3). However, most of the available data on eating out were reported in Europe and North America. A high rate of eating out is one of the specificities of the Asian food system (4–6), which is further assumed to increase alongside compressed modernization (7–9). To fill this gap, "Eating Out" is a recurrent cross-sectional survey that focuses on the spatiality, temporality, and sociality of food intakes in five Asian countries and one European country. It, thus, addresses an important data gap by allowing cross-national comparisons and quantitative assessments of movements of food between home and out of home across a large consortium. It is conducted within the framework of the chair of "Food Studies: Food, Cultures, and Health" created jointly by Taylor's University (Malaysia) and the University of Toulouse Jean Jaurès (France), in partnership with SEAMEO RECFON (Indonesia) and Ritsumeikan University (Japan), and lead by Jean-Pierre Poulain. This survey is a part of the wider Asian Food Barometer initiative and supplementary to the national Food Barometers, currently, the Malaysian (10) and the Indonesian databases. While the national surveys are including data on the food content and quantities, thus enabling analysis of the nutrient composition (11), "Eating Out" is focusing on the food day patterns.

This article briefly reviews the available data on eating out—specifically in Asia, proposes a framework, and details the methods regarding the organization of the initial data collection (2019–2020). Expected uses and limitations of the data as well as their possible contributions conclude the article.

2. Empirical data on eating out

While most agree that the prevalence of eating out in Asia is high, its analysis mainly relies on economic and nutritional data that focus, respectively, on monetary flows and nutritional intakes. In addition, the empirical analysis of the behavioral dimension is faint.[1] In addition, one of the challenges of an empirical study of eating out practices mainly lies in the polysemy of "consumption", where diverse behaviors are possibly aggregated on the same site: purchases and actual individual incorporation, which are framed by different—and at times conflicting—angles (12–14).

For the purchases, data collected by the household consumption and expenditure surveys primarily aim at deriving consumption patterns and providing input to the compilation of national accounts, from the economic perspective. Those surveys refer to the food prepared away from home, either purchased—from a commercial establishment, a canteen or cafeteria at a school or at the workplace—or received in-kind from a school or an employer, a food assistance program, or a gift from another household (15). Conversely, they typically exclude or aggregate in other expenditure food intakes within institutional care or business meals. Additional insights are provided by specialized consultants in the food service industry and generalist global market research companies (16) with the relationships between individual preferences and socio-economic variables. However, the limited access to the methodology—due to the cost incurred—constrains the identification of the objectivity of the data collected, which can vary from the declared practices to social representations.

Regarding the actual individual incorporation, numerous nutritional studies reflect the public health concern with eating out. They contribute to the analysis of individual diets in terms of nutritional composition—that is assumed to be higher in energy, fat, sugar, and salt and low in vegetables—which happens to be the reverse of dietary advice. Nonetheless, the heterogeneity in the definitions of eating out (17–19) makes difficult cross-national comparisons or attempts to quantitatively assess a change (20). Most importantly, they do not reflect movements of food between home and out of home. Globally, social scientists are interested in "Eating Out" from what it says about the character of contemporary societies and provides details on the diverse socio-cultural and socio-historical meanings of eating out and focuses on the socio-technical and institutional arrangements (21–30). However, the empirical knowledge about food habits focuses mainly on the domestic dimension (5) and concentrates mainly on Europe and America.

"Eating Out" initiative posits that food decisions are embedded within behavioral scripts, routines, or rules predefined by socio-cultural contexts (31–34). These scripts, routines, or rules allow the coordination of the social actors involved in the production, processing, distribution, preparation, and incorporation of food. Therefore, they are contributing to the synchronization (13, 35, 36) and "orchestration" (23, 33, 37) of the food practices. The food habits at home interplay with the structure of the household, the number of diners, and gender roles—to name a few. When eating out, the place, individualization of the items, relation between the client, consumer, and service provider, and policies to support eating at school or workplace—among others, are essential in the definition of food social norms and practices (38, 39). Thus, the "Eating Out" initiative considers the transformations of the societies along with compressed modernization[2] and its consequences on "food days"[3] (12, 40). It is focused on the spatiality of preparation and eating, the temporality, and the sociality of food intake. Figure 1 summarizes the research framework.

In Asia, it seems that the role of home-cooked food in practices is not as central[4] while the prevalence of eating out is high. Paradoxically, empirical social scientific studies at the national or cross-national levels are scarce, as mentioned earlier. The development of important economic factors in the food service sector in the 1970s in Europe (38, 39) has contributed to the production of data that have made this phenomenon visible. When existing, studies are mainly framed by the nutritional perspective [for example (42–46)] which, while being a matter of controversies in the West (47–49), is further applied without much consideration paid to the particularities of the Asian contexts and histories. Thus, it has been undertaken by the Asian Food Barometer initiative (6, 10, 50). "Eating Out" database contributes with a focus on the movements of food intake between home and out of home in relation to compressed modernization and provides an empirical basis to the debate of the social and public health implications of eating out in the Asian contexts. Four food spaces are identified to describe the movements of food between home and out of home: (1) home food—food prepared at home can include the use of convenience products; (2) eating out—food is consumed at an out-of-home outlet/restaurant/stall/canteen, etc., or on the go; (3) delivery/takeaway—food prepared out of home and consumed within the home; and (4) food prepared at home and eaten out of home.

3. Methods

3.1. Sample size and methods

The total sample size is over 15,000 respondents, following the geographical repartition presented in Figure 2. Locations chosen for the initial data collection of "Eating Out" were concentrated on East and Southeast Asia—with France providing an occidental comparison point—based on their geographical positions, population, and modernization dynamics, reported rates of eating out, shared and divergent histories, and public health concerns.

[1] This paradox could emerge from the absence of local stakeholders, particularly from the food service industry, that are economically significant enough to commission studies.

[2] Compressed or compacted modernization has been proposed by Chang (7) to refer to the civilizational condition where economic, political, social, and/or cultural changes occur in an extremely condensed manner in respect to both time and space.

[3] In sociology of food, the concept of "food day" characterises the concentration, time, and synchronization of food intakes during the day, as a result of socio-technical and institutional arrangements.

[4] To the extend where, based on her study of the foodscape from poor Jakarta (Indonesia) *kampungs,* Arciniegas (41) posits the need to revise the binary distinction between "home" and "out-of-home" eating behaviors.

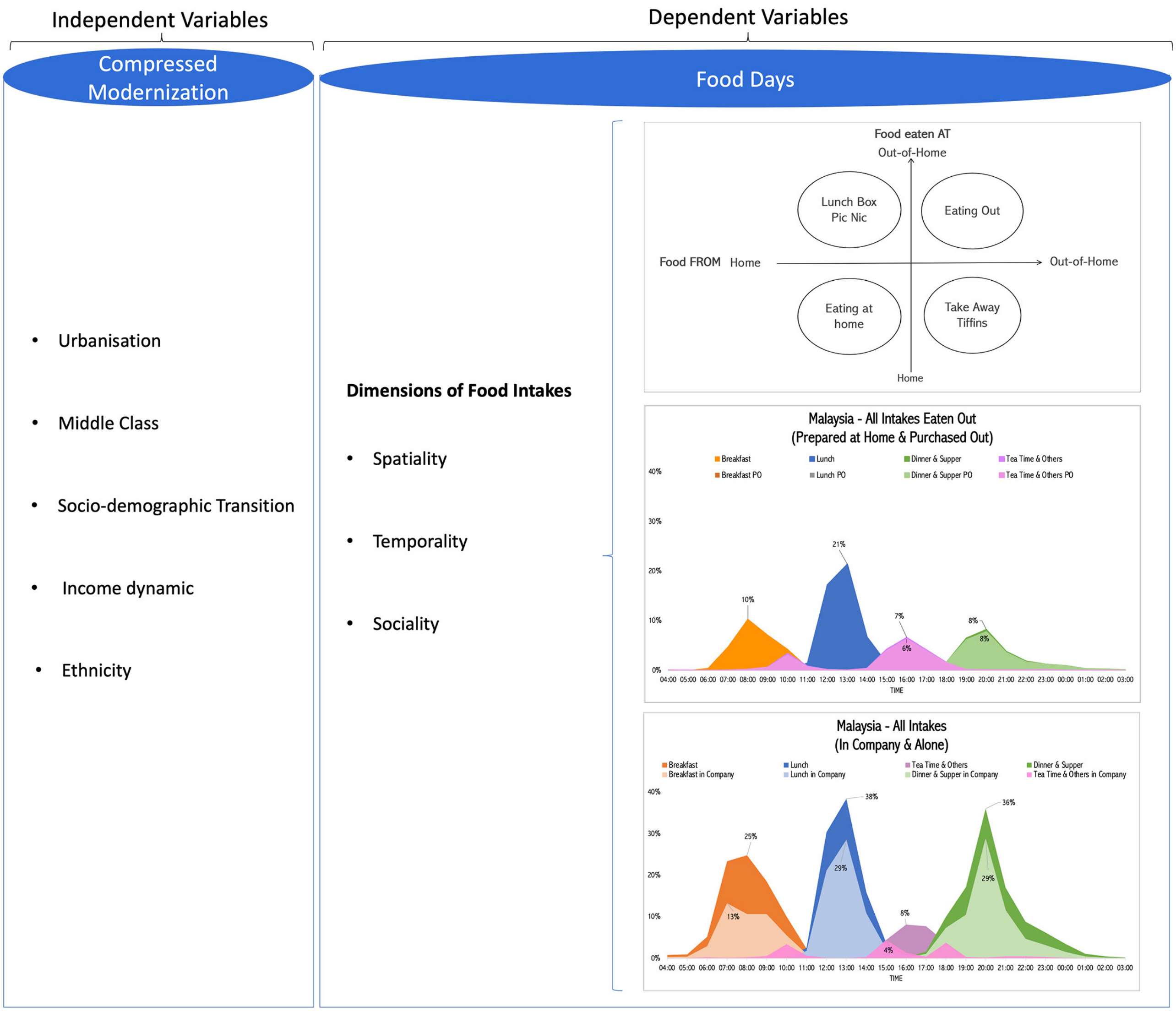


FIGURE 1
Research framework of "Eating Out" initiative.

A combination of stratified random sampling—selection from panel respondent sources—and quota sampling was applied to optimize the benefits of both sampling methods. The data collection aimed at achieving national representative samples. For China, global representativity is practically very difficult. The sample focuses on a range of medium urbanized areas. Minimum sample sizes were set for each location. Respondents were selected out of the national populations aged 18 years and older, across all income groups and rural and urban areas. Quotas were implemented in each location, so to minimize challenges in representativeness for the samples on age, gender, urbanization (except for metropolitan areas of Hong Kong and Singapore), and ethnicity where relevant (Singapore and Malaysia).

The inclusion of France in this survey is first justified by the frequency of eating out in Europe[5] that positions France as a European representative where little variations are observed compared with Asia. Second, the involvement of the research team in investing and analyzing eating out in France for approximately 30 years along with six national surveys among which one was completed with INPES (12, 13, 52). This long-term involvement constitutes both a methodological heritage and a possibility of a critical discussion with the longitudinal analysis previously developed in French data.

3.2. Research instrument

A structured self-administered online survey was deemed as an appropriate method to both collect data and manage the constraints of geography, linguistics, time, and budget. Given that eating out practices may differ across the week, the survey collects data based on a 72-h recall, a measure of food

[5] Frequency of meals eaten out is 1 out 5 in France, 1 out of 7 in Germany, 1 out of 4 in Italy, 1 out of 3 in UK, and 1 out 5 in Spain (40, 51).

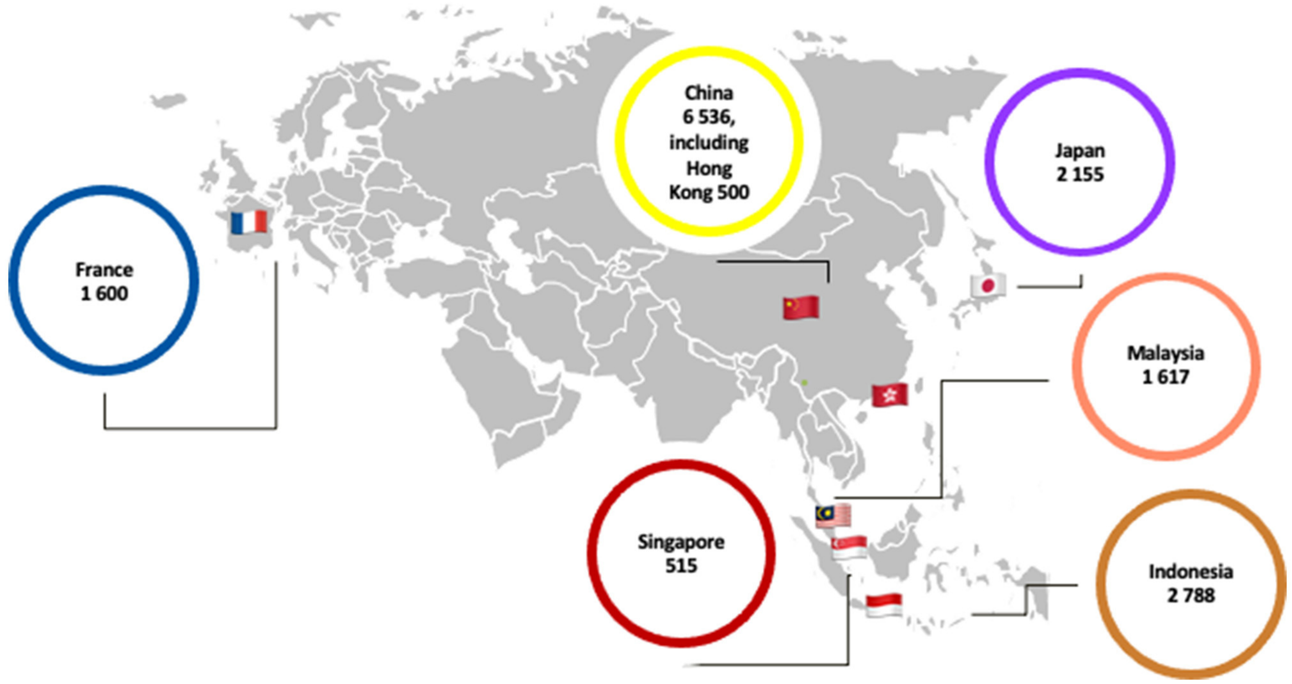


FIGURE 2
Sample of initial "Eating Out" data collection (*N* = 15,211).

intake covering typically 1 to 3 days and initially developed by Wiehl (53).

The close-ended questionnaire assisted the respondents to recall their food intakes according to their (Figure 3):

- Temporality, the recall of 3 days with the time of the food intakes was supported by a matrix breaking down each of their days into hours, starting early morning–4 a.m.–until late night–3 a.m.;
- Formality, selection of the name of the intake from a list comprising "Breakfast", "Teatime", "Lunch", "Dinner", and "Supper" and the possibility to define it as "Others";
- Spatiality, from two options, namely, "At home" vs. "Out of the Home" and source of the food from "Food prepared at home" vs. "Food Purchased outside or delivery";
- Sociality, between "Alone" vs. "With company".

Invitations to participate in the survey were staged in the week and sent out on Tuesdays with the aim to spread out 3-day recall across 2 weeks to include at least 1 day from the weekend (defined as Saturday and Sunday).

The country of the collection was automatically detected when the respondent started the questionnaire. First, questions were to filter the respondents based on quotas and collect data on age, race/ethnicity, religion, gender, number of children, number of family members living under the same roof, urbanization, education attainment, monthly household income, or Wealth Index for Indonesia. Given the potential interest of the database for public health, height and weight were also included to allow the computation of the body mass index (BMI) as an indicator of nutritional status.

The questionnaire was initially designed in English. Questions were translated into the national language(s) where needed. In that case, back-translation was performed to ensure the accuracy of the translation.[6] Typically, a respondent needed approximately 12 min to answer the questionnaire.

3.3. Data collection and procedures

The data for the initial survey were collected in partnership with a company—*Toluna*, specializing in conducting large multilocation surveys and holding large panels and affiliated networks in each of the locations of the "Eating Out" data collection. Invitations to participate were sent using email, mobile text, or on the partner company's mobile application platforms.[7] The data were collected in two windows-—6 January to 24 January 2020 and 31 January to 9 February 2020.[8]

The research team has developed along with *Toluna*, a web interface, to collect data in diverse languages and unit systems.

[6] The initial questionnaire is compared with the one obtained after back-translation. Where differences are observed between the two versions, revisions along with the translators are engaged.

[7] *Toluna* employs checks to stop double participation in the data collection phase.

[8] Festive seasons are known to modify the food intake behavior. Thus, the data collection was started after New year (6 January) and then paused for Chinese New Year (23-29 January).

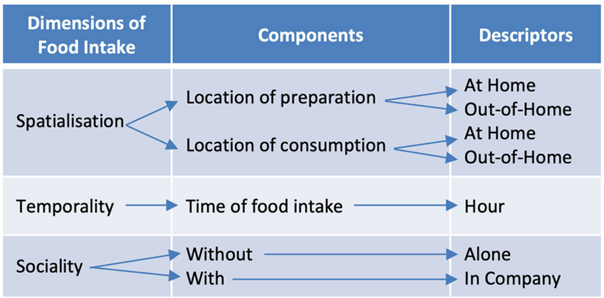
The design of the 3-day recall presents a break of each of their days into hours, starting early morning–4 a.m.–until late night–3 a.m. Any time the respondent selected a type of intake for a given time, the selection of locations of food preparation, and consumption as well as the sociality of the intake was made available.

3.4. Illustration of outcomes

The dataset provided allows analysis and comparisons of the food days across the countries based on the computation of the distribution of meals or food intakes according to the spatiality of their preparation and incorporation and their sociality as well as their temporal distribution across the time of the day. A usual statistical approach to the analysis of the variance can be applied. Figure 4 illustrates the important contrast regarding the distribution of the spatiality of food between Singapore and France—where the percentages of meals purchased out of home and either consumed at home or out of home are 50 and 13.5%, respectively.

The comparison between the temporal distribution of food intakes eaten out in Indonesia and Malaysia presented in Figure 5 shows the difference in terms of the percentage of breakfast eaten out or the synchronization of the lunch eaten out where 35% of the Indonesian population eaten out at noon while 21% does in Malaysia.

Another analysis is the temporal distribution of the food intakes according to their sociality as displayed in Figure 6. When 38% of the Malaysian population eats lunch at 1 p.m., 10% of them are eating it alone.



Dimensions of Food Intake	Components	Descriptors
Spatialisation	Location of preparation	At Home Out-of-Home
	Location of consumption	At Home Out-of-Home
Temporality	Time of food intake	Hour
Sociality	Without	Alone
	With	In Company

FIGURE 3
Dimensions, components, and descriptors of food intakes.

4. Conclusion

With its open data on the repartition of preparation and incorporation of food intakes between private/domestic and public/commercial spheres, the "Eating Out" initiative provides homogenous data across five Asian countries and one occidental country. Globally, social scientists are interested in eating out from what it says about the character of contemporary societies. Studies provide details on the diverse socio-cultural and socio-historical meanings of eating out and focuses on the socio-technical and institutional arrangements (21–30). Thus, it contributes empirically to the debates on modernization (7–9) and, more specifically, on the influence of modernization on the technical and economic organization of food habits and their social and public health consequences. In view of future data collections, the data collected

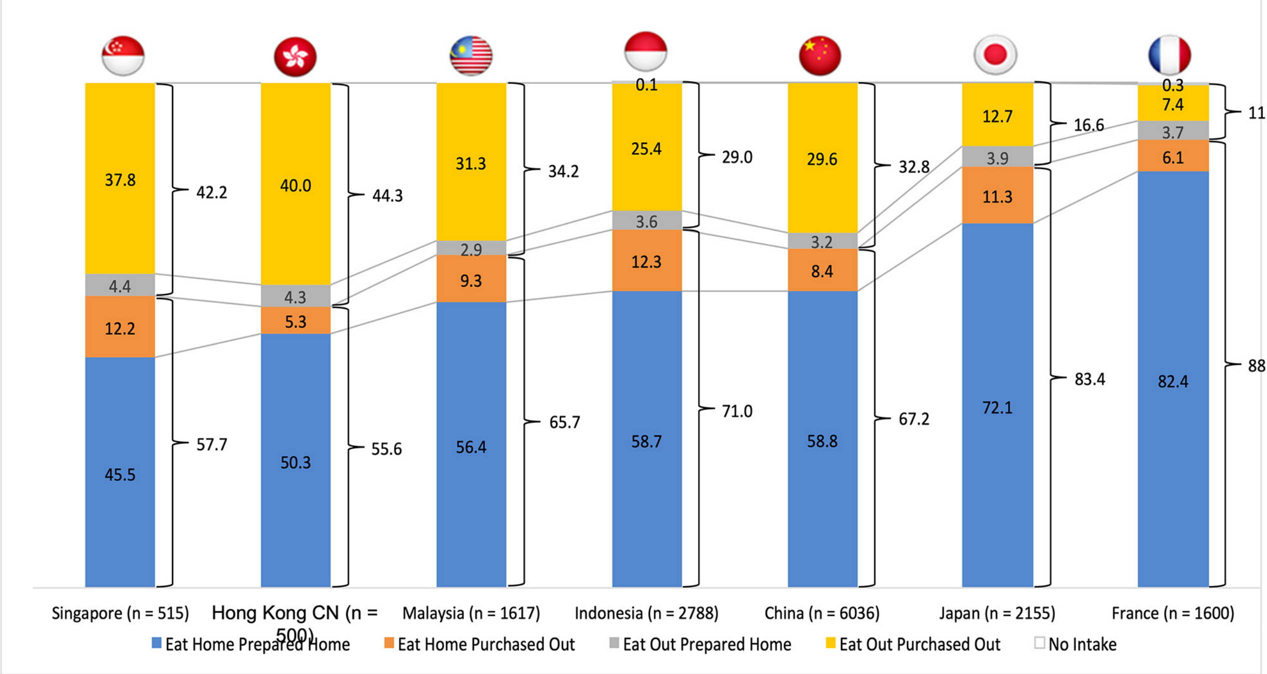


FIGURE 4
Comparison of the percentages of spatiality for preparation and consumption of meals.

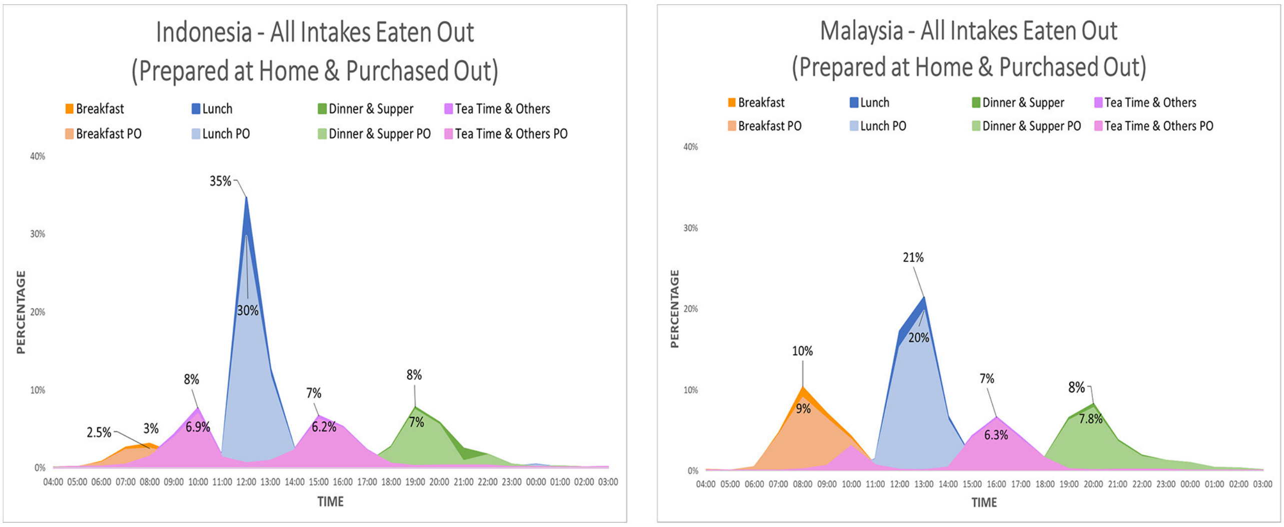


FIGURE 5
Temporal distribution of food intakes according to spatiality in Indonesia and Malaysia.

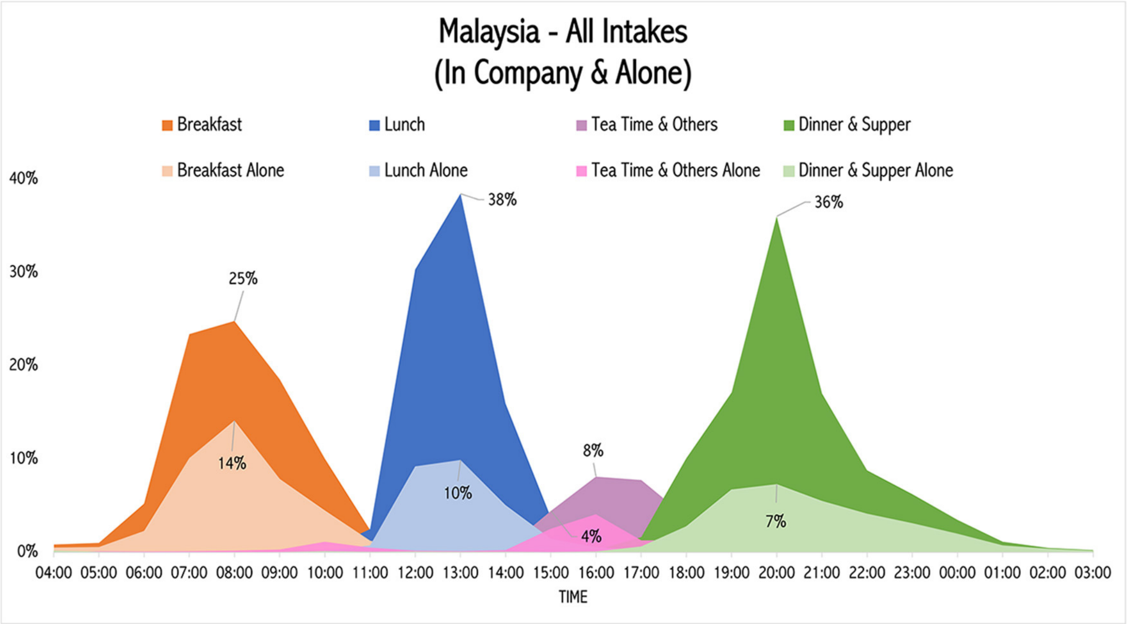


FIGURE 6
Temporal distribution of food intakes according to sociality in Malaysia.

from early 2020 offers a baseline on food practices prior to the COVID-19 pandemic and related lockdowns. Limitations could possibly be found in the representativeness of the national samples in relation to the online collection, and the invisibility of some fluctuations in one's definition of what is home and out of home, for example, in the case of eating to another's—i.e., friend, neighbor, and family—home. Nonetheless, the analysis of the "Eating Out" dataset by researchers, public and private decision-makers, and students could benefit the food (service) industry to understand the organization of the demand of the food market, public health, as a complement to nutritional surveys in designing policies focusing on the food environment.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary material.

Ethics statement

Toluna complies with the General Data Protection Regulation 2016/679 (GDPR) and ICC/ESOMAR International

Code on Market, Opinion and Social Research and Data Analytics. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

J-PP, EM, MI, NR, and PN contributed to the obtention of the funding. J-PP, EM, CL, LT, and ADu problematized and conceptualized the database. J-PP, EM, and KN designed the survey, engaged in the data collection in relation with Toluna, and drafted the manuscript. J-PP, EM, KN, and YA contributed to the construction of the constructs and data analysis. All authors contributed to manuscript revision and read and approved the submitted version.

Funding

The project was made possible with the support of international public and private supports: Taylor's University (Taylor's Flagship Research Program titled *Battle of the Bulge (Obesity): Environmental and Individuals Perspective Studies*—TUFR/2017/003/01), Toulouse University (ISTHIA-TTUC), CNRS-France, LIA 'Food, Cultures and Health, and Observatory of Food Habits of French Dairy Industry (OCHA-CNIEL). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this manuscript, or the decision to submit it for publication.

Acknowledgments

The authors are appreciative of the contibutions from Suan Zee Loh (Senior Project Manager, Toluna), Alex Wong Hoi Mun (Associate Director, Toluna) for the data collection, and from Joseph Chong and Zernaish Junaid (research assistants, Taylor's University) for the data analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2023.1066737/full#supplementary-material

References

1. Drewnowski A, Popkin BM. The nutrition transition: new trends in the global diet. *Nutr Rev.* (1997) 55:31–43. doi: 10.1111/j.1753-4887.1997.tb01593.x

2. Popkin BM, Adair LS, Ng SW. Now and then: the global nutrition transition: the pandemic of obesity in developing countries. *Nutr Rev.* (2012) 70:3–21. doi: 10.1111/j.1753-4887.2011.00456.x

3. Poulain JP. Food in transition: The place of food in the theories of transition. *Sociol Rev.* (2021) 69:702–24. doi: 10.1177/00380261211009092

4. Fournier T, Tibère L, Laporte C, Mognard E, Ismail MN, Sharif SP, et al. Eating patterns and prevalence of obesity: Lessons learned from the Malaysian Food Barometer. *Appetite.* (2016) 107:362–71. doi: 10.1016/j.appet.2016.08.009

5. Poulain JP, Laporte C, Tibère L, Mognard E, Noor IM, Ari Ragavan N. Le Malaysian Food Barometer. L'impact de la modernisation compressée sur l'alimentation. *Sci Société.* (2019) 104:10621. doi: 10.4000/sds.10621

6. Poulain JP, Tibère L, Laporte C, Mognard E, Neethiahnanthan AR, Ashrafzadeh A, et al. Malaysian Food Barometer (MFB): A study of the impact of compressed modernization on food habits. *Malays J Nutr.* (2020) 26:1–17. doi: 10.31246/mjn-2019-0042

7. Chang KS. The second modern condition? Compressed modernity as internalized reflexive cosmopolitization. *Br J Sociol.* (2010) 61:444–64. doi: 10.1111/j.1468-4446.2010.01321.x

8. Roulleau-Berger L. The fabric of Post-Western sociology: ecologies of knowledge beyond the "East" and the "West". *J Chin Sociol.* (2021) 8:10. doi: 10.1186/s40711-021-00144-z

9. Yasawa S. Civilizational encounter, cultural translation and social reflexivity: a note on history of sociology in Japan. In: Kim S guk, Li P, Yasawa S, editors. *A Quest for East Asian Sociologies*. Seoul: Seoul National University Press (2014). p. 131–67.

10. Poulain JP, Tibère L, Mognard E, Laporte C, Fournier T, Ismail MN, et al. The Malaysian food barometer open database: an invitation to study the modernization of malaysian food patterns and its economic and health consequences. *Front Nutr.* (2022) 8:800317. doi: 10.3389/fnut.2021.800317

11. Drewnowski A, Poulain JP. What lies behind the transition from plant-based to animal protein? *AMA J Ethics.* (2018) 20:987–93. doi: 10.1001/amajethics.2018.987

12. Poulain JP. *Manger aujourd'hui : attitudes, normes et pratiques.* 2nd edition. Toulouse: Privat; (2002). p. 235.

13. Poulain JP. The contemporary diet in France: "de-structuration" or from commensalism to "vagabond feeding". *Appetite.* (2002) 39:43–55. doi: 10.1006/appe.2001.0461

14. Poulain JP, Cazes-Valette G, Tibère L, Chiang D, Serra-Mallol C. *Conceptualisation d'un Observatoire Régional de l'Alimentation.* (2012). Available online at: https://hal.archives-ouvertes.fr/hal-03147893 (accessed August 30, 2022).

15. Smith LC, Dupriez O, Troubat N. *Assessment of the Reliability and Relevance of the Food Data Collected in National Household Consumption and Expenditure Surveys.* (2014). Report No.: 008.

16. Nielsen. *What's in our food and our mind: Ingredient and Dining-OutTrends Around the World.* (2016).

17. Gesteiro E, García-Carro A, Aparicio-Ugarriza R, González-Gross M. Eating out of Home: influence on nutrition, health, and policies: a scoping review. *Nutrients.* (2022) 14:1265. doi: 10.3390/nu14061265

18. Lachat C, Nago E, Verstraeten R, Roberfroid D, Van Camp J, Kolsteren P. Eating out of home and its association with dietary intake: a systematic review of the evidence. *Obes Rev Off J Int Assoc Study Obes.* (2012) 13:329–46. doi: 10.1111/j.1467-789X.2011.00953.x

19. Wellard-Cole L, Davies A, Allman-Farinelli M. Contribution of foods prepared away from home to intakes of energy and nutrients of public health concern in adults: a systematic review. *Crit Rev Food Sci Nutr.* (2021) 62:5511–5522. doi: 10.1080/10408398.2021.1887075

20. Wanjek C. *Food At Work: Workplace Solutions for Malnutrition, Obesity and Chronic Diseases*. Geneva: International Labour Organization (2005). p. 448.

21. Mennell S. *All Manners of Food: Eating and Taste in England and France from the Middle Ages to the Present*. University of Illinois Press; (1996). p. 412.

22. Warde A, Paddock J, Whillans J. *The Social Significance of Dining Out : A Study of Continuity and Change*. Manchester: Manchester University Press (2020). p. 296. doi: 10.7765/9781526134769

23. Warde A, Martens L. *Eating Out: Social Differentiation, Consumption and Pleasure*. 2000 edition. Cambridge England New York: Cambridge University Press (2000). p. 260. doi: 10.1017/CBO9780511488894

24. Lund TB, Kjærnes U, Holm L. Eating out in four Nordic countries: National patterns and social stratification. *Appetite*. (2017) 119:23–33. doi: 10.1016/j.appet.2017.06.017

25. Ray K. *The Ethnic Restaurateur*. London; New York, NY: Bloomsbury Academic (2016). p. 264. doi: 10.5040/9781474269414

26. Berris D, Sutton D. *The Restaurants Book: Ethnographies of Where we Eat*. New York: Berg (2007). doi: 10.5040/9781350044913

27. Ascher F. *Le mangeur hypermoderne : Une figure de l'individu éclectique*. Paris: Odile Jacob (2005). p. 330.

28. Fischler C. *L'Homnivore*. Édition de 1993. Paris: Odile Jacob (1990). p. 440.

29. Ferguson PP. Eating out: going out, staying in. In: *A Cultural History of Food in the Modern Age*. Bloomsbury Academic (2014). p. 111–26. doi: 10.5040/9781350044555.ch-005

30. Jacobs M, Scholliers P. *Eating Out in Europe: Picnics, Gourmet Dining and Snacks since the Late Eighteenth Century*. New York, NY: Berg Publishers (2003). p. 416. doi: 10.5040/9781350044838

31. Blake CE, Bisogni CA, Sobal J, Jastran MM, Devine CM. How adults construct evening meals. Scripts for food choice. *Appetite*. (2008) 51:654–62. doi: 10.1016/j.appet.2008.05.062

32. Corbeau JP. Reflections for a Sociological Representation of the Eater. *Soc Sci*. (2021) 10:339. doi: 10.3390/socsci10090339

33. Warde A. Consumption and theories of practice. *J Consum Cult*. (2005) 5:131–53. doi: 10.1177/1469540505053090

34. Douglas M, Gross J. Food and culture: Measuring the intricacy of rule systems. *Soc Sci Inf*. (1981) 20:1–35. doi: 10.1177/053901848102000101

35. Saint-Pol (de) T. Le dîner des français : un synchronisme alimentaire qui se maintient. *Économie Stat*. (2006) 400:45–69. doi: 10.3406/estat.2006.7111

36. Fischler C. Food habits, social change and the nature/culture dilemma. *Soc Sci Inf*. (1980) 19:937–53. doi: 10.1177/053901848001900603

37. Warde A. *The Practice of Eating*. Cambridge: Polity (2016). p. 275.

38. Laporte C, Poulain JP. Restauration d'entreprise en France et au Royaume-Uni. Synchronisation sociale alimentaire et obésité Staff Restaur Fr U K Food Synchronization. *Obes Engl*. (2014) 44:93–103. doi: 10.3917/ethn.141.0093

39. Poulain JP, Laporte C. Obesity and the proper meal at workplace: French and English at the table and (or beyond) the culturalist explanation. In: Gard M, Powell D, Tenorio J, editors. *Routledge Handbook of Critical Obesity Studies*. London: Routledge (2021). p. 177–87. doi: 10.4324/9780429344824-22

40. Poulain JP, Guignard R, Michaud C, Escalon H. Les repas, distribution journalière, structure, lieux et convivialité. In: Escalon H, Bossard C, Beck F, editors. *Baromètre Santé Nutrition 2008*. Paris: INPES (2010). p. 186–211.

41. Arciniegas L. The foodscape of the urban poor in Jakarta: street food affordances, sharing networks, and individual trajectories. *J Urban Int Res Placemaking Urban Sustain*. (2021) 14:272–87. doi: 10.1080/17549175.2021.1924837

42. Birahmatika FS, Chandra DN, Wiradnyani LAA. Determinants of diet quality among mothers of young children in an urban slum area in Jakarta: Mother's age, vegetables availability, and eating out frequency. *Malays J Nutr*. (2022) 28:177–90. doi: 10.31246/mjn-2021-0031

43. Kobayashi S, Asakura K, Suga H, Sasaki S. Diets the T generation S of W on, Group HS Living status and frequency of eating out-of-home foods in relation to nutritional adequacy in 4,017 Japanese female dietetic students aged 18–20 years: A multicenter cross-sectional study. *J Epidemiol*. (2017) 27:287–93. doi: 10.1016/j.je.2016.07.002

44. Lachat C, Khanh LNB, Khan NC, Dung NQ, Anh NDV, Roberfroid D, et al. Eating out of home in Vietnamese adolescents: socioeconomic factors and dietary associations. *Am J Clin Nutr*. (2009) 90:1648–55. doi: 10.3945/ajcn.2009.28371

45. Naidoo N, van Dam RM, Ng S, Tan CS, Chen S, Lim JY, et al. Determinants of eating at local and western fast-food venues in an urban Asian population: a mixed methods approach. *Int J Behav Nutr Phys Act*. (2017) 14:69. doi: 10.1186/s12966-017-0515-x

46. Zeng Q, Zeng Y. Eating out and getting fat? A comparative study between urban and rural China. *Appetite*. (2018) 120:409–15. doi: 10.1016/j.appet.2017.09.027

47. Julier A. Meals: eating at home and "eating out." In: Murcott A, Belasco W, Jackson P, editors. *The Handbook of Food Research*. London: Bloomsbury (2013). doi: 10.5040/9781350042261-ch-0019

48. Murcott A. Lamenting the "decline of the family meal" as a moral panic? Methodological reflections. *Rech Sociol Anthropol*. (2012) 43:97–118. doi: 10.4000/rsa.845

49. Ritzel C, Mann S. Exploring heterogeneity in meat consumption and eating out by using a latent class model. *Br Food J*. (2022) 125:132–144. doi: 10.1108/BFJ-11-2021-1183

50. Poulain JP, Smith W, Laporte C, Tibère L, Ismail MN, Mognard E, et al. Studying the consequences of modernization on ethnic food patterns: Development of the Malaysian Food Barometer (MFB). *Anthropol Food*. (2015) 6:24. doi: 10.4000/aof.7735

51. NVision, *Taylor Nelson Sofres*. Changing Lives in Europe. (2004).

52. Escalon H, Bossard C, Beck F. *Baromètre Santé Nutrition 2008*. Paris: INPES (2010).

53. Wiehl DG. Diets of a group of aircraft workers in Southern California. *Milbank Mem Fund Q*. (1942) 20:329–66. doi: 10.2307/3347838

Frontiers in Nutrition

TYPE Methods
PUBLISHED 10 February 2023
DOI 10.3389/fnut.2023.1011786

OPEN ACCESS

EDITED BY
Rikard Landberg,
Chalmers University of Technology, Sweden

REVIEWED BY
Federica Turati,
National Cancer Institute Foundation (IRCCS), Italy
Natasa Krsto Rancic,
University of Niš,
Serbia
Vasiliki Kolovou,
Hannover University of Applied Sciences, Germany

*CORRESPONDENCE
John C. Mathers
✉ john.mathers@newcastle.ac.uk

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

RECEIVED 04 August 2022
ACCEPTED 26 January 2023
PUBLISHED 10 February 2023

CITATION
Malcomson FC, Parra-Soto S, Lu L, Ho FK, Perez-Cornago A, Shams-White MM, van Zutphen M, Kampman E, Winkels RM, Mitrou P, Wiseman M, Romaguera D, Celis-Morales C, Sharp L and Mathers JC (2023) Operationalisation of a standardised scoring system to assess adherence to the World Cancer Research Fund/American Institute for Cancer Research cancer prevention recommendations in the UK biobank. *Front. Nutr.* 10:1011786.
doi: 10.3389/fnut.2023.1011786

Operationalisation of a standardised scoring system to assess adherence to the World Cancer Research Fund/American Institute for Cancer Research cancer prevention recommendations in the UK biobank

Fiona C. Malcomson[1†], Solange Parra-Soto[2,3†], Liya Lu[4], Frederick K. Ho[2], Aurora Perez-Cornago[5], Marissa M. Shams-White[6], Moniek van Zutphen[7,8], Ellen Kampman[7], Renate M. Winkels[7], Panagiota Mitrou[9], Martin Wiseman[9], Dora Romaguera[10,11], Carlos Celis-Morales[3,12], Linda Sharp[4] and John C. Mathers[1*]

[1]Human Nutrition and Exercise Research Centre, Centre for Healthier Lives, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom, [2]School of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom, [3]School of Cardiovascular and Metabolic Health, University of Glasgow, Glasgow, United Kingdom, [4]Centre for Cancer, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom, [5]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom, [6]Risk Factor Assessment Branch, Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, United States, [7]Division of Human Nutrition and Health, Wageningen University and Research, Wageningen, Netherlands, [8]Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, Netherlands, [9]World Cancer Research Fund International, London, United Kingdom, [10]Health Research Institute of the Balearic Islands (IdISBa), Palma de Mallorca, Spain, [11]CIBER Physiopathology of Obesity and Nutrition (CIBEROBN), Madrid, Spain, [12]Human Performance Lab, Education, Physical Activity and Health Research Unit, University Católica del Maule, Talca, Chile

Introduction: In 2018, The World Cancer Research Fund (WCRF)/American Institute for Cancer Research (AICR) published ten evidence-based Cancer Prevention Recommendations designed to reduce the risk of cancer *via* improved lifestyle behaviours. In 2019, Shams-White and colleagues created the "2018 WCRF/AICR Score" which aimed to standardise how adherence to these recommendations is assessed. The standardised scoring system includes seven of the recommendations concerning weight, physical activity and diet, with an optional eighth recommendation on breastfeeding. To promote transparency and reproducibility, the present paper describes the methodology for operationalisation of the standardised scoring system in the UK Biobank.

Methods: UK Biobank recruited >500,000 individuals aged 37–73years, between 2006 and 2010. In 2021, we held a workshop with experts which aimed to reach consensus on how to operationalise the scoring system using data available within UK Biobank. We used data on anthropometric measurements, physical activity and diet to calculate adherence scores. 24h dietary assessment data were used to measure adherence to the following recommendations: "Eat a diet rich in wholegrains, vegetables, fruit, and beans", "Limit consumption of "fast foods" and other processed foods high in fat, starches or sugars" and "Limit consumption of sugar-sweetened drinks"; food frequency questionnaire data were used to assess adherence to "Limit consumption of red and processed meat" and "Limit alcohol consumption".

Participants were allocated points for meeting, partially meeting or not meeting each recommendation, using cut-offs defined in the standardised scoring system.

Results: At our workshop, discussions included the use of national guidelines to assess adherence to the recommendation on alcohol consumption, as well as challenges faced including defining the adapted ultra-processed food variables. A total score was calculated for 158,415 participants (mean 3.9 points, range 0–7 points). We also describe the methodology to derive a partial 5-point adherence score using data from the food frequency questionnaire in 314,616 participants.

Conclusion: We describe the methodology used to estimate adherence to the 2018 WCRF/AICR Cancer Prevention Recommendations for participants in the UK Biobank, including some of the challenges faced operationalising the standardised scoring system.

KEYWORDS

cancer prevention recommendations, lifestyle, scoring system, cancer risk, lifestyle recommendations

1. Introduction

In 2018, the World Cancer Research Fund (WCRF)/American Institute for Cancer Research (AICR) published ten updated, evidence-based Cancer Prevention Recommendations designed to reduce the risk of cancer, *via* modifiable lifestyle behaviours including promoting healthier diets and physical activity (1). In 2019, Shams-White and colleagues created a scoring system to standardise how adherence to these Cancer Prevention Recommendations is assessed and to provide a framework to improve consistency and comparability across studies (2). The standardised scoring system includes seven of the ten 2018 WCRF/ACIR Cancer Prevention Recommendations concerning weight, physical activity and diet, with an optional eighth recommendation on breastfeeding, and is calculated for individuals. The score creators excluded the recommendation to avoid dietary supplements for cancer prevention and consume nutrients through food alone as this is largely addressed through the other five dietary recommendations, and the recommendation specific to cancer survivors as adherence to this would be derived from a composite measure of the other score components (2). Each recommendation is worth a maximum 1 point for full adherence, half a point for partially meeting the recommendation, and 0 points for not meeting the recommendation, yielding a maximum score of 7 points (8 if the optional recommendation is included).

The standardised scoring system used to assess adherence to the Cancer Prevention Recommendations has been applied, at least in part, in several studies, in countries including The Netherlands (3), Australia (4), United States (5, 6), Spain (7, 8), Italy and Switzerland (9). However, to our knowledge, it has not previously been fully applied in a UK cohort. It is important to assess adherence to lifestyle recommendations and to operationalise such scoring systems across different countries and studies because of the differences in eating patterns, lifestyle and study methods. In the Cancer Lifestyle Prevention Recommendations (CALIPER) UK Study, we aim to investigate relationships between adherence to the Cancer Prevention Recommendations and cancer risk and survival using data from the UK Biobank Study, a prospective cohort study, which recruited over half a million participants across the UK.

