Building nutritionally meaningful product groups for loyalty card data: the LoCard Food Classification process

Noora Kanerva (✉ noora.kanerva@helsinki.fi)  
University of Helsinki  https://orcid.org/0000-0001-6776-9357

Satu Kinnunen  
University of Helsinki

Jaakko Nevalainen  
Tampere University

Henna Vepsäläinen  
University of Helsinki

Mikael fogelholm  
University of Helsinki

Hannu Saarijärvi  
Tampere University

Jelena Meinilä  
University of Helsinki

Maijaliisa Erkkola  
University of Helsinki

Method Article

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

Posted Date: April 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2826970/v1

License: ☺ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Analysing customer loyalty card data is a novel method for assessing dietary 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 into appropriate categories suitable for the use of nutrition and health research. This paper depicts how such a classification is compiled and how it reflects the nutritional quality of the food classes. Healthfulness was considered the main criterion guiding the reclassification of the 3574 grocery product groups. In addition, the main ingredient of the product group, type of food and purpose of use, and carbon footprint were considered in the reclassification process. The classified food groups were linked with the national food composition database, and the nutrient profile was assessed by calculating the Nutrient Rich Food Index (NRFI) for each product group. Our four-level classification hierarchy had 38 food groups at its broadest level (Class 1). Only 1% (n=38) of the grocery product groups were left unclassified. Standard deviation in NRFI decreased from 0.21 to 0.08 from the broadest to the finest level of classification. We conclude it is possible to assign a great majority of the grocery product groups to classes based on their nutritional quality. However, the challenge is classification of product groups that lack detailed information on their contents or include main ingredients that have opposite health effects, such as products including both plant- and animal-based proteins.

Introduction

The use of food retailers’ customer loyalty card data in academic research has gained increasing attention in recent years (1). Customer loyalty card data can be used for various research purposes. For example, they provide a novel method to measure food purchase patterns and monitor the nutrition composition of food purchases over time as it reasonably well reflects the self-reported food and drink consumption of the adult population (2, 3). Further, customer loyalty card data could potentially be studied against health outcomes at the population level (4) as well as for monitoring and evaluating dietary environmental and economic sustainability (5, 6). Importantly, while grocery purchase data are about what, when, where and at what price food has been bought, customer loyalty card data also contain data on who made the purchase. Therefore, loyalty card data are especially of interest for health and health policy research as they provide an opportunity to automatically and objectively obtain huge amounts of detailed data over years on different card holders’ grocery purchases (2). This extends attention from a population-level investigation to how different consumer groups’ food purchases evolve over time.

In nutritional epidemiology, hierarchical classification of food items into broader categories plays a critical role when examining associations with food consumption and health. For these purposes, international classification systems such as the European Food Safety Authority’s FoodEx2 classification (7) or FAO/INFOODS (8) 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. Food retailers’ data systems were not, however, designed originally to serve nutrition and health research purposes but to facilitate the efficient flow of food items from suppliers to consumers and identify the profitable customers from the unprofitable ones. Food retailers commonly use classification systems that are based on logistics or product placement on the shelf, and they do not reflect products’ nutritional profiles. 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.

Only a few of the classification methods used for groceries have been transparently described, such as the Convenience Food Classification Scheme (CFCS) (9) and the NOVA classification (10). However, even though there are suitable tools available such as the Nutrient Rich Food Index (NRFI) (11) and the Grocery Purchase Quality Index-2016 (12), 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 or even discussed. Without a clear, explicit, openly available and critically evaluated grocery product classification, researchers cannot reliably examine the health and environmental impacts of grocery purchases or the literature focusing on it.

We have received a large-scale (n = 47,066) longitudinal customer loyalty card (LoCard) data set from the largest food retailer in Finland (13). While the original product grouping used by the retailer was designed for retail purposes, our main challenge was to design and compile a meaningful product grouping appropriate for nutrition and health research. Thus, the purpose of this paper is to create, describe and test a reclassification of products suitable for research purposes (LoCard Food Classification, hereafter LCFC) and to make it openly accessible. To achieve this, we demonstrate how the reclassification process was conducted and test how LCFC succeeded in reflecting the nutrient quality of the food classes by examining variation in nutritional quality within a group.

