Accepted manuscript

Building nutritionally meaningful classification for grocery product groups: the LoCard Food Classification process

Kanerva N [1*], Kinnunen S [1], Nevalainen J [2,1], Vepsäläinen H [1], Fogelholm M [1], Saarijärvi H [3], Meinilä J [1], Erkkola M [1].

[1]Department of Food and Nutrition, University of Helsinki, PO Box 66, 00014 Helsinki, Finland.

[2]Faculty of Social Sciences (Health Sciences), Tampere University, Kanslerinrinne 1, 33100 Tampere, Finland.

[3]Faculty of Management and Business, Tampere University, Kanslerinrinne 1, 33100 Tampere, Finland.

***Corresponding author:** Noora Kanerva, Department of Food and Nutrition, University of Helsinki, PO Box 66, 00014 Helsinki, Finland. Mobile: +358505243013, email: noora.kanerva@helsinki.fi

Short title: LoCard Food Classification process

This peer-reviewed article has been accepted for publication but not yet copyedited or typeset, and so may be subject to change during the production process. The article is considered published and may be cited using its DOI

10.1017/S0007114524000710

The British Journal of Nutrition is published by Cambridge University Press on behalf of The Nutrition Society

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

Abstract

Analysing customer loyalty card data is a novel method for assessing nutritional quality and changes in a population's food consumption. However, prior to its use, the thousands of grocery products available in stores must be reclassified from the retailer's original hierarchical structure into a structure that is suitable for the use of nutrition and health research. We created LoCard Food Classification (LCFC) and examined how it reflects the nutritional quality of the grocery product groups. Nutritional quality was considered the main criterion guiding the reclassification of the 3574 grocery product groups. Information on the main ingredient of the product group, purpose of use, and carbon footprint were also used at the more granular levels of LCFC. The main challenge in the reclassification was a lack of detailed information on the type of products included in each group, and that some of the groups included products that have opposite health effects. The final LCFC has four hierarchical levels and it is openly available online. After reclassification, the product groups were linked with the Finnish food composition database, and the nutrient profile was assessed by calculating the Nutrient Rich Food Index (NRFI) for each product group. Standard deviation in NRFI decreased from 0.21 of the least granular level to 0.08 of the most granular level of LCFC indicating that the most granular level of LCFC has more homogeneous nutritional quality. Studies that apply LCFC to examine loyalty card data with health and environmental outcomes are needed to further demonstrate its validity.

Keywords: food retail, food purchases, consumption, nutritional quality, food group, carbon footprint

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

Introduction

In nutritional epidemiology, hierarchical classification of food items into broader categories plays a critical role when examining associations with food consumption and health (1). For these purposes, international classification systems such as the European Food Safety Authority's FoodEx2 classification (2) or FAO/INFOODS (3) have been developed to improve the availability and reliability of dietary data obtained from traditional nutrition surveys. These classification systems also make the comparison and reproducibility of the results between different countries more feasible and easier and allow researchers to harmonise their data and food composition databases in a transparent way (1,4).

The use of grocery food purchase data by using food retailers' customer loyalty card data in academic research has gained increasing attention in recent years (5-8). Grocery purchase data are about what, when, where and at what price food has been bought, with or without a personal identifier tag (7). Customer loyalty card data always include at least some personal data of the person who made the purchase. Hence, loyalty card data provide a unique opportunity to obtain vast amounts of detailed data automatically and objectively over time on different card holders' grocery purchases (9). These data can be used to monitor the nutritional quality of food purchases (9,10) that can lead to health policy actions (for example, a sugar tax) (11) with the purpose of steering food consumption toward healthier options that could eventually improve public health nutrition (12,13). Moreover, the data can be used for monitoring and evaluating policies, as well as dietary, environmental, social, and economic sustainability (6,14). Present and future diets should reduce global health risks, like cardiovascular diseases, type 2 diabetes, and cancer, and reduce the environmental impact of the food systems (15). For evidence-based decision-making, robust and timely information on population-level health and environmental behaviours, obtainable also from loyalty card data, is essential (16).

Food retailers commonly use classification systems that are based on logistics or product placement on the shelf, and they do not essentially reflect products' nutritional profiles (17,18). Therefore, to harness the full potential of customer loyalty card data for scientific research, thousands of grocery products should be reclassified into categories that are meaningful for nutrition and health research. Although the purchase of single foods can be used for research

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

purposes, the research objectives (e.g., comparing food purchase data to individual's food consumption measured using traditional dietary assessment methods) may necessitate working on less granular levels (9). For this, we need hierarchical structures for the grocery products using a suitable theory-based approach (19,20).

Only a few of the classification methods used for groceries have been transparently described, such as the Convenience Food Classification Scheme (CFCS) (21) and the NOVA classification (22). CFCS includes three convenience categories based on the degree of processing, culinary skills required to transform the bought food into a meal, the time needed for meal preparation, the time needed after consumption (e.g., cleaning up and washing dishes) and the context in which a food or meal is consumed (e.g., snack or ready-made meal) (21). The NOVA classification assigns food products to groups based on the degree of processing: Group 1 - Unprocessed or minimally processed foods. Group 2 - Processed culinary ingredients. Group 3 - Processed foods. Group 4 - Ultra-processed foods, i.e., industrial formulations or foods prepared by the industry, packaged, ready for consumption and with a high content of salt, sugars, and fat (22). Although nutritional quality was associated with the above classification criteria (convenience and degree of processing), nutrition as such was not the starting point. Moreover, even though there are suitable tools available such as the Nutrient Rich Food Index (NRFI) (23) and the Grocery Purchase Quality Index-2016 (24), which can be used to evaluate whether a classification eventually succeeds in reflecting the nutritional quality of the grocery purchases, this type of evaluation is rarely done (25,26). We argue that a clear, explicit, openly available, and critically evaluated grocery product classification is needed to advance the research on health and environmental impacts of grocery purchases.

We have received a large-scale (n=47,066 card holders) longitudinal customer loyalty card (LoCard) data set from the largest food retailer in Finland (market share about 46%) (27). Since the original product grouping used by the retailer was designed for retail purposes, our first challenge was to design and compile a meaningful product grouping appropriate for nutrition and health research. Thus, the purpose of this paper is to compile, describe and test a reclassification of products suitable for nutrition and health research purposes (labelled as LoCard Food Classification, hereafter LCFC) and to make it openly accessible. To achieve

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

this, we demonstrate how the reclassification process was conducted and illustrate the feasibility by examining variation in nutrition quality within chosen food groups.

Methods

Hierarchy of the food retailer's original grocery product groups

The food retailer's original hierarchical structure included 3574 grocery product groups. This study describes the process of how these groups were reclassified in LCFC. Since the purpose of the LCFC was to serve nutrition and health research, a key guiding principle in the development was to design a group hierarchy that reduces variation in nutritional quality towards higher granularity. Neither this process, nor the analyses for this paper, involved any use of the customer loyalty card data or other personal data from human participants. Further, the process did not use sales data.

The retailer's product group hierarchy was based on logistics or product placement on the shelf. Consequently, the retailer's hierarchy on its most granular level included information such as flavor, form of storage and package sizing (e.g., "citrus lemonades, canned, stored in fridge", "cola-flavored drinks, 4 bottles, stored in room temperature"), and packing of the products (e.g., "cream cheese, pre-sliced" or "whole breads, pre-sliced"). From nutritional and health perspective this information was naturally irrelevant and could be excluded.