The collection of diet and nutrition information presents many challenges, including the selection of the most appropriate method to obtain the highest quality data possible whilst considering the purpose of the data collection and participant burden. The UK Biobank assessed dietary intake using two methods: a touchscreen questionnaire asking 29 diet-related questions (similar to a food-frequency questionnaire (FFQ)) and, in over 200,000 participants, used a web-based 24 h dietary assessment tool “Oxford WebQ” to collect more detailed information (10). A further complexity of the dietary data available within the UK Biobank is that, at the end of the recruitment period, participants were invited to complete another web-based dietary assessment on four occasions between February 2011 and June 2012. Thus, the number of dietary assessments completed by each participant, as well as the dates when these were completed, vary.

Therefore, the aim of this paper is to describe the methodology used to operationalise the standardised scoring system in the UK Biobank, to promote transparency and reproducibility, as encouraged by Shams-White and colleagues (2). We also describe the methodology applied to derive a partial, modified 5-point adherence score using data from the FFQ, for which we have data for a greater number of UK Biobank participants.

2. Materials and equipment

2.1. The UK Biobank study

The UK Biobank is a prospective cohort study which recruited 503,317 individuals from the general population aged 37–73 years, 56% female, from 22 recruitment (henceforth “assessment”) centres across the UK (England, Scotland and Wales) between 2006 and 2010. Full

Abbreviations: AICR, American Institute for Cancer Research; AOAC, Association of Official Analytical Chemists; aUPF, adapted ultra-processed foods; BMI, body mass index; IPAQ, International Physical Activity Questionnaire; MET, metabolic-equivalent; MVPA, Moderate to vigorous physical activity; NDNS, National Diet and Nutrition Survey; PA, physical activity; UPF, ultra-processed food; WCRF, World Cancer Research Fund.

eligibility criteria and recruitment and follow-up methods for UK Biobank are reported on the UK Biobank website (11). The UK Biobank was conducted in accordance with the Declaration of Helsinki and was approved by the North West Multi-Centre Research Ethics Committee (REC reference: 12/NW/03820). At the baseline study visit at an assessment centre, a touchscreen questionnaire was used to collect data on sociodemographic factors, diet and general health, and other participant characteristics, blood samples were collected, and anthropometric measurements were taken, as described below.

2.2. Dietary assessment within the UK Biobank

Two methods of dietary assessment were used within the UK Biobank during different periods of recruitment. Initially, a short FFQ-based approach, including 29 questions on diet and 18 on alcohol, formed part of the baseline touchscreen questionnaire and was completed by all participants at the assessment centre. The questionnaire captured information on the frequency of consumption of major food groups, including fruits and vegetables, fish, meat and cheese, in the last year.

Those participants that were recruited towards the end of the recruitment period (between 2009 and September 2010), also completed a 24 h dietary assessment, the Oxford WebQ (12), which captures information on up to 206 food and 32 drink items. In addition, between February 2011 and June 2012, there were 4 cycles, separated by 3–4 months, in which all participants who had provided a valid email address were invited to complete the 24 h dietary assessment at home. In total, 210,128 participants (42% of the total cohort) completed at least one 24 h dietary assessment and 126,096 (25% of the total cohort) completed at least two (10). Further details about the dietary assessments, including reproducibility and agreement between the two methods have been published (10). There was moderate to substantial agreement in the responses to the main food groups at baseline and approximately 4 years later in 20,348 participants, with κ Coefficients with quadratic weighting between 0.52 (for dried fruit intake) and 0.71 (for poultry intake) (κ values between 0.61–0.80 and between 0.41–0.60 represent substantial and moderate agreement, respectively) (10). Furthermore, there was reliable ranking of participants for all foods and food groups according to the touchscreen questionnaire categorisation when compared with group mean intakes from the 24 h dietary assessment (10).

In the present study, we used data from the 24 h dietary assessment (Oxford WebQ) for intakes of food groups for which there is not much variation from day to day, to assess adherence to the recommendations on the intakes of fruits and vegetables, dietary fibre, ultra-processed foods and sugar-sweetened drinks. We used FFQ data to capture the usual intake of foods not consumed daily, for operationalisation of the red meat and alcohol recommendations where the cut-offs are described as intake "per week", so as not to over or underestimate consumption of these foods.

2.3. Processing of 24h dietary assessment data

For operationalisation of the recommendations using 24 h dietary assessment data, intakes were derived by taking the mean of the completed assessments. We excluded any assessments for which the participant answered "no" to the following question "Would you say that what you ate and drank yesterday was fairly typical for you? (UK Biobank data-field 100020). We also excluded any dietary assessments with extreme energy intakes (based on the "Estimated Nutrients" data-field 100002), using the cut-offs described by Perez-Cornago et al. (13); < 3,347 or > 17,573 kJ per day (< 800 or > 4,200 kcal/per day) for men and < 2092 or > 14,644 kJ per day (< 600 or > 3,500 kcal per day) for women. Perez-Cornago et al. (13) advise that at least two 24 h dietary assessments are used, if possible, when investigating diet-disease associations, as a single dietary assessment is unlikely to reflect habitual intakes, and we will apply this for our future diet-cancer analyses.

We used the updated portion sizes assigned by Perez-Cornago et al. (13) and, where relevant, food composition tables from the UK Nutrient Databank (UKNDB), which includes food composition data most relevant to the time when UK Biobank participants completed the dietary assessments.

2.4. CALIPER UK workshop

The CALIPER UK Study team held a workshop in May 2021 with invited researchers from the WCRF, National Cancer Institute in the United States, Oxford University, Wageningen University, Radboud University and Health Research Institute of the Balearic Islands, including both those who contributed to the creation of the standardised scoring system as well as researchers applying this scoring system in cohorts worldwide. The aim of this workshop was to reach consensus on how to operationalise the scoring system using data available within UK Biobank.

3. Methods

3.1. Operationalisation of the standardised scoring system to assess adherence to the cancer prevention recommendations using UK Biobank data

A summary of the operationalisation of the standardised scoring system, including the scoring system cut-offs and the UK Biobank data used, can be found in Table 1. Operationalisation of each component of the scoring system is described in more detail below.

3.1.1. Be a healthy weight

Anthropometric data on body mass index (BMI; data-field 21001) and waist circumference (data-field 48) were used to operationalise this recommendation. These measurements were collected at the assessment centre at the baseline study visit by trained staff using standard protocols. Weight was measured to the nearest 0.1 kg using the Tanita BC-418 MA body composition analyser and height using a Seca 202 height measure. BMI was calculated from weight and height data using the formula BMI = weight (kg)/height $(m)^2$. Participants within the "normal" BMI range (18.5–24.9 kg/m^2) were classed as fully meeting this sub-recommendation and given a score of 0.5 points. Participants with a BMI classed as "overweight", who met the sub-recommendation partially, were given 0.25 points, and participants who were underweight (<18.5 kg/m^2) or obese (≥30 kg/m^2) were given 0 points.

TABLE 1 Standardised scoring system used to assess adherence to the 2018 WCRF/AICR Cancer Prevention Recommendations, as devised by Shams-White et al. (2).

2018 WCRF/AICR Recommendation	Operationalization of Recommendations	Points	Original scoring system cut-offs
1. Be a healthy weight	**BMI (kg/m^2)**		**BMI (kg/m^2)**
	18.5–24.9	0.5	18.5–24.9
	25–29.9	0.25	25–29.9
	<18.5 or ≥ 30	0	<18.5 or ≥ 30
	Waist circumference (cm (in))		**Waist circumference (cm (in))**
	Men: <94 (<37) Women: <80 (<31.5)	0.5	Men: <94 (<37) Women: <80 (<31.5)
	Men: 94–<102 (37–<40) Women: 80–<88 (31.5–<35)	0.25	Men: 94–<102 (37–<40) Women: 80–<88 (31.5–<35)
	Men: ≥102 (≥40) Women: ≥88 (≥35)	0	Men: ≥102 (≥40) Women: ≥88 (≥35)
2. Be physically active	**Total moderate-vigorous physical activity (MET min/wk)**		**Total moderate-vigorous physical activity (min/wk)[1]**
	≥600	1	≥150
	300–<600	0.5	75–<150
	<300	0	<75
3. Eat a diet rich in wholegrains, vegetables, fruit and beans	**Fruits and vegetables (g/day)**		**Fruits and vegetables (g/day)**
	≥400	0.5	≥400
	200–<400	0.25	200–<400
	<200	0	<200
	Total fibre (g/day) (AOAC definition)		**Total fibre (g/day) (AOAC definition)**
	≥30	0.5	≥30
	15–<30	0.25	15–<30
	<15	0	<15
4. Limit consumption of "fast foods" and other processed foods high in fat, starches or sugars	**Percent of total kcal from ultra-processed foods (aUPFs)**		**Percent of total kcal from ultra-processed foods (aUPFs)**
	Tertile 1 (lowest)	1	Tertile 1 (lowest)
	Tertile 2	0.5	Tertile 2
	Tertile 3 (highest)	0	Tertile 3 (highest)
5. Limit consumption of red and processed meat	**Total red meat and processed meat (g/wk)**		**Total red meat and processed meat (g/wk)**
	Red meat ≤500 and processed meat <21	1	Red meat ≤500 and processed meat <21
	Red meat ≤500 and processed meat 21–<100	0.5	Red meat ≤500 and processed meat 21–<100
	Red meat >500 or processed meat ≥100	0	Red meat >500 or processed meat ≥100
6. Limit consumption of sugar-sweetened drinks	**Total sugar-sweetened drinks (g/day):**		**Total sugar-sweetened drinks (g/day):**
	0	1	0
	>0–≤250	0.5	>0–≤250
	>250	0	>250
7. Limit alcohol consumption	**Total ethanol (UK guidelines) (units/week)**		**Total ethanol (US guidelines) (ethanol, g/day)**
	0	1	0
	≤14 units per week	0.5	>0–≤28 (2 drinks) males and ≤ 14 (1 drink) females
	> 14 units per week	0	>28 (2 drinks) males and > 14 (1 drink) females

[1]Our cut-offs in MET min/wk are equivalent to those in the standardised scoring system in min/wk.

Waist circumference was measured at the natural indent (or umbilicus if the natural indent could not be located) using a Seca 200 tape measure. The creators of the standardised scoring system derived the cut-points for the waist circumference sub-recommendation based on guidelines from the 2018 WCRF/AICR Cancer Prevention Recommendations, the Center for Disease Control and Prevention (14) and the U.S. National Heart, Lung, and Blood Institute (15). Male and female participants with waist circumferences <94 cm and < 80 cm, respectively, fully adhered to the waist circumference sub-recommendation and were given 0.5 points. Male participants with waist circumferences between 94 – 102 cm and female participants with waist circumferences between 80 and 88 cm were scored 0.25 points. Participants with waist circumferences ≥102 cm for males and ≥88 cm for females scored 0 points. The scores for the sub-recommendations on BMI and waist circumference were summed for a maximum score of 1 point for the "be a healthy weight" recommendation.

3.1.2. Be physically active

The cut-offs for this recommendation are based on the WHO and U.S. Physical Activity Guidelines which advise adults to engage in at least 150 min of moderate-intensity aerobic physical activity or at least 75 min of vigorous-intensity physical activity per week (16). These guidelines are in line with those in the UK (17) and, therefore, relevant for a UK-based cohort.

Physical activity was self-reported and data were collected at the assessment centre study visit using a validated short form of the International Physical Activity Questionnaire (IPAQ) (18). The questionnaire asked participants about the frequency, intensity and duration of walking, moderate-intensity and vigorous-intensity physical activity during last month. Time spent in moderate to vigorous physical activity (MVPA), were reported in metabolic equivalents of task per week (MET-h/week). Briefly, the number of minutes per day reported for each level of activity was multiplied by the assigned MET equivalent (4 and 8 MET hours for moderate and vigorous physical activity, respectively) and converted to MET hours per week. Participants undertaking at least 600 MET/min per week (equivalent to 150 min of MVPA per week) were given 1 point, between 300 and 600 MET/min per week (equivalent to 75–150 min of MVPA per week) were given 0.5 points, and less than 300 MET/min per week (equivalent to less than 75 min of MVPA per week) were given 0 points. It should be noted that the cut-offs used in this study, where MVPA data are expressed in MET/min per week, are equivalent to those applied in the standardised scoring system (in min/wk).

3.1.3. Eat a diet rich in wholegrains, vegetables, fruit, and beans

The wholegrains, vegetables, fruit and beans sub-score operationalises two goals pertaining to A. fruit and vegetable and B. fibre intake, described below.

3.1.3.1. Eat a diet high in all types of plant foods including at least five portions or servings (at least 400g or 15oz in total) of a variety of non-starchy vegetables and fruit every day

Data on fruit and vegetable intake in the last 24 h (obtained using 24 h dietary assessment data) were used to assess adherence to this sub-recommendation. Information on the data-fields for the included fruits and vegetables can be found in the Supplementary methods. Due to the standardised scoring system's focus on non-starchy vegetables within the fruits and vegetables sub-component (2), we excluded vegetables such as potatoes (fried, boiled/baked and mashed), sweet potatoes and butternut squash as well as beans and pulses. However, these foods were included when estimating dietary fibre intake for the fibre sub-component (please see below). Further, we did not include guacamole, found within the spreads and sauces category (data-field 20088). This is because the question simply asked whether or not items from a list of 19 spreads and sauces were consumed, so no information is available on the frequency of intake or portion size.

We used the frequency data and standard portion sizes for each food item (13) to calculate the mean intake in grams per day, and summed these to create a total intake of fruits and vegetables in grams per day. Where standard portion sizes were not defined for "Vegetable pieces" (data-field 104070), we allocated this portion as 60 g, which is the same as a standard portion of "Other vegetables" (data-field 104380). Participants who consumed at least 400 g of fruits and vegetables per day were given 0.5 points, those who consumed between 200 – 400 g were given 0.25 points, and those consuming less than 200 g per day scored 0 points.

3.1.3.2. Consume a diet that provides at least 30g/day of fibre from food sources

To operationalise the total dietary fibre intake component of the score, we used the 24 h dietary assessment nutrient data on Englyst fibre intake (data-field 100009). To estimate dietary fibre intake using the Association of Official Analytical Chemists (AOAC) method, we multiplied the dietary fibre variable, derived using the Englyst method, by a conversion factor of 1.33 as described by Lunn and Buttriss (19). Participants consuming ≥30 of dietary fibre per day were given 0.5 points, those consuming between 15 and 30 g per day were given 0.25 points and those consuming less than 15 g per day were given 0 points.

3.1.4. Limit consumption of "fast foods" and other processed foods high in fat, starches or sugars

Shams-White and colleagues captured adherence to the recommendation on "fast" and processed foods using an adapted version of the NOVA classification system, which categorises foods according to the extent and purpose of processing (20). Group 1 of the NOVA classification includes foods that are unprocessed or minimally processed such as fruits, seeds, eggs and milk. Group 2 includes processed culinary ingredients, obtained directly from group 1 foods or from nature by processes such as pressing and milling, for example salt, sugar, vegetable oils and butter. Group 3 are processed foods, for example canned vegetables, salted nuts, smoked meats and cheeses, and unpackaged freshly-made breads. Group 4 are ultra-processed foods (UPFs) and drinks, which typically have five or more ingredients and undergo ultra-processing, for example to produce products that are ready to eat and have hyper-palatability. Examples of UPFs include carbonated (fizzy) drinks, confectionery (e.g., chocolate bars), breakfast cereals, ready meals such as pizzas and chicken nuggets, instant noodles, and mass-produced packaged breads and buns.

Firstly, we categorised the food variables available for the 24 h-dietary assessment data according to the NOVA classification system. An adapted UPF (aUPF) variable was created from the foods classified as Group 4 (ultra-processed), excluding food items already accounted for in other score components (i.e., sugar-sweetened drinks, processed meats and alcohol) to avoid double penalisation as described by Shams-White and colleagues (2, 21). Further information about the foods included, and the allocated portion sizes, can be found in the Supplementary Table 1. We acquired energy values (per 100 g) for these foods from the UKNDB, taking into account the food codes that best reflected the Oxford WebQ items as updated by Perez-Cornago et al., and the percentage allocation of each food code to each Oxford WebQ food item (13). We used these data to determine energy in kcals per standard portion size. Intake frequency data were multiplied by the energy value per standard portion size for each food item, and then summed to generate a variable for total energy intake from aUPFs. The energy intake variable (data-field 100002) was used to calculate the proportion of total daily energy intake from aUPFs.

Since there are no recommended cut-offs or guidelines for the consumption of UPFs, Shams-White and colleagues applied a subjective approach awarding points according to tertiles (2, 21). Participants in the highest tertile, consuming the highest amount of energy from aUPFs, scored 0 points, those in the middle tertile were given 0.5 point and those in the lowest tertile were given 1 point. The use of tertiles (and, hence, an approach which "ranks" individuals) to score this component overcomes discrepancies due to variation in i) aUPFs consumed in different countries and cultures, ii) how different dietary assessment methods affect estimates of aUPF consumption and iii) how aUPF consumption is expressed (for example as a proportion of total energy intake or in grams per day) (2).

3.1.5. Limit consumption of red and processed meat

At our CALIPER UK workshop, we decided that for the red and processed meat recommendation, data expressed as frequency per week would be better than those obtained using the 24 h dietary assessment to capture usual intake, because red and processed meat may not be eaten on a daily basis. Therefore, data from the touchscreen FFQ-based questionnaire were used to operationalise the recommendation for red and processed meat intake.

The meat-related questions in the touchscreen questionnaire asked, "How often do you eat beef (data-field 1369)? (Do not count processed meats)," "How often do you eat lamb/mutton (data-field 1379)? (Do not count processed meats)"and "How often do you eat pork (data-field 1389)? (Do not count processed meats)." Participants were able to answer: "never," "less than once a week," "once a week," "2–4 times a week," "5–6 times a week," "once or more daily," "do not know" or "prefer not to answer." As described by Bradbury et al. (10), the following intake frequencies were applied: "never" = 0, "less than once per week" = 0.5, "once per week" = 1, "2–4 times per week" = 3, "5–6 times per week" = 5.5, "once or more daily" = 7. Data coded as – 1 (corresponding to "do not know") or – 3 (corresponding to "prefer not to answer") were recoded as missing. The intakes of beef (data-field 1369), pork (data-field 1389) and lamb/mutton (data-field 1379) in grams per week were calculated by multiplying the frequency by a standard portion size of 120 g (13). A total red meat intake (g/wk) was calculated by adding each of these meat items together.

To assess processed meat intake, the answers to the question "How often do you eat processed meats (such as bacon, ham, sausages, meat pies, kebabs, burgers, chicken nuggets)?" (data-field 1349) were used. Intake frequencies were applied as described for red meat above. To assign a portion size for processed meats, we used the portion sizes detailed by Perez-Cornago et al. (13), where available (i.e., for bacon, ham, sausages, burgers and nuggets). For chicken nuggets, it was assumed that 56% of the portion was meat, as described by Stewart et al. (22). For pies, an average of the portion sizes of the pies included by Stewart et al. was used (43 g per portion). Because the touchscreen questionnaire asked about a range of processed foods that are typically consumed in different amounts in the UK, a weighted average was calculated using data on consumption of these foods from the National Diet and Nutrition Survey (NDNS) (23). This calculated weighted mean portion size (52.5 g) is similar to the unweighted mean (50.8 g). Details of the processed meat portion size calculations can be found in the Supplementary Table 2.

Participants were classed as fully adherent to this recommendation, and allocated 1 point, if their total red meat intake was 500 g or less per week and processed meat intake was less than 21 g per week. Participants who partially adhered to this recommendation, who consumed ≤500 g red meat per week but 21 g – <100 g of processed meat per week were given 0.5 points. Zero points were given to participants who did not adhere to the recommendation and consumed either >500 g red meat per week or ≥100 g processed meat per week.

3.1.6. Limit consumption of sugar-sweetened drinks

Responses to the question "How much of the following did you drink yesterday?" and the intake of the following drinks were used to assess adherence to the recommendation on sugar-sweetened drinks: carbonated (fizzy) drinks (data-field 100170), fruit drinks, squash or cordial (data-field 100180), dairy/yoghurt-based smoothie (data-field 100230), flavoured milk (data-field 100530), hot chocolate (data-field 100550) and fruit smoothie (data-field 100220). Participants could answer the following: "none", "1/2", "1", "2", "3", "4", "5" or "6+". Values for participants who answered "none" were coded as "0", "1/2" were recoded to "0.5" and "6+" were recoded to "6".

Intakes of these drinks were summed to create a mean sugar-sweetened drink intake per week. Assuming that a standard portion (one glass/carton/250 ml) equates to 250 g, participants who drank on average > 1 sugar-sweetened drink per day were allocated 0 points, those who consumed ≤1 scored 0.5 points, and those who did not consume sugar-sweetened drinks scored 1 point.

In line with other studies that have operationalised the scoring system (19), and following agreement on this approach during our CALIPER UK Workshop, we did not include sugar added to drinks by participants (such as sugar added to tea or coffee). This was decided to avoid unnecessary penalisation for the sugar-sweetened drinks recommendation as the Oxford WebQ does not allow for an accurate calculation of total sugar added to hot drinks. For example, a participant can select that they added a "varied" amount of sugar to teas, infusions and coffees throughout the day, if they drank more than one serving per day.

3.1.7. Limit alcohol consumption

Since Shams-White and colleagues advise use of national guidelines or definitions regarding what constitutes an alcoholic drink (i.e., alcohol content and serving size) (21), we used UK national cut-offs to operationalise the alcohol recommendation (24).

The number of units of each alcoholic drink consumed per week were calculated from responses to the touchscreen questionnaire, i.e., for red wine (data-field 1568), white wine or champagne (data-field 1578), beer or cider (data-field 1588), spirits or liqueurs (data-field 1598), fortified wine (data-field 1608) and other alcoholic drinks such as alcopops (data-field 5364). The serving sizes corresponding to the question and units per serving, from the NHS website,[1] are given in Supplementary Table 3. The number of units per week were calculated by multiplying the frequency of intake per week by the number of units corresponding to each drink. If a participant answered "Do not know" (coded as – 1) or "Prefer not to answer" (coded as – 3), they were coded as missing. The total number of units of alcohol consumed per week were calculated by summing the number of units consumed per week of red wine, white wine or champagne, beer or cider, spirits or liqueurs fortified wine and other alcoholic drinks such as alcopops.

Participants who consumed more than 14 units of alcohol per week were given 0 points, those who consumed >0 – ≤14 units of alcohol per

[1] https://www.nhs.uk/live-well/alcohol-advice/calculating-alcohol-units/

week were given 0.5 points and those adhering fully to the recommendation were given 1 point. Further, participants who answered "never" (coded as "6") or "special occasions only" (coded as "5") to the question "About how often do you drink alcohol?" (data-field 1558) were allocated 1 point. Participants who answered "one to three times a month" (coded as "4") to this question were allocated 0.5 points. This is in line with the further guidance on operationalisation of the standardised scoring system by Shams-White and colleagues, which recommends that, given the limited evidence comparing non-drinkers to very rare drinkers, participants who consume up to one drink per month should be classed as non-drinkers, and those consuming more than one drink per month should fall within the 0.5 and 0 point categories, depending on the amount of alcohol consumed (21).

For future sensitivity analyses, we have also calculated a score using the cut-offs described in the standardised scoring system, based on US guidelines (28 g of ethanol (2 drinks) and 14 g of ethanol (1 drink) per day for males and females, respectively) (2).

3.1.8. Total score calculation

A total score was calculated by summing the points for each of the seven recommendations, with a range of 0–7 points. We were not able to assess adherence to the eighth optional recommendation for mothers to breastfeed their baby, if they can, as these data were not collected by the UK Biobank. A separate 5-point scoring system based on the touchscreen questionnaire was also calculated, and details of this calculation are described in the Supplementary methods.

4. Anticipated results

The methodology (described above) for fully operationalising the standardised scoring system (2) for assessing adherence to the 2018 WCRF/AICR Cancer Prevention Recommendations allows for the calculation of a "total adherence score" for participants in the UK Biobank who completed at least one 24 h dietary assessment and for whom we had data at baseline for BMI, waist circumference, physical activity and diet from the touchscreen questionnaire ($n = 158,415$). The mean total score for these 158,415 participants was 3.9 (SD 1.0) points and ranged from 0 to 7 points. The distribution of total scores for female and male participants is illustrated in Figure 1. This total score will be used to investigate relationships between adherence to the WCRF/AICR Cancer Prevention Recommendations and the risk of, and survival from, cancers, as well as other non-communicable diseases. The CALIPER UK Study will explore potential refinements to the score, such as changing the data type or cut-offs used to assess adherence to a recommendation and the weighting given to each score component in calculating the total score.

In addition, we have devised a 5-point, FFQ-based score using the baseline touchscreen questionnaire data, which allows assessment of adherence to five of the recommendations concerning (i) body weight, (ii) physical activity, (iii) fruits, vegetables and fibre intake, (iv) red and processed meats intake, and (v) alcohol consumption, in a larger subset of UK Biobank participants ($n = 314,616$). The mean FFQ-based score based on this 5-point system is 2.64 (SD 0.91) and there was a strong correlation between the full "total score" and the 5-point score (Spearman's rho = 0.796, $p < 0.0001$, $n = 127,667$). Using the modified 5-point score that is available for a larger subset of participants ($n = 314,616$) will provide greater statistical power for investigations of associations between the adherence score and health-related outcomes.

5. Discussion

We have described the methodology applied, and data used, to operationalise a standardised scoring system for assessing adherence to the 2018 WCRF/AICR Cancer Prevention Recommendations for participants in the UK Biobank prospective cohort study, with the

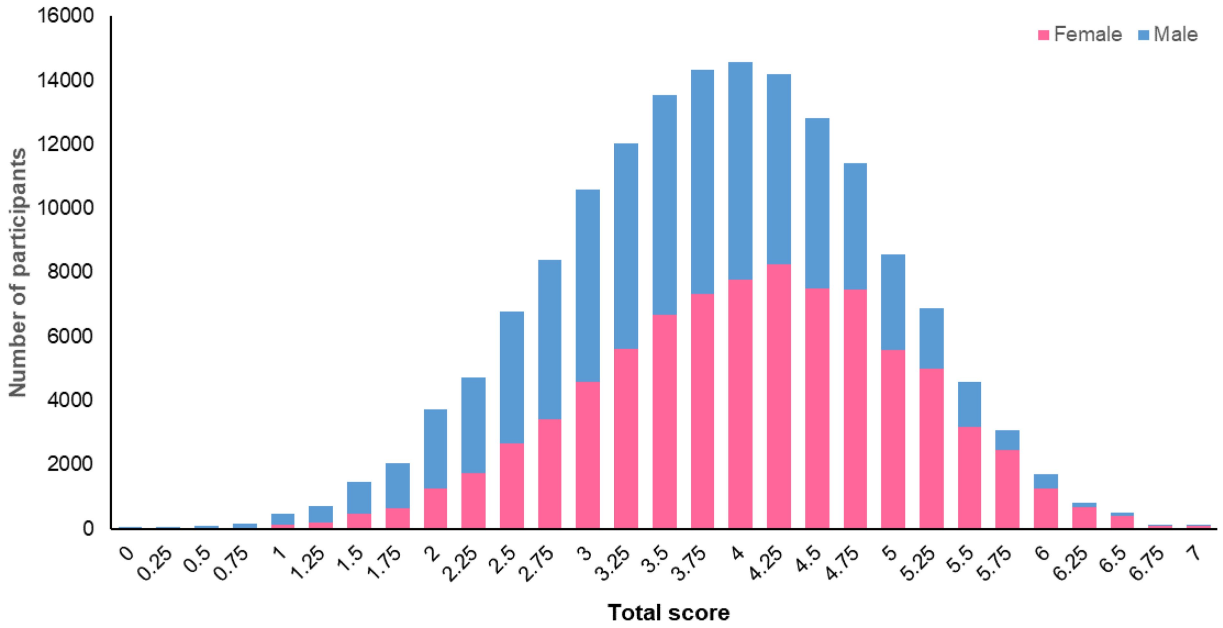


FIGURE 1
Distribution of total adherence scores for male and female UK Biobank participants.

aim of promoting transparency and enhancing reproducibility of findings. Our methodology included running a CALIPER UK Workshop with experts from across the world to allow us to identify how best to operationalise the standardised scoring system with the data available and challenges faced within the UK Biobank. These experts included creators of the standardised scoring system, researchers with substantial experience in processing and using UK Biobank dietary data, members of the WCRF who contributed to the development of the Cancer Prevention Recommendations and researchers who are operationalising the standardised scoring system in other cohorts worldwide. Discussions and decisions made at this Workshop included defining the food variables to be classed as aUPFs using the NOVA classification system and the use of alternative cut-offs based on national guidelines to assess adherence to the recommendation on alcohol consumption.

Using UK Biobank data, we operationalised all components of the score. This is in contrast with some other studies that had missing information on, for example, some of the anthropometric measurements (5, 9) or physical activity (25). As advised by Shams-White and colleagues (2), we applied country-specific guidelines and cut-offs where applicable, for example for the alcohol recommendation where, in the UK, one unit of alcohol contains 8 g of ethanol and both men and women are advised not to drink more than 14 units per week.[2] In future studies, we will explore differences in the total adherence score, including the strengths of associations with cancer incidence, when using other cut-offs including those described by Shams-White et al. (2). In addition, we have created a modified, 5-point touchscreen questionnaire-based score that will allow us to i) compare adherence scores derived from different methods of assessment of nutritional intake and ii) run investigations of associations between adherence score and health outcomes in a larger sample of UK Biobank participants ($n = 314,616$) who do not all have 24 h dietary assessment data. We found a strong and statistically significant correlation between the full "total score" and the 5-point score in 127,667 UK Biobank participants. As already described, some previous studies have also reported calculating partial or modified scores (9, 26).

A strength of this study is the alignment with other analyses of UK Biobank that have used standard portion sizes to estimate intakes of energy and of Englyst fibre from the 24 h dietary assessment data (13). Where standard portion sizes were unavailable, for example for the processed meat food items, we used data from the NDNS to estimate portion sizes. We have applied a conservative approach and minimised use of assumptions throughout. For example, because of the lack of information on intakes of specific foods, e.g., guacamole, we did not include food items from the spreads and sauces category (data-field 20088) in the "fruits and vegetables" sub-recommendation, nor did we include brown sauce and cheese sauce in the aUPF recommendation. However, inclusion of even one serving of a standard portion size of 26 g of guacamole per day is unlikely to make a substantial difference to participant scores for this sub-recommendation.

As advised by Shams-White and colleagues (2), we have considered the utility of the dietary data obtained from the two assessment methods (touchscreen questionnaire versus 24 h dietary assessment) in the UK Biobank to operationalise each score component. As a consequence, we have used a combination of the two assessment methods, with the data collected at different time points and also over time for multiple 24 h dietary assessment, which is a limitation of our study. For some foods not consumed daily, such as red and processed meats, we used data from the touchscreen questionnaire, whereas for some items such as sugar-sweetened beverages, we used the 24 h dietary assessment data because information on intake of these beverages was not collected in the touchscreen questionnaire. Nonetheless, Bradbury and colleagues have observed good agreement between the dietary data collected using the two approaches and have shown that the touchscreen questionnaire method reliably ranks participants according to the intake of main foods and food groups (10). Furthermore, there was good reproducibility between estimates of habitual diet estimated using responses to the touchscreen questionnaire at baseline and those completed 4 years later at the repeat assessment centre visit, suggesting no major long-term changes in diet during this period (10). However, participants who completed the repeat touchscreen questionnaire or at least one of the follow-up 24 h dietary assessments were more likely to be more educated and less likely to smoke compared with the full UK Biobank cohort (10).

In our future analyses we will consider adjusting for such sociodemographic factors; however, this is more of a concern for external generalisability rather than for internal validity of our findings. Although completion of two 24 h dietary assessments may not be sufficient to capture habitual intakes precisely, including participants with data from at least two 24 h dietary assessments is a reasonable compromise to avoid losing too many participants in future studies of associations with cancer and other health outcomes. When compared with the general population, participants in UK Biobank were less likely to be obese, drank less alcohol and were less likely to be smokers (27), thus our findings may not be generalisable to all adults in the UK.

Lastly, this analysis utilised self-reported data for some score components, including the dietary and physical activity data, which may be prone to recall bias or misreporting. However, a strength of this study is that the anthropometric measurements made in the UK Biobank and used to assess adherence to the recommendation to maintain a healthy body weight were collected by trained staff using standardised procedures at the assessment centre visit.

In conclusion, we have used robust methodology to apply the standardised scoring system created by Shams-White and colleagues (2) to assess adherence to the WCRF/AICR Cancer Prevention Recommendations, within the UK Biobank. Here, we are the first to describe in detail how we have operationalised the adherence scoring system in order to allow for transparency and reproducibility and aid interpretation of our future findings. Since UK Biobank is an internationally significant cohort study that is being used extensively to investigate links between lifestyle behaviours and health-related outcomes, such as cancer, we hope that this will be useful for other researchers using UK Biobank data, as well as to provide guidance on operationalising the scoring system in other studies. Our future work will investigate relationships between adherence score and cancer risk and survival within this UK cohort. In addition, as encouraged by Shams-White and colleagues (2, 21), we will explore whether assigning different weightings to each recommendation within the scoring system affects its utility. We will also investigate the impact of changes in how each component is assessed, for example using alternative measures of adiposity to assess adherence to the "be a healthy body weight" recommendation (28), on the scoring system.

[2] https://www.nhs.uk/live-well/alcohol-support/calculating-alcohol-units/

Data availability statement

The datasets presented in this article are not readily available because data are available upon request from UK Biobank (www.ukbiobank.ac.uk). Requests to access the datasets should be directed to john.mathers@newcastle.ac.uk.

Ethics statement

The studies involving human participants were reviewed and approved by The UK Biobank was conducted in accordance with the Declaration of Helsinki and was approved by the North West Multi-Centre Research Ethics Committee (REC reference: 12/NW/03820). The patients/participants provided their written informed consent to participate in this study.

Author contributions

FM, SP-S, CC-M, LS, and JM contributed to the conception and design of the study. FM, SP-S, LL, FH, AP-C, MS-W, MZ, EK, RW, GM, MW, DR, CC-M, LS, and JM attended and actively contributed to the CALIPER UK Workshop, which aided the design and execution of the study. FM wrote the first draft of the manuscript. JM, SP-S, LS, FH, CC-M, AP-C, MS-W, MZ, EK, RW, DR, and GM revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by grant number IIG_FULL_2020_032 from the Wereld Kanker Onderzoek Fonds (WKOF), as part of the World Cancer Research Fund International grant programme. SPS received financial support from the Chilean Government for their PhD (ANID-Becas Chile, project 72200012).

Acknowledgments

The authors are very grateful to the UK Biobank study participants and research team. This research has been conducted using the UK Biobank Resource under application ID 69371.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2023.1011786/full#supplementary-material

References

1. Research WCRFAIfC. 2018. Diet, nutrition, physical activity and cancer: a global perspective. Continuous update project expert report. Available from: http://dietandcancerreport.org (Accessed December 9, 2021).

2. Shams-White, MM, Brockton, NT, Mitrou, P, Romaguera, D, Brown, S, Bender, A, et al. Operationalizing the 2018 World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) cancer prevention recommendations: a standardized scoring system. *Nutrients*. (2019) 11:1572. doi: 10.3390/nu11071572

3. van Zutphen, M, Boshuizen, HC, Kenkhuis, MF, Wesselink, E, Geijsen, A, de Wilt, JHW, et al. Lifestyle after colorectal cancer diagnosis in relation to recurrence and all-cause mortality. *Am J Clin Nutr*. (2021) 113:1447–57. doi: 10.1093/ajcn/nqaa394

4. Tollosa, DN, Holliday, E, Hure, A, Tavener, M, and James, EL. Multiple health behaviors before and after a cancer diagnosis among women: a repeated cross-sectional analysis over 15 years. *Cancer Med*. (2020) 9:3224–33. doi: 10.1002/cam4.2924

5. Zhang, ZQ, Li, QJ, Hao, FB, Wu, YQ, Liu, S, and Zhong, GC. Adherence to the 2018 World Cancer Research Fund/American Institute for Cancer Research cancer prevention recommendations and pancreatic cancer incidence and mortality: a prospective cohort study. *Cancer Med*. (2020) 9:6843–53. doi: 10.1002/cam4.3348

6. Shams-White, MM, Brockton, NT, Mitrou, P, Kahle, LL, and Reedy, J. The 2018 World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) score and all-cause, cancer, and cardiovascular disease mortality risk: a longitudinal analysis in the NIH-AARP diet and health study. *Curr Dev Nutr*. (2022) 6:6006009. doi: 10.1093/cdn/nzac096

7. Olmedo-Requena, R, Lozano-Lorca, M, Salcedo-Bellido, I, Jimenez-Pacheco, A, Vazquez-Alonso, F, Garcia-Caballos, M, et al. Compliance with the 2018 World Cancer Research Fund/American Institute for Cancer Research cancer prevention recommendations and prostate cancer. *Nutrients*. (2020) 12:768. doi: 10.3390/nu12030768

8. Barrios-Rodriguez, R, Toledo, E, Martinez-Gonzalez, MA, Aguilera-Buenosvinos, I, Romanos-Nanclares, A, and Jimenez-Moleon, JJ. Adherence to the 2018 World Cancer Research Fund/American Institute for Cancer Research recommendations and breast cancer in the SUN project. *Nutrients*. (2020) 12:2076. doi: 10.3390/nu12072076

9. Turati, F, Dalmartello, M, Bravi, F, Serraino, D, Augustin, L, Giacosa, A, et al. Adherence to the World Cancer Research Fund/American Institute for Cancer Research recommendations and the risk of breast cancer. *Nutrients*. (2020) 12:607. doi: 10.3390/nu12030607

10. Bradbury, KE, Young, HJ, Guo, W, and Key, TJ. Dietary assessment in UK biobank: an evaluation of the performance of the touchscreen dietary questionnaire. *J Nutr Sci*. (2018) 7:e6. doi: 10.1017/jns.2017.66

11. UKBiobank. 2007 UK biobank. Protocol for a large-scale prospective epidemiological resource. Available from: http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf

12. Liu, B, Young, H, Crowe, FL, Benson, VS, Spencer, EA, Key, TJ, et al. Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. *Public Health Nutr*. (2011) 14:1998–2005. doi: 10.1017/S1368980011000942

13. Perez-Cornago, A, Pollard, Z, Young, H, van Uden, M, Andrews, C, Piernas, C, et al. Description of the updated nutrition calculation of the Oxford WebQ questionnaire and comparison with the previous version among 207,144 participants in UK biobank. *Eur J Nutr*. (2021) 60:4019–30. doi: 10.1007/s00394-021-02558-4

14. CDC (2022). Center for Disease Control and Prevention (CDC) healthy weight: assessing your weight. Available from: https://www.cdc.gov/healthyweight/assessing/index.html (Accessed July 7, 2022).

15. NIH (2022). National Heart, Lung, and Blood Institute: assessing your weight and health risk. Available from: https://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm (Accessed July 7, 2022).

16. U.S. Department of Health and Human Services. *Physical activity guidelines for Americans. 2nd Edn*. Washington, DC: U.S. Department of Health and Human Services. (2018).

17. Department of Health Alcohol Guidelines Review 2019. *UK Chief Medical Officers' Physical Activity Guidelines*. Available from: https://www.gov.uk/government/publications/physical-activity-guidelines-uk-chief-medical-officers-report

18. Craig, CL, Marshall, AL, Sjostrom, M, Bauman, AE, Booth, ML, Ainsworth, BE, et al. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc.* (2003) 35:1381–95. doi: 10.1249/01.MSS.0000078924.61453.FB

19. Lunn, J, and Buttriss, JL. Carbohydrates and dietary fibre. *Nutr Bull.* (2007) 32:21–64. doi: 10.1111/j.1467-3010.2007.00616.x

20. Monteiro, CA, Cannon, G, Levy, R, Moubarac, JC, Jaime, P, Martins, AP, et al. NOVA. The star shines bright. *World Nutrition*. (2016) 7:28–38.

21. Shams-White, MM, Romaguera, D, Mitrou, P, Reedy, J, Bender, A, and Brockton, NT. Further guidance in implementing the standardized 2018 World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) score. *Cancer Epidemiol Biomark Prev*. (2020) 29:889–94. doi: 10.1158/1055-9965.EPI-19-1444

22. Stewart, C, Frie, K, Piernas, C, and Jebb, SA. Development and reliability of the Oxford meat frequency questionnaire. *Nutrients*. (2021) 13:922. doi: 10.3390/nu13030922

23. NDNS. National Diet and Nutrition Survey. Headline results from years 1,2,3 and 4 (combined) of the rolling Programme (2008/2009–2011/12). Table 5.2. 2016. Available at: https://www.gov.uk/government/publications/physical-activity-guidelines-uk-chief-medical-officers-report

24. Alcohol Guidelines Review – Report from the Guidelines Development Group to the UK Chief Medical Officers 2016. *Report from the guidelines development group to the UK chief medical officers.*

25. Geijsen, A, Kok, DE, van Zutphen, M, Keski-Rahkonen, P, Achaintre, D, Gicquiau, A, et al. Diet quality indices and dietary patterns are associated with plasma metabolites in colorectal cancer patients. *Eur J Nutr*. (2021) 60:3171–84. doi: 10.1007/s00394-021-02488-1

26. Solans, M, Romaguera, D, Gracia-Lavedan, E, Molinuevo, A, Benavente, Y, Saez, M, et al. Adherence to the 2018 WCRF/AICR cancer prevention guidelines and chronic lymphocytic leukemia in the MCC-Spain study. *Cancer Epidemiol*. (2020) 64:101629. doi: 10.1016/j.canep.2019.101629

27. Fry, A, Littlejohns, TJ, Sudlow, C, Doherty, N, Adamska, L, Sprosen, T, et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am J Epidemiol*. (2017) 186:1026–34. doi: 10.1093/aje/kwx246

28. Parra-Soto, S, Malcomson, FC, Ho, FK, Pell, JP, Sharp, L, Mathers, JC, et al. Associations of a body shape index (ABSI) with cancer incidence, all-cause, and at 23 sites-findings from the UK biobank prospective cohort study. *Cancer Epidemiol Biomark Prev*. (2022) 31:315–24. doi: 10.1158/1055-9965.EPI-21-0591

frontiers | Frontiers in Nutrition

TYPE Methods
PUBLISHED 16 February 2023
DOI 10.3389/fnut.2022.1013516

Check for updates

OPEN ACCESS

EDITED BY
Alessandra Durazzo,
Council for Agricultural Research
and Economics, Italy

REVIEWED BY
Ido Toxopeus,
National Institute for Public Health
and the Environment, Netherlands
Liangzi Zhang,
Quadram Institute, United Kingdom

*CORRESPONDENCE
Mahsa Jessri
✉ mahsa.jessri@ubc.ca

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to
Nutrition Methodology,
a section of the journal
Frontiers in Nutrition

RECEIVED 07 August 2022
ACCEPTED 13 December 2022
PUBLISHED 16 February 2023

CITATION
Gilbert SZ, Morrison CL, Chen QJ, Punian J, Bernstein JT and Jessri M (2023) Algorithm-based mapping of products in a branded Canadian food and beverage database to their equivalents in Health Canada's Canadian Nutrient File. *Front. Nutr.* 9:1013516. doi: 10.3389/fnut.2022.1013516

Algorithm-based mapping of products in a branded Canadian food and beverage database to their equivalents in Health Canada's Canadian Nutrient File

Sappho Z. Gilbert[1,2†], Conor L. Morrison[3†], Qiuyu J. Chen[2†], Jesman Punian[2], Jodi T. Bernstein[2] and Mahsa Jessri[2,4*]

[1]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, United States, [2]Food, Nutrition, and Health Program, Faculty of Land and Food Systems, The University of British Columbia, Vancouver, BC, Canada, [3]Department of Statistics, Faculty of Science, The University of British Columbia, Vancouver, BC, Canada, [4]Centre for Health Services and Policy Research (CHSPR) and Health Services and Policy (HSP), Faculty of Medicine, The University of British Columbia, Vancouver, BC, Canada

Introduction: There is increasing recognition of the value of linking food sales databases to national food composition tables for population nutrition research.

Objectives: Expanding upon automated and manual database mapping approaches in the literature, our aim was to match 1,179 food products in the Canadian data subset of Euromonitor International's Passport Nutrition to their closest respective equivalents in Health Canada's Canadian Nutrient File (CNF).

Methods: Matching took place in two major steps. First, an algorithm based on thresholds of maximal nutrient difference (between Euromonitor and CNF foods) and fuzzy matching was executed to offer match options. If a nutritionally appropriate match was available among the algorithm suggestions, it was selected. When the suggested set contained no nutritionally sound matches, the Euromonitor product was instead manually matched to a CNF food or deemed unmatchable, with the unique addition of expert validation to maximize meticulousness in matching. Both steps were independently performed by at least two team members with dietetics expertise.

Results: Of 1,111 Euromonitor products run through the algorithm, an accurate CNF match was offered for 65% of them; missing or zero-calorie data precluded 68 products from being run in the algorithm. Products with 2 or more algorithm-suggested CNF matches had higher match accuracy than those with one (71 vs. 50%, respectively). Overall, inter-rater agreement (reliability) rates were robust for matches chosen among algorithm options (51%) and even higher regarding whether manual selection would be required

(71%); among manually selected CNF matches, reliability was 33%. Ultimately, 1,152 (98%) Euromonitor products were matched to a CNF equivalent.

Conclusion: Our reported matching process successfully bridged a food sales database's products to their respective CNF matches for use in future nutritional epidemiological studies of branded foods sold in Canada. Our team's novel utilization of dietetics expertise aided in match validation at both steps, ensuring rigor and quality of resulting match selections.

KEYWORDS

database mapping, nutritional surveillance and monitoring, food composition tables (FCTs), food supply, public health nutrition, fuzzy matching, Canada, Canadian Nutrient File

1. Introduction

In recent decades, food marketing and retail databases have been revisited as a largely untapped source of low-bias, high-quality data for researching trends in consumer health and nutrition (1). Such databases can also be utilized for the study of front-of-pack labeling, marketing and advertising to children, and the implementation and surveillance of dietary guidelines (2). For food and beverage manufacturing industries, these data on the nutrient content of their products and associated sales can guide healthy eating initiatives and even product reformulations (3). When used with food composition information, these data can be especially useful for health professionals and scholars, including clinicians, dietitians, and epidemiologists studying the impact of population diet and nutrition on the prevalence and incidence of certain diet-related chronic conditions (4). Innovative approaches to monitoring public health and community nutrition are particularly critical and can be enabled using these kinds of datasets (1, 5).