**Methods**

**Data**

The LoCard study explores the potential of loyalty card data for nutrition and health research (3). A more detailed description of the study design and sample was published earlier (13, 14). The present customer loyalty card data cover the period from 1 September 2016 to 31 December 2018. Each purchase was associated with the item description, time stamp, quantity (i.e. weight, volume or number of packages) and expenditure on the item, totalling about 130 million purchases. The data were received for research purposes from the S Group (a major Finnish retail co-operative) after receiving consent from the loyalty card holders. The data include 3574 grocery product groups. The food retailer’s grocery product group data did not include product-level (brand names) information.

**Reclassification of grocery product groups in LCFC**
We used a four-level hierarchical classification of product groups. Each class on the broadest level of hierarchy (Class 1) was subsequently divided into a reasonable number of finer sub-classes starting with Class 2, followed by Class 3 and, finally, Class 4, which was the most detailed level of hierarchy. An example is given in Fig. 1. The whole classification is openly available at https://doi.org/10.5281/zenodo.7781352.

We started at the broadest level of hierarchy by reclassifying the food retailer’s grocery product groups into 38 main food groups based on healthiness\(^{(15)}\) and main ingredients\(^{(16)}\) (labelled as Class 1; see Table 1). Classification to Class 2 was dictated by the type of foods in the product group, purpose of use of the product groups and food culture. At the finer levels (Class 3 and Class 4), nutritional quality and carbon footprint were used to guide the classification when reasonable.
Table 1
Principles of reclassification of a food retailer's grocery product groups in LCFC.

<table>
<thead>
<tr>
<th>Class</th>
<th>Principles of the classification</th>
<th>Product groups applied</th>
<th>Examples/additional information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1 (38 groups)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthiness</td>
<td>(nutrition recommendations)</td>
<td>All grocery product groups</td>
<td>Nordic Nutrition Recommendations (15)</td>
</tr>
<tr>
<td>Main ingredient of the product group</td>
<td>(food group classification in food composition databases)</td>
<td>All grocery product groups</td>
<td>Finnish Food Composition Database (<a href="http://www.fineli.fi">www.fineli.fi</a>) (16)</td>
</tr>
<tr>
<td>Incomplete or missing information</td>
<td>Miscellaneous</td>
<td>Not possible to define the content or main ingredient of the product group, e.g. some frozen products, some soups, delicacy basket</td>
<td></td>
</tr>
<tr>
<td>Class 2 (107 groups)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of foods in the product group</td>
<td>All classes at Class 1 (except baking product, chewing gum, cocoa, coffee, tea, desserts, dietary supplements, fruit juice, marmalade, mayonnaise, meal ingredients, miscellaneous and snacks)</td>
<td>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</td>
<td></td>
</tr>
<tr>
<td>Protein source</td>
<td>Plant protein products</td>
<td>Categorisation by their protein sources (wheat; fungal; peas, beans, lentils and soya)</td>
<td></td>
</tr>
<tr>
<td>Purpose of use</td>
<td>Nuts</td>
<td>Plain nuts were classified under ‘Dried fruits and nuts’ whereas chocolate-coated nuts were classified under ‘Sweets and chocolates’ and salted nuts under ‘Snacks’</td>
<td></td>
</tr>
<tr>
<td>National food culture</td>
<td>Pulses/legumes</td>
<td>Pea soup</td>
<td></td>
</tr>
<tr>
<td>Classes 3 (60 groups)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classes 4 (45 groups)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>Principles of the classification</td>
<td>Product groups applied</td>
<td>Examples/additional information</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------</td>
<td>------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Nutritional value/nutrient content</td>
<td>Breakfast cereals, multigrain bread, flakes and porridge, flours, rye bread and wheat bread</td>
<td>Pasta and rice</td>
<td>A cut-off of 6% for high-fibre content (17) Brown/white Cut-offs of 1% and 3% to separate skimmed, semi-skimmed and whole milk Low fat was defined as &lt; 1% of fat Cut-offs based on alcohol %: &lt;=1.2%, 1.3–2.8%, 2.9–3.5%, 3.6–4.7% and 4.8–5.5%</td>
</tr>
<tr>
<td></td>
<td>Liquid milk products</td>
<td>Yoghurts, cultured milks and curds</td>
<td>Beer, cider, long drink and wine</td>
</tr>
<tr>
<td>Carbon footprint</td>
<td>Red meat</td>
<td>Beef, pork, lamb, horse meat, game and reindeer, pork-beef, uncategorised</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fish and fish products</td>
<td>Farmed, wild, uncategorised</td>
<td></td>
</tr>
<tr>
<td>Processing</td>
<td>Fruits</td>
<td>Fresh, canned, frozen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vegetables</td>
<td>Canned, fresh, frozen, dishes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Poultry</td>
<td>Cooked, fresh, offals, patties/balls</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Processed meat</td>
<td>Canned, sausages, cold cuts, ham, jellies, pate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Red meat</td>
<td>Cooked, fresh, patties/balls, offals</td>
<td></td>
</tr>
</tbody>
</table>