For most of the product groups, the name of the product group gave adequate information to understand the nutrient quality of that product group for our research purposes. For example, even though we did not know the brand names of beverages the product group name included information if a beverage had added sugar (e.g., "cola-flavored drink, no sugar, canned"). Coca-Cola Zero and Pepsi Max are nutritionally identical (both have zero added sugar). Similarly regular Pepsi, Coke, Fanta, and Sprite are all sugar-sweetened soft drinks, which form a generic, but well-defined and nutritionally homogenous class with the essential nutrient being added sugar 10% of weight.

Our main challenge was to reclassify grocery product groups when the nutritional quality of the group was not obvious from the name, especially when the main ingredient was unclear. Examples of such groups include 'Other meat', 'Ready-made salads', 'Hamburgers' or

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

'Pizzas'. This was further complicated by not having any food-item level information on the product groups due to business secrecy. Fortunately, we received a sample of 26 000 food items including their product name, EAN code, package size, and their original product group. This information aided us in reclassifying most of the foods. Additionally, we used the retailer's online food purchasing service, which provided information about the food items, such as product name, within the grocery product groups.

Principles and selected examples of LCFC hierarchy

We built four-level hierarchical classification of product groups in the LCFC. LCFC Class 1 (LCFC-1) had the lowest granularity and was subsequently divided into subclasses of higher granularity starting with LCFC-2, followed by LCFC-3 and, LCFC-4, which had the highest granularity. An example of LCFC hierarchy is given in Figure 1. The whole LCFC is openly available at https://doi.org/10.5281/zenodo.7781352.

In LCFC-1, the grocery product groups were reclassified into 38 groups based on *healthiness* [28] and *main ingredients* [29] (see Table 1). Our approach to "healthiness" (as a proxy for nutritional quality) was based on the Nordic Nutrition Recommendations [28]. Food groups with a recommendation to limit the intake, including those high in sugar, saturated fat or salt, were to be separated from foods recommended to be included in the diet. Such foods include fruits, vegetables, and berries; whole grain cereal products; low-fat dairy; fish and seafood; plant-based meat alternatives; nuts and seeds; oils and margarine.

Compared to the retailer's original product group hierarchy, for example, the LCFC-1 separates plant-based protein products into their own main class from the meat product group where they were placed in the food retailer's original grouping. This LCFC-1 group was named 'Plant protein products' and included processed legume products such as those mimicking meat, as well as unprocessed lentils, peas, and beans. Within LCFC-1, we also formed a separate group for plant-based dairy-like products including, for instance, soy and oat milk.

Classification to LCFC-2 was dictated by *the type of foods in the product group*, *purpose of use* of the product groups and *food culture*. This means, for example, that the LFCF-1 group

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

'Milk and dairy products' was further classified to 'Cheeses', 'Ice creams' and 'Liquid milk products' (Table 1). Another example would be classification of edible fats into 'Butter and fat blends', 'Margarine', 'Vegetable oils' and 'Cooking fat'. The *purpose of use* of the product groups was also considered in the reclassification at the LCFC2- level. For example, the purpose of use for nuts may vary based on whether they are plain nuts that are often used in salads, chocolate-coated nuts which can be used as sweets or salted nuts which may resemble the use of other salty snacks. Therefore, plain nuts were classified under 'Dried fruits and nuts' whereas chocolate-coated nuts were classified under 'Sweets and chocolates' and salted nuts under 'Snacks' at the LCFC-1 level.

Classification of traditional Finnish ready-made pea soup is an example of how we considered the national *food culture* in LCFC-2. Namely, the most common pea soup contains small amounts of meat (< 5%), but the green pea is the main ingredient. Since pea soup is traditionally served on Thursdays in lunch restaurants, it is also one of the main contributors to the consumption of legumes among the Finnish population. Therefore, we decided to classify it under 'Peas, beans, lentils and soya' at the LCFC-2 level, which is under the broader 'Plant protein products' category at the LCFC1- level – not as a red-meat product.

At its most granular levels (LCFC-3 and LCFC-4), *nutritional quality* and *carbon footprint* were used to guide the classification when reasonable (Table 1). For breads and breakfast cereals, milk and dairy products and alcoholic beverages, we used their fibre, fat, and alcohol content to guide the classification process at the LCFC-3 level. To be classified as high-fibre cereal, we used a cut-off of 6% of fibre, as defined by the European Food Safety Authority [(30)]. For milk, we used 1% and 3% cut-offs to separate skimmed, semi-skimmed and whole milk. For other dairy products, low fat was defined as <1% of fat. Alcoholic beverages were classified based on the following cut-offs for their alcohol content, based on the Finnish alcohol legislation: <=1.2%, 1.3–2.8%, 2.9–3.5%, 3.6–4.7% and 4.8–5.5% [(31)]. For some foods, such as cheeses, it would be desirable to use a cut-off based on their fat content, but this would have been possible for only some of the cheeses due to the retailer's grocery product grouping. For example, the retailer grouped most of the cheeses by package size, processing, and flavouring. In addition to nutritional content, we used carbon footprint as another basis for classification when within-food-group variation in the carbon footprint of

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

the foods was large. In other words, if nutrition categorization was not detailed enough to differentiate between foods with different magnitude of carbon footprint, the categorization was more detailed. For example, because the average carbon footprint of beef is much greater than that of pork [32], in LCFC-4 we classified different types of red meat separately. The details of assigning carbon footprints have been described in prior literature [14].

For some of the retailer's grocery product groups, the LCFC remained a compromise due to the lack of detailed information on food-item level. For example, we classified pizzas under cereals and bakery products since they were originally categorised by the retailer based on whether they were fresh or frozen, or if they had thick or thin crust, but not by whether they were, for example, vegetarian or meat pizzas. Thus, we considered the main ingredient in the pizzas to be wheat (cereals). Other examples of such groups include'other canned foods', 'warm dish service', and 'other ready-made soups'. Eventually, there were only 38 grocery product groups (0.01% of all the product groups) left unclassified under 'Miscellaneous' at the LCFC-1 level.

Last, we added tobacco products as a group of its own. It is an important product group to examine along with food and alcohol products regarding health.

Examining the LCFC hierarchy in terms of nutrition quality

The retailer's grocery product groups were linked with their respective nutrient content by using the Finnish Food Composition Database Fineli$^{(R)}$, version 20 (www.fineli.fi) [29]. Fineli's open database includes 4232 food items and dishes, 1370 basic ingredients, and 55 nutrients. For each product group, we selected a food item from Fineli that best represented the product group. In most of the cases, the name of the grocery product group had enough information for us to select the food from Fineli (e.g., pineapple, oat milk, ketchup, etc.). Otherwise, we exploited the small product item-level dataset received from the retailer and decided which food in Fineli describes the group the best. If the Fineli database did not contain a food that would have described the product group sufficiently, food composition databases of other countries (e.g., Swedish and US databases) were exploited.

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

Out of the 3574 grocery product groups, 3368 were linked with nutrient content. Tobacco products and vitamin and mineral supplements (122 groups), spices and condiments (25 groups), and miscellaneous (38 groups) and 21 other product groups were left without nutrient content due to the challenge of finding a representative food in the composition databases, or the group did not include foods with relevant nutrient content.