Researchers are increasingly seeing the value for public health nutrition and nutritional epidemiology in linking food retail, manufacturing, and marketing datasets to food composition tables. Digital data on food sold in stores–such as *via* electronic point-of-sale systems–commonly include information on the quantities sold, price paid, and promotion status (4–7). These data are ubiquitous in the food retail industry and are already utilized in product development and marketing (4).

Marketing analytics databases that track trends in food sales (including for branded products) are similarly enticing for their potential use as longitudinal observation data and in scenario modeling for food policy and public health nutrition interventions; examples include multinational marketing analytics companies like Euromonitor, Kantar Worldpanel, GlobalData, Nielsen, and Gesellschaft für Konsumforschung (GfK) (8). However, realizing the potential of their sales data for population nutrition research and policy depends on the availability of corresponding food composition data. Some databases provide extensive nutrient data for their products. Nielsen (through a sister company, Brand Bank) and Kantar have nutrient data available for their tracked products (8). Meanwhile, in cases where these databases only have partial food composition data (like Euromonitor, which provides product information for energy and 7 nutrients) or none at all (such as GfK or Global Data), matching to a national, commercial, or other food composition database is likely needed (1, 8).

These marketing companies' data have recently been used to evaluate health-related interventions in diverse contexts across the globe, such as Denmark's saturated fat tax and nutrition assistance programs in the United States (1). These databases may also be used to overcome limitations of national food composition tables such as the Canadian Nutrient File (CNF). The CNF forms the basis for national health surveys in Canada and contains mostly generic aggregates of foods, is not systematically or consistently updated, and is largely based on data from the United States Department of Agriculture (USDA) (9, 10). Since some of the aforementioned companies include data on food composition of specific branded foods and beverages, their data could be a beneficial supplement to the CNF for use in population dietary surveillance. Even for the exact foods for which the brand is known, CNF information on these branded products may be outdated and not reflective of the current Canadian food supply. Additionally, many CNF foods are generic estimates of products available on the market, with some being the average of many types of that food; while this can be a boon to those seeking a more complete nutrient profile and representative nutrient data for a non-brand-specific food, it is conversely a limitation for those focused on specific branded products.

Prior studies have reported various approaches to linking food retail and marketing datasets to food group and composition data for the study of population health and nutrition, including both manual and algorithm-based approaches (1, 4, 5, 6). The traditional way of mapping a database of products to their respective food composition (nutrient content) is by manual matching; generally, this means a product is matched to a food or beverage item code in a food composition database, which then pulls in those nutrient data for use with that now-linked product (8). This can be very resource-intensive and error-prone, as it tends to require significant amounts of time and effort to match food items and/or categories manually. Brinkerhoff et al. (4) attempted automated mapping of foods in a supermarket dataset to their nutritional equivalents in the USDA's Standard Reference (USDA-SR) database but found a relatively small number of successful matches due to differences in food naming strategies and categorization conventions; as a result, manual matching was performed in full.

As long as new products enter the food supply, matching will remain a data maintenance problem–a reality that underscores the need for efficient, replicable, and adaptable matching protocols. An effective algorithm that can (at least) partially automate matching of food products to their closest equivalent in food composition tables would alleviate some of these manual matching burdens (especially in large databases). Other scholars have been able to create algorithms for foods across databases that appear to lead to good matches (8, 11, 12). However, most such articles do not report expert validation of these matches for their compositional closeness. The aim of this work was thus to design a primarily automated, dietetics expert-validated methodology for matching food and beverage products in the Euromonitor Passport Nutrition's Canada subset to their equivalents in the CNF.

2. Methods

Products in the Canada-wide Euromonitor data subset ("Euromonitor database") were linked with those in the Canadian Nutrient File (CNF) using the following methodology (10). Our Bureau of Nutritional Sciences (BNS)-integrated Nutrition and Fuzzy Match (BiNFM) algorithm was coded in R (13). Fuzzy string matching refers to a class of algorithms designed to determine the similarity of two unequal strings. In our case, we used the "partial token sort ratio" fuzzy matching algorithm that is implemented by the fuzzywuzzyR package in R (14), which ports the fuzzywuzzy package from Python (15). This particular fuzzy matching algorithm takes two strings as input and then outputs a score from 0 (indicating no similarity) to 100 (indicating near exact similarity).

Two research groups have developed algorithms to automate a similar database mapping process and reported their methods and reflections (11, 12). Elements of both of their approaches were found to be applicable to our work and were adapted to fit the nature and challenges of our matching effort. In short, motivated by Tran et al., we restricted our algorithm to only suggest matches where the Euromonitor product and the CNF food(s) shared a food category in common; inspired by Lamarine et al., we also employed fuzzy string matching in our algorithm (11, 12). Divergences from these previous works included our use of nutrient-based thresholds in the algorithm and the addition of the dietetics expert validation of final match selections, which are described in greater detail below.

2.1. Overview of databases

2.1.1. Euromonitor data subset of branded Canadian consumer food products sold between 2014 and 2018

Euromonitor International Ltd. (London, UK) is a market research company whose Passport Nutrition database offers nutrition data for products sold in different countries worldwide. We acquired a subset of this data that contained major branded foods and beverages sold in Canada from 2014 to 2018. As visualized in **Figure 1**, this dataset included 1,179 products from two main categories (Packaged Food and Soft Drinks) across 210 subcategories. A single Euromonitor product consists of two parts: (1) the subcategory it belongs to, and (2) the brand name. For example, "Children's Breakfast Cereals" (subcategory) + "President's Choice" (brand) = "Children's Breakfast Cereals, President's Choice" (= 1 Euromonitor product). Euromonitor provides definitions for each subcategory that outline the types of foods and leading market brands. For each product, data for energy and 7 nutrients are reported: carbohydrate, protein, total dietary fat, saturated fat, sugar, fiber, and salt (which was converted to sodium: grams of salt x 393 = milligrams of sodium).

2.1.2. Canadian Nutrient File (CNF)

The CNF is a national food composition database. Its latest version from 2015 was used in this work and contains 5,690 Canadian foods and data for up to 152 nutrients for each (10). The majority of CNF food names are presented as generic food descriptions (e.g., "Cheese, blue"), with a minority of foods containing brand-specific information in their names (e.g., "Cereal, ready to eat, Cheerios, General Mills"). CNF foods exist within 23 broadly named food categories (e.g., Dairy and Egg Products, Breakfast Cereals, and Nuts and Seeds).

2.1.3. Bureau of Nutritional Sciences (BNS) food groups

Bureau of Nutritional Sciences is a food category system that contains a granular classification scheme developed by Health Canada for categorizing foods (13). Due to the variability in food categorization between the Euromonitor database and the CNF, BNS food groups were utilized as a bridging tool between these two databases. Each Euromonitor product and CNF food had a BNS food group assigned to them manually by dietetics experts; while there are 78 such food groups in the BNS, only 50 were used in this project (as 28 were excluded for being dishes rather than individual foods). To optimize its efficiency and accuracy, our algorithm was designed to only offer potential

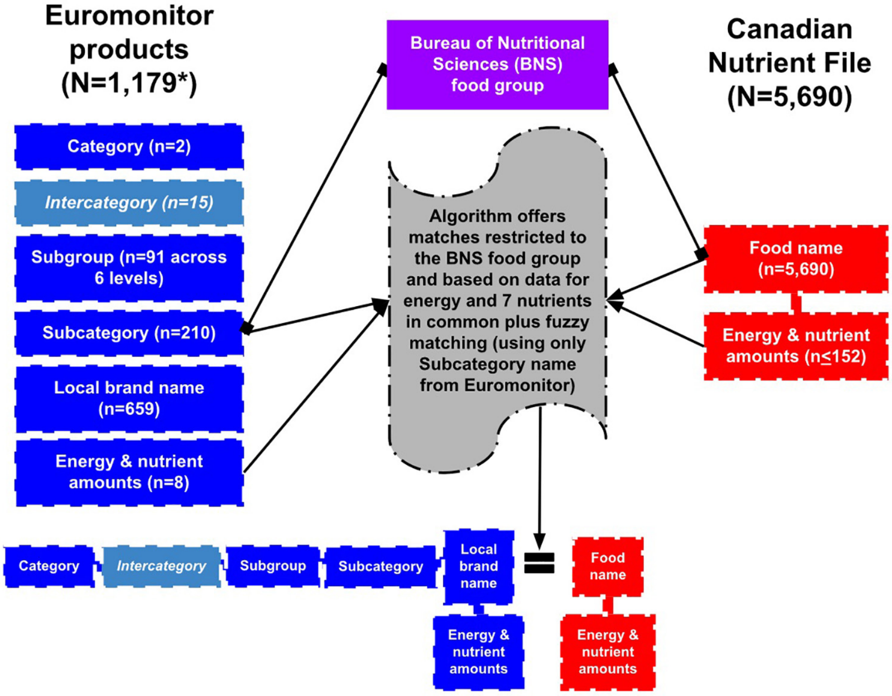

Euromonitor products (N=1,179*)
Category (n=2)
Intercategory (n=15)
Subgroup (n=91 across 6 levels)
Subcategory (n=210)
Local brand name (n=659)
Energy & nutrient amounts (n=8)
Bureau of Nutritional Sciences (BNS) food group
Algorithm offers matches restricted to the BNS food group and based on data for energy and 7 nutrients in common plus fuzzy matching (using only Subcategory name from Euromonitor)
Canadian Nutrient File (N=5,690)
Food name (n=5,690)
Energy & nutrient amounts (n≤152)
Category
Intercategory
Subgroup
Subcategory
Local brand name
Energy & nutrient amounts
=
Food name
Energy & nutrient amounts

FIGURE 1
Overview of the architectures of Euromonitor's Canadian data subset in Passport Nutrition and the Canadian Nutrient File (CNF), with an emphasis on the key variables used in their database mapping. *Of the 1,179 total Euromonitor products, only 1,111 were ultimately able to be run through the algorithm.

matches between Euromonitor products and CNF foods that share the same BNS food group.

2.2. BiNFM algorithm design

Our algorithm's matching relies on the names of Euromonitor subcategories and CNF foods, BNS food groups shared in common, and the differences in the nutrients of Euromonitor products and CNF foods. Figure 1 visually depicts our database mapping approach as it relates to algorithm design, which is described step-by-step below.

Given a particular Euromonitor product, the algorithm sifts (in rounds that we fittingly also term "sifts") through all potential CNF foods to produce a list of suggested CNF matches. Initially, this list consists of all CNF foods that share a BNS group with the Euromonitor product. The sifting process then applies up to five filters to this list of potential matches to arrive at the algorithm-suggested match options. Two of these filtering steps are marked as optional, as they were only employed for a subset of sifts in our study. Including these optional filters will generally provide a narrower, plausibly more specific list of suggested matches but may result in a lower sensitivity. The cost-benefit analyses of which optional filters to use will vary on a case-by-case basis. The five filtering steps are:

1. Only CNF foods that have macronutrient (carbohydrate, protein, or total fat) contents that differ from the Euromonitor product by an amount falling below a predefined threshold are kept as potential matches. The difference in the content of a nutrient X as a proportion of calories between foods in Euromonitor (E) and in the CNF (C) is equal to = (Nutrient X as a % of Calories in E) − (Nutrient X as a % of Calories in C). For example, if 30% of the calories in a given Euromonitor product are from carbohydrates, and 20% of the calories in a given CNF food are from carbohydrates, then the difference in carbohydrates as a proportion of calories is 10%. The thresholds we used are described later in this subsection.

Differences in nutrients as a proportion of calories are used to better account for differences in nutrient contents for foods across a large range of total caloric contents. This approach was found to be more robust than either absolute or relative differences in the grams of nutrients when simultaneously wanting to compare nutrient contents of high-calorie and very low-calorie items. As an example of absolute differences (in grams of nutrient) being less robust, consider a difference of 5 g in carbohydrates. A 5 g difference might be small for a product with 100 g of carbohydrates, but this is large for one with only 10 g of total carbohydrates. Using relative differences is non-robust for low-calorie items. For example, a product with only 1 g total carbohydrate that faces a 1 g difference with another food will equal a 100% difference—even though the absolute difference is only a mere 1 g per 100 g of product. Differences as a proportion of calories were thereby found to be more useful for both low and high nutrient foods, as the nutrients in each product are normalized by their energy content.

2. Only CNF foods that have fiber, saturated fat, and sugar contents as a proportion of calories that differ from the Euromonitor product by an amount falling below a predefined threshold are kept as potential matches.

3. If the Euromonitor product has a non-zero sodium content, the relative difference in sodium content is computed between the Euromonitor product and the CNF food and this quantity is compared to a third threshold. The relative difference in the sodium content of Euromonitor (E) vs. CNF (C) is computed as = (Sodium content of E in mg − Sodium content of C in mg) ÷ (Sodium content of E in mg).

4. (*Optional; only used in 1 of our 4 sifts*) Only CNF foods whose fuzzy match score with the Euromonitor product exceed a set threshold are kept as potential matches.

5. (*Optional; used in 3 of our 4 sifts*) Only CNF foods whose fuzzy match score with the Euromonitor product are equal to the largest fuzzy match score between the product and all potential CNF matches remaining from the previous filtering step are kept as potential matches.

Henceforth we will refer to the chosen thresholds and optional steps as "sift parameters," or more leniently as "parameters." To ensure that there were matches suggested for each Euromonitor product, several sifts were run with a variety of parameters. If a Euromonitor item had fewer suggested matches than desired, it was included in a subsequent sift with more lenient sift parameters. The parameters for each of the run sifts are detailed in **Table 1**. For example, in the *"First"* sift, the difference threshold for protein, carbohydrates, and total dietary fat as proportions of calories was set to 20%; for fiber, saturated fat, and sugar as proportions of calories was set to 10%; and for the relative difference threshold for sodium contents was set to 50%. An additional sift, "*First+*," used more lenient matching parameters, and was applied to the Euromonitor products which had single CNF matches from "*First*." This was done to increase the number of potential CNF matches and the overall sensitivity of the algorithm.

The parameters in **Table 1** were selected based on two approaches. The first approach was to minimize the possible error in suggested matches to that of the assigned threshold (e.g., at most a 20% difference in macronutrient content). The second approach was to run the matching algorithm several times with different parameters to obtain a sufficient number of suggested CNF matches for each Euromonitor product. In general, higher thresholds would result in a greater number of suggested matches, but at the cost of a diminished specificity. **Figure 2** visually demonstrates this by plotting the square roots (for easier readability) of the numbers of matches for all Euromonitor items at several candidate thresholds. Based on **Figure 2**, we heuristically decided that the difference in the number of matches suggested was most consequential when we changed the threshold for carbohydrates, proteins, and fats from 10 to 20% as well as from 50 to 100% (with thresholds for fiber, saturated fat, and sugar set to 40% and the sodium threshold set to 50%). This decision was based primarily on comparing the size of the differences in the median number and maximum number of matches for each of these thresholds. **Figure 3** shows a plot demonstrating potential sodium thresholds when the thresholds for carbohydrate, protein, and fat were set to 20% and the thresholds for fiber, saturated fat, and sugar were set to 10%. Using a sodium threshold of 50% provided more matches than smaller thresholds, and a sodium threshold of 100% greatly increased the number of matches. With a similar reasoning as before, we used **Figure 3** as a motivation for our choice of sodium threshold in the matching algorithm. This approach of parameter selection is a combination of numerical heuristics and nutrition expert judgment calls.

2.2.1. Fuzzy matching in "First+"

"First+" used more lenient matching parameters and was applied to the 591 Euromonitor products with only one CNF match from "*First*." In "*First+*," a minimum fuzzy match score of 50 (out of 100 inclusive) was required for a CNF food to be considered a potential match–in addition to satisfying the nutrient thresholds (per **Table 1**).

2.2.2. Selection among algorithm-proposed matches

All algorithm-proposed matches were nutritionally appropriate within an *a priori* error tolerance as specified by the aforementioned nutrient thresholds. Therefore, the dietetic validation of match selection among these options focused largely on the Euromonitor product's qualitative data–namely, its subcategory (including definitions) and brand–in tandem with the name(s) of CNF food(s) suggested by the algorithm. In this way, the matching algorithm acts somewhat like an advanced search engine, whereby the results of the algorithm present an expert with a narrowed list of candidates for selection. Each of the candidates for matching already meets specified nutritional criteria, which frees time for the validator to focus on the features of the Euromonitor data that cannot be so easily understood solely by a computer.

When the algorithm suggested at least one match for a given Euromonitor product, a dietetics expert team member would either choose the most accurate option (or it, if only one match was offered) or reject all suggested matches, thus sending that product for manual CNF selection. If the team member determined that multiple algorithm-proposed matches could be accurate, then the algorithm-proposed match that was deemed to have the least egregious nutritional error was selected as the most accurate (and final) match. To do this, the nutritional differences from steps 1, 2, and 3 of the algorithm (from Section

TABLE 1 Sets of thresholds as differences between Euromonitor and Canadian Nutrient File (CNF), as applied in each sift of our algorithm.

Sift			*First*	*First+*	*Second*	*Third*
Total number of Euromonitor products run through sift			1111	591	207	43
Maximum difference[a]	*All 3 conditions must be simultaneously met:*	Total dietary fat	20%	60%	40%	∞
		Carbohydrate	20%	60%	40%	∞
		Protein	20%	60%	40%	∞
	All 3 conditions must be simultaneously met:	Fiber	10%	60%	40%	∞
		Saturated fat	10%	60%	40%	∞
		Sugar	10%	60%	40%	∞
	Sodium		50%	50%	50%	∞
Minimum fuzzy matching score (0–100)[b]			0	50	0	0
Fuzzy match optimization used[c]			Yes	No	Yes	Yes

∞Indicates no maximum difference.

[a] Maximal difference thresholds for all nutrients except sodium were based on differences in those nutrients as a proportion of calories, while the maximal threshold for sodium was based on relative differences.

[b] The fuzzy matching score system was a continuum between 0 and 100 (inclusive).

[c] Fuzzy match optimization was applied to select one or more CNF foods with the highest fuzzy matching score out of the list of potential CNF matches for each Euromonitor product.

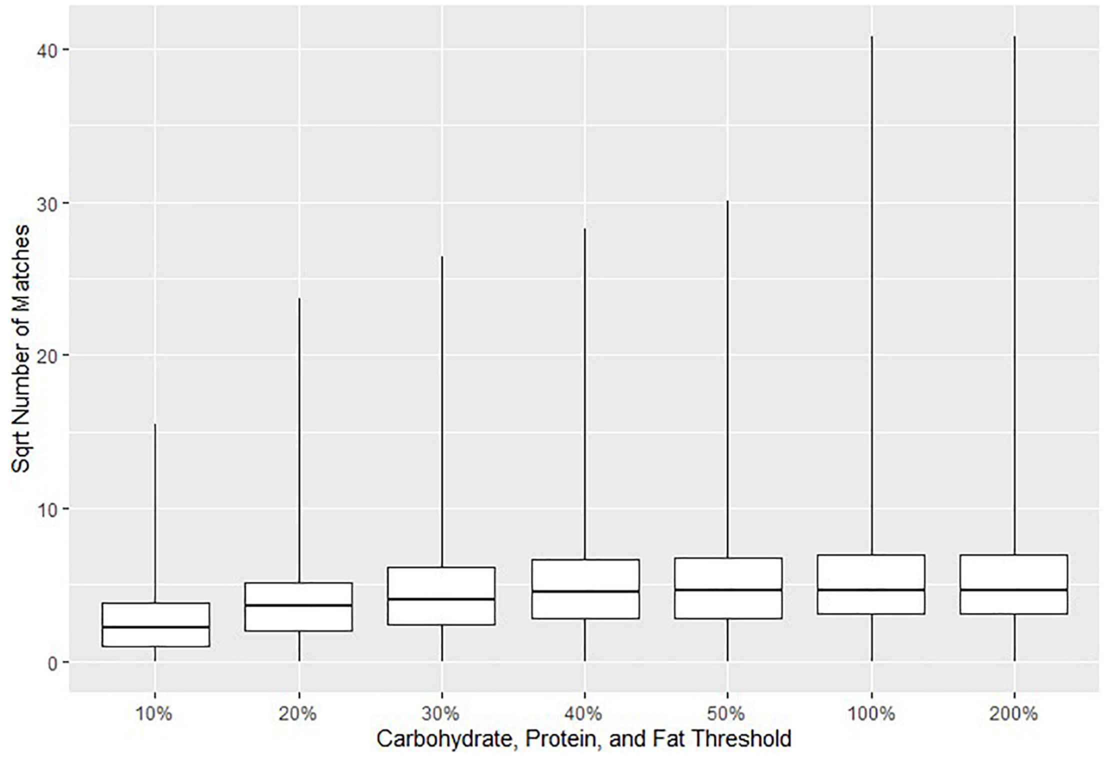


FIGURE 2
Boxplots indicating the 0th, 25th, 50th, 75th, and 100th percentiles of the square roots of the numbers of Canadian Nutrient File (CNF) suggested matches for all Euromonitor products as a function of the threshold selection for carbohydrates, proteins, and fats. The threshold for fiber, saturated fat, and sugar was set to 40% in all cases, and the sodium threshold was set to 50%. Fuzzy string matching was not used. The square root number of matches is reported due to the large numbers of matches when using higher thresholds.

“2.2 BiNFM algorithm design”) between the Euromonitor product and each suggested CNF food item were tabulated. Then, for each Euromonitor and CNF combination separately, the largest of these nutritional differences was computed. The suggested CNF matches were then listed in ascending order of this maximal nutrition difference for each Euromonitor product. The first CNF item in this order was that with the smallest nutritional error. **Figure 4** provides an example of

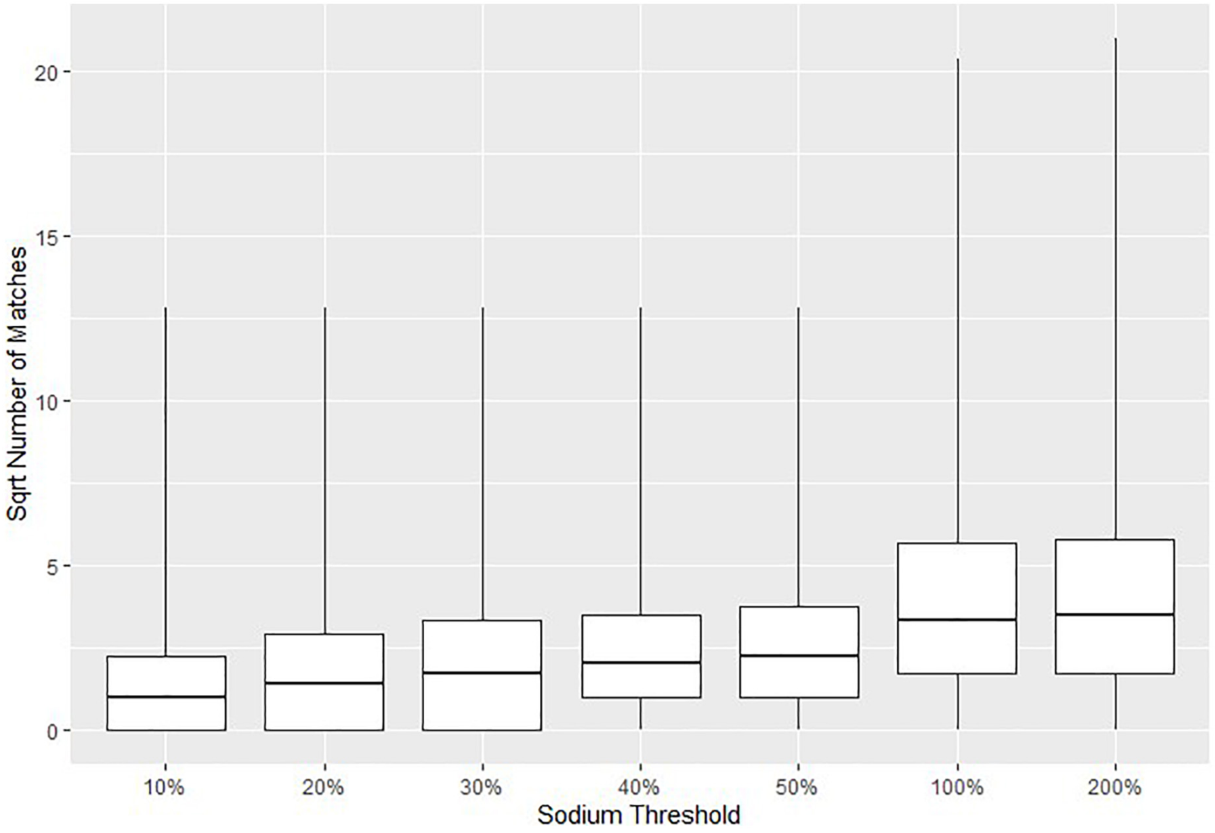


FIGURE 3
Boxplots indicating the 0th, 25th, 50th, 75th, and 100th percentiles of the square roots of the numbers of Canadian Nutrient File (CNF) suggested matches for all Euromonitor products as a function of the threshold selection for sodium. The threshold for carbohydrate, protein, and fats was set to 20% and the threshold for fiber, saturated fat, and sugar was set to 10% in all cases. Fuzzy string matching was not used. The square root number of matches is reported due to the large numbers of matches when using higher thresholds.

how the most accurate match for a Euromonitor product with multiple algorithm-suggested matches was chosen.

Selection of the most accurate match was performed by a team of dietetics experts and registered dietitians. Two team members with dietetics expertise independently worked with the same set of algorithm-proposed matches. Any discrepancies in their final match selections were reviewed and decided by a third team member (registered dietitian). Any disagreement with the registered dietitian's final selection was resolved as a full team.

2.2.3. Manual match selection

Manual selection was conducted for Euromonitor products if: (1) they were unable to be run through the algorithm, (2) the algorithm did not propose any matches, or (3) among algorithm-proposed matches, none were accurate. Just as in the algorithm-aided selection process, manual match selection also used the subcategory and brand name of each Euromonitor product, its subcategory definition, and the CNF food name(s).

To limit subjectivity as much as possible, two team members with dietetics expertise were assigned the same set of Euromonitor products for independent manual selection. Discrepancies in manual selection were assigned to a third team member with dietetics expertise, who then also independently chose the best CNF equivalent. Then, one of our team's registered dietitians reviewed the CNF matches suggested by those three team members and picked the best equivalent (which could also be a CNF food other than one of those suggested by the three colleagues). This final decision and its reasoning were reviewed together by all four of these individuals, and any lingering disagreement was discussed by the whole team until consensus was achieved.

2.3. Analyses

2.3.1. Intercategories

For the purposes of reporting results, we generated a new level of categorization by collapsing multiple Euromonitor subcategories into so-called "intercategories." This was necessary due to Euromonitor's lack of a category level that would allow for dietetically meaningful reporting of results; 210 subcategories were far too many, while the 2 categories

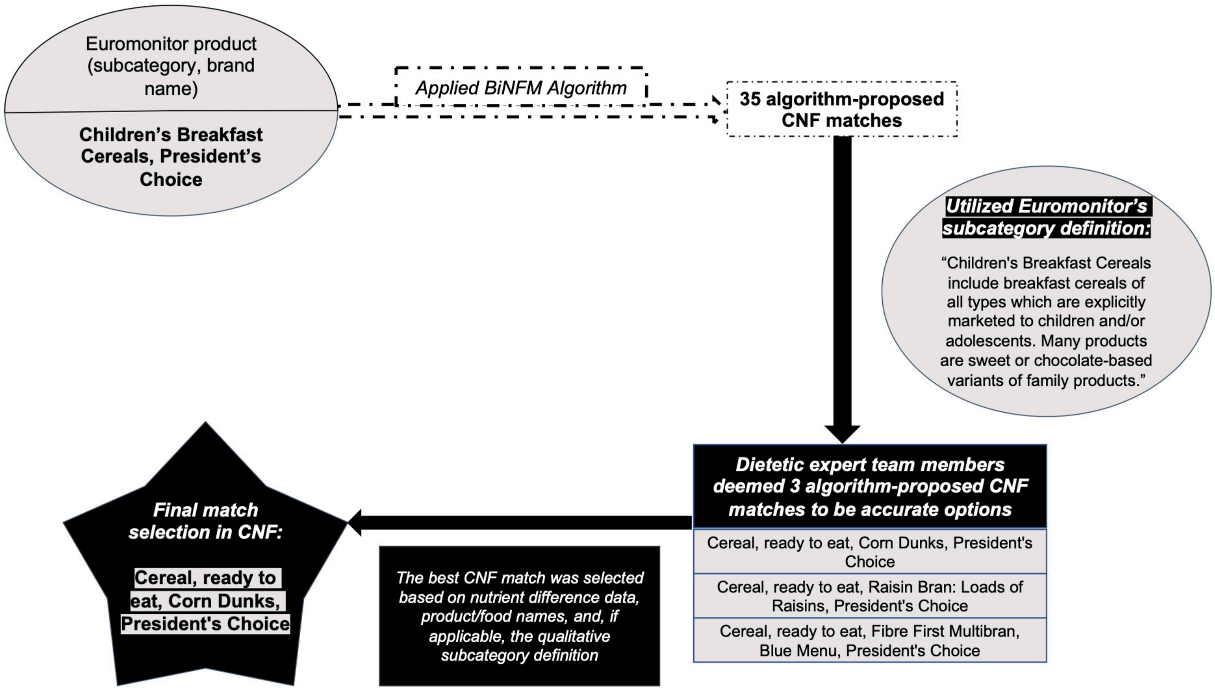

Euromonitor product (subcategory, brand name)
Children's Breakfast Cereals, President's Choice
Applied BiNFM Algorithm
35 algorithm-proposed CNF matches
Utilized Euromonitor's subcategory definition:
"Children's Breakfast Cereals include breakfast cereals of all types which are explicitly marketed to children and/or adolescents. Many products are sweet or chocolate-based variants of family products."
Dietetic expert team members deemed 3 algorithm-proposed CNF matches to be accurate options
Cereal, ready to eat, Corn Dunks, President's Choice
Cereal, ready to eat, Raisin Bran: Loads of Raisins, President's Choice
Cereal, ready to eat, Fibre First Multibran, Blue Menu, President's Choice
The best CNF match was selected based on nutrient difference data, product/food names, and, if applicable, the qualitative subcategory definition
Final match selection in CNF:
Cereal, ready to eat, Corn Dunks, President's Choice

FIGURE 4
Example of our algorithm-aided, dietetics expert-validated matching procedure.

of study were too few. Each intercategory is composed of one or multiple subgroups of Euromonitor products, as indicated in Table 2. The intercategories are: Baby Food; Dairy; Ready Meals and Soup; Sauces, Dressings, Spreads, and Dips; Sweet Snacks; Savory Snacks; Baked Goods; Cereal and Grain Products; Processed Fruit and Vegetables; Processed Meat; Meat Substitutes; Processed Seafood; Soft Drinks and Juice; Coffee and Tea; and Water and Functional Beverages.

2.3.2. Descriptive statistics

Matching accuracy (overall and by intercategory) was measured as the number of Euromonitor products with an appropriate CNF match–from the algorithm and, separately, from both the algorithm and manual matching–divided by the total number of Euromonitor products of focus. For instance, the overall accuracy of the algorithm equaled the number of Euromonitor products that had at least one accurate algorithm-proposed match divided by the total number of Euromonitor products with at least one algorithm-proposed match.

Inter-rater agreement rates were calculated as the percentage of Euromonitor products for which both team members agreed on what to select or do. This was done for the algorithm-only part of this work (for agreement in selecting the same CNF match among algorithm options or refusing them) as well as for manual selection (agreement in selecting the same CNF equivalent).

3. Results

The flow diagram in Figure 5 summarizes the number of Euromonitor products that entered and matches that resulted from each step of our procedure. At the start, the Euromonitor data subset contained 1,179 branded products. Sixty-eight of these were identified as having zero calories or missing key nutrient information and were thus sent directly for manual matching.

In total, 1,111 Euromonitor products were run through our BiNFM algorithm, with 1,070 (96%) resulting in one or more algorithm-proposed matches. Figure 6 serves as a visual aid about the process of how, through each sift, three levels of matching were possible for each Euromonitor product: zero/no algorithm-proposed CNF match (= 0), a single/one match (= 1), or multiple matches ($\geq$2). The "*First*" sift left 207 Euromonitor products without any potential CNF matches. The "*Second*" sift was applied with looser thresholds to these unmatched Euromonitor products, resulting in 43 unmatched products. Finally, these unmatched products from "*Second*" were run through "*Third*," which provided suggested matches for all 43 products. Additionally, 591 Euromonitor products were sent through "*First+*." At the end of the four sifts, 899 out of 1,111 Euromonitor products (81%) had been matched with two or more CNF foods; 171 (15%) of them matched with a single one; and 41 (4%) of them matched with none. All 41 of the ultimately unmatched Euromonitor products originally had a single CNF

TABLE 2 The 15 intercategories generated for this work, the number of products in each intercategory (N), and the contributing Euromonitor subgroup levels (and their numbers of products = *n*).

Euromonitor subgroup level 1 (*n*)	Euromonitor subgroup level 2 (*n*)	Euromonitor subgroup level 3 (*n*)	Intercategory	Total number of products (*N*)
Dairy products and alternatives (180)	Baby food (20)	—	Baby food	20
	Dairy (160)	—	Dairy	160
Cooking ingredients and meals (220)	Ready meals (42)	—	Ready meals and soup	66
	Soup (24)	—		
	Sauces, dressings, and condiments (130)	—	Sauces, dressings, spreads, and dips	154
	Sweet spreads (24)	—		
Snacks (404)	Confectionery (181)	—	Sweet snacks	324
	Ice cream and frozen desserts (65)	—		
	Sweet biscuits, snack bars, and fruit snacks (78)			
	Savory snacks (80)	—	Savory snacks	80
Staple foods (213)	Baked goods (46)	—	Baked goods	46
	Breakfast cereals (30)	—	Cereal and grain products	56
	Rice, pasta, and noodles (26)	—		
	Processed fruit and vegetables (45)	—	Processed fruit and vegetables	45
	Processed meat and seafood (66)	Processed meat (38)	Processed meat	38
		Meat substitutes (9)	Meat substitutes	9
		Processed seafood (19)	Processed seafood	19
Carbonates (55)	—	—	Soft drinks and juice	126
Concentrates (17)	—	—		
Energy drinks (12)	—	—		
Juice (37)	—	—		
Sports drinks (5)	—	—		
Ready-to-drink coffee (7)	—	—	Coffee and tea	23
Ready-to-drink tea (16)	—	—		
Bottled water (13)	—	—	Water and functional beverages	13
			Total	1,179

match in “*First*” and thus had entered “*First+*,” after which they became unmatched because their fuzzy match scores did not meet the algorithm’s threshold of 50 used in this latter sift. These by-sift numbers can also be found near the bottom of Table 3, along with the accuracy of algorithmic output for products with one (50%) or two or more (71%) match suggestions.

Figure 7 displays boxplots for the square root of the number of algorithmically suggested CNF matches for the Euromonitor products run in each of the sifts “*First+*” (591 products), “*Second*” (207 products), and “*Third*” (43 products). “*First*” is not included because every product in this sift had precisely one suggested CNF match. Square roots of the number of suggested matches are used instead of the raw numbers in order to make the plot more visually comprehensible but bears no other importance. Moving from “*First*” (excluded from the figure but equal to 1) to “*Second*” to “*Third*” sees increasing numbers of suggested matches. “*First+*” has the most suggested matches of all sifts.

Table 3 reports–by intercategory–the total number of Euromonitor products run in the algorithm and the number and percent of products accurately matched to an algorithm-suggested match. The following intercategories saw the highest percentage of products with an accurate algorithm-proposed CNF match: Meat Substitutes (89%), Processed Fruit and

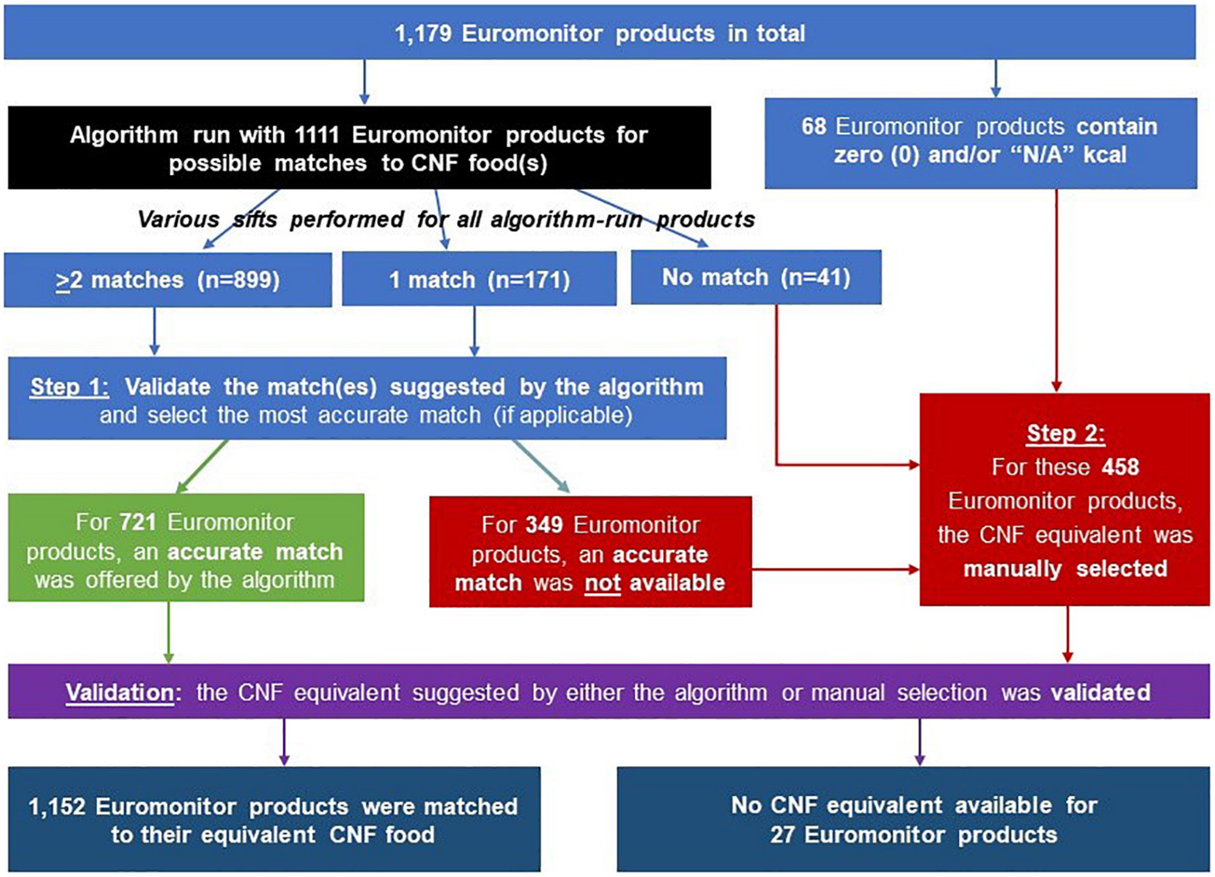

1,179 Euromonitor products in total
Algorithm run with 1111 Euromonitor products for possible matches to CNF food(s)
68 Euromonitor products contain zero (0) and/or "N/A" kcal
Various sifts performed for all algorithm-run products
≥2 matches (n=899)
1 match (n=171)
No match (n=41)
Step 1: Validate the match(es) suggested by the algorithm and select the most accurate match (if applicable)
Step 2: For these 458 Euromonitor products, the CNF equivalent was manually selected
For 721 Euromonitor products, an accurate match was offered by the algorithm
For 349 Euromonitor products, an accurate match was not available
Validation: the CNF equivalent suggested by either the algorithm or manual selection was validated
1,152 Euromonitor products were matched to their equivalent CNF food
No CNF equivalent available for 27 Euromonitor products

FIGURE 5
Flow diagram summarizing how many Euromonitor products were matched to their most accurate Canadian Nutrient File (CNF) equivalent *via* the algorithm-based and manual selection processes.

Vegetables (84%), and Dairy (81%). By contrast, Water and Functional Beverages (0%), Coffee and Tea (40%), and Processed Seafood (42%) had the lowest percentages after being run through the algorithm. Overall, of the 1,111 Euromonitor products that entered the algorithm, 721 (65%) resulted in a CNF match being selected among the algorithm suggestions.

That same table further breaks down, within each intercategory by level of algorithm matches (0, 1, or ≥2), how many algorithm-suggested matches there were, and how many of those were accurate. Out of the 15 intercategories, 13 (87%) saw the majority of their Euromonitor products end up with ≥2 algorithm-suggested CNF matches, with Processed Seafood just under half (47%). Water and Functional Beverages was the only intercategory with neither a plurality nor a majority of its products being offered ≥2 matches; instead, each of its 3 products in this intercategory had 1 algorithm-suggested CNF match. At levels 1 and ≥2, a majority of products had an accurate algorithm-proposed match, with higher accuracy observed for the latter: 50% of the 171 with 1 match option versus 71% of the 899 with ≥2 matches. The highest accuracies were observed among the following intercategories, with all at the ≥2 match level: Meat Substitutes (100%), Processed Fruit and Vegetables (90%), Dairy (88%), and Cereal and Grain Products (83%). Water and Functional Beverages was the only intercategory with 0% accuracy, as none of the single matches offered by the algorithm for its 3 products were nutritionally appropriate.

Out of the 1,179 total Euromonitor products, 1,152 (98%) were matched with a CNF equivalent either with an algorithm-suggested match or by manual selection in the CNF (Table 4). The exceptions were 3 products in Coffee and Tea (87%) and 24 in Sweet Snacks (93%); all 27 of these unmatchable products were a result of there being no nutritionally appropriate CNF match.

Inter-rater agreement rates by intercategory for both parts of the matching process–algorithm-based and manual–are reported in Table 5. The overall inter-rater agreement rate in the first step (selecting the same CNF equivalent among algorithm-suggested options or refusing all of those options) was 51%; the highest rates were in Water and Functional Beverages

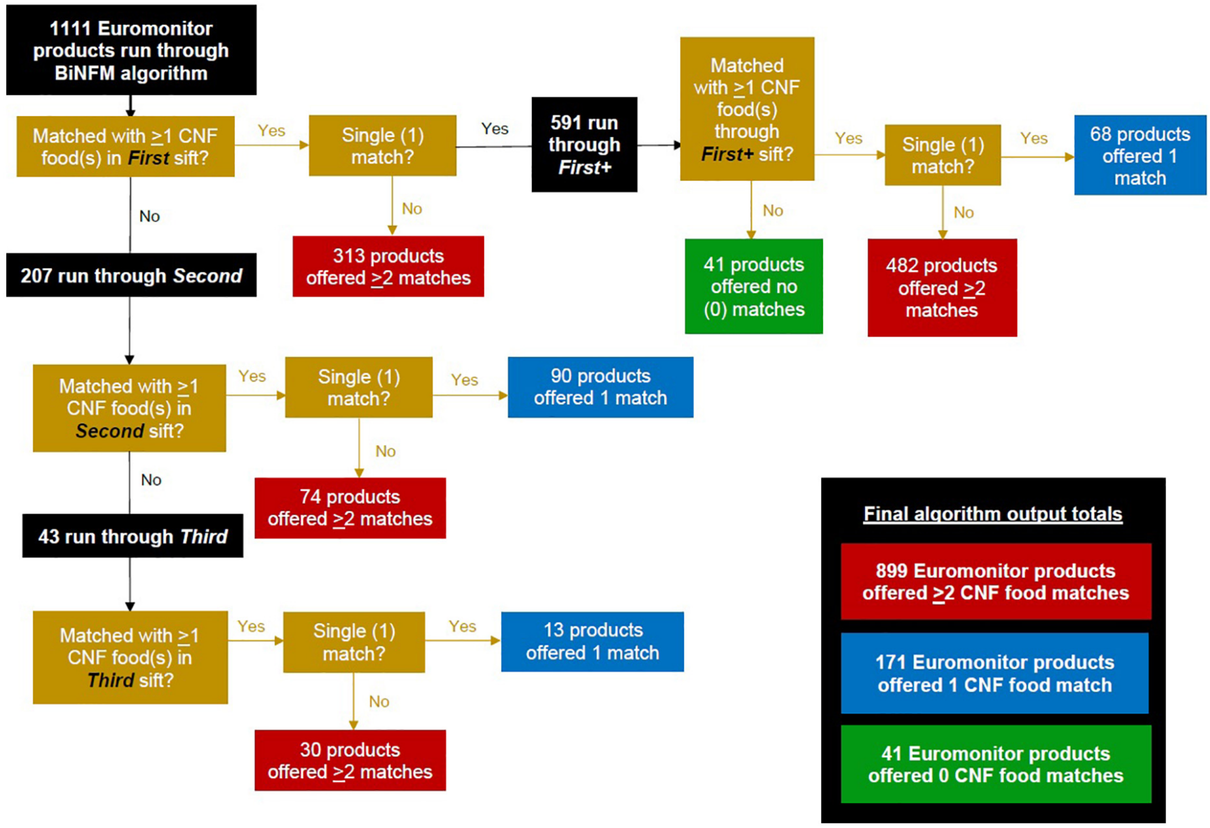

1111 Euromonitor products run through BiNFM algorithm
Matched with ≥1 CNF food(s) in *First* sift?
Yes
Single (1) match?
Yes
591 run through *First+*
Matched with ≥1 CNF food(s) through *First+* sift?
Yes
Single (1) match?
Yes
68 products offered 1 match
No
482 products offered ≥2 matches
No
41 products offered no (0) matches
No
313 products offered ≥2 matches
No
207 run through *Second*
Matched with ≥1 CNF food(s) in *Second* sift?
Yes
Single (1) match?
Yes
90 products offered 1 match
No
74 products offered ≥2 matches
No
43 run through *Third*
Matched with ≥1 CNF food(s) in *Third* sift?
Yes
Single (1) match?
Yes
13 products offered 1 match
No
30 products offered ≥2 matches
Final algorithm output totals
899 Euromonitor products offered ≥2 CNF food matches
171 Euromonitor products offered 1 CNF food match
41 Euromonitor products offered 0 CNF food matches

FIGURE 6
Overview of Euromonitor products' coursing through the algorithm's four sifts (by level of match output: 0, 1, or ≥2 matches).