For example, to enable examination of the health impacts of the customer loyalty card data, plant-based protein products were separated from the meat product group (where they were placed in the food retailer’s original grouping) into their own main class. This Class 1 category was named ‘Plant protein products’ and included processed legume products such as those mimicking meat, as well as unprocessed lentils, peas and beans. At the Class 1 level, we also formed a separate class for plant-based dairy-like products including, for instance, soy and oat milk.

Using the type of the product group as a basis for classifying grocery product groups at Class 2 meant, for example, that the Class 1 category ‘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 recategorization at the Class 2 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 Class 1 level.

Classification of traditional Finnish ready-made pea soup is an example of how we considered the national food culture. 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 Class 2 level, which is under the broader ‘Plant protein products’ category at the Class 1 level – not as a red-meat product.

At finer hierarchy levels (Class 3 and Class 4), we used nutritional quality and carbon footprint 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 Class 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 (17). 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: <=1.2%, 1.3–2.8%, 2.9–3.5%, 3.6–4.7% and 4.8–5.5% (18).

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. We used carbon footprint as a basis for classification when within-food-group variation in the carbon footprint of the foods was large. For example, because the average carbon footprint of beef is much greater than that of pork (19), in Class 4 we classified different types of red meat separately.

The main challenge in the reclassification was that, as we did not have the brand-level information, it was not always clear what foods belonged to each retailer’s grocery product groups. For example, for groups called ‘Other meat’, ‘Ready-made salads’, ‘Hamburgers’ or ‘Pizzas’, the main ingredient of the grocery product group was not clear. To solve the content of these challenging groups, we used the retailer’s online food purchasing service, which includes some information about the foods under the grocery product group. Further, we received a small data sample from the retailer that included some examples of the actual products (brand names) that were under each group. These helped us with deciding the reclassification for most of the foods.

For some of the retailer’s grocery product groups, the classification remained a compromise even after using the online food purchasing service and sample data. 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). Eventually, there were only 38 grocery product groups (0.01%) left unclassified under ‘Miscellaneous’ at
the Class 1 level; for example, for ‘Other frozen products’ and ‘Other ready-made meals’ it was not possible to define the main ingredient. Consequently, for the ‘Miscellaneous’ group, it was not possible to match any nutrient composition.

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

**Analyses of nutrient quality using the reclassified data**

Next, the reclassified food product groups were linked with nutrient content. We used the Finnish Food Composition Database Fineli**(R)**, version 20 ([www.fineli.fi](http://www.fineli.fi)) (16). Fineli’s open database includes approximately 4000 food items and dishes. For each class, the most suitable alternative from Fineli was selected. If the Fineli database did not contain the food that would have described the class, food composition databases of other countries (e.g. Swedish and US databases) were exploited. Fifty-nine grocery product groups (0.02% of the reclassified product groups) were left without nutrient content because it was impossible to select a food that would have represented the contents well enough (e.g. ‘Other canned foods’, ‘Warm side dish service’, and ‘Other ready-made soups’) or the group did not include foods that would have relevant nutrient content (e.g. spices and vitamin and mineral supplements due to the variety of different supplements in the group).