To examine how well our hierarchical reclassification reflects the nutrient quality of the grocery product groups, we calculated a Nutrient Rich Food Index (NRFI) for each LCFC level following principles of Drewnowski et al. (33,34). NRFI is a validated method of nutrient profiling aiming to provide an overall nutrient density score based on selected nutrients (33,34). We calculated the NRFI per 100 grams of product using 11 nutrients, of which eight were regarded as positive (protein, fibre, polyunsaturated fatty acids (PUFA), calcium, iron, vitamin D, vitamin C and folate) and three as negative (saturated fatty acids (SFA), saccharose and salt) in terms of anticipated health effects. Recommended values used for the 11 nutrients were from the Finnish nutrition recommendations which are the same as for the Nordic Nutrition recommendations (28), except salt which is 5000mg in the Finnish recommendations (6000 mg in the Nordic nutrition recommendations).

Among the 11 nutrients that we included in NRFI, intakes of fibre, PUFA and vitamin D have been identified as relatively low at the population level in Finland (35). In contrast, the high intake of SFA and salt have been public health concerns for decades among the Finnish population. Intakes of iron and folate have been low among Finnish women of childbearing age. Including these nutrients in the NRFI was therefore justified. It should be noted that the NRFI does not use any weights for different nutrients (33,34). The openly available LCFC hierarchy also includes values for the 11 nutrients and NRFI values.

Recommended values for calculating the percentage of daily recommendation (DR%) are:

Protein = 90g (corresponding 15% of energy in 2400 kcal diet)

Fibre = 25g

Polyunsaturated fat (PUFA)=20g (corresponding 7.5% of energy in 2400 kcal)

Calcium (Ca) = 800mg

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

Iron (Fe) = 9mg

Vitamin D = 10µg

Vitamin C = 75mg

Folate = 300mg

Sucrose = 60g (corresponding 10% of energy in 2400 kcal)

Saturated fat = 26.7g (corresponding 10% of energy in 2400 kcal)

Salt = 5000mg

Equation 1: Positive score: (DR% protein + DR% fibre + DR% PUFA + DR% Ca + DR% Fe + DR% Vit D + DR% Vit C + DR% folate) / 8

Equation 2: Negative score: (DR% sucrose + DR% SFA + DR% salt) / 3

Equation 3 (NRFI): positive score – negative score

Boxplot figures including median NRFI values (horizontal line in the box), lower and upper quartiles (outer horizontal lines of box) and expected minimum and maximum values (end of whiskers; calculated as 1.5 * inter-quartile range) for each group at different LCFC hierarchy levels were drawn using R statistical software [(36)].

Results

Figure 2 gives an overall representation of the hierarchy of the LCFC from the retailer's grocery product groups to LCFC1–3. Similar figures for all hierarchy levels including the number of the product groups at each level can be found in Supplement material 1. The whole detailed classification structure of LCFC1–4 is available at https://doi.org/10.5281/zenodo.7781352. Most of the grocery product groups were classified only at LCFC1–2, and LCFC3–4 were used when needed, for example, to distinguish foods with different nutritional or carbon footprint profiles. Therefore, not all foods were classified at the most granular levels.

The largest groups at LCFC-1 (in terms of number of grocery product groups within the class) such as ‘Alcoholic beverages’, ‘Red and processed meat’, ‘Cereals and bakery products’ and ‘Milk and dairy products’ represented about half (1509 out of 3574) of all the retailer’s grocery product groups (Figure 2). This was mostly related to the original, highly granular classification in the grocery retailer's hierarchy. The majority of the other half came from ‘Plant protein products’, ‘Sugar-sweetened beverages’, ‘Fish and seafood’, ‘Vegetables, ‘Poultry and poultry dishes’, ‘Low-sugar beverages’, ‘Sweets and chocolates’, ‘Bottled water and mineral water’ and ‘Baby foods’ (listed from the largest to smallest group). These were the next biggest LCFC-1 groups containing 100–200 grocery product groups (1249 retailer’s grocery product groups in total). The smallest 25 LCFC-1 groups included less than 100 grocery product groups each and 816 retailer’s grocery product groups in total.

To illustrate the extent to which the LCFC succeeded in reflecting the nutritional quality of the grocery product groups, Figure 3 shows the medians and the variation in the NRFI values of the food groups at LCFC-1 level. In general, when the groups at LCFC-1 were ranked by their NRFI median value, the order was logical based on the expected nutritional quality of the groups. Grocery product groups under ‘Dried fruits and nuts’ (median=0.15), ‘Fish and seafood’ (median=0.06) and ‘Eggs’ and ‘Fruit juice’ (median=0.05) were nutrient rich according to their median NRFI values. On the contrary, ‘Edible fat’ (median=-0.11), ‘Jam and marmalade’ (median=-0.12), ‘Sweets and chocolate’ (median=-0.24) and ‘Plant-based dairy-like products’ (median=-0.25) were less nutrient rich, as indicated by the negative index (Figure 3). In general, many foods high in sugar, fat and/or salt are on the lower (left side of the x-axis) side of NRFI, while foods recommended in dietary guidelines [20] tend to be positioned higher (right side).

As seen in the boxplots, the variation in NRFI of the food groups at LCFC-1 was large, as nearly all food groups expand both sides of the zero line that separates food groups that are more nutrient rich from the less nutrient rich (Figure 3). The mean standard deviation in NRFI at LCFC-1 was 0.21. ‘Edible fat’ and ‘Sauces’ had the largest standard deviation (fat: sd=0.35 index points, number of product groups n=22; sauces: sd = 0.35, n=45), followed by

‘Meal ingredients’ (sd=0.27, n=33) and ‘Red and processed meat’ (sd=0.16, n=399) (Figure 3). Less variation was found in ‘All beverages’ (sd=0.01–0.03, n=66–452), ‘Eggs’ (sd=0.01, n= 9), ‘Mayonnaise salad’ (sd=0.03, n=16) and ‘Fruits and berries’ (sd=0.06, n=60).

As an example, we examined the NRFI values closer within ‘Cereals and bakery products’ at the LCFC-2 level (Figure 4). There was less variation compared to the LCFC-1 level, and many of the food groups within ‘Cereals and bakery products’, on average, more clearly above or below the zero line. The mean variation in NRFI at LCFC-2 was 0.10. Then, we continued the example by selecting "Breakfast cereals" from the LCFC-2 food that are within ‘Cereals and bakery products’. when moving further to the LCFC-3 level in "Breakfast cereals", variation was still reduced within a single food group ('high-fibre cereal' and 'low-fibre cereal'), with the mean variation in NRFI being 0.08.

Discussion

The main purpose of this study was to compile, describe and test a reclassification of grocery product groups (LCFC) that could serve as a well-grounded basis for future examination of associations between grocery purchase data, dietary quality, sustainability, and health outcomes. The LCFC hierarchy contains four levels, of which the broadest was named LCFC-1, including food groups such as ‘Vegetables’ and ‘Alcoholic beverages’. The division to the more detailed three sub-classes were done based on the grocery product group’s type, quality (e.g., fibre or fat content), purpose of use, processing, carbon footprint and national food culture. As expected, the nutrient profiles (defined by NRFI) showed that there was more within-group variation in the nutrient quality of the food groups at LCFC-1, compared to the sub-classes LCFC-2 to LCFC-4. This indicates that the subtle sub-classes are better suited and a prerequisite for examining associations with grocery purchases and dietary quality [1,37].