(100%) and Coffee and Tea (70%), while the lowest were seen for Processed Meat and Processed Seafood (both 32%). In terms of refusal of algorithm-proposed matches, the inter-rater agreement rate was 71% overall; this was highest among products in Baby Food and Water and Functional Beverages (both 100%) and lowest for those in Processed Meat and Processed Seafood (both 32%). The highest rate of agreement for algorithm-based selection was for Water and Functional Beverages (100%), while the lowest was among Processed Fruit and Vegetables (22%). For 5 of the 15 intercategories (31%), team members were more likely to agree to refuse the algorithm's options–thus, sending those products to manual selection–than they were to agree on a specific algorithm-proposed option: Baby Food (100 vs. 65%); Dairy (85 vs. 43%); Ready Meals and Soup (86 vs. 58%); Sauces, Dressings, Spreads, and Dips (98 vs. 69%); and Sweet Snacks (81 vs. 54%). In the remaining 10 intercategories, those two agreement rates were equivalent (refusal of versus selection among algorithm-suggested options). Among the 407 Euromonitor products ultimately managed with manual selection, the overall inter-rater agreement rate of selecting the same CNF match was 33%.

4. Discussion

We developed, implemented, and documented an algorithm-assisted, expert-validated database mapping of Euromonitor Passport Nutrition's branded food and beverage products sold in Canada between 2014 and 2018 to their respective equivalents in the national food composition database, the CNF. The use of an algorithm helped optimize the efficiency of an otherwise fully manual initiative–saving time and labor. Our algorithm design is readily applicable to other contexts, as the parameters from the Euromonitor and CNF databases that we utilized are not unique in the food-related research arena. The two core requirements are a text descriptor of a food or product (for fuzzy matching) and some nutrient data; nearly all such datasets possess the former, with many also containing the latter. The use of a third food categorization system in common (the BNS) is an optional asset to further focus the algorithm's database search (in our case, of the CNF). Our approach to nutrient threshold selection combined numerical heuristics with expert judgment calls; however, one could just as well employ other parameter or threshold selection techniques to suit their needs and problem

TABLE 3 Number of Euromonitor products run through the algorithm and the number and percentage of products accurately matched overall (total and by intercategory). Additionally, by intercategory and by level of algorithm-suggested matches, the numbers and percentages of proposed matches and of accurate such matches.

Intercategory	Number of products run in algorithm	Number and percent (%) of products accurately matched	Level of algorithm-proposed matches	Number and percent (%) of algorithm-proposed matches across the levels	Number and percent (%) of accurate algorithm-proposed matches by level
Baby food	20	10 (50.0)	0	0 (0.0)	N/A
			1	0 (0.0)	N/A
			≥2	20 (100.0)	10 (50.0)
Baked goods	45	25 (55.6)	0	0 (0.0)	N/A
			1	3 (6.7)	2 (66.7)
			≥2	42 (93.3)	23 (54.8)
Cereal and grain products	55	39 (70.9)	0	1 (1.8)	0 (0.0)
			1	7 (12.7)	0 (0.0)
			≥2	47 (85.5)	39 (83.0)
Coffee and tea	20	8 (40.0)	0	2 (10.0)	0 (0.0)
			1	6 (30.0)	4 (66.7)
			≥2	12 (60.0)	4 (33.3)
Dairy	157	127 (80.9)	0	6 (3.8)	0 (0.0)
			1	16 (10.2)	8 (50.0)
			≥2	135 (86.0)	119 (88.2)
Meat substitutes	9	8 (88.9)	0	0 (0.0)	N/A
			1	4 (44.4)	3 (75.0)
			≥2	5 (55.6)	5 (100.0)
Ready meals and soup	66	51 (77.3)	0	0 (0.0)	N/A
			1	4 (6.1)	3 (75.0)
			≥2	62 (93.9)	48 (77.4)
Processed fruit and vegetables	45	38 (84.4)	0	1 (2.2)	0 (0.0)
			1	2 (4.4)	0 (0.0)
			≥2	42 (93.3)	38 (90.5)
Processed meat	38	22 (57.9)	0	1 (2.6)	0 (0.0)
			1	0 (0.0)	N/A
			≥2	37 (97.4)	22 (59.5)
Processed seafood	19	8 (42.1)	0	8 (42.1)	0 (0.0)
			1	2 (10.5)	2 (100.0)
			≥2	9 (47.4)	6 (66.7)
Sauces, dressings, spreads, and dips	145	103 (71.0)	0	4 (2.8)	0 (0.0)
			1	32 (22.1)	23 (71.9)
			≥2	109 (75.2)	80 (73.4)
Savory snacks	80	57 (71.3)	0	1 (1.3)	0 (0.0)
			1	5 (6.3)	2 (40.0)
			≥2	74 (92.5)	55 (74.3)

(Continued)

TABLE 3 (Continued)

Intercategory	Number of products run in algorithm	Number and percent (%) of products accurately matched	Level of algorithm-proposed matches	Number and percent (%) of algorithm-proposed matches across the levels	Number and percent (%) of accurate algorithm-proposed matches by level
Soft drinks and juice	96	61 (63.5)	0	2 (2.1)	0 (0.0)
			1	22 (22.9)	17 (77.3)
			≥2	72 (75.0)	44 (61.1)
Sweet snacks	313	164 (52.4)	0	15 (4.8)	0 (0.0)
			1	65 (20.8)	22 (33.9)
			≥2	233 (74.4)	142 (60.9)
Water and functional beverages	3	0 (0.0)	0	0 (0.0)	N/A
			1	3 (100.0)	0 (0.0)
			≥2	0 (0.0)	N/A
Total	1,111	721 (64.9%)	0	41 (3.7%)	N/A
			1	171 (15.4%)	86 (50.3%)
			≥2	899 (80.9%)	635 (70.6%)

Table excludes the 68 Euromonitor products sent directly to manual matching prior to the algorithm being run. Due to rounding, percentage totals may not total 100%.

context. It is important to remark that there are two processes presented in this report. One process is the algorithm for producing suggested matches; the other is the flow of the various sifts. Multiple sifts were used because, while some Euromonitor items had nutritionally appropriate CNF match suggestions in our initial sift ("*First*"), other items did not have ideal matches. Therefore, we wanted to keep the matches that were potentially good in our first run of the algorithm, but then re-run the algorithm with different sets of parameters to obtain alternative suggested matches for those Euromonitor products with poor or no suggested matches in the previous run. The integration of dietetics expertise to validate our CNF match choices ensured that these selections were appropriate based on products' nutrition information and subcategory definitions, the latter of which the algorithm was unable to leverage. Thus, despite the time and labor it added to the process, dietetics expertise was an imperative supplement to the algorithm-based matching effort, as the rigor of our planned future studies using this CNF-linked Euromonitor dataset depends on the precision of this database mapping.

In the end, 1,152 (98%) of the Euromonitor products matched with a CNF food, with the remaining 27 (2%) unmatchable products owing to a lack of an equivalent food available in the CNF. All products from the following Euromonitor subcategories were unmatchable: Lollipops, Medicated Confectionery, Power Mints, Fruit and Nut Bars, and Carbonated Ready-To-Drink Tea. Brinkerhoff et al. similarly tracked reasons for unmatchability in their manual matching effort of food sold at a supermarket, and they found 4.6% of food products were not covered by the USDA-SR (4).

Like other examples in the literature, we sought to design our algorithm in a way that would maximize both the overall quality and accuracy of matches (11, 12). We also wanted the algorithm to provide at least one match suggestion for each Euromonitor product, which we were able to achieve for nearly all products. The only reason 41 products were left without a match was due to the "*First+*" sift. In our effort to raise the number of algorithm suggestions from a single match option in "*First*," the addition of fuzzy matching with a threshold of 50 in "*First+*" may have been too stringent. Of the 591 products run through "*First+*," 41 (7%) of these products failed to meet this fuzzy match threshold and were ultimately left with no CNF match, as we did not retain the single match option from "*First*" (which instead had used fuzzy match optimization rather than a strict fuzzy matching threshold). While increasing the fuzzy matching threshold would likely have added more options to wade through (particularly for those 41 without any algorithm options), this would also have reduced the overall sensitivity of "*First+*."

The fewer the number of Euromonitor products needing manual selection after being run through the algorithm, the higher the algorithm's accuracy. As anticipated, those products with multiple algorithm suggestions had higher match accuracy versus those with only one suggestion (71 vs. 50%). Future algorithms could require multiple matches, but, like with the fuzzy matching loosening, this would then increase the possibility that thresholds would become too loose. This would render the algorithm less sensitive and increase the resource burden of choosing between multiple match options for a

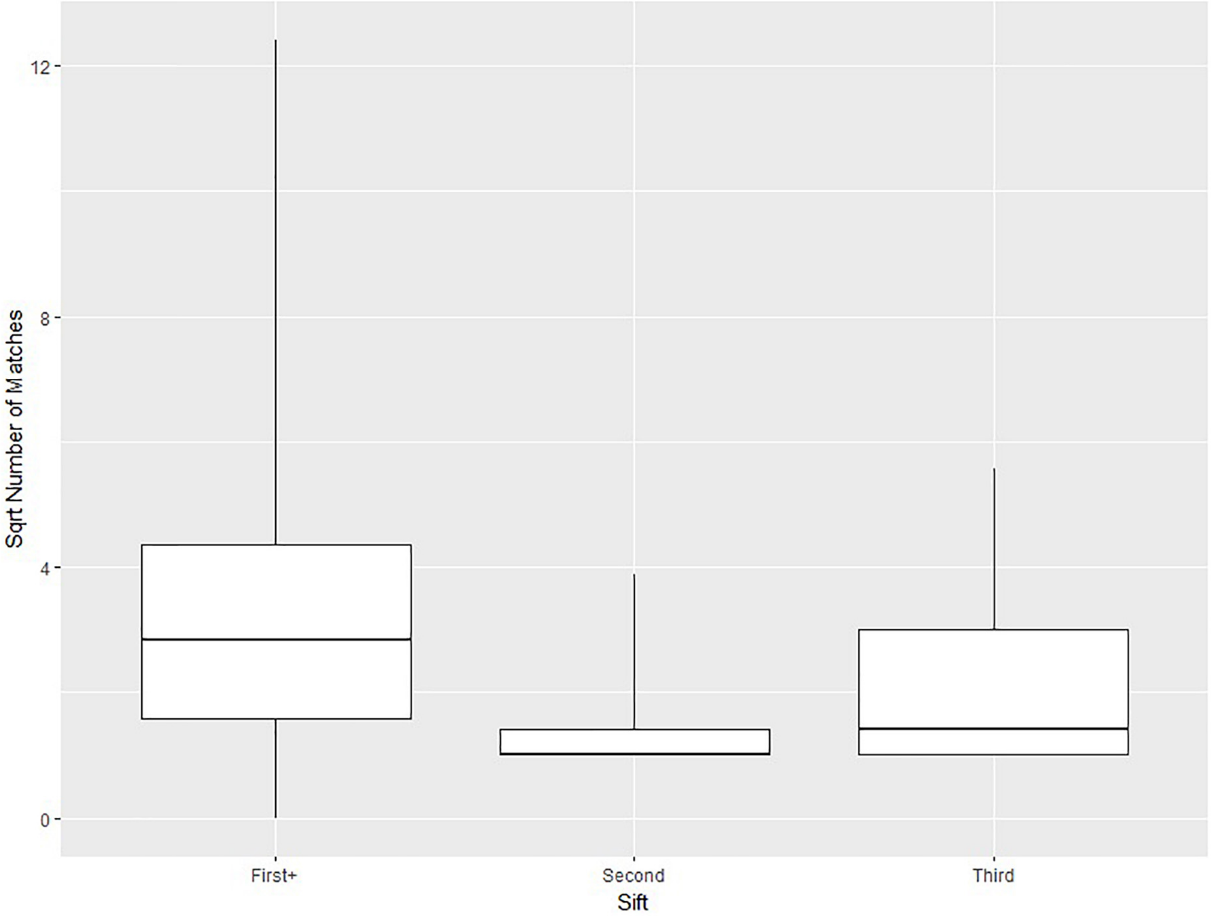


FIGURE 7
Boxplots indicating the 0th, 25th, 50th, 75th, and 100th percentiles of the square root of the number of matches across 3 of the sifts: "*First+*", "*Second*", and "*Third*".

greater proportion of the products–costing labor and time while lowering process efficiency overall.

Compared to similar published endeavors, we find that our matching experience was resonant in some ways and distinct in others. Like Thiele et al., who linked foods in GfK to their equivalents in the German food composition database, we also found that several Euromonitor products could be linked to the same (usually generic) CNF food. However, unlike their team, ours did not find the sales database possessed "extremely in-depth documentation" on food composition relative to the national food composition database (16). The semi-automated approach of Carter et al. (17) is akin in certain ways to our matching algorithm, but with some notable differences; most importantly, they appear to compare the percent difference in nutrients rather than the differences in the proportions of calories per nutrient as we have done. As we have posited, using simple percent differences is a less robust way of comparing the nutrients for low-calorie foods, and so one might expect that Carter et al.'s algorithm could have encountered issues matching these items. Another important difference with their work is our added use of fuzzy string matching to further aid our sifts in identifying the best possible algorithm-suggested matches in the CNF (17).

While this field is pushing further into fully automated approaches like artificial intelligence and natural language processing, dietetics expertise remains critically invaluable for many database mapping endeavors (18, 19). This is particularly true for datasets where the context of nutrients, food categorization systems (e.g., too-vague or too-detailed), and other heuristic aspects of matching are not easy for a computer to handle. Algorithms like ours therefore offer a pragmatic way to aid the matching process yet are not intended as a one-size-fits-all, complete solution to such matching problems. In their largely automated approach to food database mapping, Bohn et al. (20) observed that fuzzy string matching was inhibited by the non-standardized naming of food in producers' databases and had an expert manually check low-similarity potential matches. We, too, experienced

TABLE 4 Process accuracy, overall and by intercategory.

Intercategory	Number of products	Number of products after algorithm-based AND manual selection with an accurate CNF match	Number of products deemed unmatchable to CNF	Overall process accuracy as number matched (%)
Baby food	20	20	0	20 (100.0)
Baked goods	46	46	0	46 (100.0)
Cereal and grain products	56	56	0	56 (100.0)
Coffee and tea	23	20	3	20 (87.0)
Dairy	160	160	0	160 (100.0)
Meat substitutes	9	9	0	9 (100.0)
Ready meals and soup	66	66	0	66 (100.0)
Processed fruit and vegetables	45	45	0	45 (100.0)
Processed meat	38	38	0	38 (100.0)
Processed seafood	19	19	0	19 (100.0)
Sauces, dressings, spreads, and dips	154	154	0	154 (100.0)
Savory snacks	80	80	0	80 (100.0)
Soft drinks and juice	126	126	0	126 (100.0)
Sweet snacks	324	300	24	300 (92.6)
Water and functional beverages	13	13	0	13 (100.0)
Total	1179	1152	27	1152 (97.7%)

this naming quandary in Euromonitor, which we also addressed with manual effort by dietetics experts. Additionally in our case, because Euromonitor is often used for market research data, its subcategories were sometimes named for marketing and retail purposes; as a result, definitions were necessary to be used in conjunction with Euromonitor subcategory names to fully understand the products within them. For example, the subcategory "Countlines" is defined as "chocolate bars eaten as snacks," which was critical added information for matching. In future work, an expert understanding of food-related database architecture and terminologies could be used to develop appropriate text-based fuzzy strings to add to the matching algorithm without sacrificing sensitivity.

Unlike algorithm-only approaches that leverage fuzzy (or other automated text-based) matching approaches–and more akin to fully manual matching efforts–we wanted to ensure that match accuracy was not merely based on the closeness in matched food names, but that the food composition would be as nutritionally close as possible, too. Unfortunately, as previously noted, food is largely unstandardized in its terminology. This is likely owed to their distinct purposes: Euromonitor for market analyses versus CNF for federal health survey analyses. Thanks to Euromonitor and CNF entries both having data for key nutritional variables, we were able to dietetically validate final match selections using both calculated nutrient differences and food names (and, if necessary, Euromonitor subcategory definitions and brand names).

Dietetics expertise therefore played a vital role in this endeavor and was a core strength of our methodology. Instructions for validating the algorithm's proposed matches and the manual selection process were developed by team leads with extensive knowledge and clinical dietetic experience relevant to food composition, the Canadian food supply, and the implications of nutrition on health outcomes. We were able to minimize subjectivity by training team members to follow a detailed matching protocol. Other major strengths of our methodological contribution to the discipline include the low-bias and longitudinal nature of the Euromonitor dataset for Canada. We also were able to partially solve the problem faced by Lamarine et al. of nutrient variability, or "variability between different versions of the same food item. For example, 100 g portion of raw garlic would be recorded with an energy content varying between 305 and 670 kcal" (12). While they argued "data curation (including detection and correction of errors) remains a challenge and a thorough review of each composition variables cannot be performed without automated approaches," we were fortunate to be able to innovate with and incorporate nutrient

TABLE 5 Inter-rater agreement rates in the algorithm-based and manual selection processes.

Intercategory	Algorithm-based selection			Manual selection	
	Number of Euromonitor products run through algorithm	Inter-rater agreement rate of selecting same CNF food or refusing algorithm option(s) (%)	Inter-rater agreement rate of deciding that manual selection is needed (%)	Number of Euromonitor products manually managed	Inter-rater agreement rate of selecting the same CNF equivalent (%)
Baby food	20	65.0	100.0	12	0.0
Baked goods	45	35.6	35.6	2	50.0
Cereal and grain products	55	65.5	65.5	8	12.5
Coffee and tea	20	70.0	70.0	12	50.0
Dairy	157	42.7	84.7	57	10.5
Meat substitutes	9	33.3	33.3	1	0.0
Ready meals and soup	66	57.6	86.4	49	34.7
Processed fruit and vegetables	45	22.2	22.2	1	100.0
Processed meat	38	31.6	31.6	2	100.0
Processed seafood	19	31.6	31.6	0	—
Sauces, dressings, spreads, and dips	145	69.0	97.9	139	29.5
Savory snacks	80	53.8	53.8	6	50.0
Soft drinks and juice	96	39.6	39.6	45	57.8
Sweet snacks	313	54.3	80.8	60	35.0
Water and functional beverages	3	100.0	100.0	13	76.9
Total	1,111	51.2%	70.8%	407	33.2%

thresholds for matching in our algorithm, as Euromonitor had key nutritional data (12).

In terms of limitations, our algorithm design was restricted to those 7 nutrients and energy available in both databases; as such, the inclusion of fuzzy matching to draw on text-based data between the two databases proved to be a crucial addition. We were also unable to send products with missing nutrient data and/or zero calories into the algorithm, with the latter due to the non-sodium nutrient thresholds using energy as a denominator. The algorithm's BNS food group restriction was helpful in achieving a more focused set of suggested matches. However, due to product heterogeneity within some Euromonitor subcategories, this may have disadvantaged the algorithm by potentially missing out on some CNF match options that may not have fallen precisely within the preselected BNS group; we found this to be a limited concern, almost exclusively and minimally affecting the following 3 Euromonitor subcategories: Ready Meals, Processed Meat, and Processed Seafood. Brinkerhoff et al. reported a similar issue when fully manually matching their subcategories (so-called "sub-commodities") to the USDA-SR; they were unable to link 21% of them ("∼30% of the entire dataset") due to "heterogeneous sub-commodities containing nutritionally diverse food items that could not be mapped to a single [USDA-]SR item entry" (4). There is also subjectivity inherent in the evaluation of database mapping, as the algorithm can only offer us choices; we must make the final selections. We attempted to mitigate risk of bias and human error by the rigor of and fidelity to our aforementioned, standardized, expert-led match selection at each step. While our inter-rater agreement rates were only 51% among algorithm suggestions and 33% among manual CNF selections, it is important to think about the nuanced, oft-small differences between very similar options in the CNF. Our team discovered it is harder to agree on the same "best" CNF equivalent than it is to refuse all algorithm options and simply assign that Euromonitor product to manual matching. This is evidenced by the fact that no inter-rater agreement rate for algorithm-based selection was higher than that for send-off to match selection (in other words, algorithm option refusal). Most intercategories' rates were equal across these two sub-steps, with only 5 intercategories having a lower agreement rate among the former than the latter. This ties back to the value of nomenclature in a discipline, as we did not always have specific product names in the Euromonitor dataset. This is why we relied on a combination of all data at our disposal throughout the process: subcategory names (using fuzzy matching), BNS food groups (as a search restriction), and nutrient thresholds in the algorithm as well as subcategory definitions and brand names.

By choosing the nutrient thresholds we selected in our algorithm, we gave ourselves an upper bound on match quality. It only takes one nutrient beyond the threshold for the algorithm to reject a potential CNF match. In this sense, our approach was quite conservative. Multiple rounds of dietetic expert validation of the final match selection—with two independent validators plus a registered dietitian—ensured that branded products' matches were nutritionally appropriate (per our stated goal). It is possible that some of the matches to the more generic CNF foods might not be best suited for those micronutrients for which we lacked data on the Euromonitor side (e.g., vitamin D content in a particular brand of a fortified breakfast cereal versus that in its generic match in the CNF). This possible source of nutritional discrepancy limits our potential use of these matched datasets for certain population nutrition studies, as we can only be confident for those 7 nutrients and energy data from Euromonitor and that we have been able to utilize and validate in this matching effort.

With the possible exception of the BNS food group bridging, the BiNFM algorithm is flexible enough to conceivably be applied to the matching of databases other than CNF and Euromonitor. The steps of the algorithmic model we developed can be immediately applied to datasets bearing the same kinds of nutritional data (e.g., energy, carbohydrates, proteins, total fat, fiber, saturated fat, sugar, and sodium) as well as some type of string to be fuzzy-matched. The BiNFM algorithm restricted matches to products with compatible BNS food groups, but this step can be omitted or replaced with restricting matches to compatible categories from another scheme. Even the list of nutrients could be changed, or string fuzzy matching could be omitted altogether. Importantly, the BiNFM algorithm relies heavily on products having non-zero energy content (a requirement for our computation of nutrient differences) and non-missing nutritional data.

5. Conclusion

To the best of our knowledge, this paper constitutes the first algorithm-aided matching of any marketing database's branded food and beverage products sold in Canada to their nutritional equivalents in the CNF. As far as we are aware, this is also the first paper to detail the dietetic expert-driven validation of that matching process, which has now laid the groundwork for rigorous population nutrition and health research using the Euromonitor products' nutrient profiles, sales, and other variables. Indeed, the linkage of food composition data to products found in marketing databases for public health nutrition studies is still a relatively nascent and emerging field, with much of the literature in this space published within the last 15 years. As food supply, retail, marketing, and other related databases become increasingly recognized as ripe opportunities for population nutrition surveillance, methods like ours can be used to enrich analyses of Euromonitor product trends (as the CNF matches offer additional nutrient data) and to supplement national health and dietary surveys with branded food composition data (available from the now-linked Euromonitor products). Although the specific parameters and architecture of our two datasets shaped the most granular details of our matching methodology, we are confident that the overall approach (including the algorithm design) that we employed and trade-offs we weighed would be generalizable and of assistance in similar food-matching endeavors.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The Passport Nutrition dataset for Canada that was used in this work was purchased from Euromonitor International, and so access follows regulations pertaining to it. Requests to access these datasets should be directed to MJ, mahsa.jessri@ubc.ca.

Author contributions

SG contributed to the brainstorming and interpretation of the algorithm, analyzed both the algorithm-based and manual matching processes, led figure production, and led the writing of this manuscript. CM programmed and led development of the BiNFM matching algorithm, was involved in discussions of results, and participated in the writing of the manuscript and figure production. QC prepared and restructured the Euromonitor and Canadian Nutrient File datasets for the algorithm, assisted the development of the matching algorithm, co-led the algorithm validation and manual matching processes and analyses, and participated in the writing of the manuscript and figure production. JP conducted a literature review as background and assisted in drafting the manuscript. JB supervised the dietetic validation, manual matching processes, and provided editing support for the manuscript. MJ conceptualized and presented the idea of this work, and oversaw all activities related to this project, including algorithm design, execution of it as well as the manual matching process, manuscript preparation, and project/grant administration. All authors approved of the final version of this manuscript for publication.

Funding

Authors acknowledge MJ's funding from the Canadian Institutes of Health Research (CIHR #428028), Banting Foundation Discovery Grant (#2019-1406), and the Canada Research Chair program (#950–233168) for this study. SG received a Yale Fox International Fellowship that also supported this work.

Acknowledgments

The authors wish to express their gratitude for the dedicated efforts of student research assistants contributing to the matching process between summer 2020 and summer 2021: Derek Brigham, Kamila Gabitova, Mackenzie Gentek, Jacey Gowler, Sophia Jhajj, Negar Karami, and Janyn Zoccoli.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Bandy L, Adhikari V, Jebb S, Rayner M. The use of commercial food purchase data for public health nutrition research: a systematic review. *PLoS One.* (2019) 14:e0210192. doi: 10.1371/journal.pone.0210192

2. Leroy F, Rytz A, Drewnowski A, Tassy M, Orengo A, Charles VR, et al. A new method to monitor the nutritional quality of packaged foods in the global food supply in order to provide feasible targets for reformulation. *Nutrients.* (2021) 13:576. doi: 10.3390/nu13020576

3. Vlassopoulos A, Masset G, Charles VR, Hoover C, Chesneau-Guillemont C, Leroy F, et al. A nutrient profiling system for the (re)formulation of a global food and beverage portfolio. *Eur J Nutr.* (2017) 56:1105–22. doi: 10.1007/s00394-016-1161-9

4. Brinkerhoff KM, Brewster PJ, Clark EB, Jordan KC, Cummins MR, Hurdle JF. Linking supermarket sales data to nutritional information: an informatics feasibility study. *AMIA Annu Symp Proc.* (2011) 2011:598–606.

5. Jenneson VL, Pontin F, Greenwood DC, Clarke GP, Morris MA. A systematic review of supermarket automated electronic sales data for population dietary surveillance. *Nutr Rev.* (2022) 80:1711–22. doi: 10.1093/nutrit/nuab089

6. Brimblecombe J, Liddle R, O'Dea K. Use of point-of-sale data to assess food and nutrient quality in remote stores. *Public Health Nutr.* (2013) 16:1159–67. doi: 10.1017/S1368980012004284

7. Mamiya H. *Characterizing Community Dietary Patterns Using Grocery Point-of-Sales Data.* Montreal, QC: McGill University (2020).

8. World Health Organization [WHO]. *Using Third-Party Food Sales and Composition Databases to Monitor Nutrition Policies.* (2021). Available online at: https://apps.who.int/iris/bitstream/handle/10665/339075/WHO-EURO-2021-1866-41617-56855-eng.pdf?sequence=1&isAllowed=y (accessed November 24, 2022).

9. Health Canada. *Reference Guide to Understanding and using the Data.* (2017). Available online at: https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs/reference-guide-understanding-using-data-2015.html (accessed July 31, 2022).

10. Health Canada. *Canadian Nutrient File – users Guide.* (2016). Available online at: https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/nutrient-data/canadian-nutrient-file-compilation-canadian-food-composition-data-users-guide.html (accessed July 31, 2022).

11. Tran LTT, Brewster PJ, Chidambaram V, Hurdle JF. An innovative method for monitoring food quality and the healthfulness of consumers' grocery purchases. *Nutrients.* (2017) 9:457. doi: 10.3390/nu9050457

12. Lamarine M, Hager J, Saris WHM, Astrup A, Valsesia A. Fast and accurate approaches for large-scale, automated mapping of food diaries on food composition tables. *Front Nutr.* (2018) 5:38. doi: 10.3389/fnut.2018.00038

13. Government of Canada. *Bureau of Nutritional Sciences (BNS) Food Group Codes and Descriptions – Canadian Community Health Survey (CCHS) 2.2.* (n.d.). Available online at: https://www23.statcan.gc.ca/imdb-bmdi/pub/document/5049_D23_T9_V1-eng.pdf (accessed July 31, 2022).

14. Mouselimis L. *Fuzzy String Matching [R package fuzzywuzzyR version 1.0.5].* (2022). Available online at: https://CRAN.R-project.org/package=fuzzywuzzyR (accessed July 31, 2022).

15. SeatGeek. *Fuzzy string natching in Python.* (2014). Available online at: https://github.com/seatgeek/fuzzywuzzy (accessed February 14, 2020).

16. Thiele S, Peltner J, Richter A, Mensink GBM. Food purchase patterns: empirical identification and analysis of their association with diet quality, socio-economic factors, and attitudes. *Nutr J.* (2017) 16:69. doi: 10.1186/s12937-017-0292-z

17. Carter MC, Hancock N, Albar SA, Brown H, Greenwood DC, Hardie LJ, et al. Development of a new branded UK food composition database for an online dietary assessment tool. *Nutrients.* (2016) 8:480. doi: 10.3390/nu8080480

18. Carlson AC, Page ET, Zimmerman TP, Tornow CE, Hermansen S. *Linking USDA Nutrition Databases to IRI Household-Based and Store-Based Scanner Data.* Technical Bulletin Number 1952. Washington, DC: United States Department of Agriculture, Economic Research Service (2019). doi: 10.22004/ag.econ.291970

19. Eftimov T, Korošec P, Koroušić Seljak B. StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to foodEx2. *Nutrients.* (2017) 9:542. doi: 10.3390/nu9060542

20. Bohn K, Amberg M, Meier T, Forner F, Stangl GI, Mäder P. Estimating food ingredient compositions based on mandatory product labeling. *J Food Compost Anal.* (2022) 110:104508.

TYPE Review
PUBLISHED 28 February 2023
DOI 10.3389/fnut.2023.1110446

Eater-oriented knowledge framework for reducing salt and dietary sodium intake (scoping review)

Alexandra Endaltseva[1*†], Paul Coeurquetin[2†], Thierry Thomas-Danguin[3], Jean-Pierre Poulain[1,4], Laurence Tibère[1,4] and Anne Dupuy[1,4]

[1]CERTOP UMR CNRS, University of Toulouse Jean Jaurès, Toulouse, France, [2]LISST-Cers UMR, University of Toulouse Jean Jaurès, Toulouse, France, [3]Centre des Sciences du Goût et de l'Alimentation (CSGA), CNRS, INRAE, Institut Agro, Université de Bourgogne Franche-Comté, Dijon, France, [4]ISTHIA, University of Toulouse Jean Jaurès, Toulouse, France

OPEN ACCESS

EDITED BY
Alessandra Durazzo,
Council for Agricultural Research and Economics, Italy

REVIEWED BY
Solange Parra-Soto,
University of Glasgow,
United Kingdom
Lanfranco D'Elia,
University of Naples Federico II, Italy

*CORRESPONDENCE
Alexandra Endaltseva
✉ aendalts@gmail.com;
✉ alexandra.endaltseva@univ-tlse2.fr

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

RECEIVED 28 November 2022
ACCEPTED 31 January 2023
PUBLISHED 28 February 2023

CITATION
Endaltseva A, Coeurquetin P, Thomas-Danguin T, Poulain J-P, Tibère L and Dupuy A (2023) Eater-oriented knowledge framework for reducing salt and dietary sodium intake (scoping review).
Front. Nutr. 10:1110446.
doi: 10.3389/fnut.2023.1110446

Salt and dietary sodium are ubiquitously present in daily food practices and, at the same time, reducing salt intake presents an important public health issue. Given such an ambivalent position of salt in human diet, we argue that public health guidelines toward dietary sodium reduction require an eater-oriented knowledge framework. In this article we are making the first steps toward a flexible interdisciplinary database which would include nutritional, socio-economic, cultural, material, and socio-psychological determinants of salt consumption for comprehensive public health campaigns. We employ an explorative scoping review of academic articles and reports, limiting our review to the original data on salt or sodium consumption published in English or French between 2000 and 2022. We describe salt consumption as research object, identifying its representation in different research fields, data sources, methodologies, samples, and links with nutritional recommendations. We synthesize existing approaches *via* four eater-oriented categories: Socio-demographic and cultural descriptors of salt consumers; Knowledge, attitudes, and beliefs on nutritional norms; Salt practices associated with dietary or medical regimes; Salt materialities: interactions and contexts. In each category, we identify the dominant relational features, i.e., what kind of 'eater-salt' relation is being put forward. We thus build an interdisciplinary documentary base of dietary sodium consumption factors. We discuss the results, suggesting that comprehensive nutritional policies for global salt reduction require interdisciplinary eater-oriented data frameworks.

KEYWORDS

public health, dietary sodium, database, salt, interdisciplinary approach to nutrition

1. Introduction

Salt (NaCl)[1] and sodium (Na) are ubiquitously present in cooking and eating practices, and, at the same time, excessive sodium consumption presents serious dangers for human health worldwide. Such an ambivalence requires attention to the specificities of salt-in-action, as well as to the place of salt in the relations between food, nutrition, and health. While the World Health Organization

[1] In this article we will principally use 'salt' as a signifier for both salt and dietary sodium.

(WHO) campaign for dietary salt reduction shapes salt and dietary sodium overconsumption as a critical public health issue, salt keeps being a necessary element for a human body (but not more than 5 g per day) (1). Salt is attributed a manifold of meanings, depending on the ways it enters into food and culinary practices (discretionary salt vs. salt hidden inside the hyper transformed products; refined salt vs. *fleur de sel*; iodized salt, etc.). Intervening into the habits of salt consumption thus requires sensitivity to its materialities (how material properties in context influence relations with salt), food routines, techniques of application (salt shakers, apps), as well as affective positions, cultural belongings, and socio-economic status of salt eaters.

In this article we propose to rethink conventional forms of representing nutritional data for intervention (graphic results of grand cohort questionnaires, such as NutriNet-Santé (2)), and, to put it more boldly, what counts for data in salt consumption. Beyond the above-mentioned ambiguities related to regulating salt consumption, our proposal is driven by Kwong et al.'s critique of the quality of measurements of salt intake in WHO European Region, based on the systematic review of the average population daily salt intake in the 53 Member States of the WHO European Region (3). We therefore propose to rethink existing approaches to the sodium reduction *via* an eater-oriented knowledge framework, categorizing heterogeneous data on salt consumption on the basis of relations and practices at the frontier of food, diet and health. Our aim is to conduct a scoping review of existing literature on dietary sodium consumption, in order to build a flexible and interdisciplinary documentary base of salt consumption factors, exposing the multiple effects of salt for human diet and health and proposing an eater-oriented format for nutritional databases. Our objectives are, first, to identify and categorize the range of disciplinary approaches, methodologies, strategies, and objectives at the crossroads of salt consumption and public health. Second, to identify specific practices and relations through which salt is known (liking, preferences, following a dietary regime). To achieve these objectives, we use scoping review, though which we build on a documentary base of dietary sodium consumption factors. Finally, we propose guidelines for further empirical research on salt consumption for comprehensive and eater-oriented nutritional policies.

2. Reducing salt consumption: Flexible databases for public health interventions

2.1. Salt consumption as public health issue

Historically, salt has provided a great service to humanity, facilitating the development of agriculture and conservation techniques (4), taxation and exchanges of goods and services (4–6), and the increase of taste values in processed foods (6). Salt and dietary sodium enhance food taste, flavor, texture, mouthfeel, and palatability; it reduces bacterial growth in meats, cheeses, and other animal products, and it also enhances hedonic attraction to vegetables among children (5, 7). Finally, iodized salt has proved to be a successful tool for iodine deficiency prevention: due to its ubiquitous use, women and infants who use iodized salt display sufficient iodine levels (8–11). Similar logic was applied to salt fluoridation (12, 13).

An excessive intake of dietary sodium and salt, however, constitutes a major challenge to public health globally, particularly touching low-and middle-income countries (3, 14–16). Analyzing the 'Global Burden of Disease, Injuries, and Risk Factor study' (GBD) 2013 and 2017 to identify emerging public health challenges, GBD 2013 Risk Factors Collaborators observed that diets high in sodium contribute to the risk factors for multiple noncommunicable diseases, with visible increase from 2011 onwards (17), mostly in Asian countries (18). According to the authors, the minimum risk of sodium intake approaches 5 g per day, which is also the amount recommended by the WHO. However, the current mean of salt intake in the WHO member states is almost twice the recommended value despite the agreement to reduce the intake of salt by 30% by 2025 (1, 6, 19–21). Salt overconsumption is a significant factor increasing blood pressure and leading to cardiovascular and kidney disease (6, 22). It is one of the top two dietary risk factors contributing to 1.65 million cardiovascular-related deaths each year (17, 18, 23, 24). Mostly touched by hypertension are low-and middle-income countries (25), where awareness of the dangers of salt and sodium remains low (14, 15, 26, 27).

This ambivalent position of salt in human diet (both as a necessity element and as a danger if consumed in large amounts) makes it challenging to develop effective guidelines for public health interventions. The intake of dietary sodium derives from three sources: [1] processed or manufactured foods (e.g., bread, soup, snacks, and restaurant meals); [2] salt contained in foods (e.g., celery, artichoke); and [3] discretionary salt (DS) added by consumers during cooking, food preparation and/or at the table (28). Most of the research data on salt and sodium overconsumption today concerns sodium and salt in products and processed foods (5, 29) since they are major contributors of sodium intake, particularly breads, processed meats, and sauces (30). However, these data must be accompanied by an understanding of exact consumer practices of sodium intake in order to provide a basis for effective sodium reduction campaigns. For example, Blanco-Metzler et al. (31) qualitative exploratory study in Costa Rica, where cardiovascular diseases have been the leading cause of deaths, demonstrates that most of the sodium intake daily came from domestic consumption. The authors relate this phenomenon to the beliefs that food cannot be consumed without salt, to the habits of salting generously, and the lack of awareness that processed foods already contain sodium (*ibidem*).

The midpoint of the WHO public health campaign for the decrease of dietary sodium consumption shows that low-income countries are still far from reaching an international goal (3, 32). Building on this point, we propose to revisit existing approaches to reducing salt consumption from an eater-oriented perspective, taking into account social, cultural, nutritional, interactional, and hedonic influences. Our argument also follows Kwong et al.'s warning that existing methods of measuring daily salt intake (24 h urinary collections, spot urine collection, dietary recall, food frequency questionnaires (FFQs), dietary records, household budget surveys) do not produce consistent results (3). By conducting a scoping review of existing studies at the crossroads of salt consumption and public health, we categorize studies from nutrition, health management, clinics, cultural anthropology, and sociology, thus building an interdisciplinary documentary base of sodium consumption factors. This documentary base proposes an eater-oriented framework to categorize data on salt overconsumption, linking nutritional knowledge with social sciences to facilitate heterogeneous and eater-oriented knowledge and evidence production. We suggest that databases built though this framework and grounded into the context of a particular country may provide fruitful evidence for effective public health guidelines for dietary sodium reduction in the second half of the

WHO 2013 campaign aiming to decrease the amount of sodium consumed by 30% by 2025.

2.2. The effects of databases: Toward new guidelines for reducing salt intake

According to Durazzo and Lucarini, a better understanding of the frontier between diet and health requires new databases, especially in the perspective of factors contributing to chronic illnesses (33). This need is dictated, among others, by the search of adequate food policies, by promotion of nutritional knowledge among the consumers, and by the proven effects of balanced diet on the general health level. In this frame of reference, a collection of articles on the topic of databases and nutrition was published in Frontiers in March 2022 under the direction of Durazzo and Lucarini. This collection has presented new types of databases, dealing not only with nutritional components (for example, systematizing the types of low-or no-calorie sweeteners in products and beverages (34)), but also those linking nutritional and social characteristics to understand contemporary food scapes (for example, a database studying modernization of Malaysian food patterns (35)).

According to Leonelli, data is material artefacts which can be mobilized in a specific context of knowledge production (36). Databases thus are information infrastructures, which, according to most practitioners, should represent an open archive of knowledge for the use of many different scientific disciplines (37). According to Bowker, a "working archive" of a database is a tool for management (public health management, in our case), and it requires that "social, political, and organizational context is interwoven with statistics, classification systems and observational results in a generative fashion" (37). Since the distribution of data into the categories largely depends on the context, classification systems are doing social and political work (37).

We build on this perspective, arguing that the current state of affairs in the global campaign for dietary sodium reduction requires new flexible databases. In other words, nutritional classification of sodium consumption factors cannot be translated into an adequate policy, without taking into consideration the social, cultural, and psychological processes influencing salt consumption. We thus argue for the need of categorizing data on salt and sodium consumption around eaters' practices and relations, rather than around disciplinary or nutritional research questions. Such an eater-oriented approach is useful not only for building comprehensive interdisciplinary health interventions, but also for rendering nutritional knowledge accessible for non-scientific audiences, thus contributing to the increase of consumers' nutritional knowledge. Our argument resonates with Poulain et al.'s (35) observation that nutritional surveys and studies take an eating individual separately from their context, while the latter provides sociological and ethnological insights into the food habits. Leaving aside the social dimension of food consumption, however, can result in ineffective, and sometimes even contra-effective food policies and regulations. Therefore, linking nutritional knowledge with psycho-socio-cultural determinants requires new flexible knowledge frameworks.

Bowker (37) argues for flexible databases which are "as rich ontologically as the social and natural worlds they map (…)" (*ibidem*), as databases "shape the world in its image". Through the analysis of three experimental databases, Wateron (38) has demonstrated that flexible heterogeneous databases include an exposure of the intentions, reflexivity, and policy inscribed inside them, which can be heuristic. In the case of sodium consumption, an effective database would include, at the same time, the interaction of salt with human (linking, perception learning, and body effects), its material dimensions (compositions, formats, and presentation), socio-economic issues (distributions of populations groups most at risk for sodium overconsumption), cultural influences (recipes, traditions, beliefs, and symbolic significations), common salt practices and routines, and knowledge about/attitudes toward food manufacturing and nutritional messages. Our article makes the first step in this direction, building a documentary base[2] which would merge factors of salt consumption with eaters' relations and practices, as well as metadata which shapes their representation (methods, units of analysis, subject framing, and geographical distribution of existing knowledge), to suggest guiding categories for salt reduction. In the following section we will expose in detail the process of our documentary base construction.

3. Methodology

3.1. Approach

We have chosen a scoping review approach to answer to our objective of rethinking salt consumption factors through an eater-oriented approach. Scoping reviews aid to explore the range, extent, and nature of existing research approaches, to summarize evidence from multiple disciplines, and to identify research gaps or openings (39, 40), which is necessary for proposing a novel eater-oriented knowledge framework. Tricco et al. (40) and Peters et al. (41) precise that scoping reviews are useful to answer broad research questions, aiming not to compare but rather to describe existing body of work on the issue and present a type of knowledge synthesis. Therefore, scoping review is consistent with our purpose of identifying the openings for eater-oriented databases on salt consumption. Through the exploratory scoping review are able to extract existing nutritional, socio-economic, cultural, material, and socio-psychological determinants and descriptors of salt consumption, synthesizing them into the eater-oriented categories, and relating them to particular methodologies, data sources, and links with nutritional recommendations. The review was conducted within the framework of a French interdisciplinary project *Sal&Mieux: Optimizing the use of discretionary salt*, however, the scope of our reviewed was broader than only discretionary salt, with interest in the overall appearance of salt in human eating practices. To watch for the review rigor, we have adopted a PRISMA extension for scoping reviews, PRISMA-ScR (40), following the best practices for scoping reviews, identified by Peters et al. (41), as well as the use of scoping reviews in food studies (42–44).

3.2. Search strategy and initial screening

Our search strategy was to extract studies which articulate salt consumption and public health from eater's perspective across disciplines: health and epidemiology, behavioral and social sciences, nutrition and public health management. The stacked bar chart of disciplinary fields and journals selected for review represents the

[2] Presented in Supplementary Table 1.

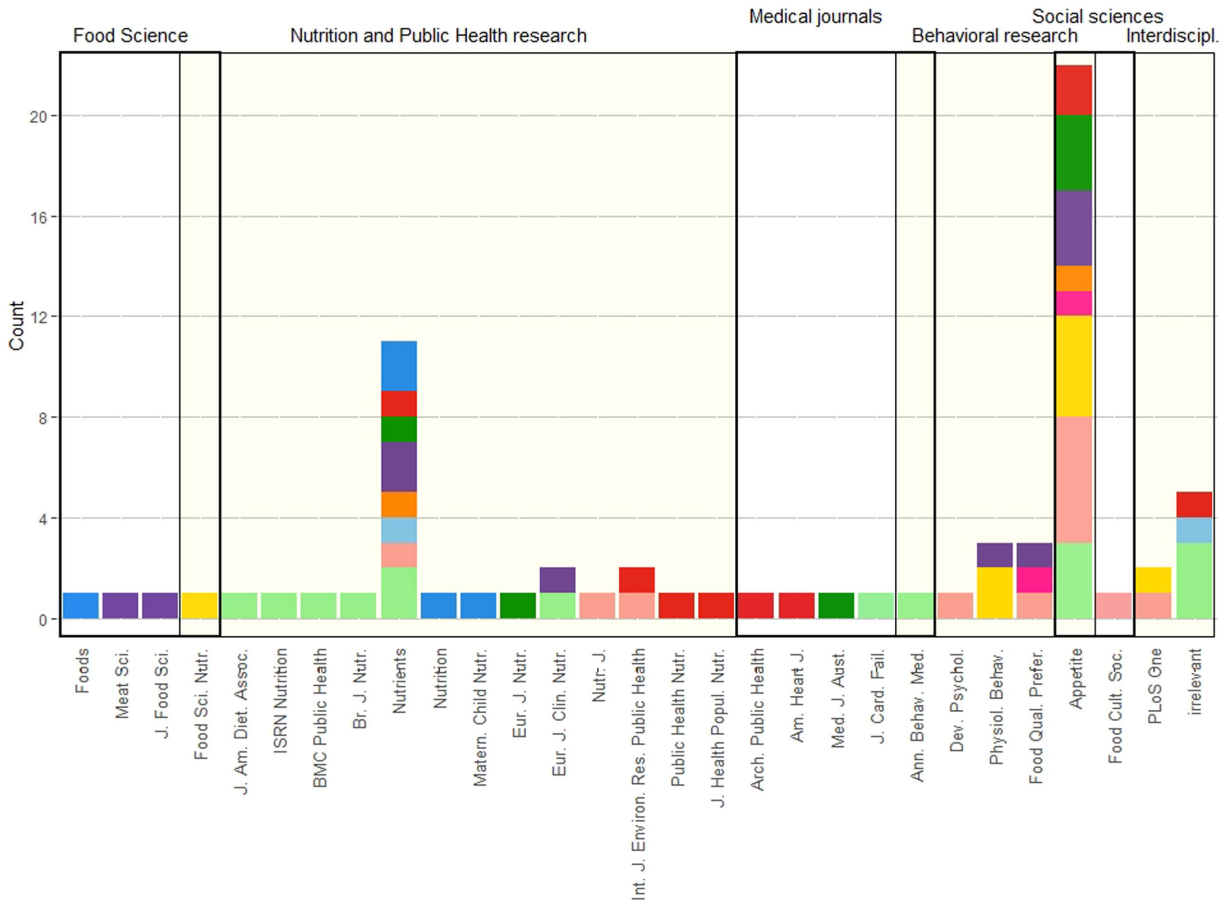


FIGURE 1
Stacked bar chart of research fields.

diversity of disciplines taken into account (Figure 1). The scale of search was limited to articles in English or French[3] published between 2000 and 2022. The search line in English was: ("Salt*" OR "Sodium") AND "Consumption"; ("Salt*" OR "Sodium") AND "Consumption" AND "Home"; "Salt*" AND (("Home*" NEAR "Table" OR "Cook*")); "Discretionary salt"; ("Salt*" OR "Sodium") AND "Consumption" AND "Health"; ("Salt*" OR "Sodium") AND "Nutrition." We have used several search engines: general (Google Scholar); specialized in health (ScienceDirect, NCBI PMC, PubMed); specialized in human and social sciences (Persée, CAIRN, JStore and ISIDORE). After the initial search, we have removed the duplicates and searched manually for studies appearing the reference lists of works already included. Eight criteria were used to facilitate the manual search and watch for the diversity of evidence: [1] search engines; [2] combinations of keywords; [3] publication dates; [4] publication outlets; [5] scientific disciplines; [6] research themes; [7] types of data and study designs; [8] the countries of residence of concerned population. We have finally compared our results to the existing reviews on salt and dietary sodium consumption, watching for the inclusion of the major works on the topic.