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 class following principles of Drewnowski et al. (20). NRFI is a method of nutrient profiling aiming to provide an overall nutrient density score on the basis of selected nutrients. The NRFI was calculated per 100 g of product using 11 nutrients, of which eight were regarded positive (protein, fibre, polyunsaturated fatty acids (PUFA), calcium, iron, vitamin D, vitamin C and folate) and three negative (saturated fatty acids (SFA), saccharose and salt) in terms of anticipated health effects. Recommended values used for the 11 nutrients are from the Finnish nutrition recommendations which are the same as for the Nordic Nutrition recommendations (15), except salt which is 5000mg in the Finnish recommendations (6000 mg in the Nordic nutrition recommendations).

First, positive and negative scores were calculated as an average of percentages of daily recommendation (DR%):

**Positive score:**

\[
\text{DR}\%\text{ protein + DR}\%\text{ fibre + DR}\%\text{ PUFA + DR}\%\text{ Ca + DR}\%\text{ Fe + DR}\%\text{ Vit D + DR}\%\text{ Vit C + DR}\%\text{ folate}} / 8
\]

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

Iron (Fe) = 9mg

Vitamin D = 10µg

Vitamin C = 75mg

Folate = 300mg

Negative score:

\[(\text{DR\% sucrose} + \text{DR\% SFA} + \text{DR\% salt}) / 3\]

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

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

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

Salt = 5000mg

Last, NRFI was calculated by subtracting the negative score from the positive 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 food class at different hierarchical levels were drawn using R statistical software \(^{(21)}\).

Results

Figure 2 gives an overall representation of the hierarchy of the reclassification from the retailer’s grocery product groups to the new Classes 1–3 (Class 4 is not shown on the figure). Most of the grocery product groups were classified only at Classes 1–2. Classes 3–4 were used when needed, for example, to distinguish foods with different nutritional and/or carbon footprint profiles. Therefore, only some foods were classified at these levels. The largest groups at Class 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’ represent about half (1509 out of 3574) of all the retailer’s grocery product groups. 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 Class 1 groups containing 100–200 grocery product groups (1249 retailer’s grocery product groups in total). The smallest 25 Class 1 groups included less than 100 grocery
product groups each (816 retailer’s grocery product groups in total). The whole detailed classification structure of Classes 1–4 is available at https://doi.org/10.5281/zenodo.7781352.

To illustrate the extent to which the reclassification succeeded in reflecting the nutritional quality of the grocery product groups, Fig. 3 shows the medians and the variation in the NRFI values of the food groups at Class 1. In general, when the groups at Class 1 were ranked by their NRFI median value, the order was logical. 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 (Fig. 3).

As seen in the boxplots, the variation in NRFI of the food groups at Class 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 (Fig. 3). The mean standard deviation in NRFI at Class 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) (Fig. 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, when examining NRFI values closer within ‘Cereals and bakery products’ at the Class 2 level (Fig. 4), there is less variation compared to the Class 1 level, and many of the food groups were, on average, more clearly above or below the zero line. The mean variation in NRFI at Class 2 was 0.10. Moreover, when moving further to the Class 3 level, variation got even lower by food group, with the mean variation in NRFI being 0.08 for Class 3. Figure 4 gives an example of food groups at Class 3 within ‘Breakfast cereals’ at Class 2.

**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 is called Class 1, including food groups such as ‘Vegetables’. 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 a lot of within-group variation in the nutrient quality of the food groups at Class 1. The variation declined in the sub-classes, and this indicates that the subtle sub-classes are more suitable and a prerequisite for examining associations with grocery purchases and dietary quality.
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 (9, 10, 22–26). The Food Price Database created by the Center of Nutrition Policy and Promotion of US for the National Food Plans (22) is one of the most extensive and oldest classification systems for grocery purchase data. The Food Plans provide representative healhful market baskets at three different cost levels in the United States, and it merges information about food consumption from the National Health and Nutrition Examination Survey (NHANES) with national data on food prices from the Nielsen Homescan™ Panels. The Panels contain the prices paid for food items by 16,821 households, reflecting the US population. The food 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 (27) 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 (23). The database contains quarterly market-level prices for food at home. The principles of food classification resemble the Food Price Database. Foods in the Nielsen Homescan™ Panels have been categorised to seven main food groups identified in the dietary guidelines for Americans: grains, vegetables, fruits, milk, meat and beans, oils and discretionary calories. The seven main food groups were further classified into 26 separate categories based on the 2005 Dietary Guidelines (28) and other factors relevant for food shopping and preparation ( premiums paid for preparation and other processing). The finest level including 52 categories defines the processing level of the food ( e.g. the food is sold fresh, canned or frozen).