To place our classification within an international context, it is essential to refer to The Classification of Individual Consumption According to Purpose (COICOP). This international reference classification of household expenditure has been developed by the United Nations Statistics [38]. Within the broadest (least granular) structure, Food and non-alcoholic beverages are one of the 15 classes (codes 01) in the least granular classification,

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

and this class is further divided hierarchically into 16 subclasses (01.x) and 68 sub-subclasses (91.x.x.). Our broadest classification LCFC-1 falls hence between the granularity of these two COICOP subclasses.

COICOP is the basis for the British Living Costs and Food Survey (LCS) (39) which uses a 5-digit scoring. The LCS covers a broad range of living costs, and "food" is one of the 2nd level groups. LCD goes then down towards more granular level from "class" (e.g., bread and cereal) to "COICOP expenditure code" (e.g., rice), and finally to "COICOP-plus code" which is close to our LCFC-4 granularity. The LCS classification has also been used in the UK to assess dietary patterns using supermarket transaction data (40). In that study, the researchers used 15 broad groups and 82 more detailed categories to identify purchase clusters, as indicators of dietary patterns. To improve the comparability of international reports and scientific research, it is crucial to openly share detailed classification descriptions when using similar hierarchical principles, but slightly different groupings.

Only a few studies have carefully described their justification and the process of classifying food purchase data for the purpose of using it for studying diet quality and health-related outcomes (19-22,31–34). The Food Price Database created by the Center of Nutrition Policy and Promotion of US for the National Food Plans (19) is one of the most extensive and oldest classification systems for grocery purchase data. The classification divides 4152 individual foods under 58 food categories and five broad food groups that are based on similarity of nutrient content, food costs, number of cup or ounce equivalents in MyPyramid (41) and use in meals. The Quarterly Food-at-Home Price Database (QFAHPD) was developed after the Food Price Database to fill the gap in available food price data and to support research on the economic determinants of diet quality and health outcomes (20). Foods were categorised to seven main food groups and further into 26 separate categories based on the 2005 Dietary Guidelines (42). The finest level of 52 categories defines the processing level (e.g., fresh, canned or frozen).

Like our classification, the Food Price Database (19) and QFAHPD (20) reflect dietary guidelines. The reports discuss the challenges of the classifications. For example, QFAHPD pointed out the difficulty of classifying foods that are composed of several ingredients. Classifying mixed foods was also one of our main challenges, and our solution was the same

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

as in QFAHPD: creating a ‘Miscellaneous’ class. However, we tried to minimise the number of grocery product groups in this class. This may have resulted in greater variation in the overall nutrient quality of the food groups at each class compared to the QFAHPD. This cannot be ascertained, as the nutritional quality of the QFAHPD has not been examined.

Despite the extensive classifications done in the Food Price Database and QFAHPD, we decided to create a new classification for our purposes. The main reasons for this were cultural and research purposes. Namely, although Finland – like the US – is a high-income economy, there are still differences in our food cultures and grocery food supply (e.g., type of bread and oil used). Moreover, the primary purpose of our LoCard grocery purchase data is to study interactions between food healthiness, environmental impact and price within the context of socio-demographic background and intentional (e.g., new taxation of foods) and sporadic (e.g., COVID-19, Ukraine crisis) transformation; hence, the new LCFC classification is needed to support this research context.

Other classifications that have been well described are the NOVA classification [(22)] and the Convenience Food Classification Scheme [(21)]. However, as explained in the Introduction, these classifications differed quite a lot from our principles, and these classifications would not have suited our purposes to link purchases primarily with health impact. The concept of UPF does not allow for nutritionally robust food grouping [(43)]: for example, industrially produced, high- and low-fibre bread are both classified as UPF, despite their different nutritional profile.

We argue that the LCFC could be directly applied in the Nordic and Baltic countries with rather similar food environments. We recommend, however, adapting the classification to the national or regional food culture when it is used in other countries or in multinational studies. The present ‘big data’ era gives many possibilities, but comparable use of data may be a challenge in international collaboration. Hence, there is a need for transparency and international classification ‘libraries’, perhaps also linked to food and diet ontologies [(44)]. Therefore, it is recommended that any new classifications are openly presented and shared among the science community.

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

Strengths and limitations

Our starting point was a grocery product grouping received from the food retailer and its original classification hierarchy. The most obvious limitation affecting our classification method was that we could not classify on the most detailed level (product level). This leads to some compromises, as well as making more assumptions of the food items under the grocery product groups. For example, frozen pizzas were classified under cereals because we had no information about whether they were meat or vegetarian pizzas.

In our evaluation of the nutrient quality of the classifications, there are possible weaknesses that need to be discussed. First, although the NRFI is a well-established method to profile groceries based on their nutrient content, it has methodological weaknesses (33,34). The choice of nutrients included in the index is subject to the researchers' discretion. Moreover, ranking of the foods by NRFI varies depending on the selection of nutrients in the index, and the equation used can also impact the outcome (34). It should also be noted that a difference in NRFI is difficult to interpret in a quantitative way, particularly when different kinds of foods are compared.

In our study, we used 11 nutrients to profile all food groups, but one could have also looked at food groups at LCFC-1 level and created separate indices for each food group with relevant nutrients included. This may have resembled the nutritional quality of the food groups better. For example, vegetables are generally perceived as very healthy, but they are not the main sources of iron, vitamin D, fibre or protein. Thus, judging vegetables by how much they include these nutrients is not relevant. Indeed, the class 'Vegetables' had relatively low NRFI, which does not resemble the true nutritional quality of this group.

In addition, NRFI does not have any upper or lower limits, meaning that the underlying assumption of the index is 'the more nutrients the better'. In practice, this is not true. Nutrient intake that exceeds the recommended value does not bring additional health benefits. This becomes relevant especially when the nutrient profiling is examined together with environmental impacts. For example, in our results, plant-based protein products received a relatively low NRFI value even though the use of these products may be advisable from an environmental perspective (45).

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

We chose to use NRFI to examine how well we succeeded in classifying the data based on dietary quality [23]. There would have been other options to use for nutrient profiling, such as the Grocery Purchase Quality Index-2016 (GPQI-2016) [24], which has been shown to associate with the Healthy Eating Index both on food group and total diet levels. There is also a scoring system developed for the QFAHPD to measure the overall quality of grocery purchases, which has been tested against the Healthy Eating Index [46]. However, NFRI is well known and widely used in nutrient profiling and allows the examination of all food groups that could be connected to the food composition database. As pointed out above, we do not claim that NRFI would be any better than other profiling systems, and whatever is chosen will always affect the results [47]. However, based on the profiling results, our classification was logical, meaning that food classes that are assumed to have relatively better nutritional quality, such as fruits and vegetables, got higher index values than foods considered to have low nutritional quality, such as sweets or chocolate. Further, our results imply that, on the more detailed levels, food classes became more homogeneous by their nutrient profiles.

Since we had the grocery purchase data on the grocery product group level, we had to select one food from the Finnish Food Composition Database to represent the nutrient content of that grocery product group. Again, since the limitation was that we did not have comprehensive knowledge on which type of grocery items were in some of the grocery product groups, the selected food from the composition database may have not always been the most optimal reference food. An improvement to this approach in the future could be selecting 3–5 of the most purchased foods that represent the grocery product group and that are also among the most consumed ones among the Finnish population and assigning the average nutrient values of those foods to represent the nutrient content of a grocery product group.