In the initial screening phase, two researchers[4] have reviewed the titles, abstracts, and keywords from the resulting documents, coupling them with two expert interviews (with a restaurant chef and his apprentice) and assembling results into a report (45). The report was discussed by a larger scientific consortium of *Sel&Mieux* to identify eligibility criteria and objectives for the current review article[5]. The further steps, performed by three researchers[6], consisted of returning to the full texts of the articles from the report and introducing publications from 2021 and 2022, obtained *via* the same research protocol. We have identified bibliometric data and bibliographic data of each document: authors, types of publication, publication year, DOI, numbers of quotes, summary, methodology, population, countries/territories of origin, institutions). The documents were exported to Zotero software (Zotero version: 6.0.13) for screening for eligibility and categorization.

3.3. Eligibility criteria and selection process

The following criteria were used for screening for eligibility: 1. original data, 2. scientific or otherwise credible data sources (such as national cohort reports), 3. citated elsewhere, 4. significant sampling

[3] The choice of languages is dictated by the authors' language capacities. Restricting the search to the works in French and English may limit the findings of this article, while opening a path for further research in other language frameworks.

[4] PC (100%) and AD (50%).

[5] PC, AD, TT-D, LT, J-PP, AE.

[6] AE, PC, and AD.

size, 5. the study articulates salt consumption and public health. We have also made sure that the retained articles have not been updated by subsequent research results published by the same authors, using a similar methodology and based on a similar sample. All documents, including those left out of the review pool, have been discussed by three researchers[7] to reach agreement on the pertinence of these works for the research objectives.

We have excluded reports from public agencies, scientific book chapters, and non-peer-reviewed articles. However, we have retained institutional reports based on nutritional epidemiological surveys in France (2, 46, 47) since they have largely informed the initial discussion within the scientific consortium of *Sal&Mieux*. Although there exists 'gray literature' on the use of salt by consumers (by agro-industrial companies or food distributors), we could not access it to include in our exploratory literature review. Our selection, therefore, includes only the sources accessible through academic databases.

We have retained 71 published peer-reviewed articles in French and English based on original data, published between 2000 and 2022. The older articles from medical and experimental sciences (prior to the year 2000) were also included if they marked an emergence of themes, sub-fields, and experimental designs still heavily referred to in contemporary research works (n = 12). Publications from the experimental sciences were selected if they provided insights into the specific salting practices and consumers' food habits. We also retained physiological works on hedonic acceptability of salt decrease, the regulation of the appetite for salty products, and on the effects of physical activity on such appetite. The studies dealing with the influence of peers or families, as well with the expression of individual food preferences in different contexts were also retained (e.g., parental control, collective influences).

3.4. Categorization process

Our next step was to synthesize and categorize retained articles from an eater-oriented perspective, that is on the basis of relations and practices at the frontier of food, diet and health. The goal was to come up with new flexible categories, speaking to everyday eaters' practices and capable to contain heterogenous multidisciplinary data.

First, we created a descriptive form to extract authors, journal type, disciplinary fields, population and sampling size, salt/sodium sources discussed, methodology and data type, and key findings for each included article. Second, this descriptive form was enriched by two *ad-hoc* columns: 1. Though what kind of relations/what practices is salt becoming an object of study (what we called "relational features"), and 2. How the study links salt consumption with nutritional recommendations. The entries in these columns were done by three authors[8] in a collective iterative discussion, after reaching a consensus, as the data was entered into the descriptive form. Finally followed the second round of reading, structured by the relational features and links with nutritional recommendations. In this round, we have synthesized six knowledge fields, ten links with nutritional recommendations, nine relational features, and four thematic categories which describe salt consumption. This was done through a discussion and consensus among the authors. The final result of our categorization process is an eater-oriented documentary base of salt consumption factors, which is accessible in Supplementary Table 1. The following section describes this documentary base in more details.

4. Results

The 71 research articles coming from 29 sources[9] were selected for this scoping review. The geography represented in our sampling included North America (United States n = 18; Canada n = 3) and South America (n = 6), European (n = 23), Asian n = 11), African (n = 8), Nordic (n = 3), and Oceanian countries (n = 10). Most of the selected articles focused on one country (n = 64), while five articles took a comparative perspective: Menyanu et al. (9, 15) involved samples from two countries, and four other studies based their analysis on a world-wide cross-country data collection (48–51). Two studies took samples from two countries without a comparative perspective (52, 53). The selected studies were based both on the sampling representing general population and specific population groups (those at risk, for example).

4.1. Eater-oriented categories to represent data on salt consumption

Five different sources of salt intake, or modalities of salt presence in eaters' diet, were identified:

- SF: salt in raw or processed food,
- TS: table salt,
- CS: cooking salt,
- CAS: controlled added salt,
- Global: sodium consumption from all sources

To obtain an overview of daily sodium intake from the consumers, the data was mostly collected *via* dietary recalls before converting food intakes into nutrients, using standard food composition tables[10] (n = 7). Other methodologies included 24-h urine collection tests[11] (n = 10) or both dietary recalls and urine collection tests (54–56). One study used data modelling to assess the potential impact of reformulated products on the population salt intake (48). Out of the 39 studies that included table salt use, 21 used frequency scales (from never to always), binary yes/no questions about salt use in general or self-perceived quantities (from too little to too much) to determine salting habits. Some studies only focused on the use of table salt (TS) and cooking salt (CS; n = 7) (51). Some studied baby-food seasoning (57); table salt compensation (28, 58), and the use of a salt shaker (59). Six groups of authors focused

[7] AE, PC, and AD.

[8] AE, PC, and AD.

[9] Sources other than scientific journals are labelled "irrelevant" [scientific book chapters (n= 2), institutional reports (n= 4)].

[10] ANSES-CIQUAL food composition table (n= 1); Arnault et al.'s "Table de composition des aliments NutriNet-Santé" (2013) (n= 3); Nutrition Data Systems for Research software version 5.0_35 (n= 1); Ethiopian and Tanzanian food composition tables (n= 1); United States, United Kingdom and Dutch food composition table (n= 1); unknown conversion method (55,566).

[11] In this section whenever the number of articles is 3 or more, we do not repeat authors' names leaving just the number of articles. The exact names can be found in the Supplementary Table 1.

on salty foodstuffs only (SF; e.g., industrial food, salty snacks, bread, spreads and dressings) (60–65).

We have extracted from the retained articles nine non-exclusive relational features, that is through what type of relations or practices does the problem of salt consumption emerge:

- Intake: concerns consumption or absorption levels during meal preparation and consumption.
- Knowledge: concerns nutritional and culinary dimensions.
- Awareness: concerns health-related risks.
- Beliefs: concerns values and representations of the benefits or harms of salt.
- Attitudes: concerns form of justification, "relation to," "attitude toward" recommendations, prescriptions or communications.
- Behaviors/social practices: concerns routines and habits.
- Liking or Preference: concerns hedonic dimension.
- Taste perception: concerns registration of salty tastes.
- Consumer practices: purchasing salt and salt containing products.

In some instances, we have accompanied each relational feature with a more specific connotation, such as 'caregiving practices', 'DS use', 'commercial value'.

Finally, four *ad hoc* categories synthesize existing knowledge into the dietary sodium consumption factors: [1] "Socio-demographic and cultural descriptors of salt consumers"; [2] "Knowledge, attitudes, and beliefs on nutritional norms"; [3] "Salt practices associated with dietary or medical regimes"; [4] "Salt materialities: interactions and contexts." The categories are formulated from the articles' keywords, titles, research questions and findings. Some formulations, such as 'descriptors of salt consumers' or 'knowledge, attitudes, beliefs' are taken directly from keywords, titles or results. Others are formulated after reading into summary by three researchers[12]. In the latter case, 'salt materialities' refer to the translation of salt performances and figuration of salt as both material and symbolic object configured in practice. This is inspired by Bowker and Star's analysis of classification as a product of action and relations, creating boundaries between communities of practice (66).

The Table 1 synthesizes our corpus through the categories resulting from the scoping review. It includes [1] The four thematic categories. [2] Sampling (representative or randomized). [3] The sources of sodium intake investigated. [4] The type of data or methodology. [5] The relational features with references. The columns and rows of this table were adjusted to map the studies that have similarities to each other.

4.2. Salt consumption/health articulation

In each category, we have identified six knowledge fields which articulate salt consumption and health:

- Food Science and Nutrition;
- Nutrition and Public Health;
- Medical and Behavioral Studies;
- Behavioral Research and Social Science;
- Social Science;
- Interdisciplinary.

We have also identified ten different strategic and methodological approaches to the issue of excessive sodium consumption, which we called 'links with nutritional recommendations':

- Assessment of salt reduction intervention ($n = 5$).
- Awareness and knowledge as strategies for salt reduction ($n = 9$).
- Compliance with nutritional guidelines ($n = 6$).
- Consumer acceptance ($n = 10$).
- Identification of barriers to salt reduction ($n = 2$).
- Impact of material environment on salt usage ($n = 2$).
- Physiology-biological determinants ($n = 8$).
- Socio-cultural determinants and awareness and knowledge as strategies for salt reduction ($n = 2$).
- Socio-cultural determinants ($n = 12$).
- Sodium reduction ($n = 15$).

The Table 2 demonstrates the intersections between the research fields represented by specific journals, links with nutritional recommendations, and types of population concerned in the thematic category [1] "Socio-demographic and cultural descriptors of salt consumers."

4.3. Salt and dietary sodium consumption factors as guidelines for salt reduction

In this section we will briefly discuss each thematic category of salt consumption factors, providing an overview of existing knowledge. As the aim of this article is a scoping review to propose an interdisciplinary and eater-oriented knowledge framework, we will not discuss each article individually. The findings of the individual articles can be found in the documentary base presented in the Supplementary Table 1.

4.3.1. Socio-demographic and cultural descriptors of salt consumers

Eighteen works have been classified in this category. Ten of them were national demographic studies, and six were based on population-representative samples. Nine studies used declarative data to measure the overall salt intake or specific salty food consumption (e.g., snacks) among particular population. Four institutional reports presented declarative data obtained from the representative samples of French population (2, 46, 47, 67). None of these national surveys used 24-h urine collection as methodology. Studies using objective measures were rare in this first category. Some of them aimed to evaluate compliance with nutritional guidelines: Piovesana, Sampaio, and Gallani, for example, focused on a random sample of 108 hypertensive and normotensive participants (54). Iacone et al. in their study of the influence of iodized salt consumption on the iodine levels among the pediatric population evaluated compliance with nutritional guidelines regarding sodium and iodine (8). The study of Huggins et al. (68) characterized table salt consumption with a cohort sample of 784 Australians.

Five articles were based on data from controlled experiments focused on the preferences for salt. Two articles evaluated gustatory perception thresholds, highlighting socio-demographic, cultural, and biological descriptors of salt preferences and perception patterns (54, 69). Beauchamp and Cowart (70), meanwhile, treated salt consumption in perspective with socio-economic and racial characteristics. Finally, this category also contains articles presenting cultural descriptions of salt preferences. Kerrihard et al.'s (71) work, for example, explored the effect of acclimation in the United States on hedonic evaluation of salt.

[12] PC, AD, and AE.

TABLE 1 Corpus synthesis through the eater-oriented categories.

Thematic category				*Sampling*		*Salt sources*					*Data type*					
Socio-demographic and cultural...	Knowledge, attitudes, and beliefs...	... dietary or medical regimes	Salt materialities...	Representative	Random	SF	TS	CS	CAS	Global	Declarative data	Objective measures	Collected data using experiments	Qualitative data	*n*=	Relational features (and article reference in the References)
			X		X	X							X		1	• Consumer practices, commercial values (58)
			X		X	X							X		1	• Behaviours (60)
X				X		X					X				1	• Intake (56)
	X				X	X					X				1	• Awareness, knowledge, attitude (57)
			X		X	X					X				1	• Intake (43)
X					X	X	X	X			X				1	• Liking (102)
	X				X	X	X	X			X				5	• Attitude (105, 106) • Knowledge, attitude, behaviours, beliefs (27, 44, 106)
		X			X	X	X	X			X				1	• Attitude (78)
			X		X	X	X	X			X		X		1	• Preference (111)
	X				X	X	X	X			X	X			1	• Awareness, knowledge, attitude (68)
	X				X	X	X	X						X	3	• Attitude, opinions, practices (13) • Intake, caregiving practices (70) • Knowledge, beliefs (71)
			X		X	X	X	X						X	1	• Attitude, practices (31)
		X			X	X	X	X				X		X	1	• Attitude, behaviours (75)
X				X		X	X	X			X				2	• Intake, awareness, knowledge, behaviours (32, 104)
	X			X		X	X	X			X				3	• Attitude (73) • Intake, beliefs, behaviours (47) • Knowledge, beliefs, attitude (45)
X					X	X					X				1	• Liking, behaviours, attitude (61)
			X		X	X		X			X				1	• Attitude (89)
			X		X	X							X		1	• Behaviours (59)
X					X		X	X			X				2	• Attitude (100) • Behaviours (46)
		X			X		X	X			X				1	• Intake (107)
		X		X			X	X			X				1	• Intake, caregiving practices (79)
			X		X		X	X						X	1	• Intake, caregiving practices (53)
			X		X		X						X		1	• Behaviours (54)
			X		X		X		X				X		2	• Behaviours, preference (28, 55)
X					X				X				X		3	• Preference (65, 66, 103)

(Continued)

TABLE 1 (Continued)

Thematic category				*Sampling*		*Salt sources*					*Data type*				*n=*	Relational features (and article reference in the References)
Socio-demographic and cultural…	Knowledge, attitudes, and beliefs…	… dietary or medical regimes	Salt materialities…	Representative	Random	SF	TS	CS	CAS	Global	Declarative data	Objective measures	Collected data using experiments	Qualitative data		
			X		X				X				X		11	• Behaviours (88) • Preference (82, 84, 86, 108, 109, 111) • Preference, taste perception (85, 87, 110)
X					X				X		X		X		1	• Attitude, preference (101)
			X		X				X		X		X		1	• Preference (90)
		X			X					X	X				2	• Intake, attitude (77, 80)
X					X					X	X				1	• Intake (2)
X					X					X	X	X			1	• Intake (63)
X					X					X	X	X	X		1	• Intake, taste perception (50)
	X				X					X	X				1	• Intake, attitude , behaviours (49)
	X				X					X	X	X			3	• Intake, caregiving practices (69) • Intake, knowledge, attitude, behaviours (26, 51)
		X			X					X	X	X			1	• Awareness, attitude (48)
			X		X					X	X	X			2	• Intake (8, 81)
	X			X						X	X	X			2	• Awareness, beliefs, attitude (67, 52)
		X		X						X	X	X			1	• Intake, attitude (74)
X				X						X		X			1	• Intake (7)
X				X						X	X				2	• Intake (40, 62)

TABLE 2 Salt consumption/health articulation in the thematic category [1] "Socio-demographic and cultural descriptors of salt consumers".

Research fields (n =)	*Populations targeted by nutritional recommendations (n =)*	*Links with nutritional recommendations*	*Journals (n =)*
Food Science and Nutrition (4)	*General population* (4)	Assessment of salt reduction intervention	*Foods* (1)
		Consumer acceptance	*Meat Science* (1)
		Sodium reduction	*J Food Sci* (1)
		Physiology-biological determinants	*Food Sci Nutr* (1)
Nutrition and Public Health (25)	*General population* (17)	Assessment of salt reduction intervention	*Nutrients* (2) *Nutrition* (1)
		Awareness and knowledge as strategies for salt reduction	*Int J Environ Res Public Health* (1) *J Health Popul Nutr* (1) *Public Health Nutr* (1) *Nutrients* (1)
		Consumer acceptance	*Eur J Clin Nutr* (1) *Nutrients* (2)
		Identification of barriers to salt reduction	*Nutrients* (1)
		Socio-cultural determinants & Awareness and knowledge as…	*Nutrients* (1)
		Socio-cultural determinants	*Int J Environ Res Public Health* (1) *Nutrients* (1) *Nutr J* (1)
		Sodium reduction	*Nutrients* (1) *Eur J Clin Nutr* (1)
	Specific group of the population (8)	Assessment of salt reduction intervention	*Matern Child Nutr* (1)
		Compliance with nutritional guidelines	*Eur J Nutr* (1) *Nutrients* (1)
		Sodium reduction	*J Am Diet Assoc* (1) *ISRN Nutrition* (1) *BMC Public Health* (1) *Br J Nutr* (1) *Nutrients* (1)
Medical & Behavioural Studies (5)	*General population* (1)	Compliance with nutritional guidelines	*Med J Aust* (1)
	Specific group of the population (4)	Awareness and knowledge as strategies for salt reduction	*Am Heart J* (1) *Arch Public Health* (1)
		Sodium reduction	*J Card Fail* (1) *Ann Behav Med* (1)
Behavioural Research & Social Science (28)	*General population* (17)	Awareness and knowledge as strategies for salt reduction	*Appetite* (2)
		Compliance with nutritional guidelines	*Appetite* (2)
		Consumer acceptance	*Appetite* (3) *Food Qual Prefer* (1) *Physiol Behav* (1)
		Identification of barriers to salt reduction	*Appetite* (1)
		Impact of material environment on salt usage	*Appetite* (1) *Food Qual Prefer* (1)
		Physiology-biological determinants	*Appetite* (2)
		Socio-cultural determinants	*Appetite* (2)
		Sodium reduction	*Appetite* (1)
	Specific group of the population (11)	Compliance with nutritional guidelines	*Appetite* (1)
		Physiology-biological determinants	*Appetite* (2) *Physiol Behav* (2)
		Socio-cultural determinants	*Appetite* (3) *Dev Psychol* (1)
		Sodium reduction	*Appetite* (2)
Social Science (1)	*Specific group of the population* (1)	Social and cultural determinants	*Food Cult Soc* (1)
Interdisciplinary (7)	*General population* (2)	Awareness and knowledge as strategies for salt reduction	Irrelevant (1)
		Socio-cultural determinants & Awareness and knowledge as…	Irrelevant (1)
	Specific group of the population (5)	Physiology-biological determinants	*Plos One* (1)
		Socio-cultural determinants	*Plos One* (1)
		Sodium reduction	Irrelevant (3)

Drewnowski et al. (69) focused on the relationship between age, gender, perception of salt taste, and actual sodium consumption.

4.3.2. Knowledge, attitudes, and beliefs on nutritional norms

This category includes 21 works. Like in the previous category, declarative data was used in studies relying on questionnaires or food records ($n = 14$). Six publications combined questionnaires and objective measures (24-h urinary sodium excretion). Four works were based on the qualitative material obtained through the semi-structured interviews and focus groups; one relied on intervention to reduce added salt during cooking (56).

The choice to include knowledge, attitudes, and beliefs in the same category was consistent with the combination of keywords chosen by the articles' authors (15, 49, 55, 56, 61, 72). However, some works demonstrate the relevance to distinguish the roles of knowledge, attitudes, and beliefs, since practices and recommendations in the three cases may differ. In some cases, beliefs and knowledge linked to chronic illnesses have been studied to investigate how knowledge structures diet in diabetes management (73). Other studies treated the transmission of knowledge and everyday skills from dieticians as compared to the self-help literature for people with hypertension (74). Rhodes et al. (64), also addressed the transmission of knowledge within the multicultural and intergenerational families. Family, therefore, can be considered as an institution for socialization to salt consumption (75, 76).

The barriers to salt reduction among the general population also appear in this category ($n = 5$), often pointing to knowledge about the official dietary sodium recommendations (50). Another mention is the difficulty of differentiating between sodium and salt, as well as calculating the ratio of one to another (61, 77). Other issues addressed in this category include knowledge about the sources of sodium intake ($n = 8$), awareness about the salt-related health risks ($n = 10$), beliefs about the nutritional or symbolic values of certain salts (e.g., sea salt, iodized salt) (52, 56).

This category, finally, includes socio-psychological factors which play a role in salt consumption, reinforcing Ahn et al.'s (78) invitation for a tailored intervention approach based on case-by-case public health campaigns and addressing different stages of behavioral change. Outcome expectancy, barriers, knowledge, purchasing skills also deserve to be seen as marketing issues, as salt reduction is best achieved and maintained with the concrete goals and rewards (79).

4.3.3. Salt practices associated with dietary or medical regimes

This thematic category includes eight works on adherence and attitudes toward nutritional and medical guidelines, as well as on the impact of some dietary regimes on salt intake. All records in this category used reported data collection in the form of questionnaires and diet recalls. Four studies went further, measuring excreted sodium from participant's urine or comparing declarative data to the data obtained *via* objective biomedical measurements. In these cases, objective measures allowed: (a) to evaluate the effects of a multi-faceted and population-wide salt reduction intervention (80); (b) to assess the impact of a controlled experiment where healthy adults would follow an appropriate amount of daily salt intakes (81); (c) to assess the impact of a controlled experiment where people with heart failure would follow an appropriate low-sodium diet (53); (d) to assess the role of the absence of the table salt or the use of salt substitutes on hypertension and stroke (82). Articles by Chung and collaborators, as well as those by Adriouch et al. and Henson et al. (53, 83, 84), put salt consumption in perspective with cardiometabolic diseases, arguing for the necessity of a particular dietary regime. We have finally included in this category the articles on dietary regimes: Bournez et al.'s (85) work on children's dietary regimes; Dyett et al.'s (86) paper on the effect of daily vegan diet on sodium and other nutrients' intake.

4.3.4. Salt materialities: Interactions and contexts

Our fourth category assembles studies ($n = 24$) which regard sodium as belonging to or interacting with different matters, including the influence of the environment. The predominant methodology here is experimental protocols ($n = 18$); some have relatively small study samples (participants are less than 83 in 55% of cases). Experiments generally rely on pre-salted preparations or solutions (called CAS for "controlled added salt"; $n = 16$). A significant proportion of reviewed articles ($n = 12$) deal with the physiological processes related to salt consumption and iodine intake (9, 87). The study of Frye and Demolar (88), for example, attempted to relate sodium intake to women's menstrual cycle; however, the results did not reveal a dependency. A number of works dealt with the physico-chemical composition of the meal, physiological interactions, and preferences (89, 90); with the table salt compensation strategies (59) or personal acceptability of salt reduction (91).

The second common feature is that the articles in this category pay attention to how material elements or specific environment influence perception and preferences (31, 92, 93), behaviors (58, 94), attitudes (95), caregiving (57), and consumer practices (62, 63). This is an important relational factor in understanding salt consumption. Materialities identified in this category also include variations of the widths of holes in salt shakers (58); noise variations during tasting (94); the influence of summer heat on sodium loss and salt avidity (96), and low-salt food alternatives (63). Finally, one article in this category approached salt materialities through data modelling methodology (48). The authors conducted dietary impact modelling to demonstrate that product reformulation by the food industry has the potential to contribute substantially to salt-intake reduction without jeopardizing products' taste values, yet this process should be supported by a multi-stakeholder approach (48).

5. Discussion and further research directions

Salt and its excessive intake have been approached from manyfold research positions: scholarly works on salt reduction can be found in the fields of nutrition and food sciences, public health and health management, biomedical and clinical sciences, cultural anthropology, social and behavioral sciences, and interdisciplinary works. In this section, we discuss the results of our scoping review through the lens of other reviews conducted on the subject of salt consumption, identifying the dominant approaches and the novelty of our eater-oriented knowledge framework for building effective public health interventions.

The reviews of salt consumption and reduction can be divided into three approaches. The first is an *evidence-providing approach* for regulating high rates of salt consumption. The examples of such approach are: Moschonis and Karatzi's (97) study of dietary approaches for decreasing hypertension (the major factor contributing to cardiovascular diseases); Campbell and Train's (98) argument for labelling the dangers of salt on the packages and shakers; Wong et al.'s (21) assessment of the studies on dietary salt in relation to health outcomes. This approach is also present is the systematic reviews of interventionist scientific studies. For instance, Tsirimiagkou et al. (99) reviewed the scientific literature on the relationships between sodium intake and three cardiovascular diseases (arteriosclerosis, arterial remodeling and atheromatosis), concluding that the issue requires further interventional scientific studies.

The second approach is *managerial approach to salt reduction*. Through this approach, He et al. (6) reviewed different strategies of salt consumption management. The authors also evoked Asaria et al.'s (100) argument of cost-effectiveness of salt intake decrease for combatting the epidemics of cardiovascular diseases in the developing countries, where salt intake is extremely high. Jaenke et al. (101) have taken a further step in researching salt overconsumption management, conducting a systematic literature review to understand how products can be reformulated for lesser salt without losing consumer acceptability. Their results have shown that a < 40% salt reduction in breads and approximately 70% in processed meats (obtained as a result of sodium compensation and/or replacement) would not significantly impact consumer acceptability. Some other examples of a managerial approach to reviews on salt overconsumption include Regan et al.'s (29) review of the current reformulation strategies in regard to consumer behavior or Eyles et al.'s (19) study on the use of smartphone apps for salt reduction.

Finally, some reviews on salt reduction take a *nutri-behavioural, socio-demographic and cultural approach* to the problem of salt overconsumption. Darmon and Drewnowski's review, for example, associates higher sodium intake with lower socioeconomic status as the latter supposes lower quality diet due to the limited economic resources (102). In another study, Laisney (103) noted that teenagers from lower socio-economic backgrounds are especially at risk for salt consumption. This approach also reveals regional specificities of sodium intake.

In this article, we aimed for the development of a new, *eater-oriented approach* at the intersection of salt consumption and public health. This approach challenges the boundaries between the different communities of practice (nutritional science, sociology, public health, etc.) and assembles heterogeneous knowledge on salt intake around the figure of an eater. Such an approach serves as a model for flexible and eater-oriented databases on dietary sodium consumption for effective public health interventions. Our eater-oriented categories synthesize a wide array of relational dynamics involved in salt consumption: socio-demographic, methodological, contextual, technical and technological, affective, communicative, and deliberative. These dynamics can translate as interdisciplinary guidelines for decreasing salt intake.

We have identified four non-exclusive thematic categories which help to understand salt consumption: [1] Socio-demographic and cultural descriptors of salt consumers; [2] Knowledge, attitudes, and beliefs on nutritional norms; [3] Salt practices associated with dietary or medical regimes; [4] Salt materialities: interactions and contexts. Each thematic category gravitates toward particular relational feature: the second block deals mostly with attitudes, beliefs, knowledge; the first block concerns mostly intake and preference. The third category presents a wide relational scale, from caregiving practices to intake and preferences. Finally, salt materialities are associated mostly with preferences and taste perceptions. We propose that the variety of salt sources, techniques, and technologies of salting, as well as different modalities of sodium appearances in the human diet should be further researched. Taking salt materialities and interactions seriously is important for understanding compensation strategies during salt reduction programs (48, 59) and also for a coherent coexistence of different public health initiatives. The latter point has been raised by Iacone et al. and Menyanu et al. (8, 9) in relation with the usage of iodized salt for preventing iodine deficiencies.

Our scoping review shows that while there are numerous reports relying on declarative or experimental data (column G in the Supplementary Table 1), there is little qualitative data on the practices of salt consumption. Most studies in our review drew data from dietary recalls or 24-h urine collection tests. However, Blanco-Metzler et al.'s (31) explorative qualitative study of the food practices and perceptions related to excessive consumption of salt/sodium in Costa Rica shows the benefit of approaching salt overconsumption as a complex phenomenon across the different thematic categories. Blanco-Metzler et al. relied on ethnography and ethnology to understand salt-related practices from the participants' perspective, watching out at the same time for different environmental contexts (different regions, eating out/at home), different socio-demographic and cultural profiles (age, sex, cultural roots, etc.), and individual knowledge, beliefs, and perceptions. We propose that future works employ cross-thematic qualitative approach from eaters' perspective, and we argue that our categorization system can be used as framework for conducting qualitative interviews and observations. The documentary base presented in this article (Supplementary Table 1) can serve as a guide for this endeavor.

6. Conclusion

In this article, we argue that in order to achieve the WHO goal for decreasing the amount of dietary sodium consumed by 30% by 2025 there is a need for interdisciplinary and eater-oriented knowledge framework, which would include nutritional knowledge, as well as the dominant beliefs, attitudes, and practices of salt consumption. This knowledge framework would serve as guidelines for building flexible databases, informing public health campaigns. As health-and-diet databases are performative, that is to say that they both emerge from and perpetuate certain practices (38, 104), they have a considerable implication for the social and natural orders. For example, a relational database of rare diseases in France (assembling knowledge across communities of patients, health practitioners, and institutions) have challenged the production of knowledge on rare diseases (105). New eater-related databases of salt and sodium consumption, we argue, may not only inform more effective public health measures, but also make a subject more accessible for the general population.

We therefore have built an eater-oriented documentary base (available in the Supplementary Table 1), which would serve as knowledge framework for further databasing the factors of dietary sodium consumption. For this, we conducted a scoping review of existing academic literature on dietary sodium consumption, following a PRISMA-ScR checklist and best practices for scoping reviews (40, 41). We have selected 71 studies published in English and French and falling into the nexus of salt consumption and public health for a detailed review and categorization. The selected works presented a heterogeneous pool of disciplines, methodologies, geographies, salt sources, population samples, and data types. Through the two steps of categorization, we have designed a knowledge database around eater-oriented interdisciplinary categories: [1] Socio-demographic and cultural descriptors of salt consumers; [2] Knowledge, attitudes, and beliefs on nutritional norms; [3] Salt practices associated with dietary or medical regimes; [4] Salt materialities: interactions and contexts. We have also categorized each article according to the dominant relational features (how salt becomes an issue): Intake; Knowledge; Awareness; Beliefs; Attitudes; Behaviors/social practices; Liking or Preference; Taste perception; Consumer practices. Finally, we have extracted ten different strategic and methodological approaches to the issue of excessive sodium consumption, which we called 'links with nutritional recommendations'. The synthesis of the resulted knowledge framework is presented in the Table 1, and the full documentary base—in the Supplementary Table 1. This documentary base, we argue, can serve as a framework to classify empirical and contextualized data in order to design an adequate public health response to the issue of dietary sodium consumption. Following this perspective, we have proposed guidelines for further research on salt and sodium consumption, as well as for effective public health interventions. These guidelines accentuate the contexts of food intake, eaters' knowledge, habits and practices, cultural predispositions, meal preparation routines, and consumption environment.

Author contributions

AE: conceptualization, data collection, investigation, analysis, writing—original draft, and writing—review and editing. PC (equal contribution with AE): data collection, investigation, software, visualization, writing—original draft, and writing—review and editing. TT-D: funding acquisition, project administration, and writing—review and editing. J-PP: writing—review and editing and resources. LT: writing—review and editing. AD: project administration, investigation, methodology, writing—original draft, writing—review and editing, and supervision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the French National Research Agency (ANR) as part of the Sal&Mieux project under grant (ANR-19-CE21-0009).

Acknowledgments

We would like to thank Christophe Lavelle and the chefs Lionel Beylat and Vincent Gayraud for accepting an expert interview. We also thank Christelle Abraham for the administrative support of this research project, as well as CNRS and Université Toulouse II JeanJaurès for facilities and hosting the project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2023.1110446/full#supplementary-material

References

1. WHO. Global action plan for the prevention and control of noncommunicable diseases 2013–2020. (2013). Available at: https://www.who.int/publications/i/item/9789241506236 (Accessed November 25, 2022).

2. NutriNet-Santé. État d'avancement et résultats préliminaires 18 mois après le lancement. (2010). Available at: http://media.etude-nutrinet-sante.fr/download/dossier_presse_nutrinet_22_11_10.pdf (Accessed November 25, 2022).

3. Kwong, EJL, Whiting, S, Bunge, AC, Leven, Y, Breda, J, Rakovac, I, et al. A systematic review. *Public Health Nutr.* (2022) 2022:1–14. doi: 10.1017/S136898002200218X

4. Contreras, J. Sel In: JP Poulain, editor. *Dictionnaire des cultures alimentaires. 1st* ed. Paris: Presses universitaires de France (2012). 1233–7.

5. Durack, E, Alonso-Gomez, M, and Wilkinson, M. Salt: a review of its role in food science and public health. *Curr Nutr Food Sci.* (2008) 4:290–7. doi: 10.2174/157340108786263702

6. He, FJ, Jenner, KH, and MacGregor, GA. WASH—world action on salt and health. *Kidney Int.* (2010) 78:745–53. doi: 10.1038/ki.2010.280

7. Bouhlal, S, Issanchou, S, and Nicklaus, S. The impact of salt, fat and sugar levels on toddler food intake. *Br J Nutr.* (2011) 105:645–53. doi: 10.1017/S0007114510003752

8. Iacone, R, Iaccarino Idelson, P, Campanozzi, A, Rutigliano, I, Russo, O, Formisano, P, et al. Relationship between salt consumption and iodine intake in a pediatric population. *Eur J Nutr.* (2021) 60:2193–202. doi: 10.1007/s00394-020-02407-w

9. Menyanu, E, Corso, B, Minicuci, N, Rocco, I, Zandberg, L, Baumgartner, J, et al. Salt-reduction strategies may compromise salt iodization programs: learnings from South Africa and Ghana. *Nutrition.* (2021) 84:111065. doi: 10.3390/nu9090939

10. Zimmermann, MB. The impact of iodised salt or iodine supplements on iodine status during pregnancy, lactation and infancy. *Public Health Nutr.* (2007) 10:1584–95. doi: 10.1017/S1368980007360965

11. Skeaff, SA, and Lonsdale-Cooper, E. Mandatory fortification of bread with iodised salt modestly improves iodine status in schoolchildren. *Br J Nutr.* (2013) 109:1109–13. doi: 10.1017/S0007114512003236

12. Bergmann, KE, and Bergmann, RL. Salt fluoridation and general health. *Adv Dent Res.* (1995) 9:138–43. doi: 10.1177/08959374950090021401

13. Nath, SK, Moinier, B, Thuillier, F, Rongier, M, and Desjeux, JF. Urinary excretion of iodide and fluoride from supplemented food grade salt. *Int J Vitam Nutr Res.* (1992) 62:66–72. PMID: 1587711. PMID: 1587711

14. Pesantes, M, Diez-Canseco, F, Bernabé-Ortiz, A, Ponce-Lucero, V, and Miranda, J. Taste, salt consumption, and local explanations around hypertension in a rural population in northern Peru. *Nutrients.* (2017) 9:698. doi: 10.3390/nu9070698

15. Menyanu, E, Charlton, K, Ware, L, Russell, J, Biritwum, R, and Kowal, P. Salt use Behaviours of Ghanaians and south Africans: a comparative study of knowledge, attitudes and practices. *Nutrients.* (2017) 9:939. doi: 10.3390/nu9090939

16. Elorriaga, N, Gutierrez, L, Romero, I, Moyano, D, Poggio, R, Calandrelli, M, et al. Collecting evidence to inform salt reduction policies in Argentina: identifying sources of sodium intake in adults from a population-based sample. *Nutrients.* (2017) 9:964. doi: 10.3390/nu9090964

17. Forouzanfar, MH, Alexander, L, Anderson, HR, Bachman, VF, Biryukov, S, Brauer, M, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *Lancet.* (2015) 386:2287–323. doi: 10.1016/S0140-6736(15)00128-2

18. Stanaway, JD, Afshin, A, Gakidou, E, Lim, SS, Abate, D, Abate, KH, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet.* (2018) 392:1923–94. doi: 10.1016/S0140-6736(18)32225-6

19. Eyles, H, McLean, R, Neal, B, Jiang, Y, Doughty, RN, McLean, R, et al. A salt-reduction smartphone app supports lower-salt food purchases for people with cardiovascular disease: findings from the SaltSwitch randomised controlled trial. *Eur J Prev Cardiol.* (2017) 24:1435–44. doi: 10.1177/2047487317715713

20. Powles, J, Fahimi, S, Micha, R, Khatibzadeh, S, Shi, P, Ezzati, M, et al. Global, regional and national sodium intakes in 1990 and 2010: a systematic analysis of 24 h urinary sodium excretion and dietary surveys worldwide. *BMJ Open.* (2013) 3:e003733. doi: 10.1136/bmjopen-2013-003733

21. Wong, MMY, Arcand, J, Leung, AA, Thout, SR, Campbell, NRC, and Webster, J. The science of salt: a regularly updated systematic review of salt and health outcomes (December 2015-march 2016). *J Clin Hypertens.* (2017) 19:322–32. doi: 10.1111/jch.12970

22. He, FJ, and MacGregor, GA. Reducing population salt intake worldwide: from evidence to implementation. *Prog Cardiovasc Dis*. (2010) 52:363–82. doi: 10.1016/j.pcad.2009.12.006

23. Batuman, V. Salt and hypertension: why is there still a debate? *Kidney Int Suppl*. (2013) 3:316–20. doi: 10.1038/kisup.2013.66

24. Mozzaffarian, D, Singh, G, and Powles, J. Sodium and cardiovascular disease. *N Engl J Med*. (2014) 371:2134–9. doi: 10.1056/NEJMc1412113

25. Abarca-Gómez, L, Abdeen, ZA, Hamid, ZA, Abu-Rmeileh, NM, Acosta-Cazares, B, Acuin, C, et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *Lancet*. (2017) 390:2627–42. doi: 10.1016/S0140-6736(17)32129-3

26. Johnson, C, Mohan, S, Rogers, K, Shivashankar, R, Thout, S, Gupta, P, et al. The association of knowledge and behaviours related to salt with 24-h urinary salt excretion in a population from north and South India. *Nutrients*. (2017) 9:144. doi: 10.3390/nu9020144

27. Wentzel-Viljoen, E, Steyn, K, Lombard, C, De Villiers, A, Charlton, K, Frielinghaus, S, et al. Evaluation of a mass-media campaign to increase the awareness of the need to reduce discretionary salt use in the south African population. *Nutrients*. (2017) 9:1238. doi: 10.3390/nu9111238

28. De Kock, HL, Zandstra, EH, Sayed, N, and Wentzel-Viljoen, E. Liking, salt taste perception and use of table salt when consuming reduced-salt chicken stews in light of South Africa's new salt regulations. *Appetite*. (2016) 96:383–90. doi: 10.1016/j.appet.2015.09.026

29. Regan, Á, Kent, M, Raats, M, McConnon, Á, Wall, P, and Dubois, L. Applying a consumer behavior lens to salt reduction initiatives. *Nutrients*. (2017) 9:901. doi: 10.3390/nu9080901

30. Webster, JL, Dunford, EK, and Neal, BC. A systematic survey of the sodium contents of processed foods. *Am J Clin Nutr*. (2010) 91:413–20. doi: 10.3945/ajcn.2009.28688

31. Blanco-Metzler, A, Núñez-Rivas, H, Vega-Solano, J, Montero-Campos, MA, Benavides-Aguilar, K, and Cubillo-Rodríguez, N. Household cooking and eating out: food practices and perceptions of salt/sodium consumption in Costa Rica. *Int J Environ Res Public Health*. (2021) 18:1208. doi: 10.3390/ijerph18031208

32. Santos, JA, Tekle, D, Rosewarne, E, Flexner, N, Cobb, L, Al-Jawaldeh, A, et al. A systematic review of salt reduction initiatives around the world: a midterm evaluation of Progress towards the 2025 global non-communicable diseases salt reduction target. *Adv Nutr*. (2021) 12:1768–80. doi: 10.1093/advances/nmab008

33. Durazzo, A, and Lucarini, M. Editorial: databases and nutrition. *Front Nutr*. (2022) 9:853600. doi: 10.3389/fnut.2022.853600

34. Samaniego-Vaesken, MDL, González-Fernández, B, Partearroyo, T, Urrialde, R, and Varela-Moreiras, G. Updated database and trends of declared low-and no-calorie sweeteners from foods and beverages marketed in Spain. *Front Nutr*. (2021) 8:670422. doi: 10.3389/fnut.2021.670422

35. Poulain, JP, Tibère, L, Mognard, E, Laporte, C, Fournier, T, Noor, IM, et al. The Malaysian food barometer open database: an invitation to study the modernization of Malaysian food patterns and its economic and health consequences. *Front Nutr*. (2022) 8:800317. doi: 10.3389/fnut.2021.800317

36. Leonelli, S. What counts as scientific data? A relational framework. *Philos Sci*. (2015) 82:810–21. doi: 10.1086/684083

37. Bowker, GC. Biodiversity Datadiversity. *Soc Stud Sci*. (2000) 30:643–83. doi: 10.1177/030631200030005001

38. Waterton, C. Experimenting with the archive: STS-ers as analysts and co-constructors of databases and other archival forms. *Sci Technol Hum Values*. (2010) 35:645–76. doi: 10.1177/0162243909340265

39. Paré, G, Trudel, MC, Jaana, M, and Kitsiou, S. Synthesizing information systems knowledge: a typology of literature reviews. *Inf Manag*. (2015) 52:183–99. doi: 10.1016/j.im.2014.08.008

40. Tricco, AC, Lillie, E, Zarin, W, O'Brien, KK, Colquhoun, H, Levac, D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. (2018) 169:467–73. doi: 10.7326/M18-0850

41. Peters, MDJ, Godfrey, C, McInerney, P, Khalil, H, Larsen, P, Marnie, C, et al. Best practice guidance and reporting items for the development of scoping review protocols. *JBI Evid Synth*. (2022) 20:953–68. doi: 10.11124/JBIES-21-00242

42. Truman, E, Lane, D, and Elliott, C. Defining food literacy: a scoping review. *Appetite*. (2017) 116:365–71. doi: 10.1016/j.appet.2017.05.007

43. Azevedo Perry, E, Thomas, H, Samra, H, Edmonstone, S, Davidson, L, Faulkner, A, et al. Identifying attributes of food literacy: a scoping review. *Public Health Nut*. (2017) 20:2406–15. doi: 10.1017/S1368980017001276

44. Vonthron, S, Perrin, C, and Soulard, CT. Foodscape: a scoping review and a research agenda for food security-related studies. *PloS ONE*. (2020) 15:e0233218. doi: 10.1371/journal.pone.0233218

45. Dupuy, A, Coeurquetin, P, Poulain, JP, Tibère, L, and Zancanaro, F. Rapport de recherche task 1.1 WP1—Systematic review of the use of table and cooking salt at home, ANR SalEtMieux (ANR-19-CE21-0009). (2020) Available at: https://hal.archives-ouvertes.fr/hal-03251571 (Accessed November 25, 2022).

46. ANSES. Étude individuelle nationale des consommations alimentaires 3 (INCA 3). Avis de l'ANSES, Rapport d'expertise scientifique. Édition scientifique. (2017). Available at: https://www.anses.fr/fr/system/files/NUT2014SA0234Ra.pdf (Accessed November 25, 2022).

47. DGAL. Rapport du groupe PNNS/PNA sur le sel. (2013). Available at: https://agriculture.gouv.fr/telecharger/44214?token=dfd53c2f7426d27754391f687ad8d9c8638e8fa2614525757491dca86bc1d0b6 (Accessed November 25, 2022).

48. Dötsch-Klerk, M, Goossens, W, and Meijer, GW. Van het Hof KH. Reducing salt in food; setting product-specific criteria aiming at a salt intake of 5 g per day. *Eur J Clin Nutr*. (2015) 69:799–804. doi: 10.1038/ejcn.2015.5

49. Marakis, G, Katsioulis, A, Kontopoulou, L, Ehlers, A, Heimberg, K, Hirsch-Ernst, KI, et al. Knowledge, attitude and behaviour of university students regarding salt and iodine: a multicentre cross-sectional study in six countries in Europe and Asia. *Arch Public Health*. (2021) 79:68. doi: 10.1186/s13690-021-00593-5

50. Newson, RS, Elmadfa, I, Gy, B, Cheng, Y, Prakash, V, Rust, P, et al. Barriers for progress in salt reduction in the general population. An international study. *Appetite*. (2013) 71:22–31. doi: 10.1016/j.appet.2013.07.003

51. Wardle, J, Haase, AM, Steptoe, A, Nillapun, M, Jonwutiwes, K, and Bellisie, F. Gender differences in food choice: the contribution of health beliefs and dieting. *Ann Behav Med*. (2004) 27:107–16. doi: 10.1207/s15324796abm2702_5

52. Brockman, C. Dietary salt: consumption, reduction strategies and consumer awareness In: C Beeren, K Groves and PM Titoria, editors. *Reducing Salt in Foods*. *2nd* ed. Cambridge: Woodhead Publishing (2019). 97–110.