The principles and solutions of the Food Price Database (22) and QFAHPD (23) are closest to our classification. They reflect healthy or unhealthy eating by categorisation based on dietary guidelines. The reports openly discuss many of the challenges of the classifications. For example, QFAHPD pointed out the difficulty of classifying foods that are composed of several ingredients, which many researchers trying to classify food purchases are facing. Classifying mixed foods was also one of our main challenges, and our solution was the same 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 classification needed to support this research context.

Other classifications that have been well described are the NOVA classification \(^{(10)}\) and the Convenience Food Classification Scheme \(^{(9)}\). However, these classifications differed quite a lot from our principles, and these classifications would not have suited our purposes to link purchases primarily with health.

We chose to use NRFI to examine how well we succeeded in classifying the data based on dietary quality \(^{(11)}\). There would have been other options to use for nutrient profiling, such as the Grocery Purchase Quality Index-2016 (GPQI-2016) \(^{(12)}\), 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 \(^{(29)}\). However, NRFI 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. We do not claim that NRFI would be any better than other profiling systems, and whatever is chosen will always affect the results. 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.

Of the 11 nutrients that we included in NRFI, intakes of fibre, PUFA and vitamin D have been identified to be relatively low at the population level in Finland \(^{(30)}\). 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 who are in fertile age. Including these nutrients in the NRFI was therefore justified. An interesting question is whether protein should be in the equation. Although protein per se is needed for health, most of the protein in the Finnish diet comes from animal sources \(^{(31)}\). Hence, the environmental impact is not optimal \(^{(32)}\).

**Strengths and limitations**

Our starting point was automatically collected customer loyalty card data, which were provided to us on a grocery product group level. Thus, the most obvious limitation affecting our classification method was that we could not classify on the most detailed level (brand level). This leads to some compromises, as well as making more assumptions of the grocery 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 study, we used 11 nutrients to profile all food groups, but one could have also looked at food groups at Class 1 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. The class ‘Vegetables’ had relatively low NRFI, which does not resemble the true nutritional quality of this group.

In our evaluation of the nutrient quality of the classifications, there are possible weaknesses that need to be mentioned. First, although the NRFI is a well-established method to profile groceries based on their nutrient content, it has its methodological weaknesses. The nutrients that are included in the index are subjectively selected by the researchers who choose to use them. Moreover, ranking of the foods by NRFI varies depending on which nutrients have been selected in the index.

Second, 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, NRFI does not have any upper or lower limits, meaning that the underlying assumption of the index is ‘the more nutrients the better’. This in practice 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 (32).

Conclusion And Recommendations

Our group is one of the few that has carefully described the categorisation of grocery purchase data and among the first ones who have used customer loyalty card data as the basis for the classification. Our classification is freely available to use by other researchers and retailers (https://doi.org/10.5281/zenodo.7781352). Based on nutrient profiling and using the NRFI, the nutritional quality of the food classes was logical. The decrease of variation in nutritional quality in the more detailed classes 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.

We believe 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. In addition, it is recommended that any new classifications are openly presented and shared among the science community. 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 (33).

Declarations

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 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


19. 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).


**Figures**

---

Page 16/20
Figure 1

The LoCard reclassification process. Grocery product groups were reclassified first at the broadest hierarchy level called Class 1. Each food class on the Class 1 level was then subsequently divided to finer sub-classes on the Class 2 level, followed by Class 3 and, finally, Class 4, which was the most detailed level of hierarchy. The number of food classes at each level is given in the blue boxes on the left.

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 Class 1. The middle circle represents Class 2, and the outer circle represents Class 3. Class 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 Class 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.
Figure 4

Variation in Nutrient Rich Food Index values for grocery product groups (grey dots) at different LoCard Food Classification levels. The figure shows an example of the food class ‘Cereal and bakery products’ from Class 1 and the food classes that are located in it at Class 2. Further, the figure shows an example of the food class ‘Breakfast cereals’ from Class 2 and the food classes that are located in it at Class 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.