Last, the Finnish food retail market is very concentrated, since the two largest chains account for more than 80% market share. The reclassification was based on data from a single food retailer. However, although there are some differences in how the product categories are designed and managed, the overall selection is very similar across the major food chains. Therefore, we have been working with a food selection that is representative of the entire Finnish food market, and do not consider this as a major weakness of the LCFC. The selection of foods is likely to be different in other countries, but still the principles for grouping described in this paper apply.

Conclusion and recommendations

We have shown the multiple steps and amount of work needed to hierarchically classify grocery product groups for nutritional, health and environmental impact research. Based on nutrient profiling and using the NRFI, the nutritional quality of the LCFC was logical from a health viewpoint. The decrease of variation in nutritional quality in LCFC classes with higher granularity was reassuring and indicates good possibility to use the classification in studies linking food purchase data with health, environment, sociodemographic variables, and expenditure (price). Hence, we have shown that even without brand-level information, food purchase data can be classified in a meaningful way.

Customer loyalty card data holds manyfold potential for enhancing understanding of individuals' food purchase profiles and motives. Furthermore, it can contribute to enhanced means to design food systems that promote healthy food selection (48). Retail stores are core environments in such a food system, with a potential to promote both healthy and unhealthy food selection. The UK government has launched a ground-breaking new legislation to change retailers’ food marketing strategies, starting from October 2022 (49). Loyalty card data has a clear and strong potential to evaluate the immediate and long-term effects of such a policy action. As also concluded by Clark et al. (40), loyalty card data offers exceptional possibilities for multiple aspects of research related to grocery food selection, with broad societal implications.

Author contributions

NK did the analysis and had the main responsibility of writing and finalising the manuscript. SK had the main responsibility of the classification of the foods and linking the foods with a food composition database. ME, HV, JM and MF were part of the group of nutrition experts contributing to the creation of the classification method. JM supervised the linking of the purchase and food composition data. Further, MF, JN, HS and ME participated in data acquisition and curation, project administration and obtaining resources. All authors

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

participated in commenting and modifying the manuscript, and all have read and approved the final version of the manuscript.

Conflict of interest

MF is a member of the S Group societal responsibility advisory board. Membership does not include any sort of compensation. HV has received a fee from the S Group. The collaboration included offering professional advice to influencers and writing a blog post with regard to interpretation of the nutrition calculator in S Group’s mobile app.

Other authors have nothing to declare.

Funding

This study was funded by the Juho Vainio foundation (grant to ME #202200480). The LoCard study is funded by Academy of Finland (research grant to MF and JN #350852). Linking the purchase data with the food composition database was funded by Juho Vainio and Yrjö Jahnsson foundations (grants to JM #202100202 and #20207300, respectively).

References

1. Fardet A, Rock E, Bassama J, et al. (2015) Current food classifications in epidemiological studies do not enable solid nutritional recommendations for preventing diet-related chronic diseases: the impact of food processing. *Adv Nutr* **6**, 629-38.
2. European Food Safety Authority (2015) The food classification and description system FoodEx 2 (revision 2). *EFSA Supporting Publications* **12**, 804E.
3. Charrondiere U, Stadlmayr B, Haytowitz D, et al. (2012) Guidelines for Checking Food Composition Data Prior to the Publication of a User Table/Database Version 1.0. Rome, Italy; FAO.
4. Finglas PM, Berry R, Astley S. (2014) Assessing and improving the quality of food composition databases for nutrition and health applications in Europe: the contribution of EuroFIR. *Adv Nutr* **5**, 608S-614S.
5. Bandy L, Adhikari V, Jebb S, et al. (2019) The use of commercial food purchase data for public health nutrition research: A systematic review. *PLoS One* **14**, e0210192.

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

6. Jenneson VL, Pontin F, Greenwood DC, et al. (2022) A systematic review of supermarket automated electronic sales data for population dietary surveillance. *Nutr Rev* **80**, 1711–1722.
7. Vuorinen AL, Erkkola M, Fogelholm M, et al. (2020) Characterization and Correction of Bias Due to Nonparticipation and the Degree of Loyalty in Large-Scale Finnish Loyalty Card Data on Grocery Purchases: Cohort Study. *J Med Internet Res* **22**, e18059.
8. Sørensen KK, Nielsen EP, Møller AL, et al. (2021) Food purchases in households with and without diabetes based on consumer purchase data. *Prim Care Diabetes* **16**, 574-580.
9. Vepsäläinen H, Nevalainen J, Kinnunen S, et al. (2021) Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study. *Br J Nutr*, 1–24.
10. Lintonen T, Uusitalo L, Erkkola M, et al. (2020) Grocery purchase data in the study of alcohol use - A validity study. *Drug Alcohol Depend* **214**, 108145.
11. Erkkola M, Kinnunen SM, Vepsäläinen HR, et al. (2022) A slow road from meat dominance to more sustainable diets: An analysis of purchase preferences among Finnish loyalty-card holders. *PLOS Sustainability and Transformation* **1**, e0000015. Public Library of Science.
12. Teng AM, Jones AC, Mizdrak A, Signal L, et al. (2019) Impact of sugar-sweetened beverage taxes on purchases and dietary intake: Systematic review and meta-analysis. *Obes Rev* **20**,1187-1204. doi: 10.1111/obr.12868.
13. Vall Castelló J, Lopez Casasnovas G. (2020) Impact of SSB taxes on sales. *Econ Hum Biol* **36**, 100821.
14. Meinilä J, Hartikainen H, Tuomisto HL, et al. (2022) Food purchase behaviour in a Finnish population: patterns, carbon footprints and expenditures. *Public Health Nutr* **25**, 3265–3277.
15. Willett W Rockström J, Loken B, et al. (2019) Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. *Lancet* **39**, 447-492.
16. Beaglehole R, Bonita R, Horton R, et al. (2011) Priority actions for the non-communicable disease crisis. *Lancet* **377**, 1438-47.

17. Pauler G, Dick A (2006). Maximizing profit of a food retailing chain by targeting and promoting valuable customers using Loyalty Card and Scanner Data. *European Journal of Operational Research* ***174***, 1260-1280.
18. Zhong R, Xu X, Wang L (2017), Food supply chain management: systems, implementations, and future research. *Industrial Management & Data Systems* ***117***, 2085-2114.
19. Carlson A, Lino M & Fungwe TV (2007) *The Low-Cost, Moderate-Cost, and Liberal Food Plans, 2007. CNPP Reports 45850*. United States Department of Agriculture, Center for Nutrition Policy and Promotion.
20. Todd JE, Mancino L, Leibtag E, et al. Methodology Behind the Quarterly Food-at-Home Price Database. http://www.ers.usda.gov/publications/pub-details/?pubid=47567 (accessed June 2022).
21. Peltner J & Thiele S (2018) Convenience-based food purchase patterns: identification and associations with dietary quality, sociodemographic factors and attitudes. *Public Health Nutr* **21**, 558–570.
22. Monteiro CA, Levy RB, Claro RM, et al. (2010) A new classification of foods based on the extent and purpose of their processing. *Cad Saude Publica* **26**, 2039–2049.
23. Drewnowski A (2010) The Nutrient Rich Foods Index helps to identify healthy, affordable foods. *Am J Clin Nutr* **91**, 1095S-1101S.
24. Brewster PJ, Durward CM, Hurdle JF, et al. (2019) The Grocery Purchase Quality Index-2016 Performs Similarly to the Healthy Eating Index-2015 in a National Survey of Household Food Purchases. *J Acad Nutr Diet* **119**, 45–56.
25. Vadiveloo MK, Juul F, Sotos-Prieto M, et al. (2022) Perspective: Novel Approaches to Evaluate Dietary Quality: Combining Methods to Enhance Measurement for Dietary Surveillance and Interventions. *Adv Nutr* **13**, 1009-1015.
26. Wu J, Fuchs K, Lian J, Haldimann ML, Schneider T, Mayer S, Byun J, Gassmann R, Brombach C, Fleisch E. Estimating Dietary Intake from Grocery Shopping Data-A Comparative Validation of Relevant Indicators in Switzerland. Nutrients. 2021 Dec 29;14(1):159. doi: 10.3390/nu14010159.
27. Nevalainen J, Erkkola M, Saarijärvi H, et al. (2018) Large-scale loyalty card data in health research. *Digit Health* **4**, 2055207618816898.