53. Chung, ML, Moser, DK, Lennie, TA, Worrall-Carter, L, Bentley, B, Trupp, R, et al. Gender differences in adherence to the sodium-restricted diet in patients with heart failure. *J Card Fail*. (2006) 12:628–34. doi: 10.1016/j.cardfail.2006.07.007

54. Piovesana, Pde M, Sampaio, K D L, and Gallani, M C B J. Association between taste sensitivity and self-reported and objective measures of salt intake among hypertensive and normotensive individuals. *Int Sch Res Not Nutr*. (2013) 2013: 1–7, doi: 10.5402/2013/301213

55. Saje, SM, Endris, BS, Nagasa, B, Ashebir, G, and Gebreyesus, SH. Dietary sodium and potassium intake: knowledge, attitude and behaviour towards dietary salt intake among adults in Addis Ababa. *Ethiopia Public Health Nutr*. (2021) 24:3451–9. doi: 10.1017/S1368980020003663

56. Silva-Santos, T, Moreira, P, Pinho, O, Padrão, P, Norton, P, and Gonçalves, C. Salt-related knowledge, attitudes and behavior in an intervention to reduce added salt when cooking in a sample of adults in Portugal. *Foods*. (2022) 11:981. doi: 10.3390/foods11070981

57. Brembeck, H, and Fuentes, M. Convenient food for baby: a study of weaning as a social practice. *Food Cult Soc*. (2017) 20:569–86. doi: 10.1080/15528014.2017.1357950

58. Farleigh, CA, Shepherd, R, and Wharf, SG. The effect of manipulation of salt pot hole size on table salt use. *Food Qual Prefer*. (1990) 2:13–20. doi: 10.1016/0950-3293(90)90026-Q

59. Shepherd, R, Farleigh, CA, and Wharf, SG. Limited compensation by table salt for reduced salt within a meal. *Appetite*. (1989) 13:193–200. doi: 10.1016/0195-6663(89)90012-3

60. Dunford, E, Poti, J, and Popkin, B. Emerging disparities in dietary sodium intake from snacking in the US population. *Nutrients*. (2017) 9:610. doi: 10.3390/nu9060610

61. Grimes, CA, Riddell, LJ, and Nowson, CA. Consumer knowledge and attitudes to salt intake and labelled salt information. *Appetite*. (2009) 53:189–94. doi: 10.1016/j.appet.2009.06.007

62. McMahon, E, Webster, J, and Brimblecombe, J. Effect of 25% sodium reduction on sales of a top-selling bread in remote indigenous Australian community stores: a controlled intervention trial. *Nutrients*. (2017) 9:214. doi: 10.3390/nu9030214

63. Payne Riches, S, Aveyard, P, Piernas, C, Rayner, M, and Jebb, SA. Optimising swaps to reduce the salt content of food purchases in a virtual online supermarket: a randomised controlled trial. *Appetite*. (2019) 133:378–86. doi: 10.1016/j.appet.2018.11.028

64. Rhodes, K, Chan, F, Prichard, I, Coveney, J, Ward, P, and Wilson, C. Intergenerational transmission of dietary behaviours: a qualitative study of Anglo-Australian, Chinese-Australian and Italian-Australian three-generation families. *Appetite*. (2016) 103:309–17. doi: 10.1016/j.appet.2016.04.036

65. Ton Nu, C, Mac Leod, P, and Barthelemy, J. Effects of age and gender on adolescents' food habits and preferences. *Food Qual Prefer*. (1996) 7:251–62. doi: 10.1016/S0950-3293(96)00023-7

66. Bowker, GC, and Star, SL. *Sorting Things Out: Classification and Its Consequences*. *8th* ed. Cambridge: MIT Press (1999). 377 p.

67. Esteban. Étude de santé sur l'environnement; la biosurveillance, l'activité physique et la nutrition (ESTEBAN, 2014-2016). Volet Nutrition Chapitre Consommations alimentaires (2018). Available at: https://www.santepubliquefrance.fr/determinants-de-sante/nutrition-et-activite-physique/documents/rapport-synthese/etude-de-sante-sur-l-environnement-la-biosurveillance-l-activite-physique-et-la-nutrition-esteban-2014-2016-chapitre-consommations-alimentair (Accessed November 25, 2022).

68. Huggins, CE, O'Reilly, S, Brinkman, M, Hodge, A, Giles, GG, English, DR, et al. Relationship of urinary sodium and sodium-to-potassium ratio to blood pressure in older adults in Australia. *Med J Aust*. (2011) 195:128–32. doi: 10.5694/j.1326-5377.2011.tb03239.x

69. Drewnowski, A, Henderson, SA, Driscoll, A, and Rolls, BJ. Salt taste perceptions and preferences are unrelated to sodium consumption in healthy older adults. *J Am Diet Assoc*. (1996) 96:471–4. doi: 10.1016/S0002-8223(96)00131-9

70. Beauchamp, GK, and Cowart, BJ. Preference for high salt concentrations among children. *Dev Psychol*. (1990) 26:539–45. doi: 10.1037/0012-1649.26.4.539

71. Kerrihard, AL, Khair, MB, Blumberg, R, Feldman, CH, and Wunderlich, SM. The effects of acclimation to the United States and other demographic factors on responses to salt levels in foods: an examination utilizing face reader technology. *Appetite*. (2017) 116:315–22. doi: 10.1016/j.appet.2017.05.004

72. Cheong, SM, Ambak, R, Othman, F, He, FJ, Salleh, R, Mohd Sallehudin, S, et al. Knowledge, perception, and practice related to sodium intake among Malaysian adults: findings from the Malaysian community salt study (MyCoSS). *J Health Popul Nutr*. (2021) 40:5. doi: 10.1186/s41043-021-00231-4

73. Gray, KL, Petersen, KS, Clifton, PM, and Keogh, JB. Attitudes and beliefs of health risks associated with sodium intake in diabetes. *Appetite*. (2014) 83:97–103. doi: 10.1016/j.appet.2014.08.004

74. Arcand, JAL, Brazel, S, Joliffe, C, Choleva, M, Berkoff, F, Allard, JP, et al. Education by a dietitian in patients with heart failure results in improved adherence with a sodium-restricted diet: a randomized trial. *Am Heart J*. (2005) 150:716.e1–5. doi: 10.1016/j.ahj.2005.02.016

75. Cuadrado-Soto, E, Peral-Suarez, Á, Rodríguez-Rodríguez, E, Aparicio, A, Andrés, P, Ortega, RM, et al. The Association of Parents' behaviors related to salt with 24 H urinary sodium excretion of their children. A Spanish cross-sectional study. *PLoS One*. (2019) 14:e0227035. doi: 10.1371/journal.pone.0227035

76. Hoeft, KS, Guerra, C, Gonzalez-Vargas, MJ, and Barker, JC. Rural Latino caregivers' beliefs and behaviors around their children's salt consumption. *Appetite*. (2015) 87:1–9. doi: 10.1016/j.appet.2014.11.031

77. Kenten, C, Boulay, A, and Rowe, G. Salt. UK consumers' perceptions and consumption patterns. *Appetite*. (2013) 70:104–11. doi: 10.1016/j.appet.2013.06.095

78. Ahn, S, Hyun, KJ, Kim, K, and Kim, HK. Stages of behavioral change for reducing sodium intake in Korean consumers: comparison of characteristics based on social cognitive theory. *Nutrients*. (2017) 9:808. doi: 10.3390/nu9080808

79. Mørk, T, Lähteenmäki, L, and Grunert, KG. Determinants of intention to reduce salt intake and willingness to purchase salt-reduced food products: evidence from a web survey. *Appetite*. (2019) 139:110–8. doi: 10.1016/j.appet.2019.04.018

80. Pillay, A, Trieu, K, Santos, J, Sukhu, A, Schultz, J, Wate, J, et al. Assessment of a salt reduction intervention on adult population salt intake in Fiji. *Nutrients*. (2017) 9:1350. doi: 10.3390/nu9121350

81. Lofthouse, C, Te Morenga, L, and McLean, R. Sodium reduction in New Zealand requires major behaviour change. *Appetite*. (2016) 105:721–30. doi: 10.1016/j.appet.2016.07.006

82. Li, Z, Hu, L, Rong, X, Luo, J, Xu, X, and Zhao, Y. Role of no table salt on hypertension and stroke based on large sample size from National Health and nutrition examination survey database. *BMC Public Health*. (2022) 22:1292. doi: 10.1186/s12889-022-13722-8

83. Adriouch, S, Lelong, H, Kesse-Guyot, E, Baudry, J, Lampuré, A, Galan, P, et al. Compliance with nutritional and lifestyle recommendations in 13,000 patients with a Cardiometabolic disease from the Nutrinet-Santé study. *Nutrients*. (2017) 9:546–64. doi: 10.3390/nu9060546

84. Henson, S, Blandon, J, Cranfield, J, and Herath, D. Understanding the propensity of consumers to comply with dietary guidelines directed at heart health. *Appetite*. (2010) 54:52–61. doi: 10.1016/j.appet.2009.09.006

85. Bournez, M, Ksiazek, E, Charles, MA, Lioret, S, Brindisi, MC, de Lauzon-Guillain, B, et al. Frequency of use of added sugar, salt, and fat in infant foods up to 10 months in the Nationwide ELFE cohort study: associated infant feeding and caregiving practices. *Nutrients*. (2019) 11:733. doi: 10.3390/nu11040733

86. Dyett, PA, Sabaté, J, Haddad, E, Rajaram, S, and Shavlik, D. Vegan lifestyle behaviors. An exploration of congruence with health-related beliefs and assessed health indices. *Appetite*. (2013) 67:119–24. doi: 10.1016/j.appet.2013.03.015

87. Siro, SS, Zandberg, L, Ngounda, J, Wise, A, Symington, EA, Malan, L, et al. Iodine status of pregnant women living in urban Johannesburg, South Africa. *Matern Child Nutr*. (2022) 18:e13236. doi: 10.1111/mcn.13236

88. Frye, CA, and Demolar, GL. Menstrual cycle and sex differences influence salt preference. *Physiol Behav*. (1994) 55:193–7. doi: 10.1016/0031-9384(94)90031-0

89. Leshem, M. Salt preference in adolescence is predicted by common prenatal and infantile Mineralofluid loss. *Physiol Behav*. (1998) 63:699–704. doi: 10.1016/S0031-9384(97)00525-8

90. Leshem, M, Abutbul, A, and Eilon, R. Exercise increases the preference for salt in humans. *Appetite*. (1999) 32:251–60. doi: 10.1006/appe.1999.0228

91. Girgis, S, Neal, B, Prescott, J, Prendergast, J, Dumbrell, S, Turner, C, et al. A one-quarter reduction in the salt content of bread can be made without detection. *Eur J Clin Nutr*. (2003) 57:616–20. doi: 10.1038/sj.ejcn.1601583

92. Bobowski, N, Rendahl, A, and Vickers, Z. A longitudinal comparison of two salt reduction strategies: acceptability of a low sodium food depends on the consumer. *Food Qual Prefer*. (2015) 40:270–8. doi: 10.1016/j.foodqual.2014.07.019

93. Shepherd, R, Farleigh, CA, and Wharf, SG. Preferences for salt in different foods and their relationship to availability of sodium. *Hum Nutr Food Sci Nutr*. (1987) 41:173–81. doi: 10.1080/09528954.1987.11904113

94. Ferber, C, and Cabanac, M. Influence of noise on gustatory affective ratings and preference for sweet or salt. *Appetite*. (1987) 8:229–35. doi: 10.1016/0195-6663(87)90022-5

95. Odom, E, Whittick, C, Tong, X, John, K, and Cogswell, M. Changes in consumer attitudes toward broad-based and environment-specific sodium policies—summer styles 2012 and 2015. *Nutrients*. (2017) 9:836. doi: 10.3390/nu9080836

96. Leshem, M. Salt appetite is not increased in summer heat. *Appetite*. (2017) 108:28–31. doi: 10.1016/j.appet.2016.09.017

97. Moschonis, G, and Karatzi, K. Novel dietary approaches for controlling high blood pressure. *Nutrients*. (2020) 12:3902. doi: 10.3390/nu12123902

98. Campbell, N, and Train, E. A systematic review of fatalities related to acute ingestion of salt. A need for warning labels? *Nutrients*. (2017) 9:648. doi: 10.3390/nu9070648

99. Tsirimiagkou, C, Basdeki, ED, Argyris, A, Manios, Y, Yannakoulia, M, Protogerou, AD, et al. Current data on dietary sodium, arterial structure and function in humans: a systematic review. *Nutrients*. (2019) 12:5. doi: 10.3390/nu12010005

100. Asaria, P, Chisholm, D, Mathers, C, Ezzati, M, and Beaglehole, R. Chronic disease prevention: health effects and financial costs of strategies to reduce salt intake and control tobacco use. *Lancet*. (2007) 370:2044–53. doi: 10.1016/S0140-6736(07)61698-5

101. Jaenke, R, Barzi, F, McMahon, E, Webster, J, and Brimblecombe, J. Consumer acceptance of reformulated food products: a systematic review and meta-analysis of salt-reduced foods. *Crit Rev Food Sci Nutr*. (2017) 57:3357–72. doi: 10.1080/10408398.2015.1118009

102. Darmon, N, and Drewnowski, A. Does social class predict diet quality? *Am J Clin Nutr*. (2008) 87:1107–17. doi: 10.1093/ajcn/87.5.1107

103. Laisney, C. Disparités sociales et alimentation. (2013). Available at: https://agriculture.gouv.fr/disparites-sociales-et-alimentation-document-de-travail-ndeg9 (Accessed November 25, 2022).

104. Star, SL. The ethnography of infrastructure. *Am Behav Sci*. (1999) 43:377–91. doi: 10.1177/00027649921955326

105. Dagiral, É, and Peerbaye, A. Making knowledge in boundary infrastructures: inside and beyond a database for rare diseases. *Sci Technol Stud*. (2016) 29:44–61. doi: 10.23987/sts.55920

106. Antúnez, L, Vidal, L, Giménez, A, Curutchet, MR, and Ares, G. Age, time orientation and risk perception are major determinants of discretionary salt usage. *Appetite*. (2022) 171:105924. doi: 10.1016/j.appet.2022.105924

107. Guàrdia, MD, Guerrero, L, Gelabert, J, Gou, P, and Arnau, J. Consumer attitude towards sodium reduction in meat products and acceptability of fermented sausages with reduced sodium content. *Meat Sci*. (2006) 73:484–90. doi: 10.1016/j.meatsci.2006.01.009

108. Lampuré, A, Deglaire, A, Schlich, P, Castetbon, K, Péneau, S, Hercberg, S, et al. Liking for fat is associated with sociodemographic, psychological, lifestyle and health characteristics. *Br J Nutr*. (2014) 112:1353–63. doi: 10.1017/S0007114514002050

109. Mennella, JA, Finkbeiner, S, Lipchock, SV, Hwang, LD, and Reed, DR. Preferences for salty and sweet tastes are elevated and related to each other during childhood. Meyerhof W, éditeur. *PLoS One*. (2014) 9:e92201. doi: 10.1371/journal.pone.0092201

110. Purdy, J, and Armstrong, G. Dietary salt and the consumer: reported consumption and awareness of associated health risks In: D Kilcast and F Angus, editors. *Reducing Salt in Foods: Practical Strategies*. Boca Raton, FL: CRC Press (2007). 99–123.

111. Huang, Z, and Zeng, D. Factors affecting salt reduction measure adoption among Chinese residents. *IJERPH*. (2021) 18:445. doi: 10.3390/ijerph18020445

frontiers | Frontiers in Nutrition

TYPE Original Research
PUBLISHED 28 August 2023
DOI 10.3389/fnut.2023.1170084

Check for updates

OPEN ACCESS

EDITED BY
Massimo Lucarini,
Council for Agricultural Research and Economics, Italy

REVIEWED BY
Evangelia Katsouri,
Hellenic Food Authority, Greece
Saurav Mallik,
Harvard University, United States

*CORRESPONDENCE
Yuzhong Zheng
✉ zhengyuzhong@gmail.com

†These authors have contributed equally to this work

RECEIVED 20 February 2023
ACCEPTED 16 August 2023
PUBLISHED 28 August 2023

CITATION
Liu Y, Zhang Z, Lin W, Liang H, Lin M, Wang J, Chen L, Yang P, Liu M and Zheng Y (2023) A novel FCTF evaluation and prediction model for food efficacy based on association rule mining.
Front. Nutr. 10:1170084.
doi: 10.3389/fnut.2023.1170084

A novel FCTF evaluation and prediction model for food efficacy based on association rule mining

Yaqun Liu[1†], Zhenxia Zhang[1†], Wanling Lin[1], Hongxuan Liang[1], Min Lin[1,2], Junli Wang[2], Lianghui Chen[2], Peikui Yang[1], Mouquan Liu[1] and Yuzhong Zheng[1,2*]

[1]School of Food Engineering and Biotechnology, Hanshan Normal University, Chaozhou, Guangdong, China, [2]School of Laboratory Medicine, Youjiang Medical University for Nationalities, Baise, Guangxi, China

Introduction: Food-components-target-function (FCTF) is an evaluation and prediction model based on association rule mining (ARM) and network interaction analysis, which is an innovative exploration of interdisciplinary integration in the food field.

Methods: Using the components as the basis, the targets and functions are comprehensively explored in various databases and platforms under the guidance of the ARM concept. The focused active components, key targets and preferred efficacy are then analyzed by different interaction calculations. The FCTF model is particularly suitable for preliminary studies of medicinal plants in remote and poor areas.

Results: The FCTF model of the local medicinal food Laoxianghuang focuses on the efficacy of digestive system cancers and neurological diseases, with key targets ACE, PTGS2, CYP2C19 and corresponding active components citronellal, trans-nerolidol, linalool, geraniol, α-terpineol, cadinene and α-pinene.

Discussion: Centuries of traditional experience point to the efficacy of Laoxianghuang in alleviating digestive disorders, and our established FCTF model of Laoxianghuang not only demonstrates this but also extends to its possible adjunctive efficacy in neurological diseases, which deserves later exploration. The FCTF model is based on the main line of components to target and efficacy and optimizes the research level from different dimensions and aspects of interaction analysis, hoping to make some contribution to the future development of the food discipline.

KEYWORDS
association rule mining, medicinal food, components, target, function

1. Introduction

During the battle against coronavirus disease (COVID-19), the role of the medicinal food concept in the prevention and control of pandemics has attracted widespread attention, involving mostly local foods with medicinal value (1). Slogans such as “Food as Medicine, Medicine as Food” have driven the development of functional foods into a trendy form (2). Unfortunately, medicinal foods around the world possess national characteristics, and efficacy studies mostly rely on the inheritance of traditional experiences, which are mainly prevalent in the local area (3). Influenced by factors such as national character, traditional habits and environmental isolation, the medicinal effects of local specialties remain relatively independent and lag behind in development, thus presenting a blind or random process.

Most modern systematic and comprehensive food efficacy studies are based on genomics (4), proteomics (5), metabolomics (6), lipidomics (7), glycomics (8), and other methods and techniques, which are time-consuming, cost substantially, require expensive equipment, and less friendly to traditional specialty foods from poverty-stricken areas. Association rule mining (ARM) is a rule-based machine learning algorithm that can discover hidden patterns and interesting relationships in large databases. Recently, ARM has become a promising technique in multiple fields including biomedical, educational, and social sciences, such as predicting COVID-19 cases and symptom patterns (9, 10), the application of multiresource for MOOC teaching (11), the study of improving English achievement analysis (12) and the investigation of the relationships between shifts in digital skills and cybersecurity awareness (13). Common analysis algorithms and evaluation approaches for ARM include the Apriori algorithm (14), entropy weight method (EWM) (15), technique for order preference by similarity to ideal solution (TOPSIS) (16), support vector machine (SVM) and random forest (RF) (17). This study innovatively developed the ARM and its algorithms to the poorly understood field of medicinal foods and can be applied to other foods as well. By mining the identified food ingredients for their targets and related functions on a big data platform, the active components, key targets and preferred functions are inferred through multidimensional interactions and cross commonalities. To the best of our knowledge, this may be the first application of this model in the field of food efficacy research, which we define as food-components-target-function (FCTF) association rule mining. Data association analysis provides exciting research opportunities and contemporary themes for known food components, which can not only validate traditional empirical medicinal efficacy but also build initial theoretical platforms for future in-depth research.

The FCTF model was carried out on the example of Laoxianghuang, a characteristic medicinal food from Chaozhou, Guangdong Province, China. Laoxianghuang is obtained by fermenting *Citrus medica* L. var. *Sarcodactylis Swingle* for more than several years through a complex process of salting, desalting, sugaring, cooking and drying. Compared to the bitter and spicy raw material *Citrus medica* L. var. *Sarcodactylis Swingle*, the fermented Laoxianghuang not only enhances its edibility but also expands its efficacy as revealed in empirical pharmacology (18–20). It is a local cultural symbol because of its aromatic taste and excellent efficacy. However, due to the remoteness of the region and the limitations of scientific conditions, research on Laoxianghuang is still in the initial stage. Previously, although we explored the components of Laoxianghuang through different methods (18–20), there were obstacles to a more in-depth study under poor and weak scientific research conditions. Therefore, we established an FCTF model to perform a deeper exploration of Laoxianghuang by searching the correlations between components, targets and efficacy. Meanwhile, proposing the FCTF model is expected to help enhance the research connotation and denotation of featured products in the future.

2. Materials and methods

2.1. Food components library construction

We have previously detected the components of Laoxianghuang and will not repeat it here (18–20). The components of Laoxianghuang were also searched in various literature databases as much as possible, and a component library was obtained by removing the overlap (Supplementary Table S1). The relative contents of the components detected by different methods varied, and we subjectively selected the components with relative contents greater than 1% under each detection method. If a component is represented under different methods, as long as its relative content is higher than 1%, it will be taken into account.

2.2. Food-components-target framework construction

The components were used as entry points to query their simplified molecular-input line-entry system (SMILES) numbers on PubChem[1] (21). The SMILES numbers were imported into SwissTargetPrediction[2] to retrieve the targets of each component (22), and targets with a probability greater than 0 were selected as the study objects (23) (Supplementary Table S2).

2.3. Food-components-target-function model building and analysis

Rough functional enrichment: Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed by Metascape[3] after de-duplication of all potential targets (24). The top 20 items with $p < 0.01$ were selected for advanced bubble mapping. GO analysis included molecular function (MF), cellular component (CC) and biological process (BP).

Function Refined Searching: Potential target-related functions and preferred efficacy-related targets were obtained by searching the Comparative Toxicogenomics Database (CTD, http://ctdbase.org/) (25), Online Mendelian Inheritance in Man Database (OMIM, http://www.omim.org) (26), Therapeutic Target Database (TTD, http://db.idrblab.net/ttd/) (27) and GeneCards Database[4] (28). Accessed on all of the above servers on December 30th, 2022 (Supplementary Tables S3, S4).

2.4. Food-components-target-function model verification

The protein crystal structures of the key targets and the corresponding 3D structures of the active components were retrieved from the Protein Data Bank (PDB, http://www.rcsb.org/pdb/) (29) and PubChem (See footnote 1) (21), respectively (PDB IDs are shown in Supplementary Table S5). Application of Cavity-Detection Guided Blind Docking (CB-Dock; http://clab.labshare.cn/cb-dock/php/; accessed on January 9th, 2023) (30) and Mcule 1-Click Docking (https://mcule.com/apps/1-click-docking/; accessed on January 9th, 2023) (31) for evaluation and comparison of molecular docking. The affinity of the docked key target and active components is expressed as binding energy (kcal/mol). Among the four different binding fractions given, the results with more negative values were considered (30).

[1] https://pubchem.ncbi.nlm.nih.gov

[2] http://www.swisstargetprediction.ch/

[3] https://metascape.org/gp/index.html

[4] http://www.genecards.org/

2.5. Data analysis

Descriptive analysis: Microsoft Office Excel 2019 (Microsoft Corporation) was used to perform statistics on the frequency of components, targets and functions, including duplicate and position distribution.

Association rule mining analysis: To obtain high frequency items, analysis was performed using Cytoscape 3.9.0[5] (32) and a node-weighting scheme using Degree Centrality (33). The association rules are expressed in the form of components → target, target → efficacy, efficacy → target, etc. The front is the basis, and the back is the mining object. The support for association is measured by the Degree value, i.e., the frequency of nodes crossing each other, which can reflect their importance and dependency. This makes the association rule valuable only when the Degree value is greater than 1. Interactions between targets were analyzed using the Search Tool for the Retrieval of Interacting Genes (STRING, https://cn.string-db.org/) with confidence score >0.9 (34). Overlapping targets in different projects were visualized by a Venn diagram (35). The Apriori algorithm further mines the set of frequent items of association rules, with components and functions as index items and targets as analysis items, and assesses its reliability by support, confidence, lift, leverage, and conviction (36):

Support: the probability that the target set Tx, Ty occurs in all items, or the probability that the two target sets Tx → Ty occur in all items. The formula is as follows:

$$\mathit{Support}\left(\mathrm{Tx}\right)=\mathrm{P}\left(\mathrm{Tx}\right)/\,\mathrm{N}$$

$$\mathit{Support}\left(\mathrm{Ty}\right)=\mathrm{P}\left(\mathrm{Ty}\right)/\,\mathrm{N}$$

$$\mathit{Support}\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)=\mathrm{P}\left(\mathrm{Tx}\cup\mathrm{Ty}\right)/\,\mathrm{N}$$

P: number of target occurrences; N: total number of items.

Confidence: the frequency of Ty in the set of items containing Tx. The higher the confidence level, the more relationship between Tx and Ty is considered. The formula is as follows:

$$\mathit{Confidence}\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)=\mathrm{P}\left(\mathrm{Tx}\cup\mathrm{Ty}\right)/\,\mathrm{P}\left(\mathrm{Tx}\right)$$

Lift: measures how much more often the Tx and Ty occur together rather than them occurring independently. Lift >1: Tx and Ty associated positively; Lift <1: Tx and Ty associated negatively; Lift = 1: Tx and Ty are independent of each other. The formula is as follows:

$$\mathit{Lift}\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)=\mathit{Confidence}\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)/\,\mathit{Support}\left(\mathrm{Ty}\right)$$

Leverage: the proportion of additional examples covered by both the Tx and Ty above those expected if the premise and consequence were independent of each other. Tx and Ty are independent when the leverage is 0; the greater the leverage, the closer A and B are. The formula is as follows:

$$\mathit{Leverage}\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)=\mathit{Support}\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)-\mathit{Support}\left(\mathrm{Tx}\right)\times\mathit{Support}\left(\mathrm{Ty}\right)$$

Conviction: another measure of departure from independence. The greater the conviction, the closer A and B are. The formula is as follows:

$$\mathit{Conviction}\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)=\left[1-\mathrm{Support}\left(\mathrm{Ty}\right)\right]/\left[\begin{matrix}1-\mathit{Confidence}\\\left(\mathrm{Tx}\rightarrow\mathrm{Ty}\right)\end{matrix}\right]$$

The screening conditions for the association rules were set as follows: support >20%, confidence >35%, lift >1, leverage >0, conviction >0. Information entropy and weight (%) analysis of the above evaluation parameters based on the EWM was performed, and the performance of each target was ranked according to the distance from positive ideal solution, distance from negative ideal solution and the composite score index by the TOPSIS algorithm (37). SVM and RF machine learning algorithms were constructed using the DALEX R package for bidirectional targets, and the diagnostic performance of both models was evaluated by receiver operating characteristic (ROC) curves and associated area under the curve (AUC) (38).

3. Results

3.1. Component library construction and analysis

We systematically studied and summarized the results of different methods to determine the components of Laoxianghuang (Supplementary Table S1). The results showed that there are 156 components in Laoxianghuang, including 32 terpenes, with relatively high contents of limonene, γ-terpinene, trans-β-ocimene, and p-cymene. Twenty-nine alcohols, including linalool, trans-nerolidol, α-terpineo, and terpinen-4-ol were relatively high. Twenty aldehydes reflecting high levels are citronellal and 2-furaldehyde. In addition there are 18 esters, 9 acids, 12 ketones, 16 amino acids, and 20 others in the Laoxianghuang component categories (Figure 1).

3.2. Prediction and analysis of potential targets

First, 35 components with higher relative content were screened out from numerous components of Laoxianghuang according to the screening rules. Then, their SMILES numbers were determined by PubChem, and the target genes corresponding to individual components were retrieved in the SwissTargetPrediction platform according to the SMILES numbers (Supplementary Table S2). A total of 454 predicted targets closely related to components were retrieved, among which trans-nerolidol showed the highest number of related targets, including SQLE, BACE1, PER2 and 82 others. This was followed by citronellal with 52 relevant targets, such as FAAH, CYP19A1, and TRPV1, etc. α-Terpineol, terpinen-4-ol, linalool, geraniol and ethyl valerate were next with 51, 37, 36, 33, and 29 relevant targets, respectively (Figure 2A). The results of focusing on related targets

[5] https://cytoscape.org/

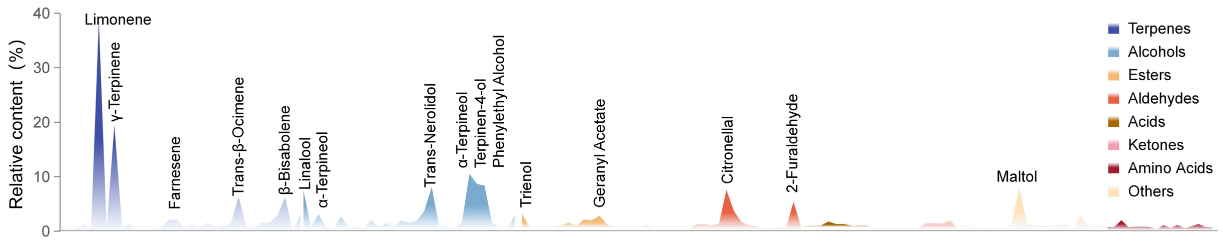


FIGURE 1
Profile chart of the relative content of Laoxianghuang. Different categories of components are shown in different colors. The actual relative content values of the components in the different methods are plotted, and the values of the same components detected in the different methods are averaged. The specific names of some components with significantly higher content are indicated in the figure.

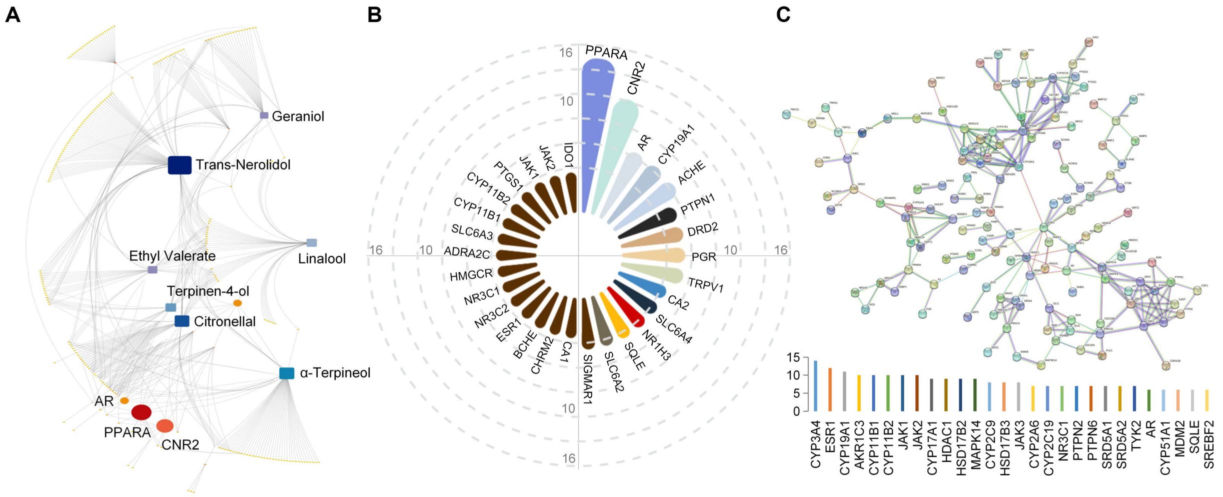


FIGURE 2
Network construction for Laoxianghuang components and targets. **(A)** Component-target linkage network diagram of Laoxianghuang. Focused elements are highlighted. **(B)** The number of potential target intersections. **(C)** Network diagram of target interactions derived from the STRING input library. The nodes and edges in the network indicate the targets and target–target associations, respectively. The following histogram shows the corresponding Degree values.

showed that the intersection target with the most components was PPARA, with 14 components associated with it, such as trans-β-ocimene, terpinolene, and cadinene. Subsequently, CNR2 and AR followed in order, with 11 and 7 intersecting components, respectively (Figure 2B). After removing duplicates from the above 454 targets, 236 targets were finally obtained. A high confidence correlation analysis was performed on these duplicate-free targets to identify the target–target interactions. A total of 230 nodes and 244 edges were found in the network excluding unconnected nodes. It indicated 230 interacting targets, with the more interacting target being CYP3A4 (Degree = 14). ESR1 and CYP19A1 followed closely, showing higher target–target interactions with Degree values of 12 and 11, respectively (Figure 2C).

3.3. Functional enrichment of potential targets

GO and KEGG enrichment analyses were performed on the above predicted potential targets of Laoxianghuang to roughly describe their possible biological functions and signaling pathways. Biological processes (BP) in GO enrichment analysis are mainly involved in regulation of secretion, circulatory system process, response to hormone, etc., and molecular functions (MF) are associated with oxidoreductase activity, inorganic cation transmembrane transporter activity and hydrolase activity, acting on ester bond, etc. Cellular components (CC) are related to the synaptic membrane, membrane raft, receptor complex, etc. The KEGG enrichment results showed that Laoxianghuang was mainly associated with cancer, inflammation, immunity and nervous system, including neuroactive ligand–receptor interaction, glutamatergic synapse, pathways in cancer and steroid hormone biosynthesis (Figure 3).

3.4. Refined mining and analysis of functions

A deep mining of the efficacy corresponding to each target was conducted, identifying 1,976 relevant diseases with potential associations (Supplementary Table S3). One target can carry different disease profiles; for example, AR has been associated with both cancer

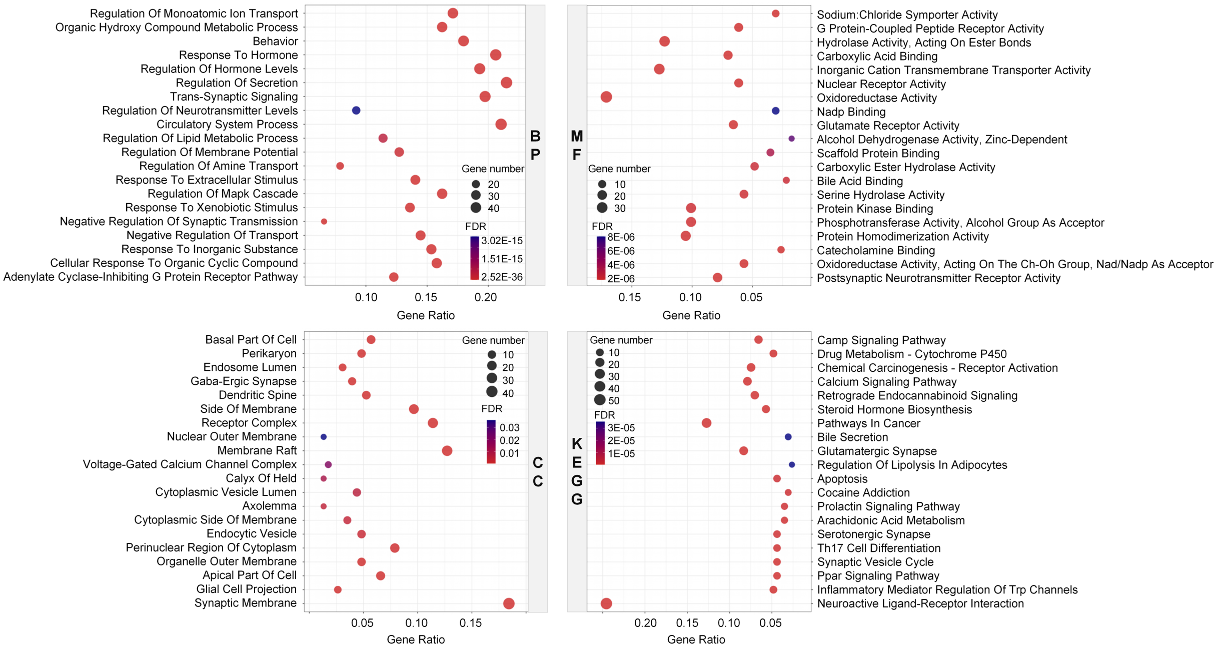


FIGURE 3
Enrichment of potential targets of Laoxianghuang using GO (BP, MF, and CC) and KEGG analyses. The larger the circles in the figure, the more genes are included. Higher FDR values are indicated with a stronger blue color.

and digestive system disorders. PTGS2 is associated with the largest number of diseases, including adenocarcinoma, diabetes mellitus, fever and 108 others (Figure 4A). Among disease categories, cancer presented the strongest association with all targets, appearing 470 times, followed by nervous system diseases and digestive system diseases with 311 and 264, respectively. One disease also matches different targets; for example, nervous system disease can involve CYP19A1, TRPM8, CHRM2, etc.

In this work, the first five associated diseases were selected for further analysis, and each disease was deduplicated to reveal 123, 61, 140, 64, and 79 diseases under the categories of cancer, digestive system diseases, nervous system diseases, mental disorders and pathological processes, respectively. Of these, cancer overlaps with digestive system diseases to a high level, with 15 items belonging to both categories, including stomach neoplasms, colonic neoplasms and intestinal neoplasms. Nervous system diseases and mental disorders also share a high Degree of disease, with 19 diseases such as Alzheimer's disease, learning disabilities and Tic disorders (Figure 4B). Among all specific diseases, prostatic neoplasms correlated the most with the target (Degree = 34), followed by breast neoplasms, liver cirrhosis, hepatocellular carcinoma, liver injury, schizophrenia, etc. (Figure 4C).

3.5. Analysis of association rules for the FCTF model

The Apriori correlation algorithm was used to analyze the crucial link targets using components and functions as index items. Thresholds were set according to support, confidence, lift, leverage and conviction, and a total of 279 sets of target sets were filtered, with a support interval of 20–45%, confidence interval of 35–90%, lift interval of 1.0–4.3, leverage interval of 0.01–0.19, and conviction interval of 1.0–6.6. (CNR2 → PPARA), (AR → (CYP19A1), and (ACHE → CYP19A1) performed better in items of support, all at 43%. (DRD2 → AR), (ESR1 → BCHE), and (PGR → DRD2) showed better confidence levels, all at 90%. (PARP1 → PPARG), (PARP1 → ACE), and (CYP17A1 → ESR2) showed higher lift, all at 4.23. (ESR1 → BCHE), (AR → ESR1) and (ACHE→ESR1) possessed better leverage, all at 0.18. (ESR1 → BCHE), (HMOX1 → TYK2), and (SLC6A4 → CHRM2) displayed better conviction, with (ESR1 → BCHE) at 6.6 and the other two at 6.13 (Figure 5A). All the targets appearing in the antecedent item were duplicated in the consequent item, and the consequent item showed more targets such as (CYP2C19), (ESR2), and (PPARG) compared to the antecedent item (Figure 5B). The results of the EWM showed that the maximum value of indicator weight was support (64.637%), followed by conviction (17.4515%), and the lowest was leverage (2.931%). The best performing information entropy value was leverage (0.991), and the lowest was support (0.797; Figure 5C). The prioritization rankings of the target set obtained using the TOPSIS models demonstrate that (CNR2 → PPARA), (PPARA → CNR2), and (AR → CYP19A1) had the highest priority ranks with scores of 0.708, 0.664, and 0.556, respectively (Figure 5D). The targets that excelled in both antecedent and consequent items were AR, CYP19A1, ACE, etc. (Figure 5E).

3.6. Evaluation of the FCTF model

The target is a key link in the design of the FCTF model, so it is necessary to first screen for bidirectional targets (which are associated with more than two components as well as more than two functions) in a large dataset. Finally, 80 bidirectional target sets such as PTGS2 (Degree$_{\text{CT}}$ = 3, Degree$_{\text{TF}}$ = 108), HMOX1 (Degree$_{\text{CT}}$ = 3,

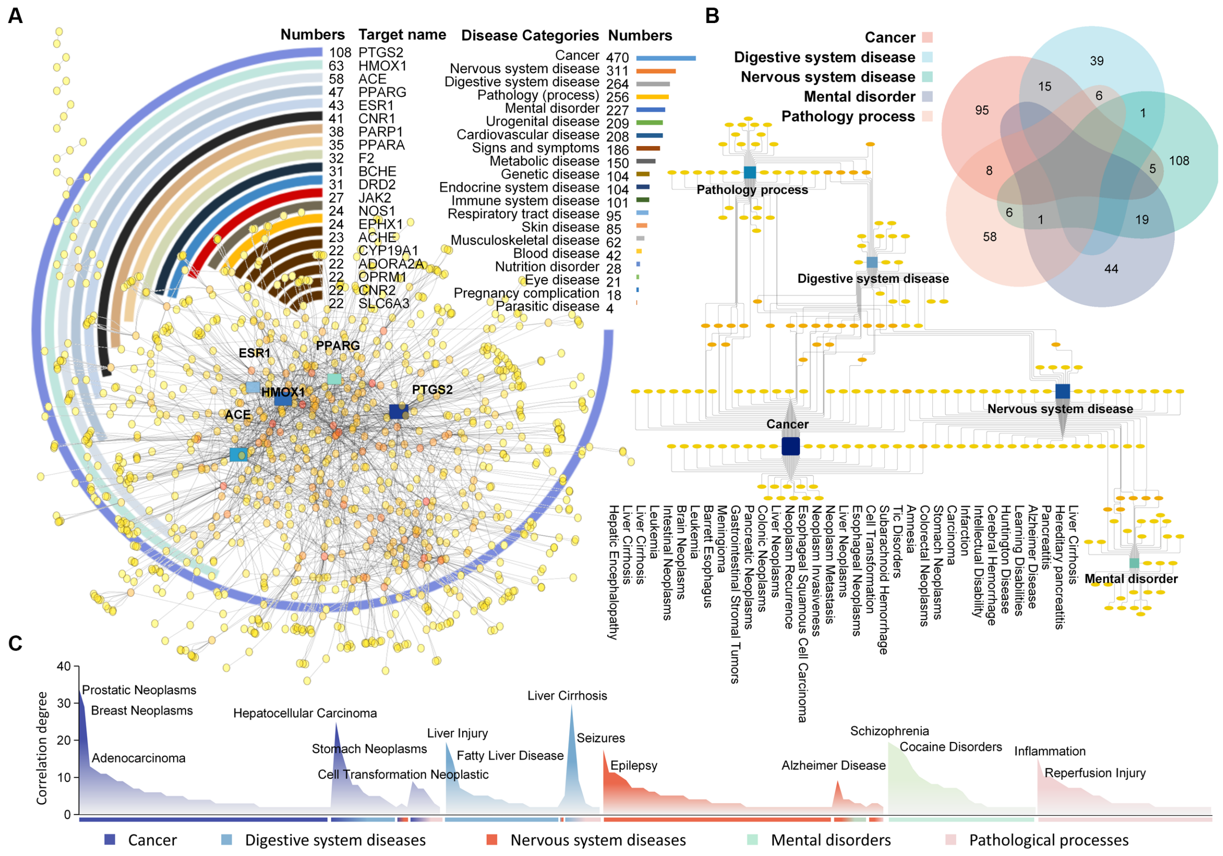


FIGURE 4
Analysis diagram of Laoxianghuang-related function mining. **(A)** The target-function diagram of Laoxianghuang. The circle diagram shows the Degree of target action. The bar chart shows the functional category Degree of relevance. The network diagram shows the target-function interaction connections, and targets with strong interactions are highlighted. **(B)** Relationships among the five main functions of Laoxianghuang. The multidimensional Venn diagram shows the Degree of overlap among the five functions of Laoxianghuang. The network diagram specifically shows the association between diseases, some of which are listed in the diagram. **(C)** Importance Degree value of specific functions. Some highly correlated disease names are marked in the figure, with different colors representing different disease categories and mixed colors representing cross-linked disease categories.

Degree $_{TF}$ = 63), ESR1 (Degree $_{CT}$ = 4, Degree $_{TF}$ = 43) were screened (Degree $_{CT}$: Degree of component and target; Degree $_{TF}$: Degree of target and function). Classification models were constructed for this target set, and the results showed that the RF model (AUC = 1.000) achieved higher separation accuracy compared to the SVM model (AUC = 0.856). The importance ranking of the targets in the RF model was filtered according to the Gini coefficient, and PRARA was particularly important, followed by PTGS2, CNR2, ACE, CYP2C19, etc. (Figure 6).

3.7. Verification and analysis of the FCTF model

We previously focused on relevant functions by targets to digestive system cancers and neurological diseases (Supplementary Table S4). To further verify this speculation, digestive system cancer and neurological disease genes were retrieved and intersected with the Laoxianghuang target. The results showed that the focus genes for target and digestive system cancers were ACE, PTGS2, CYP2C19, and CYP2A6, while the number of intersections with neurological diseases was 73, and the three shared a focus on ACE, PTGS2 and CYP2C19 (Figure 7, top left). The Laoxianghuang inverse association component search of these three genes revealed that the main relationships were citronellal-ACE, trans-nerolidol-PTGS2, linalool-PTGS2, geraniol-PTGS2, α-terpineol-CYP2C19, cadinene-CYP2C19 and α-pinene-CYP2C19. Notably, these targets were also ranked high in the evaluation of the FCTF model. Molecular docking was performed for the above target component relationships, a 3D diagram between each target and component (detailed information and coordinate locations are shown in Supplementary Table S5) and the specific binding sites are shown in the simulation model (Figure 7). It is now generally accepted that the Vina score is considered to represent the binding activity between the protein and the ligand, with lower compound-target binding free energy indicating more stable binding between the two, and binding energy <−5.0 kcal/mol indicating better binding of the compound to the target site. In addition, the accuracy of docking is improved if the size of the cavity is close to or larger than that of the compound (30). The results of the docking of the synthetically screened compounds and targets in this study showed that their Vina scores were less than−5.0 kcal/mol, and the cavity sizes also displayed strong interactions between the target and components (Figure 7, bottom right).

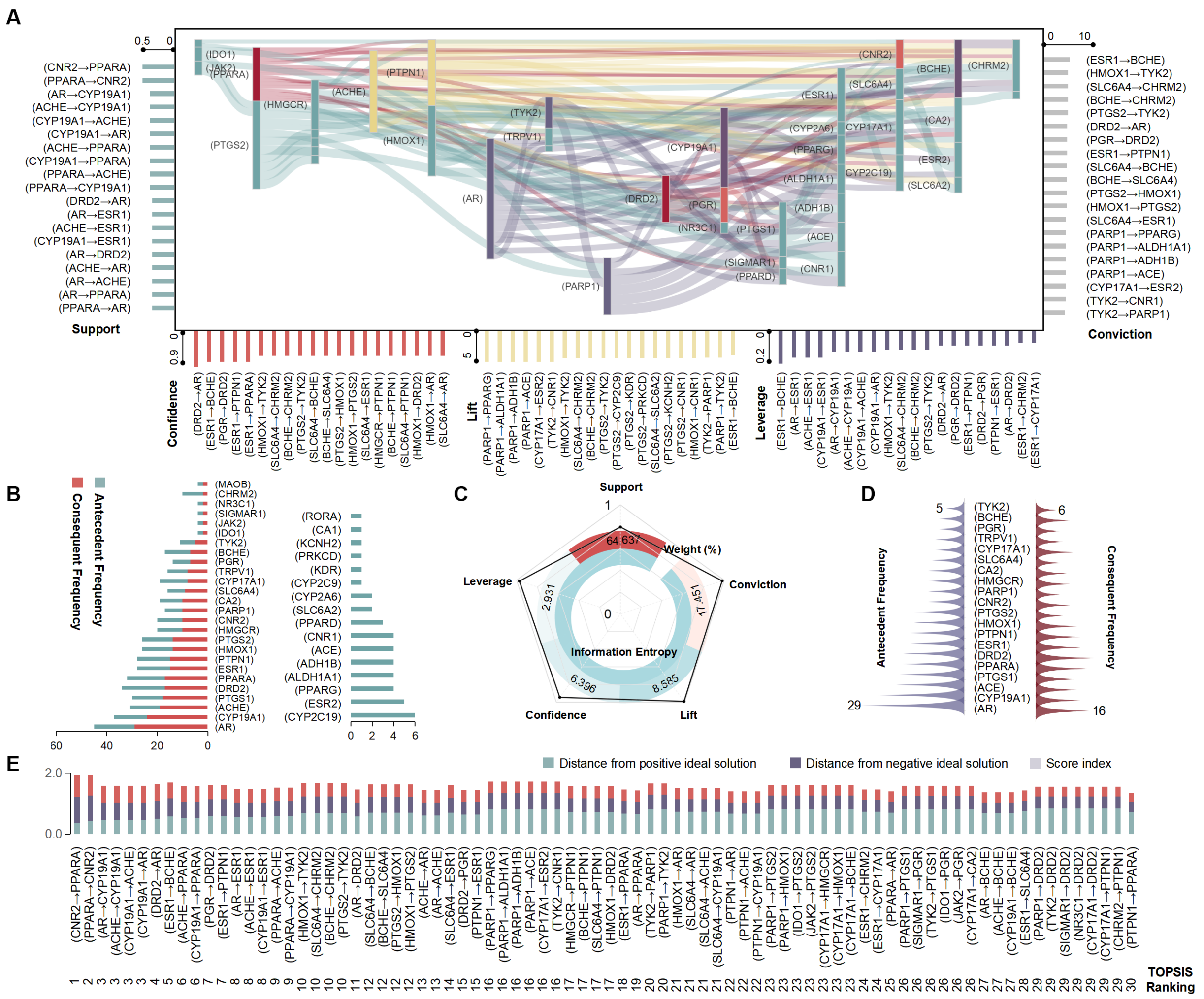


FIGURE 5
Analysis of association rules for the Laoxianghuang FCTF model. **(A)** Association parameters: left: support; right: conviction, bottom (from left to right): confidence, lift and leverage, all presented in the top 20. Middle: Sankey diagram of association rules. **(B)** Target frequency of antecedent and consequent items by Apriori algorithm. **(C)** Analysis of association rule parameters by EWM. **(D)** Target frequency of antecedent and consequent items by TOPSIS. **(E)** Ranking of target sets by TOPSIS, presented in the top 30.

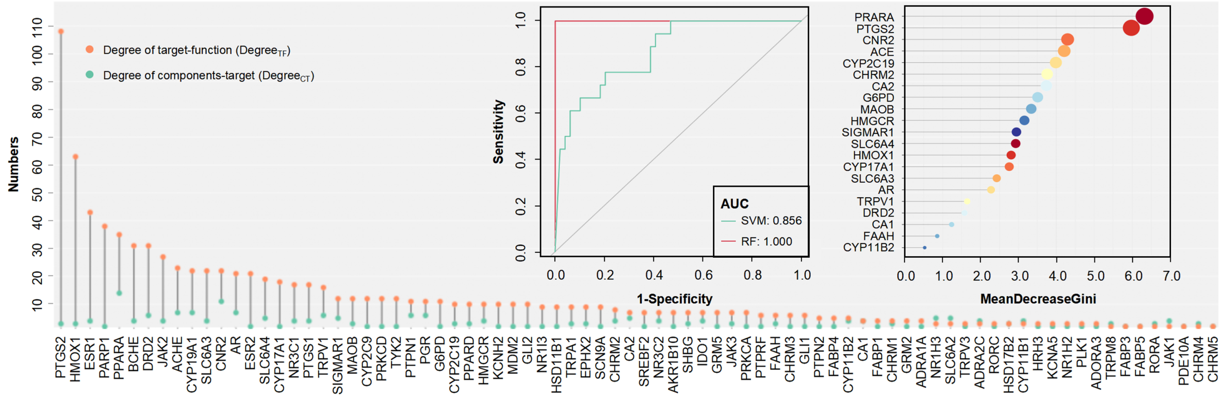


FIGURE 6
Numbers, RF and SVM classifier of bidirectional targets. Top left: AUC of the two models on the bidirectional targets. Top right: The order of importance of bidirectional targets.

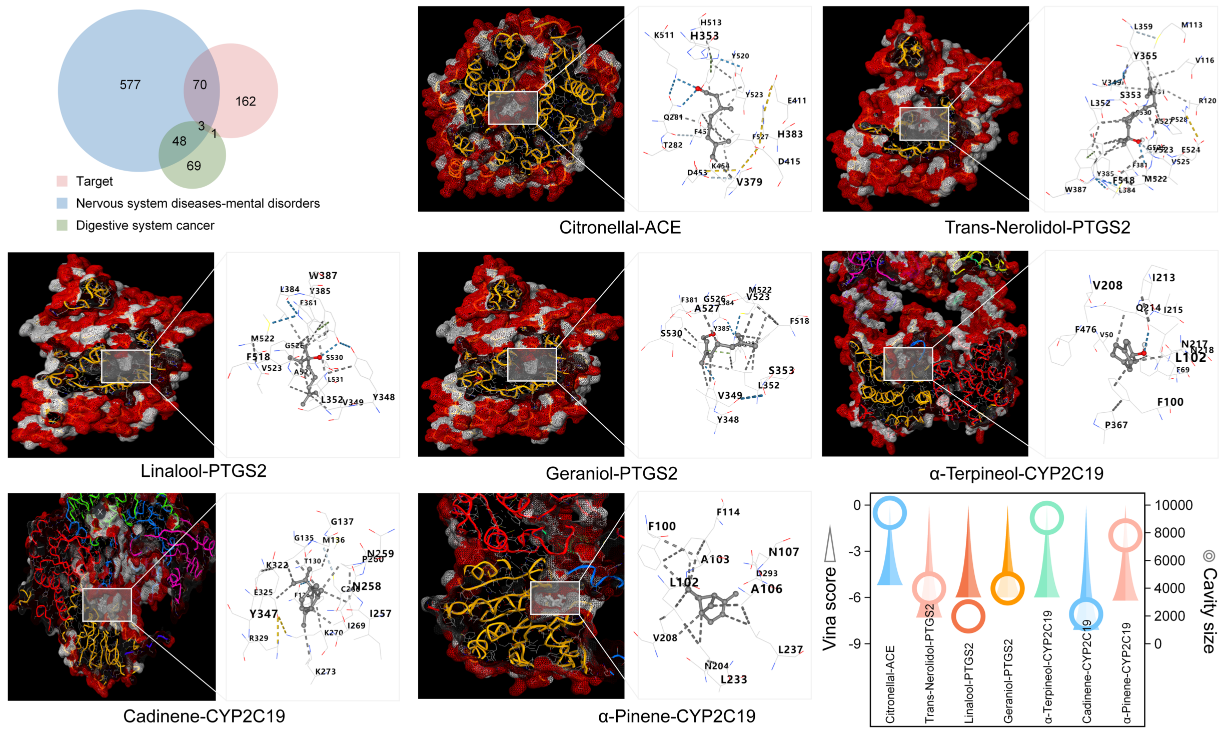


FIGURE 7
Key target screening and target-component binding model construction. Top left: Venn diagram of targets and associated disease genes. Bottom right: Vina scores and cavity information of the docking simulation pose.

4. Discussion

Association rule mining was introduced as a powerful approach to explore interesting but tangential relationships among components, targets, and effectives of medicinal foods. An FCTF evaluation and prediction model was developed to visualize the association rules of the three and capture the active components, key target and preferred efficacy. Finally, the degree of match between the preferred efficacy-related target and the active components was demonstrated by molecular docking. FCTF analysis identifies anchor targets that link components to efficacy, providing new opportunities for purposeful validation of traditional empirical medicinal efficacy and initiation of future research programs. The model can be applied not only to medicinal foods but also to preliminary studies of other foods.

Laoxianghuang was chosen because this medicinal food has been handed down in the region for hundreds of years and is one of the most ethnically distinctive medicinal foods. Locals believe that it possesses excellent functions, such as soothing the liver, regulating gas, relieving pain in the stomach, eliminating dampness and resolving phlegm (39), and is respected as the first of the "Three Treasures of Chaozhou," which is a cultural symbol. However, due to geographical factors, economic underdevelopment and human culture, this medicinal food is only prevalent in the local area and has rarely been studied in depth.

We have assayed and validated the components by different detection methods and constructed a components database with as many components as possible to avoid losing the retrieval of targets and efficacy later. In ARM analysis, the primary condition is to obtain "A" information as a way to mine "B," "C" or more information and explore the interaction relationship between them. The FCTF evaluation and prediction model first obtains the components of food products and then retrieves the corresponding targets of each component and the corresponding efficacy of the targets. The interactions and commonalities between components and targets, targets and targets, and targets and efficacy were also analyzed. In addition to the usual data analysis in Microsoft Excel, Cytoscape network visualization intersection analysis was applied here, and the Degree algorithm was used to filter out key targets and preferred efficacy. It is a common software dedicated to the visualization of interaction network data and is proficient in identifying central objects and subnetworks from complex blind interaction sets, mostly used in bioinformatics (40) and network pharmacology (41). With the help of this analysis software, we can easily focus on the preferred effects of Laoxianghuang on digestive system cancers and neurological diseases, which coincides with the traditional proposal of Laoxianghuang as an adjuvant treatment for digestive diseases (39), while also pointing to its possible adjunctive therapeutic potential for neurological diseases. The raw material for the production of Laoxianghuang is *Citrus medica* L. var. *Sarcodactylis Swingle* of the family *Rutaceae*, whose neolignan derivatives possess hepatoprotective and neuroprotective activities (42). Therefore, future studies on the efficacy of Laoxianghuang could cover neurological diseases in addition to digestive system diseases. Finally, virtual molecular docking, which demonstrates the mode of action of a component to a target is most commonly used in pharmacology (43), drug design (44) and traditional Chinese medicine (45) and is also an important part of computational chemistry and biology, computer science, structural biology, and molecular biology (46). The interaction processes between targets and components are studied from the atomic level by computer simulation techniques to illustrate the availability of the predicted targets and components from the side.

In addition, the association rule for the FCTF models was evaluated by the Apriori algorithm, EWM, TOPSIS, SVM and RF for the FCTF model. Apriori is a classical algorithm for ARM techniques that is widely applied in pharmaceutical and biological analysis (47). The EWM is commonly used in combination with TOPSIS for application assessment in different fields such as management (48), medicine and biology (49). SVM and RF classifiers are also popularly adopted for model evaluation of pharmaceuticals (50). In this study, we combined the multiple methods mentioned above to show that the (CNR2 → PPARA), (PPARA → CNR2), and (AR → CYP19A1) target sets were ranked high, and the targets with more frequent contributions were AR, CYP19A1 and ACE, etc. The association rule and its algorithm is a common technique in the field of data mining to discover correlations and patterns between items in a dataset. Currently, data mining and association rule analysis are used extensively in biomedical research (51–54), but less in the research on medicinal food or other food, thus there exists a potential prospect and wide space in the food field. As the FCTF model constructed by this research, by mining a large amount of food data, and analyzing and exploring the correlations and patterns between components, targets and functions, more valuable information about medicinal foods, health foods or green foods can be revealed to provide a scientific basis for their promotion and development.

The FCTF model proposed in this study is still in the early stage of establishment, and there are several limitations that need to be considered. First, the library of components included in the study needs to be continuously updated, the components determine the later targets and efficacy, and changes in components will lead to potential bias in the model construction. Second, the databases and online platforms for target and efficacy searches are also constantly being updated and need to be researched and updated in time to present more convincing models. In addition, the evaluation index of the FCTF model in this study is mostly based on Degree Centrality, i.e., the higher the value of intersecting nodes is taken into consideration, which is a common weighting scheme for commonality analysis (33). However, it is more subjective and limited, and other rules can be added to refine and improve it according to specific situations in the future. Here we only propose a new theoretical direction for food research, which is the result of data integration, and further experimental validation is needed to explore this model. The composition information of the example Laoxianghuang was not specifically recorded in any of the databases and was mainly obtained by our own detection and the literature, so the FCTF model of Laoxianghuang needed to be re-analyzed and re-established when new compositions appeared.

The FCTF model not only uses ARM theory but also combines analytical tools from systems biology and computational biology, which is a major breakthrough in the interdisciplinary and innovative ideas of food science, and we suggest taking a place for it in modern expensive and time-consuming research. Cross-fertilization of disciplines is an important driver for accelerating science and technology innovation, and strengthening interdisciplinarity and seeking new research paradigms are important ways to promote science and technology innovation (55). The proposed FCTF research model breaks away from the inertia of research in the food field and facilitates its continuous integration with different disciplines to achieve complementary strengths, with a view to promoting the development of the food discipline to a new level.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YL and YZ: conceptualization, formal analysis. ZZ and WL: methodology. YL and HL: software, visualization. MiL, JW, PY, and MoL: investigation. YL, HL, and LC: data curation. YL: writing—original draft preparation. YZ: writing—review and editing, supervision, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Guangdong Key Laboratory of Functional Substances in Medicinal Edible Resources and Healthcare Products (grant number 2021B1212040015), Scientific projects of key disciplines in Guangdong Province (grant numbers 2021ZDJS042 and 2022ZDJS070), Doctor Initiating Project of the Hanshan Normal University (grant numbers QD202125 and QD20190527), Special Focus Areas for General Universities of Guangdong Province (grant number 2022ZDZX2043), Education Department Project of Guangdong Province (grant numbers 2019-GDXK-0032 and 2020KZDZX1146), and Eastern Guangdong Technological Engineering Research Center (grant number P19004).

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2023.1170084/full#supplementary-material

References

1. Yang F, Zhang Y, Tariq A, Jiang X, Ahmed Z, Zhihao Z, et al. Food as medicine: a possible preventive measure against coronavirus disease (COVID-19). *Phytotherapy Res.* (2020) 34:3124–36. doi: 10.1002/ptr.6770

2. Hyman M, Bradley E. Food, medicine, and function: food is medicine part 2. *Phys Med Rehabil Clin N Am.* (2022) 33:571–86. doi: 10.1016/j.pmr.2022.04.002

3. Joshi VK, Joshi A. Rational use of Ashwagandha in Ayurveda (traditional Indian medicine) for health and healing. *J Ethnopharmacol.* (2021) 276:114101. doi: 10.1016/j.jep.2021.114101

4. Rist MJ, Wenzel U, Daniel H. Nutrition and food science go genomic. *Trends Biotechnol.* (2006) 24:172–8. doi: 10.1016/j.tibtech.2006.02.001

5. Almeida AM, Bassols A, Bendixen E, Bhide M, Ceciliani F, Cristobal S, et al. Animal board invited review: advances in proteomics for animal and food sciences. *Animal.* (2015) 9:1–17. doi: 10.1017/S1751731114002602

6. Yuliana ND, Hunaefi D, Goto M, Ishikawa YT, Verpoorte R. Measuring the health effects of food by metabolomics. *Crit Rev Food Sci Nutr.* (2022) 62:6359–73. doi: 10.1080/10408398.2021.1901256

7. Sun T, Wang X, Cong P, Xu J, Xue C. Mass spectrometry-based lipidomics in food science and nutritional health: a comprehensive review. *Compr Rev Food Sci Food Saf.* (2020) 19:2530–58. doi: 10.1111/1541-4337.12603

8. Huang YP, Robinson RC, Barile D. Food glycomics: dealing with unexpected degradation of oligosaccharides during sample preparation and analysis. *J Food Drug Anal.* (2022) 30:62–76. doi: 10.38212/2224-6614.3393

9. Somyanonthanakul R, Warin K, Amasiri W, Mairiang K, Mingmalairak C, Panichkitkosolkul W, et al. Forecasting COVID-19 cases using time series modeling and association rule mining. *BMC Med Res Methodol.* (2022) 22:281. doi: 10.1186/s12874-022-01755-x

10. Tandan M, Acharya Y, Pokharel S, Timilsina M. Discovering symptom patterns of COVID-19 patients using association rule mining. *Comput Biol Med.* (2021) 131:104249. doi: 10.1016/j.compbiomed.2021.104249

11. Jia N, Madina Z. An association rule-based multiresource mining method for MOOC teaching. *Comput Math Methods Med.* (2022) 2022:6503402–7. doi: 10.1155/2022/6503402

12. Hu L. Research on English achievement analysis based on improved CARMA algorithm. *Comput Intell Neurosci.* (2022) 2022:8687879–11. doi: 10.1155/2022/8687879

13. Pawlicka A, Tomaszewska R, Krause E, Jaroszewska-Choraś D, Pawlicki M, Choraś M. Has the pandemic made us more digitally literate?: innovative association rule mining study of the relationships between shifts in digital skills and cybersecurity awareness occurring whilst working remotely during the COVID-19 pandemic. *J Ambient Intell Humaniz Comput.* (2022) 1–11:1–11. doi: 10.1007/s12652-022-04371-1

14. Lu PH, Keng JL, Tsai FM, Lu PH, Kuo CY. An Apriori algorithm-based association rule analysis to identify Acupoint combinations for treating diabetic gastroparesis. *eCAM.* (2021) 2021:6649331–9. doi: 10.1155/2021/6649331

15. Wu RMX, Zhang Z, Yan W, Fan J, Gou J, Liu B, et al. A comparative analysis of the principal component analysis and entropy weight methods to establish the indexing measurement. *PLoS One.* (2022) 17:e0262261. doi: 10.1371/journal.pone.0262261

16. Damle M, Krishnamoorthy B. Identifying critical drivers of innovation in pharmaceutical industry using TOPSIS method. *Methods X.* (2022) 9:101677. doi: 10.1016/j.mex.2022.101677

17. Qin Z, Xi Y, Zhang S, Tu G, Yan A. Classification of Cyclooxygenase-2 inhibitors using support vector machine and random Forest methods. *J Chem Inf Model.* (2019) 59:1988–2008. doi: 10.1021/acs.jcim.8b00876

18. Guo S, Zheng Y, Guo R, Zeng X, Liang H, Chen Y, et al. Quantitative analysis and chemical pattern recognition of Lao-Xiang-Huang preserved in different years. *J Instrum Anal.* (2021) 40:10–8. doi: 10.3969/j.issn.1004-4957.2021.01.002

19. Liu Z, Zhang Z, Lai X, Yang Q, Lu Y, Huang Q, et al. Analysis on HPLC fingerprints and index content determination of Lao-Xiang-Huang of Chaozhou. *World Sci Technol.* (2017) 19:1370–4. doi: 10.11842/wst.2017.08.020

20. Yaqun L, Hanxu L, Wanling L, Yingzhu X, Mouquan L, Yuzhong Z, et al. SPME-GC-MS combined with chemometrics to assess the impact of fermentation time on the components, flavor, and function of Laoxianghuang. *Front Nutr.* (2022) 9:915776. doi: 10.3389/fnut.2022.915776

21. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* (2021) 49:D1388–95. doi: 10.1093/nar/gkaa971

22. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* (2014) 42:W32–8. doi: 10.1093/nar/gku293

23. Cai J, Zhao J, Gao P, Xia Y. Patchouli alcohol inhibits GPBAR1-mediated cell proliferation, apoptosis, migration, and invasion in prostate cancer. *Transl Androl Urol.* (2022) 11:1555–67. doi: 10.21037/tau-22-667

24. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* (2019) 10:1523. doi: 10.1038/s41467-019-09234-6

25. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative Toxicogenomics database (CTD): update 2021. *Nucleic Acids Res.* (2021) 49:D1138–43. doi: 10.1093/nar/gkaa891

26. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM. org: online Mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* (2015) 43:D789–98. doi: 10.1093/nar/gku1205

27. Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* (2020) 48:D1031–41. doi: 10.1093/nar/gkz981

28. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards version 3: the human gene integrator. *Database.* (2010) 2010:baq020. doi: 10.1093/database/baq020

29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* (2000) 28:235–42. doi: 10.1093/nar/28.1.235

30. Liu Y, Grimm M, Dai WT, Hou MC, Xiao ZX, Cao Y. CB-dock: a web server for cavity detection-guided protein-ligand blind docking. *Acta Pharmacol Sin.* (2020) 41:138–44. doi: 10.1038/s41401-019-0228-6

31. Odhar HA, Rayshan AM, Ahjel SW, Hashim AA, Albeer AAMA. Molecular docking enabled updated screening of the matrix protein VP40 from Ebola virus with millions of compounds in the MCULE database for potential inhibitors. *Bioinformation.* (2019) 15:627–32. doi: 10.6026/97320630015627

32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303

33. Tadaka S, Kinoshita K. NCMine: Core-peripheral based functional module detection using near-clique mining. *Bioinformatics.* (2016) 32:3454–60. doi: 10.1093/bioinformatics/btw488

34. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* (2021) 49:D605–12. doi: 10.1093/nar/gkaa1074

35. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinform.* (2011) 12:35. doi: 10.1186/1471-2105-12-35

36. Li Z, Li X, Tang R, Zhang L. Apriori algorithm for the data Mining of Global Cyberspace Security Issues for human participatory based on association rules. *Front Psychol.* (2021) 11:582480. doi: 10.3389/fpsyg.2020.582480

37. Abdel-Basset M, Manogaran G, Gamal A, Smarandache F. A group decision making framework based on Neutrosophic TOPSIS approach for smart medical device selection. *J Med Syst.* (2019) 43:38. doi: 10.1007/s10916-019-1156-1

38. Chen H, Jiang R, Huang W, Chen K, Zeng R, Wu H, et al. Identification of energy metabolism-related biomarkers for risk prediction of heart failure patients using random forest algorithm. *Front Cardiovasc Med.* (2022) 9:993142. doi: 10.3389/fcvm.2022.993142

39. Chen X, Lin L, Cai H, Gao X. Identification and analysis of metabolites that contribute to the formation of distinctive flavour components of Laoxianghuang. *Foods.* (2023) 12:425. doi: 10.3390/foods12020425

40. Fan T, Qu R, Yu Q, Sun B, Jiang X, Yang Y, et al. Bioinformatics analysis of the biological changes involved in the osteogenic differentiation of human mesenchymal stem cells. *J Cell Mol Med.* (2020) 24:7968–78. doi: 10.1111/jcmm.15429

41. Li T, Zhang W, Hu E, Sun Z, Li P, Yu Z, et al. Integrated metabolomics and network pharmacology to reveal the mechanisms of hydroxysafflor yellow a against acute traumatic brain injury. *Comput Struct Biotechnol J.* (2021) 19:1002–13. doi: 10.1016/j.csbj.2021.01.033

42. Ma QG, Wei RR, Yang M, Huang XY, Wang F, Dong JH, et al. Isolation and characterization of neolignan derivatives with hepatoprotective and neuroprotective activities from the fruits of *Citrus medica* L. var. Sarcodactylis Swingle. *Bioorg Chem.* (2021) 107:104622. doi: 10.1016/j.bioorg.2020.104622

43. Gupta M, Sharma R, Kumar A. Docking techniques in pharmacology: how much promising? *Comput Biol Chem.* (2018) 76:210–7. doi: 10.1016/j.compbiolchem.2018.06.005

44. Torres PHM, Sodero ACR, Jofily P, Silva-Jr FP. Key topics in molecular docking for drug design. *Int J Mol Sci.* (2019) 20:4574. doi: 10.3390/ijms20184574

45. Jiao X, Jin X, Ma Y, Yang Y, Li J, Liang L, et al. A comprehensive application: molecular docking and network pharmacology for the prediction of bioactive constituents and elucidation of mechanisms of action in component-based Chinese medicine. *Comput Biol Chem.* (2021) 90:107402. doi: 10.1016/j.compbiolchem.2020.107402

46. Kaushik AC, Sahi S, Wei DQ. Computational methods for structure-based drug design through system biology. *Methods Mol Biol Clifton.* (2022) 2385:161–74. doi: 10.1007/978-1-0716-1767-0_9

47. Zheng Y, Chen P, Chen B, Wei D, Wang M. Application of Apriori improvement algorithm in asthma case data mining. *J Healthc Eng.* (2021) 2021:9018408–7. doi: 10.1155/2021/9018408

48. Li M, Sun H, Singh VP, Zhou Y, Ma M. Agricultural water resources management using maximum entropy and entropy-weight-based TOPSIS methods. *Entropy*. (2019) 21:364. doi: 10.3390/e21040364

49. Chen JM, Wang T, Guo QS, Li HW, Zuo L, Zou QJ, et al. Comprehensive antioxidant and anti-inflammatory activity of alcohol extracts from *Chrysanthemum indicum* in different areas based on entropy weight and TOPSIS methodology. *Zhongguo zhongyao zazhi*. (2021) 46:907–14. doi: 10.19540/j.cnki.cjcmm.20201122.102

50. Nayarisseri A, Khandelwal R, Tanwar P, Madhavi M, Sharma D, Thakur G, et al. Artificial intelligence, big data and machine learning approaches in precision medicine & drug discovery. *Curr Drug Targets*. (2021) 22:631–55. doi: 10.2174/1389450122999210104205732

51. Pradhan GN, Prabhakaran B. Association rule Mining in Multiple, multidimensional time series medical data. *J Healthc Inform Res*. (2017) 1:92–118. doi: 10.1007/s41666-017-0001-x

52. Martínez-Romero M, O'Connor MJ, Egyedi AL, Willrett D, Hardi J, Graybeal J, et al. Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases. *Database*. (2019) 2019:baz059. doi: 10.1093/database/baz059

53. Bandyopadhyay S, Mallik S. Integrating multiple data sources for combinatorial marker discovery: a study in tumorigenesis. *IEEE/ACM Trans Comput Biol Bioinform*. (2018) 15:673–87. doi: 10.1109/TCBB.2016.2636207

54. Mallik S, Mukhopadhyay A, Maulik U. RANWAR: rank-based weighted association rule mining from gene expression and methylation data. *IEEE Trans Nanobiosci*. (2015) 14:59–66. doi: 10.1109/TNB.2014.2359494

55. Vári Á, Podschun SA, Erős T, Hein T, Pataki B, Iojă IC, et al. Freshwater systems and ecosystem services: challenges and chances for cross-fertilization of disciplines. *Ambio*. (2022) 51:135–51. doi: 10.1007/s13280-021-01556-4

frontiers | Frontiers in Nutrition

TYPE Original Research
PUBLISHED 13 October 2023
DOI 10.3389/fnut.2023.1186221

OPEN ACCESS

EDITED BY
Alessandra Durazzo,
Council for Agricultural Research and Economics, Italy

REVIEWED BY
M. Graça Dias,
Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA), Portugal
Kathleen L. Hefferon,
Cornell University, United States
Yan Bai,
World Bank Group, United States
Christian Napoli,
Sapienza University of Rome, Italy

*CORRESPONDENCE
Yusentha Balakrishna
✉ yusentha.balakrishna@mrc.ac.za

RECEIVED 14 March 2023
ACCEPTED 28 September 2023
PUBLISHED 13 October 2023

CITATION
Balakrishna Y, Manda S, Mwambi H and van Graan A (2023) Determining classes of food items for health requirements and nutrition guidelines using Gaussian mixture models. *Front. Nutr.* 10:1186221.
doi: 10.3389/fnut.2023.1186221

Determining classes of food items for health requirements and nutrition guidelines using Gaussian mixture models

Yusentha Balakrishna[1,2]*, Samuel Manda[2,3], Henry Mwambi[2] and Averalda van Graan[4,5]

[1]Biostatistics Research Unit, South African Medical Research Council, Durban, South Africa, [2]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa, [3]Department of Statistics, University of Pretoria, Pretoria, South Africa, [4]Biostatistics Research Unit, SAFOODS Division, South African Medical Research Council, Cape Town, South Africa, [5]Division of Human Nutrition, Department of Global Health, Stellenbosch University, Cape Town, South Africa

Introduction: The identification of classes of nutritionally similar food items is important for creating food exchange lists to meet health requirements and for informing nutrition guidelines and campaigns. Cluster analysis methods can assign food items into classes based on the similarity in their nutrient contents. Finite mixture models use probabilistic classification with the advantage of taking into account the uncertainty of class thresholds.

Methods: This paper uses univariate Gaussian mixture models to determine the probabilistic classification of food items in the South African Food Composition Database (SAFCDB) based on nutrient content.

Results: Classifying food items by animal protein, fatty acid, available carbohydrate, total fibre, sodium, iron, vitamin A, thiamin and riboflavin contents produced data-driven classes with differing means and estimates of variability and could be clearly ranked on a low to high nutrient contents scale. Classifying food items by their sodium content resulted in five classes with the class means ranging from 1.57 to 706.27 mg per 100 g. Four classes were identified based on available carbohydrate content with the highest carbohydrate class having a mean content of 59.15 g per 100 g. Food items clustered into two classes when examining their fatty acid content. Foods with a high iron content had a mean of 1.46 mg per 100 g and was one of three classes identified for iron. Classes containing nutrient-rich food items that exhibited extreme nutrient values were also identified for several vitamins and minerals.

Discussion: The overlap between classes was evident and supports the use of probabilistic classification methods. Food items in each of the identified classes were comparable to allowed food lists developed for therapeutic diets. This data-driven ranking of nutritionally similar classes could be considered for diet planning for medical conditions and individuals with dietary restrictions.

KEYWORDS

food composition database, nutrient table, mixture model, clustering, classification, nutritional content

1. Introduction

The study of single nutrients in food items has played an important role in our understanding of the basic causes and treatment strategies of nutrition-related diseases (1). Establishing the relationships between specific nutrients and food items and determining the association between specific nutrient intakes and diseases, may help with the interpretation of dietary patterns found in a population and the explanation of the association between dietary patterns and disease (2). In addition, a reasonable first step toward the development of food-based dietary guidelines (FBDGs) is identifying the food sources of the nutrient of interest. This information can be ascertained from food composition databases (FCDBs) and understanding food items and their nutrients promotes a basic knowledge of nutrition amongst the population (3). The analysis of dietary patterns is dependent on the categorization of food items but the rules determining this categorization, which are based on conceptual and compositional similarity, are not always well-defined (2).

The need to group foods by nutritional content was recognized by Khan (4) who proposed categorizing foods as having a either a low, medium or high specific nutrient content to assist dietitians with food recommendations. However, the proposed category thresholds were suggestive and a more rigorous method of determining the thresholds was needed. More recently, a more suitable, data-driven categorization was proposed, using *k*-means clustering to group foods by nutrient content (5). Other methods that have been used to classify food items are hierarchical clustering, principal component analysis (PCA), factor analysis and fuzzy clustering (6). Thus, employing statistical clustering methods to food composition data can produce objectively determined classes. A previous study (7) applied PCA to food composition data to identify nutritionally similar groups. However, evaluating similar food items through PCA does not account for the uncertainty in assigning food items to classes. In addition, while food items were able to be grouped by overall nutritional similarity, food items were unable to be ranked by the level of a specific nutrient content. The ability to rank food items by the level of nutrients is essential for creating food lists for therapeutic diets. Some common therapeutic diets that involve nutrient modification are renal diets for the management of chronic kidney disease (8) and low carbohydrate diets for the management of diabetes (9).

A recent review has shown that mostly centroid-based and hierarchical clustering techniques have been applied to food composition data (6) but mixture models have yet to be investigated in this context. The application of mixture models to identify dietary patterns in food consumption studies has shown advantages over nonparametric approaches (10, 11). Nonparametric approaches result in classes wherein each food item belongs exclusively to one class, thus assuming that the classification uncertainty is zero. However, if arbitrary thresholds existed to separate low and high nutrient content foods, there is a weak separation between food items containing nutrient levels that are near the threshold. Mixture models accommodate for this uncertainty by measuring the probability of class membership, which takes values between zero and one (10). With probabilistic clustering, the focus is not on whether a food is in a class, but rather to what extent it is associated with that class (12). The consideration of the uncertainty in determining nutritional classes allows for greater precision and reduced allocation bias (13).

Probabilistic clustering or distribution-based clustering assumes that the nutrient values are generated by a mixture of probability distributions and that each distribution forms a class. Each food item is assigned a probability of class membership (these being the posterior probabilities), thus supporting multiple class membership and also the assignment of outliers to classes. The most popular algorithm of this approach is the Gaussian mixture model (GMM). For a dataset of n food items that one wants to classify into k compositionally similar groups, the GMM assumes that the overall nutrient content distribution consists of a mixture of k Gaussian (normal) distributions. In this study, we apply univariate GMMs to food composition data to determine classes that contain similar levels of specific nutrients and to allow for the estimation of the class membership probabilities for each food item.

2. Materials and methods

2.1. Data

The 2017 SAFCDB (11) contains nutritional information for 1,667 food items and 169 food components (hereon termed ‘nutrients’). The compilation of food composition data for the SAFCDB comprises various number of data sources ranging from national projects involving direct methods and indirect methods, to the sourcing of scientific literature, certificate of analyses and product nutritional information from various data generators.

Of the 169 food components, we selected the most common nutrients with the least amount of missing values for inclusion. We also considered nutrients that were non-collinear. For example, since total carbohydrate is the sum of available carbohydrate and dietary fibre, available carbohydrate and dietary fibre were included instead of total carbohydrate. Using these criteria, we selected 28 nutrients (nine macronutrients, nine minerals and ten vitamins) for analysis and included food items ($n = 971$) which had non-missing nutrient information for all 28 nutrients.

For each of the 28 nutrients, each of the 971 food items had either a known nutrient value, a zero nutrient value or a trace value. Food items with a zero nutrient value for a particular nutrient are excluded from the univariate GMM analysis since we are interested in classifying only food items known to have the nutrient of interest. Trace values were imputed with half the limit of detection for each nutrient (14). Thus, only food items containing either a known nutrient value or trace value are included in the analysis. Extreme nutrient values were retained in the dataset. Raw food items, cooked food items and combined dishes (where nutrient composition has been calculated using standard recipes) from various food groups were included in the analysis (Table 1). All nutrient values were expressed per 100 g edible part.

2.2. Methods

2.2.1. Univariate Gaussian mixture model

In the case of food composition data, the univariate Gaussian mixture model assumes that the nutrient content values arise from a

TABLE 1 Number of food items analyzed by food group.

Food group	*n* (%)
Cereals and cereal products	195 (20.08)
Vegetables	245 (25.23)
Fruit	132 (13.59)
Legumes and legume products	26 (2.68)
Nuts and seeds	20 (2.06)
Milk and milk products	41 (4.22)
Eggs	27 (2.78)
Meat and meat products	120 (12.36)
Fish and seafood	36 (3.71)
Fats and oils	26 (2.68)
Sugar, syrups and sweets	17 (1.75)
Soups, sauces, seasonings and flavorings	30 (3.09)
Beverages	27 (2.78)
Infant and paediatric feeds and foods	10 (1.03)
Therapeutic/special/diet products	7 (0.72)
Miscellaneous	12 (1.24)
Total	971 (100)

mixture of two or more Gaussian distributions. Each Gaussian distribution represents a class of food items. Since Gaussian distributions can be described by the mean and variance, the means and variances for each class of food items can be estimated.

The means and variances for each class of food items can be estimated via an iterative process called the Expectation–Maximization (EM) algorithm (15). Since we do not know the means and variances for each class of food items beforehand, we begin with an initial guess for each and iterate between an expectation step (E-step) and a maximization step (M-step). In the E-step, we calculate the probability that a food item belongs to a specific class. In the M-step, we update the mean and variances for each class, based on the probabilities calculated in the expectation step. The steps are repeated until there are no significant changes in either the means and variances or the log-likelihood (how well the model fits the data). The mathematical definitions of the univariate GMM follow.

Suppose that x_{ij} is the amount of nutrient j for food item i ($i = 1,2,\ldots,971; j = 1,2,\ldots,28$). We assume that the nutrient value x_{ij} arises from a mixture composed of k unobserved classes. Formally, x_{ij} is a sum of class-specific nutrient distributions as

$$p\left(x_{ij}\right) = \sum_{k=1}^{K} p\left(x_{ij}|z_{ij} = k\right) p\left(z_{ij} = k\right)$$

$$= \sum_{k=1}^{K} \pi_k \, p_k\left(x_{ij}\right)$$

where K is the number of classes, $z_{ij} = 1,2,\ldots,k,\ldots,K$ indicates the class for x_{ij}, $p_k\left(x_{ij};\boldsymbol{\theta}_k\right)$ is the probability distribution for class k with parameter vector $\boldsymbol{\theta}_k$ and π_k is the proportion of food items that belong to class k such that

$$0 \le \pi_k \le 1$$

and

$$\sum_{k=1}^{K} \pi_k = 1$$

Assuming $p_k\left(x_{ij}\right) = N\left(\mu_k, \sigma_k^2\right)$, then $p_k\left(x_{ij}\right)$ follows a Gaussian distribution and $p\left(x_{ij}\right)$ becomes a Gaussian mixture distribution. Thus, for the univariate GMM

$$z_{ij} \sim Cat\left(\boldsymbol{\pi}\right)$$

$$x_{ij} \mid z_{ij} = k \sim N\left(\mu_k, \sigma_k^2\right)$$

where $\boldsymbol{\pi}$ is the vector of proportions, μ_k is the mean nutrient content for class k and σ_k is the associated standard deviation for class k.

The EM algorithm can be utilized when we need to conduct a maximum likelihood estimation of parameters in the presence of missing data or latent variables. The E- and M-steps for the univariate GMM are outlined below.

2.2.2. The E-step

Calculate the responsibilities γ_{iz} (posterior probabilities) for the *i*th food item and *z*th class:

$$\gamma_{iz} = \frac{\pi_z\left(\frac{1}{\sigma_z\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_z^2}\left(x_i - \mu_z\right)^2\right)\right)}{\sum_{k=1}^{K}\pi_k\left(\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_k^2}\left(x_i - \mu_k\right)^2\right)\right)}$$

2.2.3. The M-step

Calculate the new parameters μ_z^*, σ_z^*, and π_z^* via maximization using

$$\mu_z^* = \frac{\sum_{i=1}^{N}\gamma_{iz}x_i}{\sum_{i=1}^{N}\gamma_{iz}}$$

$$\sigma_z^* = \sqrt{\frac{1}{\sum_{i=1}^{N}\gamma_{iz}}\sum_{i=1}^{N}\gamma_{iz}\left(x_i - \mu_z^*\right)^2}$$

$$\pi_z^* = \frac{\sum_{i=1}^{N}\gamma_{iz}}{N}$$

The EM algorithm begins with initialization and is iterated until convergence of the parameters or log-likelihood is reached (16).

2.2.4. Statistical analysis

After examining the distributions for each nutrient, we aimed to fit a univariate GMM for each natural log-transformed nutrient. 'Moisture' was kept on the original scale. We used the 'mclust' (17) and 'mixtools' (18) R packages to fit the models. The steps followed are outlined below. For each of the 28 nutrients:

1. We determined the optimal number of classes to fit using quantiles to initialize the EM algorithm. Ten GMMs were fitted in succession for *k* (the number of classes) ranging from 1 to 10 and the Bayesian Information Criterion (BIC) (19) was computed for each model. The *k* that minimized the BIC was selected as the optimal number of classes.
2. We used the EM algorithm with random initialisation to fit the GMM with the optimal *k*. To avoid local optima, the model was fitted 10 times and the model with the highest log-likelihood was selected. Convergence was declared when the change in the observed log-likelihood increased by less than 10^{-8} .
3. The parameter estimates for the mean (μ), standard deviation (σ) and proportion (π) from the selected model were recorded and the GMM density function was plotted.
4. Food items were assigned to classes based on their highest estimated probability of class membership. The class validity of the GMM solutions was assessed using the Davies-Bouldin (DB) (20) index and silhouette coefficient (21).

The DB index measures the average separation between each class and its next nearest class. The index is bounded between zero and infinity with values closer to zero indicating a better partitioning. The silhouette coefficient measures how similar an observation is to observations in its own class (compactness) compared to observations in other classes (separation). The silhouette coefficient is bounded between −1 and 1, where negative values indicate incorrect classifications, values close to 1 indicate highly dense classifications and scores around zero indicate overlapping classifications (observations lying between two classes). Scores greater than 0.5 are generally desirable for good classifications (22).

3. Results

3.1. Model selection

The BIC was compared for the 1- to 10-class GMMs. The most frequent model selected was the two-class model (n = 14/28) followed by the four-class model (n = 6/28). The highest number of classes was found when food items were grouped based on sodium content and niacin content with five and seven classes, respectively. Plant protein, calcium and vitamin B_6 were best described by a single class, that is, the univariate normal model.

3.2. Identified classes

The parameter estimates corresponding to the classes are presented in Table 2. Figures 1–3 depict each nutrient-based classification, which can be described as a mixture of Gaussian distributions. Hence, each Gaussian distribution on the plots represents a class.

Five classes of food items were identified when classifying food items by sodium content and the mean sodium content of the classes ranged from 1.57 mg to 706.27 mg per 100 g (Table 2). Food items identified as having the highest sodium content were bread, potato crisps, breakfast cereals, canned vegetables, dehydrated potato mash, milk powders, processed meat, canned/cured/smoked fish, butter, margarine, mayonnaise and packaged soup mix (Table 3). Grouping foods by their available carbohydrate content resulted in four identified classes (Table 2). Class 4 contained foods with the highest mean available carbohydrate content of 59.15 g per 100 g and consisted of baked goods, starchy vegetables, and sugar and sweets (Table 4). Food items grouped by their fatty acid content were found to consist of two classes for each of the fatty acids, suggesting that food items could naturally be grouped into having either a low or a high fatty acid content. Food items associated with having a high fatty acid content were baked goods, fried foods, nuts and seeds, dairy products, eggs, meat products, caviar, high-fat fish and fats and oils (Table 5). Three classes of food items were identified when the grouping was based on iron content. Class 2 had the highest mean iron level of 1.46 mg per 100 g and contained mainly wheat products, dehydrated raw vegetables, green vegetables, beetroot, mushroom, dried fruit, legumes, nuts and seeds, milk powder with added iron, eggs, meat (excluding white meat chicken and veal) and certain seafood (Table 6).

The study found that the classification of food items using moisture (Supplementary Table 1), animal protein (Table 7) and sodium (Table 3) content could be described by low, moderately-low, moderately-high and high nutrient content classes. Based on the saturated, mono-unsaturated and polyunsaturated fatty acid content, food items could be described by low- and high-content classes (Table 5). When examining their available carbohydrate content, food items could be described as having an extremely low, low, moderate and high available carbohydrate content (Table 4). Low, moderate and high nutrient content classes of food items were also identified based on vitamin A (RE) and thiamin content (Supplementary Table 3). While most food items exhibited a clear belonging to classes, a few food items exhibited multiclass membership. For example, for vitamin A (RE) content, raw leaves other than amaranth had an approximately equal probability of belonging to either the moderate content or high content class while amaranth leaves had a clear belonging to the high content class. Other classes of interest are shown in Supplementary Tables 2–5. Food items with a high copper content were identified by class 2 and consisted of wheat flour, maize meal, leafy greens, mushrooms, potatoes, beans, lentils, nuts and seeds, organ meat, shellfish and chocolate (Supplementary Table 2). The distributions for cholesterol and manganese did not display classes that could be intuitively ranked (Supplementary Table 6).

Classes capturing foods exhibiting extreme values of nutrients were also identified. These classes contained both foods having low or extremely low nutrient content and foods having a high or extremely high nutrient content. This class was present in the distributions for magnesium, potassium, sodium, copper, riboflavin and pantothenic acid. The distribution of phosphorous also contained two classes with class 1 describing foods with extremely low phosphorous content such as marrow squash, tomato juice, butter ghee, margarine, tea and baking powder.

TABLE 2 Parameter estimates for the univariate Gaussian mixture model[§].