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

28. Nordic Council of Ministers (2013) *Nordic nutrition recommendations. Part 1. Summary, principles and use.* Nord 2013/009, 5th ed. Denmark: Norden.
29. Reinivuo H, Hirvonen T, Ovaskainen ML, et al. (2010) Dietary survey methodology of FINDIET 2007 with a risk assessment perspective. *Public health nutrition* **13**, 915–919.
30. European Parliament and the Council of the European Union (2006) Regulation (EC) No 1924/2006 of the European Parliament and of the Council of 20 December 2006 on nutrition and health claims made on foods. http://data.europa.eu/eli/reg/2006/1924/oj/eng (accessed June 2022).
31. Uusitalo L, Nevalainen J, Rahkonen O, et al. (2022) Changes in alcohol purchases from grocery stores after authorising the sale of stronger beverages: The case of the Finnish alcohol legislation reform in 2018. *Nordic Studies on Alcohol and Drugs* **39**, 589-604.
32. Hartikainen H & Pulkkinen H (2016) *Summary of the chosen methodologies and practices to produce GHGE-estimates for an average European diet*. Finland; Natural Resources Institute Finland (Luke).
33. Drewnowski A & Fulgoni VL (2014) Nutrient density: principles and evaluation tools. *Am J Clin Nutr* **99**, 1223S–8S.
34. Fulgoni VL, Keast DR, Drewnowski A (2009) Development and validation of the nutrient-rich foods index: a tool to measure nutritional quality of foods. *J Nutr* **139**, 1549-1554.
35. Valsta L, Kaartinen N, Tapanainen H, et al. (2019) Ravitsemus Suomessa – FinRavinto 2017 -tutkimus (Nutrition in Finland – The National FinDiet 2017 Survey). Report 12/2018. Helsinki, Finland; Institute for Health and Welfare (THL).
36. R: The R Project for Statistical Computing. https://www.r-project.org/ (accessed October 2022).
37. Astrup A, Monteiro CA (2022) Does the concept of "ultra-processed foods" help inform dietary guidelines, beyond conventional classification systems? NO. *Am J Clin Nutr* **116**, 1482-1488.
38. United Nations (2018) Classification of Individual Consumption According to Purpose (COICOP) 2018. Department of Economic and Social Affairs, Statistical Papers Series M No 99. New York, USA; UN. https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf (accessed September 2023).

39. Rafferty A, Walthery P (2014) Introductory guide to the living cost and food survey. UK Data Service, University of Essex and University of Manchester. https://ukdataservice.ac.uk/app/uploads/livingcostsfoodsurvey.pdf (acessed September 2023).
40. Clark SD, Shute B, Jenneson V, et al. (2021) Dietary Patterns Derived from UK Supermarket Transaction Data with Nutrient and Socioeconomic Profiles. *Nutrients* **13**, 1481.
41. Britten P, Lyon J, Weaver CM, et al. (2006) MyPyramid food intake pattern modeling for the Dietary Guidelines Advisory Committee. *J Nutr Educ Behav* **38**, S143-152.
42. U.S. Department of Health and Human Services & U.S. Department of Agriculture (2005) *Dietary guidelines for Americans, 2005*. Washington DC, US; Government Printing Office.
43. Braesco V, Souchon I, Sauvant P, et al. (2022) Ultra-processed foods: how functional is the NOVA system? *Eur J Clin Nutr* **76**, 1245–53.
44. Andrés-Hernández L, Blumberg K, Walls RL, et al. (2022) Establishing a Common Nutritional Vocabulary - From Food Production to Diet. *Front Nutr* **9**, 928837.
45. Clark M, Springmann M, Rayner M, et al. (2022) Estimating the environmental impacts of 57,000 food products. *Proc Natl Acad Sci U S A* **119**, e2120584119.
46. Volpe R, Okrent A (2012) *Assessing the Healthfulness of Consumers' Grocery Purchases*. Economic Information Bulletin 262129, United States Department of Agriculture, Economic Research Service.
47. Drewnowski & Fulgoni. (2008) Nutrient profiling of foods: creating a nutrient-rich food index. *Nutrition Reviews* **66**, 23–39.
48. Muir S, Dhuria P, Roe E, et al. (2023) UK government's new placement legislation is a 'good first step': a rapid qualitative analysis of consumer, business, enforcement and health stakeholder perspectives. *BMC Med* **21**, 33.
49. UK Government (2021) The Food (Promotion and Placement) (England) Regulations 2021. London: UK government. https://www.legislation.gov.uk/uksi/2021/1368/contents (accessed September 2023).

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

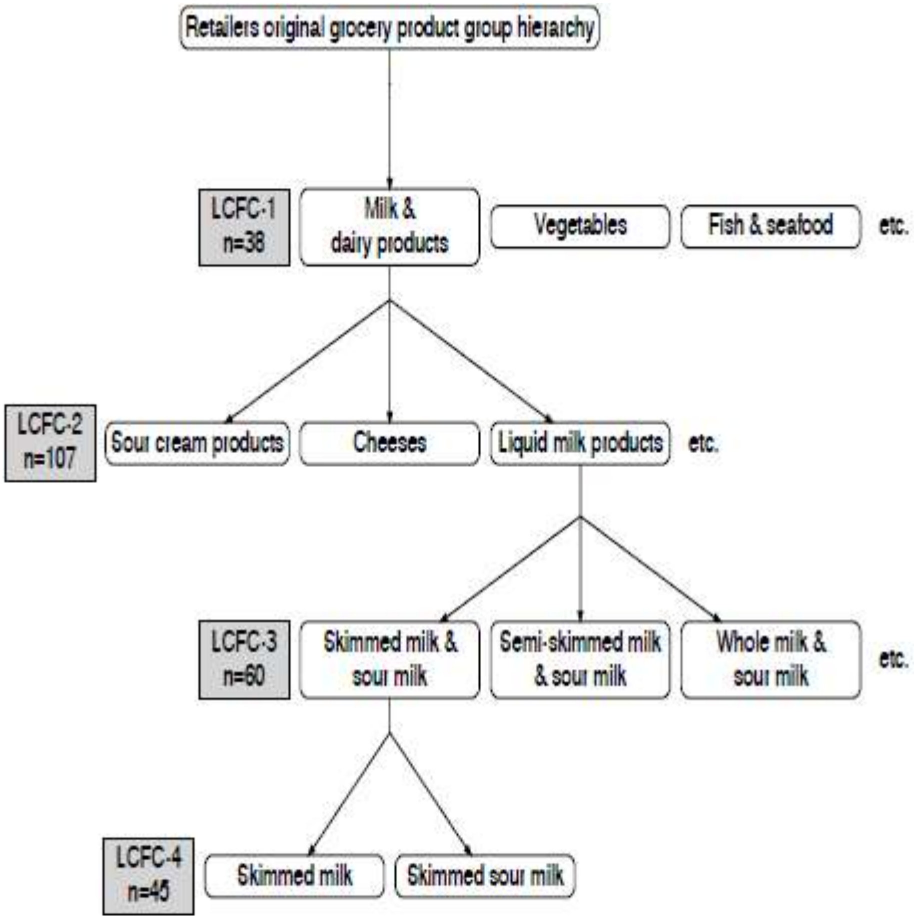

Figure 1. The LoCard Food Classification (LCFC) process. Grocery product groups were reclassified first at the least granular hierarchy level called LCFC-1. Each food group on the LCFC-1 level was then subsequently divided to finer sub-classes on the LCFC-2 level, followed by LCFC-3 and, finally, LCFC-4, which was the most granular level of our hierarchy. The number of food groups at each level is given in the blue boxes on the left.

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

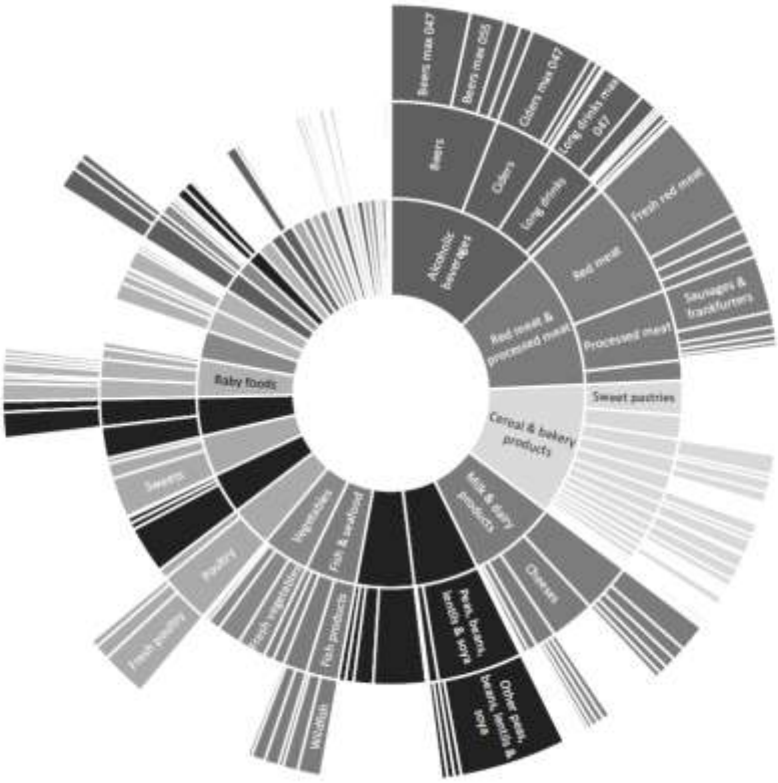

Figure 2. Illustration of the hierarchical structure of the reclassification of the 3574 original grocery product groups received from the retailer. The inner circle represents LoCard Food Classification level 1 (LCFC-1). The middle circle represents LCFC-2, and the outer circle represents LCFC-3. LCFC-4 is not shown in the figure due to the small number. Further, some of the boxes are missing labels due to lack of space.

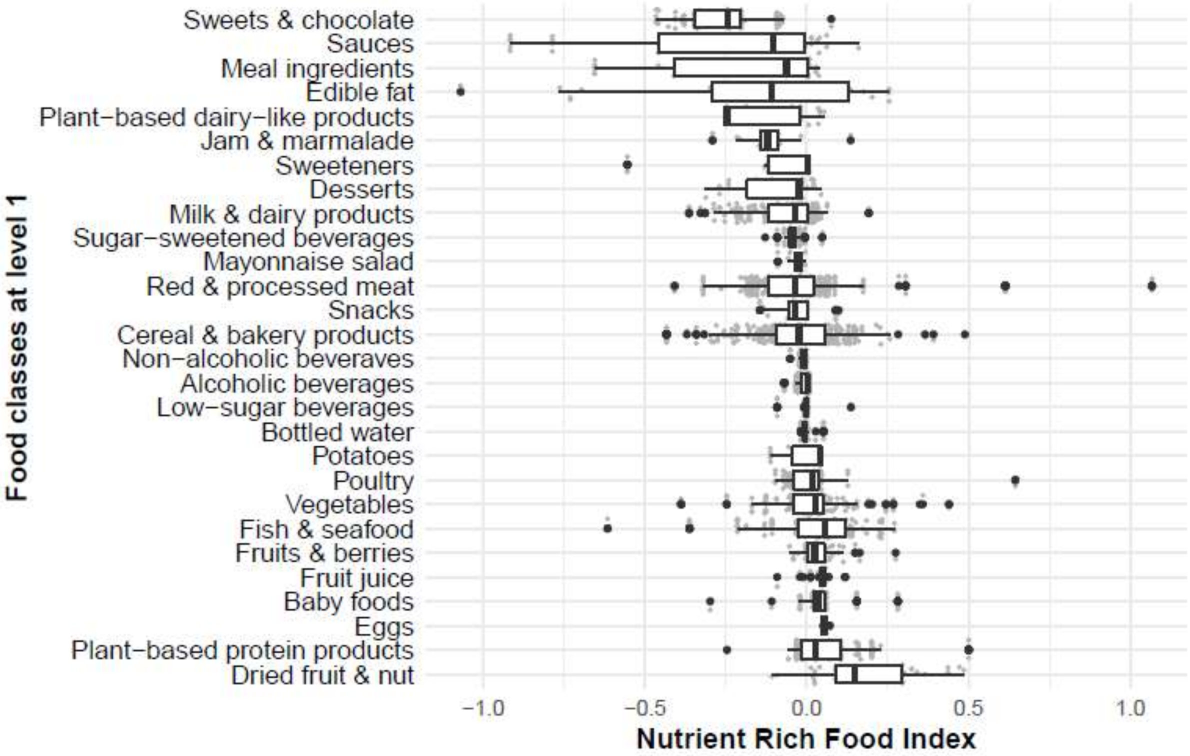

Figure 3. Variation in Nutrient Rich Food Index values for grocery product groups (grey dots) at LoCard Food Classification level 1 (LCFC-1). Positive values indicate food classes that are more nutrient rich whereas negative values indicate food classes that are less nutrient rich. The boxplot illustrates the median index value (middle vertical line) and 25th and 75th percentiles (outer lines of the box). The upper/lower whisker extends from the outer box to the largest/smallest value no further than 1.5 * inter-quartile range (for the lower whisker -1.5 * inter-quartile range) from the box. Data beyond the end of the whiskers are 'outlying' points and are plotted individually using black dots.

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

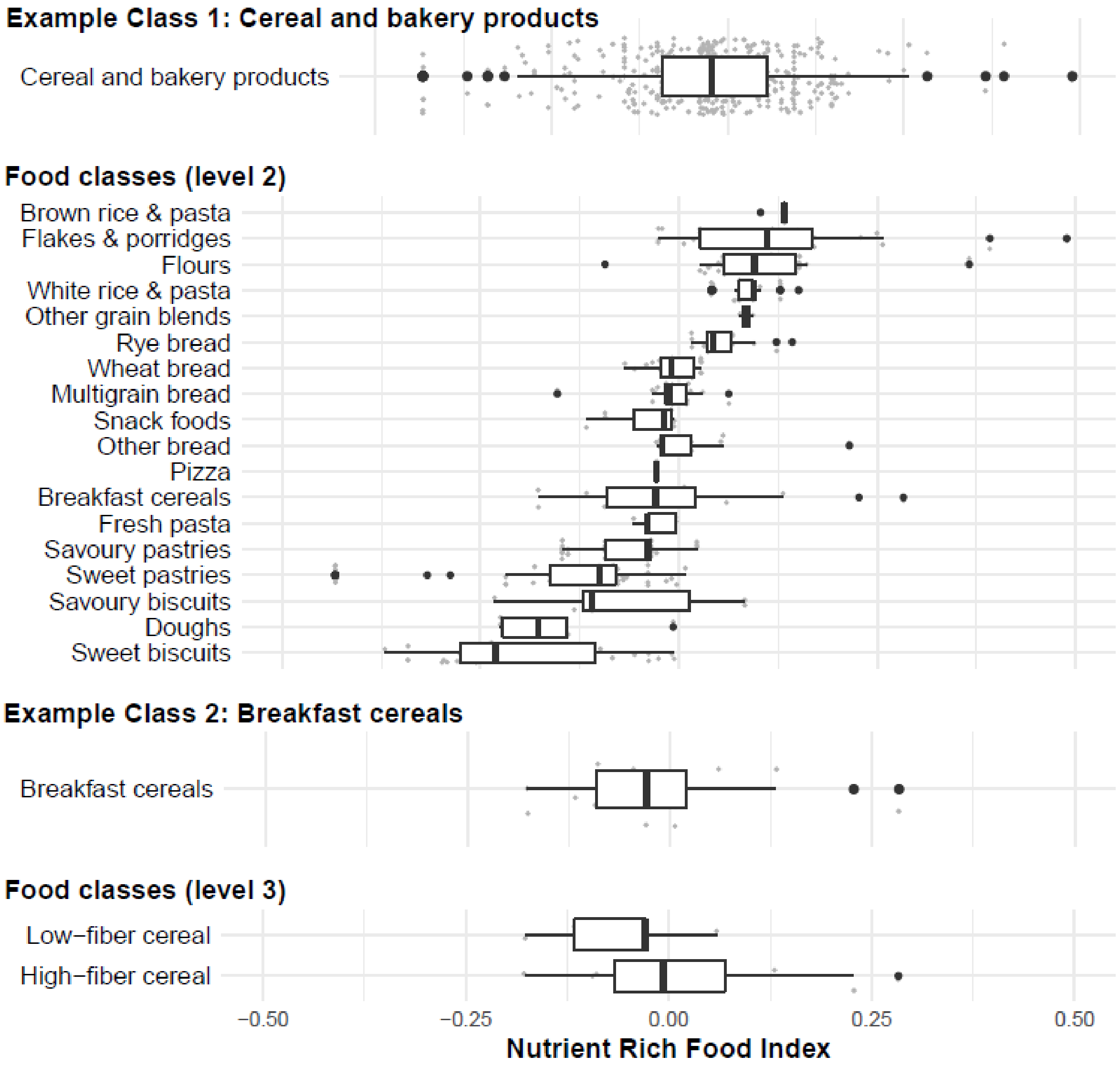

Figure 4. Variation in Nutrient Rich Food Index values for grocery product groups (grey dots) at different LoCard Food Classification (LCFC) levels. The figure shows an example of the food class ‘Cereal and bakery products’ from LCFC-1 and the food classes that are located in it at LCFC-2. Further, the figure shows an example of the food class ‘Breakfast cereals’ from LCFC-2 and the food classes that are located in it at LCFC-3. The boxplots in the figure illustrate the median index value (middle vertical lines) and 25th and 75th percentiles (outer lines of the boxes). The upper/lower whiskers extend from the outer line of the box to the largest/smallest value no further than 1.5 * inter-quartile range (for the lower whiskers -1.5 * inter-quartile range). Data beyond the end of the whiskers are ‘outlying’ points and are plotted individually using black dots.

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press

Table 1. Principles of reclassification of a food retailer's grocery product groups in LoCard Food Classification (LCFC).

Class	**Principles of the classification**	**Product groups applied**	**Examples/additional information**
LCFC-1 (38 groups)			
	Healthiness (nutrition recommendations)	All grocery product groups	Nordic Nutrition Recommendations [28]
	Main ingredient of the product group (food group classification in food composition databases)	All grocery product groups	Finnish Food Composition Database (www.fineli.fi) [29]
	Incomplete or missing information	Miscellaneous	Not possible to define the content or main ingredient of the product group, e.g. some frozen products, some soups, delicacy basket
LCFC-2 (107 groups)			
	Type of foods in the product group	All classes at LCFC-1 (except baking product, chewing gum, cocoa, coffee, tea, desserts, dietary supplements, fruit juice, marmalade, mayonnaise, meal ingredients, miscellaneous and snacks)	For example: sugar-sweetened beverages were categorised further to energy drinks, juices and soft drinks; alcoholic beverages were categorised to beers, ciders, long drinks and wines; sweeteners were categorised to honey, sugars and syrups; cereal and bakery products were categorised to different types of breads, rice, pasta, cereals, pastries, biscuits and pizza; edible fats were categorised to cooking fat, butter, vegetable oils and margarine; plant-based dairy-like products were categorised to plant-based drinks, ice creams, plant-based puddings and yoghurts and curds
	Protein source	Plant protein products	Categorisation by their protein sources (wheat; fungal; peas, beans, lentils and soya)
	Purpose of use	Nuts	Plain nuts were classified under 'Dried fruits and nuts' whereas chocolate-coated nuts were classified under 'Sweets and chocolates' and salted nuts under 'Snacks'
	National food culture	Pulses/legumes	Pea soup
LCFC-3 (60 groups) and 4 (45 groups)			
	Nutritional value/nutrient content	Breakfast cereals, multigrain bread, flakes and porridge, flours,	A cut-off of 6% for high-fibre content [30]

		rye bread and wheat bread	
		Pasta and rice	Brown/white
		Liquid milk products	Cut-offs of 1% and 3% to separate skimmed, semi-skimmed and whole milk
		Yoghurts, cultured milks and curds	Low fat was defined as <1% of fat
		Beer, cider, long drink and wine	Cut-offs based on alcohol % (29): <=1.2%, 1.3–2.8%, 2.9–3.5%, 3.6–4.7% and 4.8–5.5%
	Carbon footprint	Red meat	Beef, pork, lamb, horse meat, game and reindeer, pork-beef, uncategorised
		Fish and fish products	Farmed, wild, uncategorised
	Processing	Fruits	Fresh, canned, frozen
		Vegetables	Canned, fresh, frozen, dishes
		Poultry	Cooked, fresh, offal, patties/balls
		Processed meat	Canned, sausages, cold cuts, ham, jellies, pate
		Red meat	Cooked, fresh, patties/balls, offal

https://doi.org/10.1017/S0007114524000710 Published online by Cambridge University Press