Nutrient	*N*	Class 1			Class 2			Class 3			Class 4			Class 5			Class 6			Class 7		
		%	Mean	SD	%	Mean	SD	%	Mean	SD	%	Mean	SD	%	Mean	SD	%	Mean	SD	%	Mean	SD
Moisture (g)	964	12	6.68	4.33	15	33.8	14.35	43	71.1	10.65	29	87.2	5.03									
Plant protein (g)	749	100	1.49	1.17																		
Animal protein (g)	487	13	0.02	0.81	55	2.86	1.1	11	13.87	0.24	20	25.53	0.14									
Saturated fatty acids (g)	956	37	0.03	0.96	63	2.2	1.17															
Mono-unsaturated fatty acids (g)	955	39	0.03	1.06	61	2.83	1.12															
Polyunsaturated fatty acids (g)	957	47	0.08	1.23	53	2.1	1.2															
Cholesterol (mg)	434	65	30.3	1.59	35	70.81	0.34															
Carbohydrate, available (g)	879	4	0.48	1.64	22	2.75	0.56	58	13.07	0.64	15	59.15	0.22									
Fibre, total (g)	752	38	0.73	1.67	62	2.23	0.77															
Calcium (mg)	961	100	27.39	1.28																		
Iron (mg)	965	18	0.34	1.31	58	1.46	0.89	25	0.48	0.46												
Magnesium (mg)	960	62	14.01	0.48	38	24.29	1.35															
Phosphorous (mg)	959	2	6.11	3.55	98	66.69	1.1															
Potassium (mg)	963	23	156.02	1.53	77	186.79	0.56															
Sodium (mg)	959	9	1.57	0.45	27	6.23	0.79	3	13.07	3.75	54	78.26	0.83	8	706.27	0.48						
Zinc (mg)	962	72	0.44	0.99	6	0.39	0.06	13	0.34	1.87	9	3.6	0.32									
Copper (mg)	958	44	0.09	0.45	56	0.1	1.34															
Manganese (µg)	957	75	93.69	1.81	25	165.67	0.53															
Vitamin A (RE) (µg)	817	18	1.51	1.25	78	48.42	1.33	4	1844.57	0.76												
Thiamin (mg)	954	77	0.06	0.87	2	0.003	0.29	20	0.31	0.7												
Riboflavin (mg)	960	24	0.02	0.45	26	0.08	0.59	22	0.11	1.68	28	0.2	0.33									
Niacin (mg)	954	1	0.003	0.33	5	0.1	0.02	14	0.48	0.33	33	0.36	0.94	10	5.26	0.28	33	1.62	0.67	3	12.06	0.24
Vitamin B_6 (mg)	952	100	0.08	1.14																		
Vitamin B_{12} (µg)	487	10	0.005	0.22	32	0.34	0.48	45	0.45	1.97	13	1.7	0.29									
Pantothenic acid (mg)	954	45	0.29	1.49	55	0.28	0.57															
Vitamin C (mg)	721	68	2.03	1.93	32	11.7	0.93															
Vitamin D (µg)	471	13	0.03	1.14	87	0.85	1.19															
Vitamin E (mg)	924	98	0.51	1.5	2	0.005	1.05															

[§] Mean estimates are presented on its original scale per 100 g. Standard deviation (SD) estimates are presented on the natural-log scale. The percentage (%) of food items belonging to the class is also reported.

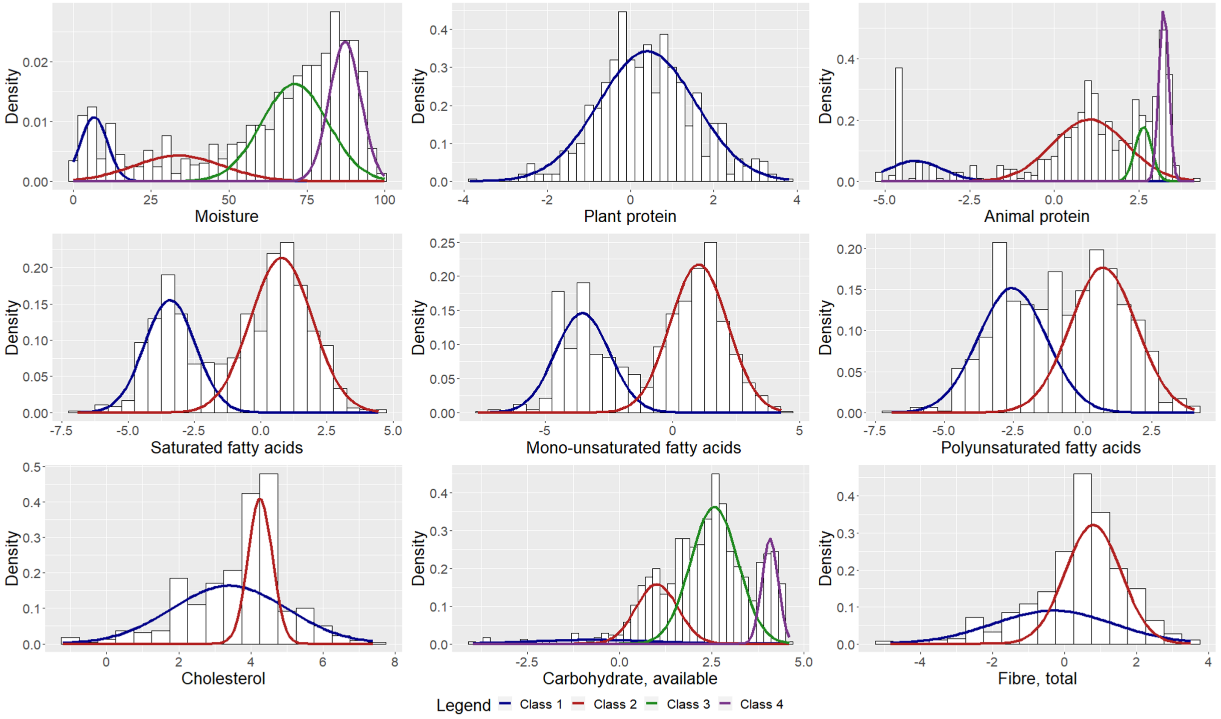


FIGURE 1
Univariate Gaussian mixture model for macronutrients.

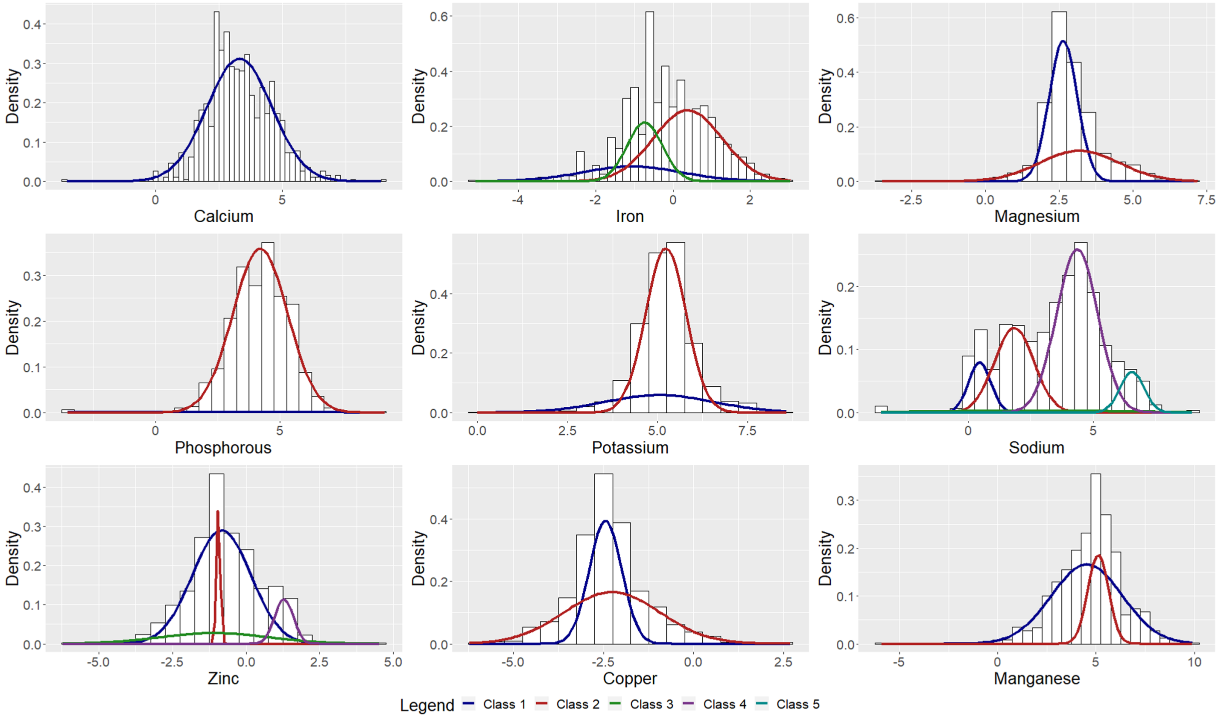


FIGURE 2
Univariate Gaussian mixture model for minerals.

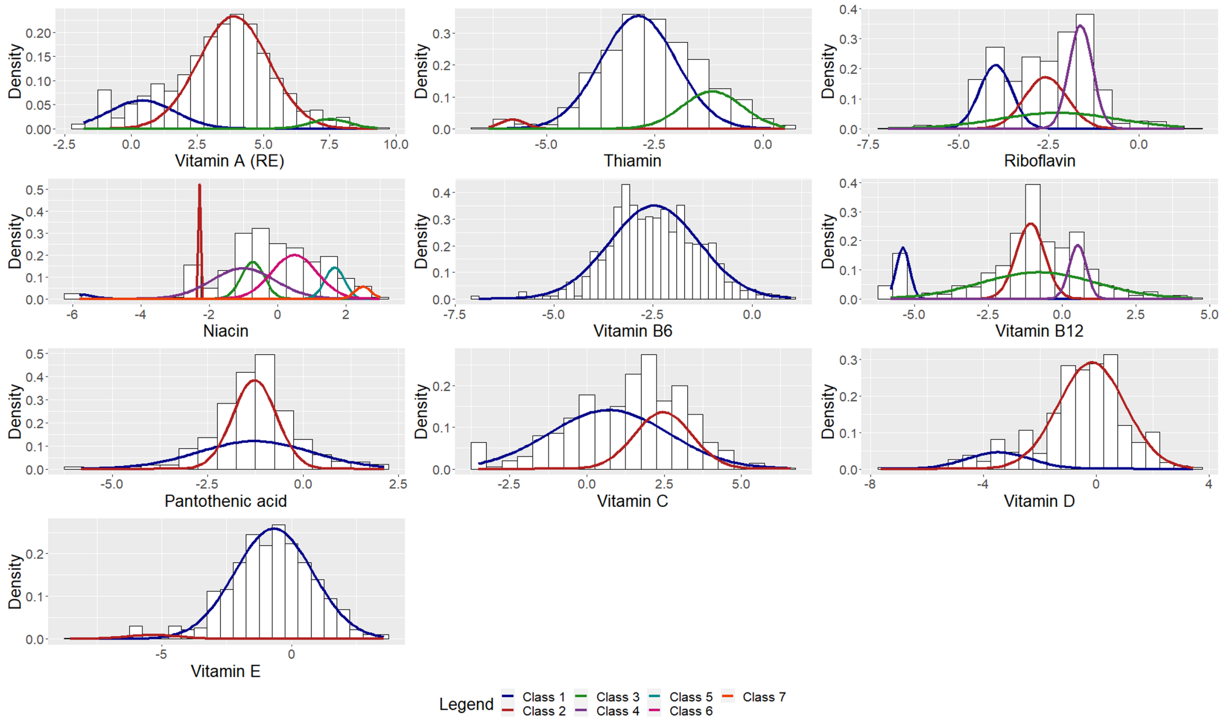


FIGURE 3
Univariate Gaussian mixture model for vitamins.

3.3. Class validity

The internal class validity indices are presented in Figure 4. Nutrient-based classifications with good DB indices and silhouette coefficients are indicated by green shading. The median DB index was 0.78 (IQR 0.45–4.08), suggesting that the GMM resulted in good classification. The minimum score was 0.3 for the vitamin E classifications and the highest scores, ranging from 4.65 to 9.7, were found for the potassium, sodium, zinc, copper and vitamin B_{12} classifications. An outlying score of 151.06 was found for the classification by pantothenic acid. Each of the classifications that were found to have a high DB score, contained a class that simultaneously captured foods with extremely high nutrient levels and foods with extremely low nutrient levels. For example, class 2 of the copper classification accounted for 56% of the food items and contained foods with both extremely low and extremely high levels of copper. Thus, classifications that contained such a class, tended to have the most overlap of classes and were congruent with having high DB scores.

The median silhouette coefficient was 0.5 (IQR 0.39–0.62), also suggesting that the GMM resulted in good classification. Negative silhouette coefficients were found for the zinc and manganese classifications, both of which had a significant overlap of classes. When examining the coefficients for each class, both the zinc and manganese classifications had some classes with high coefficients, suggesting that observations within these classes displayed good cohesion. Again, individual class coefficients were low for classes that captured both extremely high and extremely low values. For example, the first class of cholesterol accounted for 65% of the food items and described foods with either an extremely low cholesterol value or an extremely high cholesterol value. This class had a silhouette coefficient of −0.12 compared with the second cholesterol class which scored 0.81. This similar pattern was also seen for the potassium, sodium, zinc, copper, manganese, riboflavin, niacin and vitamin B_{12} classifications. Classes that had a significant overlap with other classes tended to have a negative silhouette coefficient.

4. Discussion

In this paper, we have applied Gaussian mixture models to the South African Food Composition Database to evaluate the application of probabilistic classification to food composition data. The classification of food items into nutritionally similar food groups is a common objective of studies that apply statistical methods for the analysis food composition data. Traditional food groupings are not enough to describe the nutritional landscape of food and compositionally similar food groups also need to be investigated. Identifying compositionally similar food groups can be achieved through clustering algorithms which are simple to employ. However, most of the clustering algorithms applied thus far assign food items exclusively to one class and the indistinct thresholds that may exist between food groups, based on nutritional content, needs to be considered. The application of probabilistic clustering can account for this uncertainty.

An important application of FCDBs is its role in the design of therapeutic diets (23). Renal disease, diabetes mellitus and anaemia are some examples of health conditions that require the monitoring of specific nutrients. Allowed food lists and food exchange lists are a useful tool for health practitioners and patients when managing such conditions. They are also useful for healthy individuals to improve

TABLE 3 Food items within the identified sodium classes.

Class	Class 1	Class 2	Class 3		Class 4	Class 5
Class description	Low content	Moderately-low content	Extremely low content	Extremely high content	Moderately-high content	High content
Food group						
Cereals and cereal products	Cooked maize meal porridges, cooked white rice, cooked oats, wheat flour, cooked pasta, uncooked semolina, roti	Cooked wheat, cooked egg noodles, cooked, brown rice, brown rice flour, cooked barley, raw maize meal, wheat germ, wheat flour, cooked wholewheat pasta			Baked goods, pasta dishes	Bread, potato crisps, breakfast cereals, self-raising wheat flour
Vegetables	Squash, potato, melon, boiled pumpkin	Bamboo shoots, green beans, tomato, baby marrow squash, brinjal, leaves, peas, mushroom, Brussels sprouts, onion, white-fleshed sweet potato, cauliflower, cabbage	Asparagus soup and boiled mangetout		Beetroot, vegetables cooked with margarine, dehydrated raw vegetables, carrots, leaves, baby sweetcorn, celery, canned vegetables	Canned baby sweetcorn, canned asparagus, canned sauerkraut, dehydrated potato mash, spinach, dehydrated cauliflower, canned olives
Fruit	Apple, banana, gooseberry, grapes, grapefruit juice, guava, lemon juice, mango, naartjie juice, orange juice, orange, pineapple, sour plum, prickly pear, raspberry, rhubarb, youngberry, date, granadilla, kiwifruit, lime, marula, medlar, mineola, nectarine, dried peach, dried prune	Canned fruit, stewed fruit, dried fruit, prunes, dates, pawpaw, figs, cherries, plums, peaches, rhubarb stems, strawberry, watermelon, kumquat, avocado, grapefruit, lemon, litchi, pear			Melon, raisins, fruit mincemeat, dried apple, candied orange/lemon peel, glazed cherry	
Legumes and legume products	Dried beans, cooked split peas, cooked lentils	Cooked rice and lentils dish, raw lentils, raw split peas, tofu, cooked beans, cooked chickpeas			Bean dishes, raw chickpeas, lentil dishes	
Nuts and seeds	Almonds (unsalted, blanched), pistachios, chestnuts, coconut, pine nuts, walnut	Unsalted peanuts, macadamia nuts, sunflower seeds, Brazil nuts, cashew nuts, unblanched almonds			Sesame seeds, desiccated coconut, salted peanuts	
Milk and milk products					Milk, yoghurt, custard, cottage cheese	Cheese, milk powders (low-fat, skim, added vitamins)
Eggs					Eggs	Dried egg
Meat and meat products					Meat	Processed meat
Fish and seafood				Fish biltong, anchovy	Fish, oyster, tuna, crab, mussels	Shrimp/prawn, rollmop/pickled herring, caviar, smoked fish, canned sardine
Fats and oils	French salad dressing, butter ghee, olive oil	Pressurized cream			Salad dressing, cream	Butter, margarine, mayonnaise
Sugar, syrups and sweets	Sugar	Honey, dark chocolate, jam/marmalade, jelly (with fruit)			Chocolate, icing, molasses	
Soups, sauces, seasonings and flavorings		Curry sauce, soup mix (with beef and vegetables)			Sauces and soups	Soup (packet mix)
Beverages		Fruit juices, fruit nectars			Milk beverages	
Infant and paediatric feeds and foods					Infant feeds	
Therapeutic/special/diet products					Therapeutic powders	
Miscellaneous	Tea, spirits	Vinegar, wine, liqueur, sherry, tea		Baking powder	Liqueur with cream	

TABLE 4 Food items within the identified available carbohydrate classes.

Class	Class 1	Class 2	Class 3	Class 4
Class description	Extremely low content	Low content	Moderate content	High content
Food group				
Cereals and cereal products			Milk tarts, white rice, pancakes, puddings, pasta dishes, soft and stiff maize meal porridge, scones	Raw maize meal, rice flour, wheat flour, potato flour, oats, cookies, cakes, bread, breakfast cereals
Vegetables	Rhubarb	Cucumber, leaves, marrow squash, spinach, broccoli, cabbage, cauliflower, Brussels sprouts, mushroom, brinjal, tomato, avocado, rhubarb stems, olives	Potato, white-fleshed sweet potato, butternut squash, parsnip, sweetcorn, carrot, peas, tomato paste, tomato purée, onion	Raw dehydrated starchy vegetables (carrot, onion, peas, potato)
Fruit		Grapefruit, melon, youngberry	Canned fruit, stewed fruit, raw fruit	Dried fruit
Legumes and legume products		Raw tofu, cooked soybeans	Beans, rice and lentil dishes, lentils, raw soybeans, cooked chickpeas	Dried beans, dried chickpeas
Nuts and seeds	Sesame seeds	Coconut, pecan nuts	Nuts and seeds	Chestnuts
Milk and milk products	Some cheeses (medium/reduced fat, Leicester, Gouda)	Cheese, sour milk	Milk, yoghurt, custard	Skim and low-fat milk powders
Eggs	Raw chicken egg (omega-3 enriched), raw quail egg	Eggs	Soufflé	
Meat and meat products	Offal, mutton, beef heart, beef kidney, beef patty	Frankfurter, pastrami, offal, luncheon meat, bacon, sausage, meatball, schnitzel, liver, ham, steak and kidney, chicken giblets	Commercial meat pies, meat spread, biltong, pâté, stews with meat and vegetables, corned beef	
Fish and seafood	Baked/fried fish	Boiled shrimp/prawn, baked kipper, oyster, caviar	Battered/crumbed fish, mussel, rollmop/pickled herring	
Fats and oils	Butter ghee, margarine	Cream	Salad dressing, peanut butter	
Sugar, syrups and sweets				Sugar and sweets
Soups, sauces, seasonings and flavorings		Cucumber soup, meat gravy, snakehead soup	Sauces	Caramel sauce
Beverages	Coffee, tea		Fruit juices, milk beverages	Malted milk powder, drinking chocolate powder
Infant and paediatric feeds and foods			Reconstituted infant feeds	Infant feed powders
Therapeutic/special/diet products			Reconstituted therapeutic products	Therapeutic powders
Miscellaneous		Wine	Baking powder, liqueur with cream, sherry	Liqueur

their nutrition education (24). Applying clustering methods to food composition data provides a data-driven method of establishing foods with similar nutritional content, for the development of allowed food lists.

Classifications based on cholesterol, total fibre, magnesium, potassium, copper and pantothenic acid content, indicated a clear overlap of two classes, supporting the use of probabilistic classification methods. The differing class variances also suggest that the *k*-means clustering algorithm may be less suitable when applied to food items since the *k*-means algorithm separates items into groups of equal variance.

The classes obtained from the GMMs provided greater detail when compared to the groupings identified in a previous study that applied principal component analysis to the SAFCDB to identify compositionally similar food items (7). While the PCA groupings identified the 'meat and meat products' food category as a whole being

TABLE 5 Food items within the identified fatty acid classes.

Class	Class 1	Class 2
Class description	Low content	High content
Food group		
Cereals and cereal products	Maize, wheat, barley	Baked goods
Vegetables	All vegetables	
Fruit	All fruit	
Legumes and legume products	Beans, lentils	
Nuts and seeds		Nuts and seeds
Milk and milk products	Skim milk, fat-free cottage cheese	Other milk and milk products
Eggs		Eggs
Meat and meat products		All meat and meat products
Fish and seafood	Tuna, crab, haddock, low-fat fish	Caviar, high-fat fish
Fats and oils		Fats and oils, fried foods
Sugar, syrups and sweets	Molasses	Chocolate, icing
Soups, sauces, seasonings and flavorings		Sauces
Beverages		Milk beverages
Infant and paediatric feeds and foods		Infant feeds
Therapeutic/special/diet products		Some therapeutic powders

high in animal protein, the identified GMM classes based on animal protein was able to further separate this food category into three subclasses. Specific food items, such as red meat and oily/fatty fish were identified to be high in animal protein. Similarly, while the PCA groupings identified leaves such as lambs quarters and sow thistle leaves as containing a high vitamin A content, our analysis has shown that only amaranth leaves exhibit a higher than average vitamin A content. Thus, within broad food categories, our classification provides detailed subcategories with a focus on the individual food items. In addition, regarding the vitamin A content of leaves other than amaranth leaves, other leaves had an approximately equal probability of belonging to either the moderate content class or the high content class. This finding emphasizes the uncertain thresholds between clusters in food composition data and is possible to quantify through evaluating the class membership probabilities, available with probabilistic classification and is an advantage over PCA.

Although there was a discernible link between the identified classes and the SAFCDB food groups, the identified classes included food items from various SAFCDB food groups. This suggests that compositional similarity cannot be completely described by traditional food groups such as grains, vegetables and dairy, which was a similar finding in other studies (25–27). This also supports the nutritional practice of disease specific food exchange lists in diet therapy, such as renal exchange lists, that are informed by the nutrients of concern. Individuals with kidney disease are advised to follow the renal diet (8) which limits particular nutrients, such as protein, sodium, phosphate and potassium. Our analysis classified food items such as rice, pasta, marrow and peach and pear nectars as low potassium foods. Food items such as potatoes, dried raw vegetables, some nuts, milk powder, fish biltong and molasses were found to have a high potassium content. This is consistent with the recommended list of foods to consume and avoid when controlling potassium intake according to the renal diet (28).

Limiting sodium is also necessary for both kidney disease and hypertension (29). Foods identified as having the highest sodium content were bread, potato crisps, canned vegetables, processed meat and instant soups which is consistent with the recommended foods to avoid (30). Foods with the lowest sodium content were mostly fruit and vegetables with some fruit and vegetables containing less sodium than others, an aspect which was easily identifiable from our results and consistent with the recommendations of the DASH diet (31). Since this is data from before the current salt regulations (32) were implemented, future work could explore the impact of the salt regulations on the sodium content of foods using an updated version of the SAFCDB.

Carbohydrate content is also often monitored as part of a healthy diet to control type 2 diabetes and metabolic syndrome (33). Foods identified in the high available carbohydrate class, such as baked goods, starchy vegetables and sweets, are often considered as a source of low-quality carbohydrates (34, 35) and individuals can use this ranking as a guide on foods to monitor when following a low-carbohydrate diet. The foods found in our carbohydrate classes align with the classification of foods by GI (36). Low GI foods such as non-starchy vegetables, fruit and protein-rich foods were grouped together as foods with a low carbohydrate content. In addition, milling was a common processing method in the high available carbohydrate content group and this is known to increase the glycaemic index (GI) of certain foods (finer food particles increase absorption contributing to a higher GI) (29). Using our results, similar food lists can be developed for anaemia and hemochromatosis (requires the control of iron intake), Wilson's disease (requires the control of copper intake), coronary heart disease (requires the control of fatty acid and dietary cholesterol intake), and gut health (impacted by total fibre intake). Using GMM to classify food items for the development of food lists provides objective rankings of food items while also accounting for the structure of food composition data. Since GMM is a data-driven method, the process of ranking food items using this method reduces the need for manual categorization and food groups can easily be reassessed with the addition of more or updated data.

Food composition data has similar methodological challenges to that of food consumption data such as right-skewness and a large proportion of food items having zero content of a particular nutrient (37). Using a log-transform before applying the GMM adjusted for the skewness and enabled the patterns of each nutrient distribution to become discernible. This also revealed that the distribution of nutrients could be modeled as mixture of Gaussians and foods with a zero nutrient content could be easily excluded from the univariate analysis. This is a desirable property since we are only interested in classifying foods known to have a particular nutrient. The separation of zero nutrient content foods from foods known to have the nutrient was also advocated for by Khan (4). In addition, classes capturing food items with either an extremely low nutrient content or an extremely high nutrient content were also identified. This facilitates outlier

TABLE 6 Food items within the identified iron classes.

Class	Class 1	Class 2	Class 3
Class description	Low content	High content	Moderate content
Food group			
Cereals and cereal products	Super/special soft maize meal porridge (unfortified), low-fat milk and whole milk pudding (blancmange, instant)	Wheat flour, oats, semolina, baked goods, pasta dishes, raw maize meal, bread	Stiff and crumbly maize meal porridge, fortified soft maize meal porridge, rice, rice flour
Vegetables	Squash, tomato, asparagus	Leaves, dehydrated raw vegetables, peas, spinach, broccoli, Brussels sprouts, green beans, beetroot, baby marrow squash, mushroom, tomato juice	Brinjal, cabbage, sweetcorn, squash, sweet potato, tomato, onion, potato, parsnip, cauliflower, carrots
Fruit	Apple, lemon juice, grapefruit, naartjie, pawpaw, watermelon, cherries, nectarine, canned peaches, canned pears, rhubarb	Dried fruit, prune juice	Apricot, avocado, guava, canned fruit, figs, pears, prunes, granadilla, dates, grapes, peaches, plums, pineapple, fruits nectars (apricot, pear), fruit juices (grapefruit, pineapple, grape)
Legumes and legume products		Legumes and legume products	
Nuts and seeds		Nuts and seeds	
Milk and milk products	Milk, yoghurt, custard, reconstituted skim milk powder	Milk powder with added iron, cheese (feta, cottage, Gouda)	Milk powders, evaporated milk, custard
Eggs		Eggs	
Meat and meat products		Meat and meat products	Chicken (white meat), veal, chicken stew
Fish and seafood		Anchovy, oyster, sardines, mussels, tuna, fried fish, shrimp/prawn	Low-fat fish, shrimp/prawn, crab, salmon, sole
Fats and oils	Vegetable oil, cream, French salad dressing, butter ghee, butter and hard margarine (mixed), coconut oil, soybean oil	Peanut butter, canned cream	Olive oil, salad dressing
Sugar, syrups and sweets	Icing, sugar	Chocolate, jam/marmalade, molasses	Honey
Soups, sauces, seasonings and flavorings		Curry sauce, soups with meat and vegetables	Sauces
Beverages	Malted milk beverages, coffee, tea	Malted milk powder, drinking chocolate powder	Malted milk beverages, drinking chocolate powder
Infant and paediatric feeds and foods		Infant feeds	
Therapeutic/special/diet products		Therapeutic powders	
Miscellaneous	Spirits, liqueur, vinegar	Baking powder	Wine, sherry

detection which could represent foods with an actual extreme nutrient content, foods with added components such as added sugar or added salt, or foods with erroneous values for a specific nutrient. Using extreme values to identify errors was also previously investigated (38).

Overall, the class validity indices indicated that application of the GMM resulted in good classification. Classes with a substantial overlap between them were shown to have poorer internal validity scores than classes that were more separable. Since internal indices focus on separability as one of the criteria for class validity, these indices are unsuitable when the data displays mixed class membership. Further research is needed on appropriate internal class validity indices in the presence of overlapping classes obtained through GMM clustering and on the stability of the identified classes.

The univariate GMM provided useful results but multiple nutrients are present in food items and thus multiple nutrients are consumed simultaneously. While it is important to know which foods may have a relatively low or high nutrient content, consuming a food high in particular nutrient may also unknowingly increase the intake of other nutrients. Thus, it is important to consider the multivariate GMM as future work. However, this can be challenging in the case of high-dimensional data such as food composition data. GMMs often fit extra classes to capture the outliers and can result in poor data fit. Future work could investigate the mixture of multivariate t-distributions (39) to account for the long tails and outliers seen in our data and incorporating the structural zeroes into the clustering algorithm using a zero-inflation model could also be explored (40). Alternatively, Lo and Gottardo (41) proposed a multivariate t-distribution with Box-Cox transformation that could simultaneously address data transformation and outlier detection which are characteristics pertinent to the analysis food composition data.

TABLE 7 Food items within the identified animal protein classes.

Class	Class 1	Class 2	Class 3	Class 4
Class description	Low content	Moderately-low content	Moderately-high content	High content
Food group				
Cereals and cereal products	Rice cooked with margarine, pastry/crust made with margarine	Puddings, baked goods, pasta dishes	Tuna pie	
Vegetables		Vegetables coated in batter		
Legumes and legume products	Beans cooked with margarine	Lentils with egg		
Milk and milk products		Milk, yoghurt	Cottage cheese, feta	Milk powder, cheese
Eggs		Scrambled egg, soufflé	Raw egg, fried egg	
Meat and meat products		Commercial meat pies, pork/beef sandwich spread, ham and tongue loaf, offal	Chicken, ham, meat stews/curries, duck, Frankfurters, sausage, pâté, luncheon meat, corned beef	Beef, pork, veal, mutton, turkey, goose, pork sausage, salami
Fish and seafood		Fish biltong, oyster, low-fat fish cakes, fish fingers	Crab, fatty fish, baked/crumbed fish, sole, rollmop/pickled herring	Anchovy, tuna, fish, haddock, sardines, caviar, kipper, mussels, salmon, shrimp/prawn
Fats and oils	Butter ghee	Cream, butter, margarine, homemade salad dressing		
Sugar, syrups and sweets	Icing, dark chocolate	Chocolate, jelly, cottage cheese icing		
Soups, sauces, seasonings and flavorings		Sauces, soups with beef		
Beverages		Milk beverages, eggnog		
Infant and paediatric feeds and foods		Reconstituted infant feeds	Whey-predominant infant feed powder	
Therapeutic/special/diet products		Reconstituted therapeutic products		Some therapeutic powders
Miscellaneous		Liqueur with cream		

In conclusion, this study has explored the application of univariate Gaussian mixture models to examine the classification of food items within the South African Food Composition Database. The identified classes exhibited overlap, supporting the use of probabilistic classification methods to account for the uncertainty of nutrient thresholds between classes. Classifying food items by moisture, animal protein, fatty acid, available carbohydrate, total fibre, sodium, vitamin A, thiamin and riboflavin content produced classes with differing means and estimates of variability and could be clearly ranked on a low to high nutrient content scale. Our results highlight that classifications within the broader, traditional food groups exist and our method focuses on identifying the individual food items within these subclasses. The results can be used to inform the development of nutrient profiling indices, allowed food lists and food-based dietary guidelines. The identified classes could also be incorporated into food composition databases to provide an additional level of classification and understanding of food items, thus promoting nutrition education for the user. Since we included processed and manufactured food items in our analysis, manufacturers can use these findings to inform product formulation as well.

Data availability statement

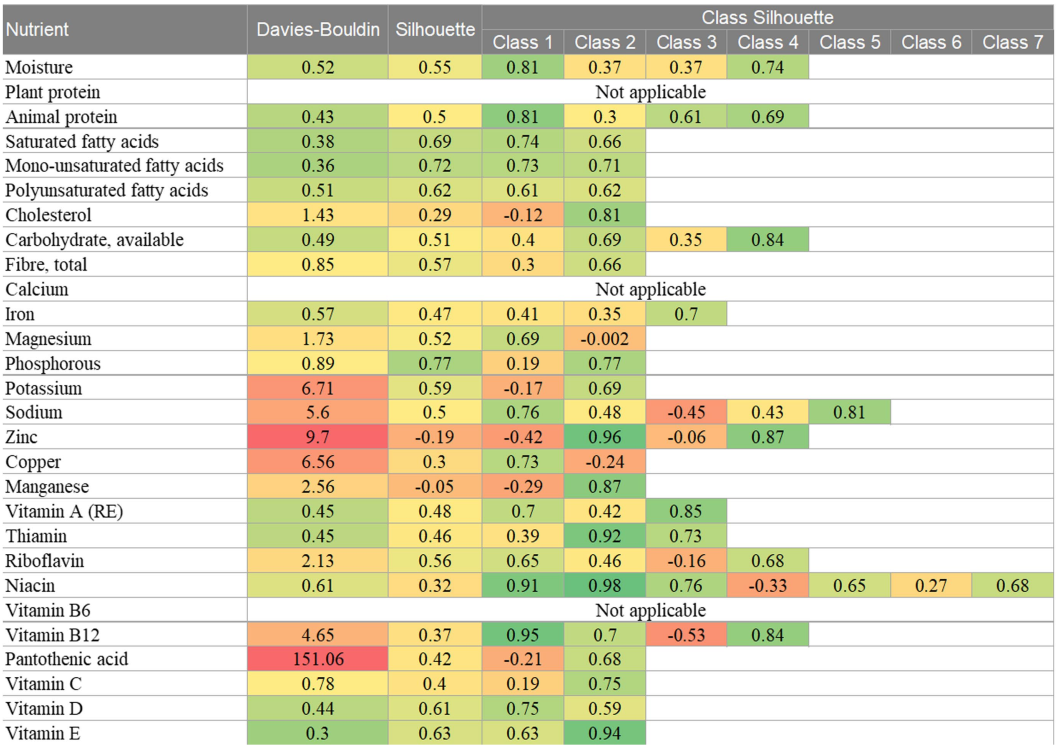
The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YB and SM contributed to the conception and design of the study. AG provided access to the database. YB performed the statistical analysis and wrote the first draft of the manuscript. SM, HM, and AG provided supervision. All authors contributed to manuscript revision and approved the submitted version.

Funding

YB and AG time on this research was funded by the South African Medical Research Council.



Nutrient	Davies-Bouldin	Silhouette	Class Silhouette						
			Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Moisture	0.52	0.55	0.81	0.37	0.37	0.74			
Plant protein	Not applicable								
Animal protein	0.43	0.5	0.81	0.3	0.61	0.69			
Saturated fatty acids	0.38	0.69	0.74	0.66					
Mono-unsaturated fatty acids	0.36	0.72	0.73	0.71					
Polyunsaturated fatty acids	0.51	0.62	0.61	0.62					
Cholesterol	1.43	0.29	-0.12	0.81					
Carbohydrate, available	0.49	0.51	0.4	0.69	0.35	0.84			
Fibre, total	0.85	0.57	0.3	0.66					
Calcium	Not applicable								
Iron	0.57	0.47	0.41	0.35	0.7				
Magnesium	1.73	0.52	0.69	-0.002					
Phosphorous	0.89	0.77	0.19	0.77					
Potassium	6.71	0.59	-0.17	0.69					
Sodium	5.6	0.5	0.76	0.48	-0.45	0.43	0.81		
Zinc	9.7	-0.19	-0.42	0.96	-0.06	0.87			
Copper	6.56	0.3	0.73	-0.24					
Manganese	2.56	-0.05	-0.29	0.87					
Vitamin A (RE)	0.45	0.48	0.7	0.42	0.85				
Thiamin	0.45	0.46	0.39	0.92	0.73				
Riboflavin	2.13	0.56	0.65	0.46	-0.16	0.68			
Niacin	0.61	0.32	0.91	0.98	0.76	-0.33	0.65	0.27	0.68
Vitamin B6	Not applicable								
Vitamin B12	4.65	0.37	0.95	0.7	-0.53	0.84			
Pantothenic acid	151.06	0.42	-0.21	0.68					
Vitamin C	0.78	0.4	0.19	0.75					
Vitamin D	0.44	0.61	0.75	0.59					
Vitamin E	0.3	0.63	0.63	0.94					

FIGURE 4
Internal class validity indices for the univariate GMM classifications. Key: green = good score, yellow = moderate score, red = poor score.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2023.1186221/full#supplementary-material

References

1. Jacobs DR Jr, Steffen LM. Nutrients, foods, and dietary patterns as exposures in research: a framework for food synergy. *Am J Clin Nutr*. (2003) 78:508s–13s. doi: 10.1093/ajcn/78.3.508S

2. Tapsell LC, Neale EP, Probst Y. Dietary patterns and cardiovascular disease: insights and challenges for considering food groups and nutrient sources. *Curr Atheroscler Rep*. (2019) 21:9. doi: 10.1007/s11883-019-0770-1

3. Sandström B. A framework for food-based dietary guidelines in the European Union. *Public Health Nutr*. (2001) 4:293–305. doi: 10.1017/s1368980001001550

4. Khan AS. Processes in ranking nutrients of foods in a food data base. *Nutr Health*. (1996) 11:59–72. doi: 10.1177/026010609601100105

5. Nikitina MA, Chernukha IM, Uzakov YM, Nurmukhanbetova DE. Cluster analysis for databases typologization characteristics. *News Natl Acad Sci Repub Kaz Ser Geol Tech Sci*. (2021) 2:114–21. doi: 10.32014/2021.2518-170X.42

6. Balakrishna Y, Manda S, Mwambi H, van Graan A. Statistical methods for the analysis of food composition databases: a review. *Nutrients*. (2022) 14:2193. doi: 10.3390/nu14112193

7. Balakrishna Y, Manda S, Mwambi H, van Graan A. Identifying nutrient patterns in south African foods to support national nutrition guidelines and policies. *Nutrients*. (2021) 13:3194. doi: 10.3390/nu13093194

8. Hershey K. Renal diet. *Nurs Clin N Am*. (2018) 53:481–9. doi: 10.1016/j.cnur.2018.05.005

9. Meng Y, Bai H, Wang S, Li Z, Wang Q, Chen L. Efficacy of low carbohydrate diet for type 2 diabetes mellitus management: a systematic review and meta-analysis of randomized controlled trials. *Diabetes Res Clin Pract*. (2017) 131:124–31. doi: 10.1016/j.diabres.2017.07.006

10. Fahey MT, Ferrari P, Slimani N, Vermunt JK, White IR, Hoffmann K, et al. Identifying dietary patterns using a normal mixture model: application to the epic study. *J Epidemiol Community Health*. (2012) 66:89–94. doi: 10.1136/jech.2009.103408

11. Fahey MT, Thane CW, Bramwell GD, Coward WA. Conditional Gaussian mixture modelling for dietary pattern analysis. *J R Stat Soc A Stat Soc*. (2007) 170:149–66. doi: 10.1111/j.1467-985X.2006.00452.x

12. Windham CT, Windham MP, Wyse BW, Hansen RG. Cluster-analysis to improve food classification within commodity groups. *J Am Diet Assoc*. (1985) 85:1306–14. doi: 10.1016/S0002-8223(21)03795-0

13. Luke JN, Schmidt DF, Ritte R, O'Dea K, Brown A, Piers LS, et al. Nutritional predictors of chronic disease in a central Australian aboriginal cohort: a multi-mixture modelling analysis. *Nutr Metab Cardiovasc Dis*. (2016) 26:162–8. doi: 10.1016/j.numecd.2015.11.009

14. Greenfield H, Southgate DAT. *Food composition data. Production management and use. 2nd* ed. Rome, Italy: Food and Agriculture Organization of the United Nations (2003).

15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the Em algorithm. *J R Stat Soc B*. (1977) 39:1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

16. Abbi R, El-Darzi E, Vasilakis C, Millard P, editors. Analysis of stopping criteria for the Em algorithm in the context of patient grouping according to length of stay. *Vol. 3*. Proceedings of the 4th International IEEE Conference on Intelligent Systems IS'08. Varna, Bulgaria (2008), 9–14.

17. Scrucca L, Fop M, Murphy TB, Raftery AE. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The {R}Journal*. (2016) 8:289–317. doi: 10.32614/RJ-2016-021

18. Benaglia T, Chauveau D, Hunter DR, Young D. Mixtools: an R package for analyzing finite mixture models. *J Stat Softw*. (2009) 32:1–29. doi: 10.18637/jss.v032.i06

19. Schwarz G. Estimating the dimension of a model. *Ann Stat*. (1978) 6:461–4. doi: 10.1214/aos/1176344136

20. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. (1979) PAMI-1:224–7. doi: 10.1109/TPAMI.1979.4766909

21. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. (1987) 20:53–65. doi: 10.1016/0377-0427(87)90125-7

22. Saranya S, Poonguzhali S, Karunakaran S. Gaussian mixture model based clustering of manual muscle testing grades using surface electromyogram signals. *Phys Eng Sci Med*. (2020) 43:837–47. doi: 10.1007/s13246-020-00880-5

23. Elmadfa I, Meyer AL. Importance of food composition data to nutrition and public health. *Eur J Clin Nutr*. (2010) 64:S4–7. doi: 10.1038/ejcn.2010.202

24. Russolillo-Femenías G, Menal-Puey S, Martínez JA, Marques-Lopes I. A practical approach to the management of micronutrients and other nutrients of concern in food exchange lists for meal planning. *J Acad Nutr Diet*. (2018) 118:2029–41. doi: 10.1016/j.jand.2017.07.020

25. Atsaam DD, Oyelere S, Balogun OS, Wario R, Blamah NV. K-means cluster analysis of the west African species of cereals based on nutritional value composition. *African J Food Agric Nutr*. (2021) 21:17195–212. doi: 10.18697/ajfand.96.19775

26. SBR DP, Giuntini EB, Grande F, de Menezes EW. Techniques to evaluate changes in the nutritional profile of food products. *J Food Compost Anal*. (2016) 53:1–6. doi: 10.1016/j.jfca.2016.08.007

27. Phanich M, Pholkul P, Phimoltares S, eds. Food recommendation system using clustering analysis for diabetic patients. 2010 international conference on information science and applications; 21–23 April 2010 (2010).

28. National Kidney Foundation. *Potassium and your Ckd diet [11 September 2023]*. Available from: https://www.kidney.org/atoz/content/potassium

29. Grillo A, Salvi L, Coruzzi P, Salvi P, Parati G. Sodium intake and hypertension. *Nutrients*. (2019) 11:1970. doi: 10.3390/nu11091970

30. National Kidney Foundation. *Sodium and your Ckd diet; how to spice up your cooking* (2023). Available from: https://www.kidney.org/atoz/content/sodiumckd

31. National Heart Lung and Blood Institute. *Dash eating plan: U.S. Department of Health and Human Services* (2023). Available from: https://www.nhlbi.nih.gov/education/dash-eating-plan

32. South African Government. *Government gazette: No. R. 214 foodstuffs, cosmetics and disinfectants act, 1972 (act 54 of 1972) regulations relating to the reduction of sodium in certain foodstuffs and related matters*. (2013).

33. Via MA, Mechanick JI. Nutrition in type 2 diabetes and the metabolic syndrome. *Med Clin N Am*. (2016) 100:1285–302. doi: 10.1016/j.mcna.2016.06.009

34. Drewnowski A, Maillot M, Papanikolaou Y, Jones JM, Rodriguez J, Slavin J, et al. A new carbohydrate food quality scoring system to reflect dietary guidelines: an expert panel report. *Nutrients*. (2022) 14:1485. doi: 10.3390/nu14071485

35. Hou W, Gao J, Jiang W, Wei W, Wu H, Zhang Y, et al. Meal timing of subtypes of macronutrients consumption with cardiovascular diseases: Nhanes, 2003 to 2016. *J Clin Endocrinol Metabol*. (2021) 106:e2480–90. doi: 10.1210/clinem/dgab288

36. Diabetes Canada. *Glycemic index food guide – diabetes Canada* (2023). Available at: https://guidelines.diabetes.ca/docs/patient-resources/glycemic-index-food-guide.pdf

37. Carriquiry AL. Understanding and assessing nutrition. *Annu Rev Stat Appl*. (2017) 4:123–46. doi: 10.1146/annurev-statistics-041715-033615

38. Chu C-M, Lee M-S, Hsu Y-H, Yu H-L, Wu T-Y, Chang S-C, et al. Quality assurance with an informatics auditing process for food composition tables. *J Food Compost Anal*. (2009) 22:718–27. doi: 10.1016/j.jfca.2009.03.005

39. Peel D, McLachlan GJ. Robust mixture modelling using the T distribution. *Stat Comput*. (2000) 10:339–48. doi: 10.1023/A:1008981510081

40. Thanataveerat A. *Clustering algorithm for zero-inflated data*. New York, NY: Columbia University (2020).

41. Lo K, Gottardo R. Flexible mixture modeling via the multivariate *t* distribution with the box-cox transformation: an alternative to the skew-*t* distribution. *Stat Comput*. (2012) 22:33–52. doi: 10.1007/s11222-010-9204-1

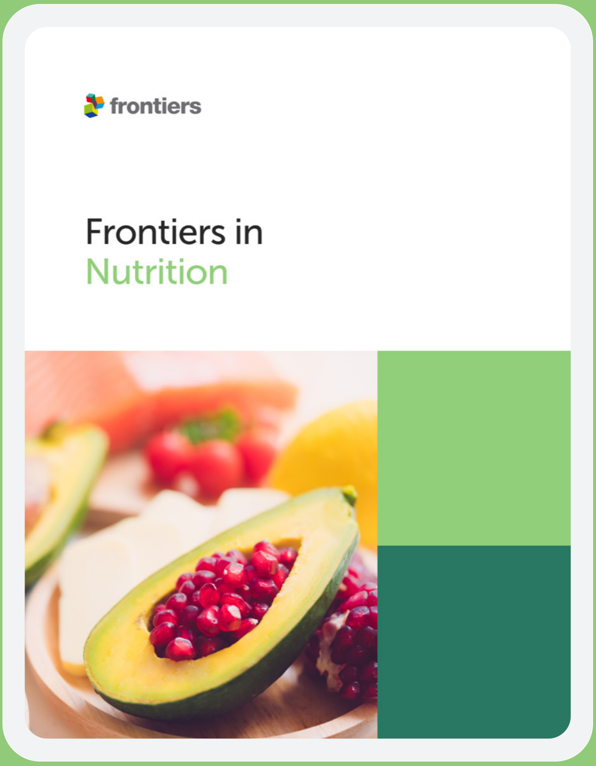
Frontiers in Nutrition

Explores what and how we eat in the context of health, sustainability and 21st century food science

A multidisciplinary journal that integrates research on dietary behavior, agronomy and 21st century food science with a focus on human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact