BMJ Open

Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure

Richard Andrew Harrington, Vyas Adhikari, Mike Rayner, Peter Scarborough

To cite: Harrington RA, Adhikari V, Rayner M, *et al.* Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. *BMJ Open* 2019;**9**:e026652. doi:10.1136/bmjopen-2018-026652

► Prepublication history and additional material for this paper are available online. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2018-026652).

Received 14 September 2018
Revised 1 April 2019
Accepted 2 April 2019

Nuffield Department of Population Health, University of Oxford Nuffield Department of Population Health, Oxford, UK

Correspondence to
Dr Richard Andrew Harrington; richard.harrington@ndph.ox.ac.uk

ABSTRACT

Objectives Traditional methods for creating food composition tables struggle to cope with the large number of products and the rapid pace of change in the food and drink marketplace. This paper introduces foodDB, a big data approach to the analysis of this marketplace, and presents analyses illustrating its research potential.

Design foodDB has been used to collect data weekly on all foods and drinks available on six major UK supermarket websites since November 2017. As of June 2018, foodDB has 3 193 171 observations of 128 283 distinct food and drink products measured at multiple timepoints.

Methods Weekly extraction of nutrition and availability data of products was extracted from the webpages of the supermarket websites. This process was automated with a codebase written in Python.

Results Analyses using a single weekly timepoint of 97 368 total products in March 2018 identified 2699 ready meals and pizzas, and showed that lower price ready meals had significantly lower levels of fat, saturates, sugar and salt ($p<0.001$). Longitudinal analyses of 903 pizzas revealed that 10.8% changed their nutritional formulation over 6 months, and 29.9% were either discontinued or new market entries.

Conclusions foodDB is a powerful new tool for monitoring the food and drink marketplace, the comprehensive sampling and granularity of collection provides power for revealing analyses of the relationship between nutritional quality and marketing of branded foods, timely observation of product reformulation and other changes to the food marketplace.

Strengths and limitations of this study

- foodDB is a new database with greater temporal granularity than any other food composition database in the UK.
- foodDB collects information on over 100 000 products per week.
- Price and promotional information is collected alongside nutritional composition.
- foodDB does not account for geographical availability of foods within the UK.
- foodDB only collects information on products sold on the websites of the UK's major supermarkets.

INTRODUCTION

Much of nutritional epidemiology is dependent on the conversion of questionnaire-based data on food consumption into nutrient consumption by the use of food composition tables.[1] Such methods are often used to estimate the association between nutrient consumption and health outcomes in observational studies[2] and to monitor the nutritional quality of food consumed by a population.[3] However, food composition tables are expensive to construct and maintain, only cover a small sample of the foods available for consumption and are frequently based on old data, which is problematic for processed foods where formulations change regularly.[4] Such limitations can potentially lead to misclassification bias in the observational studies that rely on food composition tables.

A potential way to increase the number of foods included in food composition tables is to include nutritional data taken from food packaging on a large sample of foods.[4] Such a method could collate up-to-date nutritional data on a comprehensive set of foods that are purchased within a specific setting, although limited to only the nutritional data that are provided on food packaging. There have been various attempts to collate such databases: by crowdsourcing food label data using mobile phones (eg, FoodSwitch[5]) or web applications (eg, Open Food Facts[6]); by collecting data through contact with food manufacturers,[7] industry or by periodic audits of foods on the market[8]; and by public–private partnerships aimed at extending national food composition tables.[9–13] However, these databases are limited for research purposes as they do not regularly update nutritional data on products[5 6 9–12]; do not achieve comprehensive coverage of targeted foods[5 6 10–13]; require high levels of resources to maintain and update[5]; do not have transparent methods or adequate audit trails[8]; or rely on ongoing contributions from the food industry.[9–13]

This paper introduces foodDB, a terabyte-scale, weekly updated database that collects data on a comprehensive sample of food and drink products available for purchases in all major UK supermarkets, using big data techniques for collection, processing, storage and analysis. Now commonplace in many fields across the business, public, non-profit and scientific sectors, big data refers to any data exhibiting unusual features of any of five dimensions: volume, variety, velocity, volatility and veracity.[14] foodDB uses big data techniques to address limitations of other food composition tables and allow for a number of types of research including in-depth investigations of correlations between nutritional and commercial variables (eg, price, promotions) at a single point in time, as well as monitoring the food and drink marketplace longitudinally. Such a database could be used to identify important levers for promoting healthy diets, evaluate the impact of population-level public health policies such as the Soft Drink Industry Levy[15] and support monitoring of the nutritional quality of the UK diet by dietary survey, so that changes in nutrient intake can be measured that are due to both change in behaviour and change in the food supply.

The three objectives of this paper are to describe the foodDB database and data collection method; to provide an example of the potential use of foodDB in cross-sectional analyses by analysing the nutritional content of all ready meals and pizzas that were available for purchase in March 2018 in the six major UK supermarkets; and to provide an example of the potential use in longitudinal analyses by analysing the change in availability and healthiness of pizzas over the course of the first 6 months of full data collection.

METHODS

Data collection, processing and storage

foodDB consists of a relational database populated by custom-built software to collect, process and store data on food and drink products available to buy online in the UK, including alcoholic beverages, and not including supplements. A full list of the main categories from which data are collected is provided in the online supplementary material. Data are collected weekly by the foodDB software on all products available in the online offering of UK supermarkets, and are stored in the foodDB database. Each instance of data collection from each supermarket is referred to as a 'snapshot'; for each snapshot, foodDB collects data in a staged way, as illustrated in figure 1. Starting at the supermarket's main webpage, category trees are built, generating and storing a list of category hierarchies. Category hierarchies in all supermarkets can be grouped into four levels of classification, which approximate to 'department', 'section', 'aisle' and sometimes 'shelf'. For example, in Sainsbury's, chilled pizzas are all found under the 'Pizza' aisle: *Chilled (department) -> Pizzas and Garlic Bread (section) -> Pizza (aisle)*. All chilled ready meals in Asda are found under the 'Ready Meals' aisle, although this aisle additionally has a number of shelves to distinguish between different ready meal types: *Chilled Food (department) -> Ready Meals & Soup (section) -> Ready Meals (aisle) -> Indian/Italian & Mediterranean/Traditional/Vegetarian etc. (shelf)*. For each snapshot, foodDB collects detailed category information; each category's product pages are then used to create a list of all product names and URLs (website addresses) for that category; the lists of product–category pairs, and unique product names and product URLs are collected and stored; the URL for each unique product is then used to load each product page, and product data are extracted, processed and stored. The data collected for each product include the following, where available: product name; price; serving size; product size; promotion details; supermarket's own product code (a numeric identifier for that product, different to the barcode and embedded within the webpage); product image; front-of-pack nutrition labelling data (% reference intake and nutrition traffic light labels); nutrient declaration data (referred to as a nutrition table on each website); ingredients; dietary information (eg, 'suitable for vegetarians'); allergen information; storage information; brand; manufacturer; and the date and time of data collection. Each product's supermarket 'product code' is used for identification of and tracking unique products over time, supplemented by product name, URL and image where necessary. Within foodDB, internal unique identifiers are used to identify individual instances of each product, combinations of the product code for a particular snapshot. In addition to these data, the full HTML text of each product page with date and time of data collection is stored separately for audit and data verification purposes, and to provide a mechanism for re-extracting data in the event of errors at the time of data collection, for example, in the case of an unexpected change to the supermarket web page structure.

All collection and processing code is written in Python V.2.7[16] and the foodDB software is object-oriented and modular. A set of core classes and a helper library provide the main functionality for data collection, processing and storage. Subclassing[17] (class inheritance) allows for the same core code to be used for different supermarkets that have different structures and page loading mechanisms, with only small unique sections of code required to handle the specific requirements of each supermarket, rather than an entire custom codebase being required for each. The object-oriented design means that foodDB can easily adapt to and handle changes in individual supermarket websites as they are updated over time, and also allows new data sources (eg, other supermarkets or food suppliers) to be added as required. The well-established open source Python libraries requests[18] and selenium[19] allow foodDB to make large numbers of sequential calls to website servers, each receiving HTML text for processing

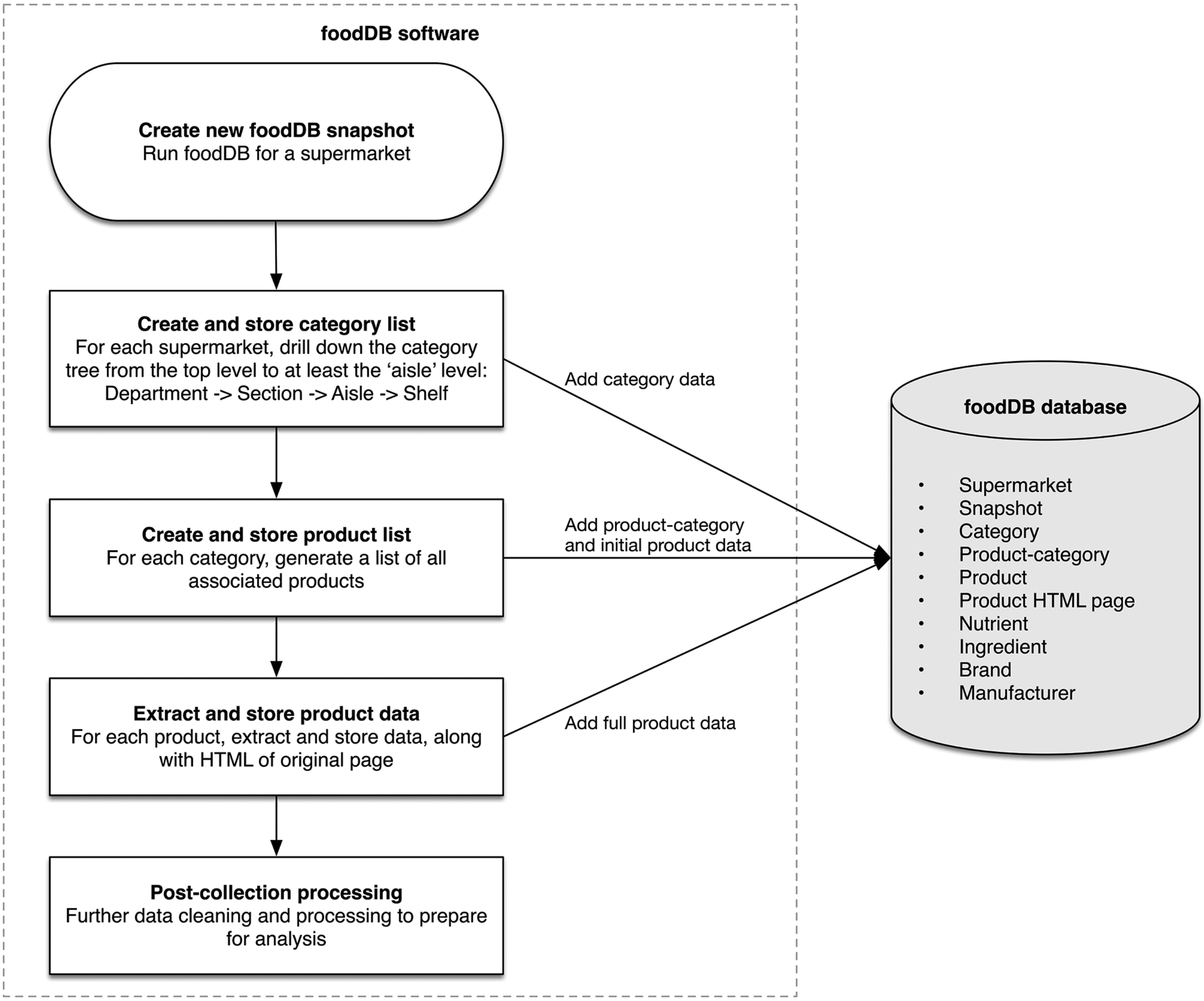
foodDB software
Create new foodDB snapshot
Run foodDB for a supermarket
Create and store category list
For each supermarket, drill down the category tree from the top level to at least the 'aisle' level: Department -> Section -> Aisle -> Shelf
Create and store product list
For each category, generate a list of all associated products
Extract and store product data
For each product, extract and store data, along with HTML of original page
Post-collection processing
Further data cleaning and processing to prepare for analysis
Add category data
Add product-category and initial product data
Add full product data
foodDB database
Supermarket
Snapshot
Category
Product-category
Product
Product HTML page
Nutrient
Ingredient
Brand
Manufacturer

Figure 1 Dataflow of snapshot data generation in foodDB.

and data extraction: requests is used for websites that return plain HTML from a call to a server, while selenium facilitates this process for dynamically generated supermarket webpages, simulating the in-browser page-generation. Errors, for example, caused by pages failing to load, scheduled supermarket website maintenance or individual products not being available to purchase at the time of loading, are caught and reruns are carried out to try to fill in these data gaps. Data collection has inbuilt pauses to stop any excess load on the website servers, which has the added benefit of reducing errors. To reduce errors from occurring, and to allow for efficient maintenance of the foodDB software, a comprehensive testing suite has been created using pytest[20] which identifies errors caused by changes to the website structures, and ensures that new features and fixes perform as required, and do not have any unintended consequences. The foodDB database uses the open source relational database management system MySQL V.5.7[21] with the InnoDB storage engine[22] to allow transactions, row-level locking, foreign key relationships and increased data integrity.

Data analysis

For cross-sectional analysis of ready meals and pizzas, a single weekly snapshot of the six major UK supermarkets (Tesco, Asda, Sainsbury's, Morrisons, Waitrose, Ocado) taken in March 2018 was used. These six supermarkets account for over 75% of the grocery market in the UK.[23] All products available within supermarket categories of ready meals and pizzas were manually curated to exclude products that would not be eaten as a standalone meal (eg, pasta sauces, burgers without buns, garlic bread and so on). In order to define product groups for the example analyses, we first examined the existing supermarket hierarchies to select all potential ready meal and pizza products, and then filtered these using regular expressions to include/exclude specific subcategories and products. The ready-meal product categories selected required particular care due to the fact that they contain a diverse range of products, including whole meals (eg, vegetable masala and rice), components of such meals (eg, a portion of precooked rice) and supplementary items (often within meal deals, eg, small pots of chutney, soft drinks and so on). Duplicate products (ready meals or pizzas of a particular brand and type that appear in more than one supermarket) were removed. Using data extracted from the websites, a traffic light 'healthiness' score was generated for all ready meals and pizzas. This score ranks the perceived healthiness of foods based only on the front-of-pack traffic light colours for total fat, saturated fat, total sugar and salt, based on the results of a choice experiment conducted with supermarket shoppers.[24] We assessed the distribution of each traffic light nutrient and the traffic light healthiness score across the products.

To test the relationship between price and levels of fat, saturated fat, sugar and salt, the full dataset without removal of duplicates was used (since the same product in different supermarkets can have different prices). We categorised the dataset into the lowest 50% and highest 50% of products by price (£ per 100g) and looked at differences in traffic light healthiness scores between the two categories using Mann-Whitney tests.

For the longitudinal analyses, we used data on all pizzas appearing in the six supermarkets between 30 November 2017 and 1 June 2018, which included 27 weekly snapshots (NB: the Tesco dataset contained only 26 snapshots in this time period, due to errors in data collection on the week of 30 December 2018). To evaluate the stability of the pizza market over this time period, all products were categorised into four groups depending on their availability: products present in every collected snapshot; products usually present (defined as being available in all snapshots except gaps of 1 or 2weeks); products with line change (defined either as products entering the market—first time available after the initial snapshot—or leaving the market—last time available before the final snapshot); and products with any other pattern of availability.

To evaluate the degree of product reformulation over this time period, the percentage of pizzas that changed their content of at least one of the four traffic light nutrients (total fat, saturated fat, total sugars and salt) were calculated, as well as the percentage where the change in nutritional content was enough to prompt a change in a traffic light colour for the front-of-pack label.

Analyses were carried out using MySQL and Python V.2.7 using the Python libraries numpy V.1.14[25] and Pandas V.0.22,[26] and visualisations created using Matplotlib V.2.2.[27]

Patient and public involvement

There was no patient or public involvement in this analysis.

RESULTS

Descriptive statistics

Between 30 November 2017 and 5 June 2018, foodDB collected full snapshots for 27weeks of full data collection for six UK supermarkets, consisting of 3 193 171 food or drink records. In one single time point (all snapshots starting on 2 March 2018) foodDB obtained 97368 product records (table 1). Across the course of data collection, we have collected 128283 distinct products measured at multiple time points (in this paper, a 'distinct product' is defined as a unique combination of product and supermarket). Product page data were collected for over 99.5% of products identified from the supermarkets

Table 1 Summary of data collected in a single foodDB snapshot

Supermarket	Aisle categories, N	Products, N	Product pages with data available, N (%)*	Products with ingredients data, N (%)†	Products with a nutrient declaration, N (%)†	Ready meals, N (%)†	Pizzas, N (%)†
Tesco	641	16051	16020 (99.8)	13284 (82.9)	14237 (88.9)	328 (2.0)	90 (0.6)
Sainsbury's	765	15597	15191 (97.4)	12588 (82.9)	12986 (85.5)	350 (2.2)	89 (0.6)
Ocado	750	23956	23948 (100.0)	18929 (79.0)	18308 (76.5)	354 (1.5)	104 (0.4)
Morrisons	648	12946	12932 (99.9)	10339 (80.0)	10467 (80.9)	354 (2.7)	94 (0.7)
Waitrose	713	13614	13614 (100.0)	10839 (79.6)	13509 (99.2)	294 (2.2)	68 (0.5)
Asda	286	15677	15663 (99.9)	12789 (81.7)	14080 (89.9)	387 (2.5)	187 (1.1)
Total	3803	97841	97368 (99.5)	78768 (80.9)	83587 (85.9)	2067 (2.1)	632 (0.6)

Summary of numbers of aisle categories and products, and individual product data types collected by foodDB in the first week of March 2018.
*Base for the percentage is the number of products listed in categories.
†Base for the percentage is the available product pages.

at the single timepoint. Five of the six online supermarkets contained 13000–16 000 food and drink products. The exception was Ocado (an online only retailer) that held just under 24000 products. Data on ingredients were captured for >80% of the food and drinks, and nutrient declaration tables[28] were captured in over 85% of products. Data for energy, protein, carbohydrate, fat, sugar, salt and saturates were present in over 90% of these tables, data for fibre in 70%, while data for other nutrients were present in fewer than 7% of tables. A full list of nutrients reported and stored in the foodDB snapshot at this single timepoint can be found in the online supplementary material.

A total of 140 aisle categories across both fresh and frozen sections of the six supermarkets were identified as either 'Ready Meal' or Pizza in foodDB for the 2 March 2018 snapshots, which contained 2699 ready meals and pizzas, comprising between 1.9% and 3.7% of each supermarket's product range. Excluding products without complete nutrition and price data, and clustering to remove duplicate resulted in a dataset of 2139 unique ready meals and pizzas. One thousand five hundred and eighty-four (74.1%) of these clustered products were supermarket own-brand, and 555 (25.9%) were branded products. All products had unique supermarket product codes, while 7% of products had minor name changes, and 9% had URL changes. A dataflow with the numbers of products at each stage of filtering and classification is available the online supplementary material.

Nutritional analyses

Figure 2 shows the distribution of total fat, saturated fat, total sugars and salt per 100 g across all ready meals and all pizzas. Boundaries for classification of traffic light levels[29] are illustrated with vertical green and red lines superimposed on the graphs, with the proviso that a product may also have a 'red' classification for a particular nutrient due to the per-serving value.

Figure 3 shows the calculated traffic light colours associated with ready meals and pizzas and quantifies what might be expected of those products: for example, that most ready meals and pizzas are low in sugar (91.6% (90.5%–92.7%) of these products have sugar levels that would qualify for a green traffic light). The derived traffic light healthiness scores are scaled from 0 (least healthy, for a product with four red lights) to 1 (most healthy, for a product with four green lights)—ready meals and pizzas showed a reasonably uniform distribution over this score (see online supplementary material) indicating that

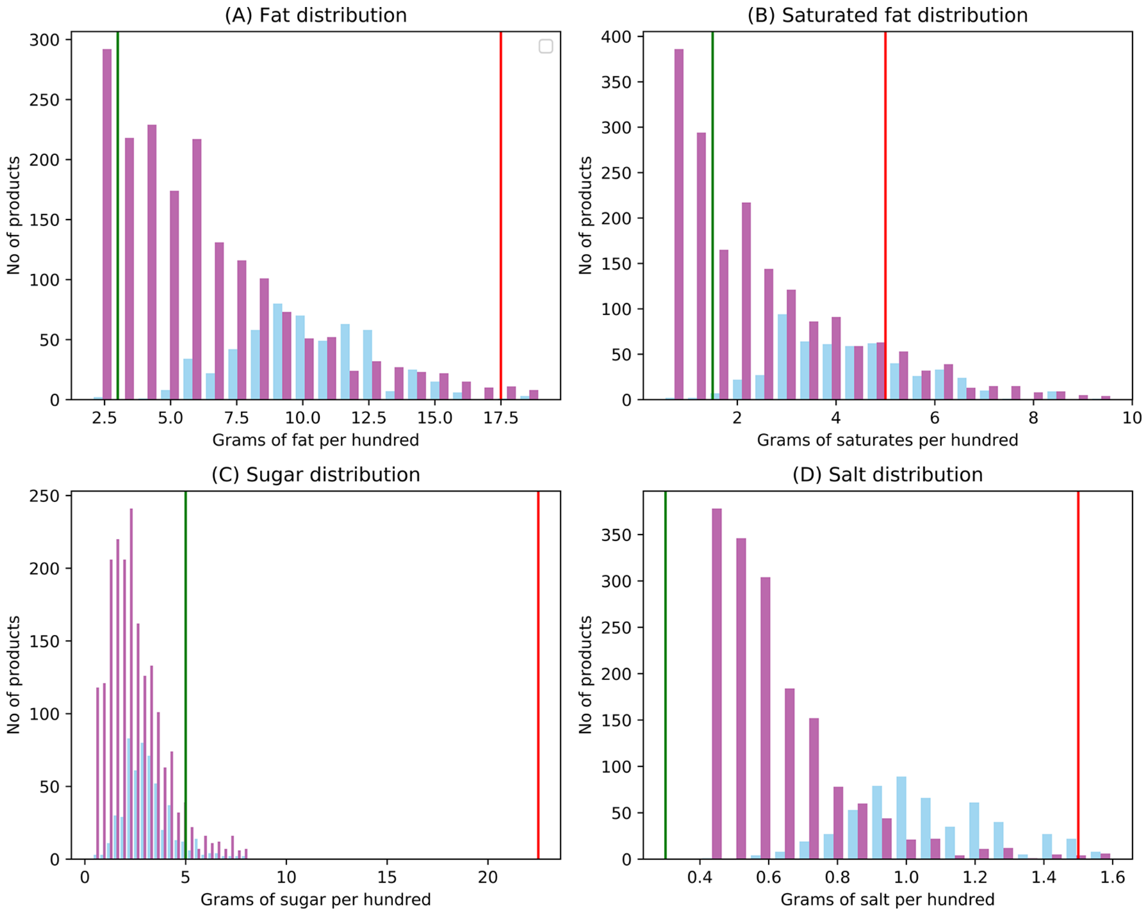

Figure 2 Distribution of nutrients across all ready meals and pizzas at a single timepoint. Distribution of grams of (A) fat, (B) saturated fat, (C) total sugars, (D) salt per 100 g across all ready meals and pizzas in a single week of foodDB snapshots. Vertical lines illustrate the 100 g value limits for calculation of green and red traffic light labels.

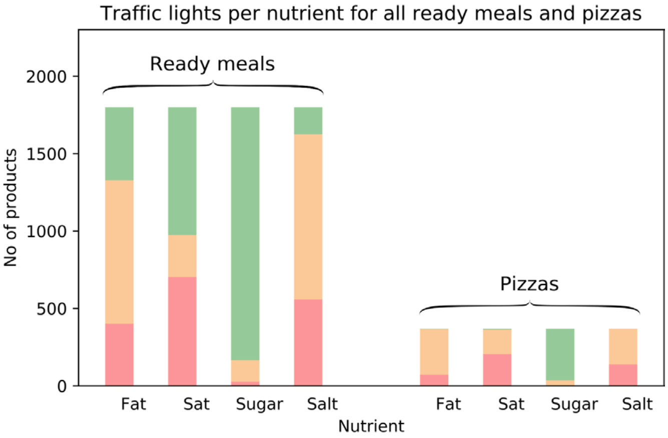

Figure 3 Distribution of traffic light colours across all ready meals and pizzas at a single timepoint. Distribution of traffic light colours across all ready meals and pizzas in a single week of foodDB snapshots.

there is substantial choice available to purchasers in this food category.

Correlations between price and levels of fat, saturated fat, sugar and salt

The median price for ready meals was 75p per 100g and for pizzas was 65p per 100g. These values were used to classify products into two groups, lower cost and higher cost. The distribution of price is shown in the online supplementary material. Table 2 compares the levels of fat, saturated fat, sugar and salt between low-cost and high-cost ready meals and pizzas, using all 2033 ready meals and 534 pizzas that had both nutrition and price data (2567 products in total, including duplicate products from different supermarkets). For ready meals, the lower price products had significantly lower quantities of all four traffic light nutrients (p<0.001). For pizzas, there was no difference in fat, saturated fat and sugar levels, but lower price pizzas tended to have lower salt levels (p<0.001).

Longitudinal analysis of healthiness of all pizzas

The dataset contained 903 distinct pizzas over all six supermarkets, of which 43.3% (40.1%–46.5%) were found to be available in every week over the 6months. There was considerable churn in the pizza marketplace, shown in table 3, with 3 in 10 (29.9% (26.9%–32.9%)) pizzas over the 6months either being discontinued or introduced as a new (or returning) product. Figure 4 illustrates changes in product availability over time using heatmaps for each supermarket, with rows representing products and columns representing weekly snapshots; each square represents either presence (sand-coloured) or absence (black) of a product in a snapshot.

As shown in table 4, changes to the nutritional composition of 10.8% (8.6%–13.0%) of pizzas were observed over 6months. Over a third of the changes resulted in a change to the (calculated) front-of-pack label traffic light colours for the product.

DISCUSSION

Using automated techniques to collect data from online supermarkets can result in food composition tables with far greater coverage and temporality than have been achieved in the past, allowing for more detailed evaluation of the grocery marketplace. Such granularity can reveal insights about the constantly changing set of products available in the UK marketplace (our example case of pizzas showed that 3 in 10 products were either discontinued or introduced to product ranges over just 6months), and the rapid rate of reformulation within the marketplace (over 1 in 10 pizzas changing nutritional content within the space of 6months). The greater

Table 2 Relationship between price (£ per 100g) and levels of fat, saturated fat, sugar and salt for 2567 ready meals and pizzas

		Low price*, median (IQR)	High price*, median (IQR)	P for difference†
Ready meals (n=2033)	Total fat (g per 100g)	3.9 (2.4–5.9)	6.0 (3.5–9.0)	<0.001
	Saturated fat (g per 100g)	1.4 (0.7–2.5)	2.1 (0.9–4.0)	<0.001
	Sugar (g per 100g)	2.1 (1.5–3.0)	2.4 (1.5–3.7)	<0.001
	Salt (g per 100g)	0.49 (0.4–0.6)	0.56 (0.44–0.70)	<0.001
	Healthiness score‡	0.70 (0.40–0.90)	0.65 (0.40–0.80)	<0.001
Pizzas (n=534)	Total fat (g per 100g)	10.0 (8.5–12.0)	9.9 (8.5–11.8)	0.152
	Saturated fat (g per 100g)	4.0 (3.1–5.1)	4.3 (3.2–5.1)	0.138
	Sugar (g per 100g)	3.0 (2.3–3.7)	2.9 (2.3–3.7)	0.405
	Salt (g per 100g)	1.0 (0.89–1.18)	1.1 (0.94–1.23)	<0.001
	Healthiness score‡	0.55 (0.40–0.70)	0.55 (0.40–0.70)	0.416

*Low-price and high-price ready meals and pizzas are split at the median price of 75p per 100g for ready meals and 69 p per 100g for pizzas.
†Derived from Mann-Whitney tests.
‡Index derived directly from the colours attached to the traffic light label for total fat, saturated fat, sugar and salt.

Table 3 Changes in pizza ranges for each supermarket between 30 November 2017 and 1 June 2018

Supermarket	Observations, N	Distinct pizza products, N	Pizzas always present, % (95% CI)*	Pizzas usually present, % (95% CI)*	Pizzas with line changes, % (95% CI)*	Pizzas with other availability pattern, % (95% CI)*
Tesco	2798	143	27.3 (20.0% to 34.6%)	24.5 (17.4% to 31.5%)	38.5 (30.5% to 46.4%)	9.8 (4.9% to 14.7%)
Sainsbury's	2411	132	18.2 (11.6% to 24.8%)	18.2 (11.6% to 24.8%)	25.8 (18.3% to 33.2%)	37.9 (29.6% to 46.2%)
Ocado	3133	123	61.8 (53.2% to 70.4%)	5.7 (1.6% to 9.8%)	31.7 (23.5% to 39.9%)	0.8 (–0.8% to 2.4%)
Morrisons	3294	143	35.0 (27.1% to 42.8%)	22.4 (15.5% to 29.2%)	24.5 (17.4% to 31.5%)	18.2 (11.9% to 24.5%)
Waitrose	2442	83	62.7 (52.2% to 73.1%)	12.0 (5.0% to 19.1%)	24.1 (14.9% to 33.3%)	1.2 (–1.1% to 3.6%)
Asda	6040	279	53.8 (47.9% to 59.6%)	5.4 (2.7% to 8.0%)	31.2 (25.7% to 36.6%)	9.7 (6.2% to 13.1%)
All	20118	903	43.3 (40.1% to 46.5%)	13.6 (11.4% to 15.9%)	29.9 (26.9% to 32.9%)	13.2 (11.0% to 15.4%)

*Base for the percentage is the number of distinct pizza products. 'Usually present' defined as being available in all snapshots except for gaps of 1 or 2 weeks; 'line change' defined either as products entering the market—first time available after the initial snapshot—or leaving the market—last time available before the final snapshot.

coverage allows for a clearer description of the nutritional quality of foods available in the marketplace, and an assessment of the association between nutritional quality and key variables that affect purchasing behaviour, such as price. Analyses of this large and dynamic dataset can reveal insights such as the differences in level fat, saturated fat, sugar and salt between lower-priced and higher-priced ready meals, and of the variability of available products, and changes in their composition over time, as illustrated here.

The automated data collection process and the regularity of data capture provides foodDB with distinct analytical advantages compared with other large food composition databases of branded foods. The two FoodSwitch databases, which were initially primed by a commercial database and then supplemented with data collected from crowdsourcing currently contain ~100 000 and 60 000 food and drink items for the UK and Australia, respectively.[5] Also using crowdsourcing, Open Food Facts, is a free-to-use web application that contains data on over 700 000 products worldwide since 2012, including over 300 000 from France and 150 000 from the USA.[6] foodDB collects data on over 90 000 food and drink items every week, collating >125 000 distinct food and drink items available at some point over the 6-month data collection period (and over three million observations of food or drink items in total). A significant advantage of foodDB over crowdsourcing is that foodDB is updated systematically and consistently. This ensures that new products, discontinued products or product changes (eg, to formulation, price or promotion) are captured.

Another way to compile branded food datasets is to build partnerships with the food industry: such datasets are available in the USA[9–12] and Belgium[13] and rely on co-operation from the food industry to update data occasionally (eg, the USDA Branded Food Dataset is updated annually[9]). The granularity of such datasets is not sufficient to capture the dynamic food market, where products are discontinued, new products emerge and old products are reformulated frequently, for example, in the USA it has been estimated that 20 000 new food products emerge each year, of which only 10% are still available after 3 years[4] and we show that 30% of pizzas available at some point in a 6-month period are new or discontinued products. This churn in pizza ranges is also reflected in the numbers of total distinct products collected; while we have collected 128 283 distinct products over 6 months, each single weekly snapshot contains only ~95 000 products. A further advantage of foodDB over such datasets is that it is capable of capturing data on other variables that define the food purchasing environment, such as price, promotion and labelling.

foodDB is not the first project automatically to collect and store data from UK online supermarkets. This approach is also used by price comparison websites (eg, MySupermarket[30]), market research companies (eg, BrandView[31]) and by Internet Archive[32]—a non-profit organisation that collects and stores webpages, currently holding over 279 billion pages including many from online supermarkets between 2011 and the present day. The datasets built by price comparison websites are not available to researchers, and while market research databases are made available commercially, datasets may be incomplete, contain inconsistent data, lack audit trails and details on their compilation are unclear. The Internet Archive has a substantial archive of supermarket

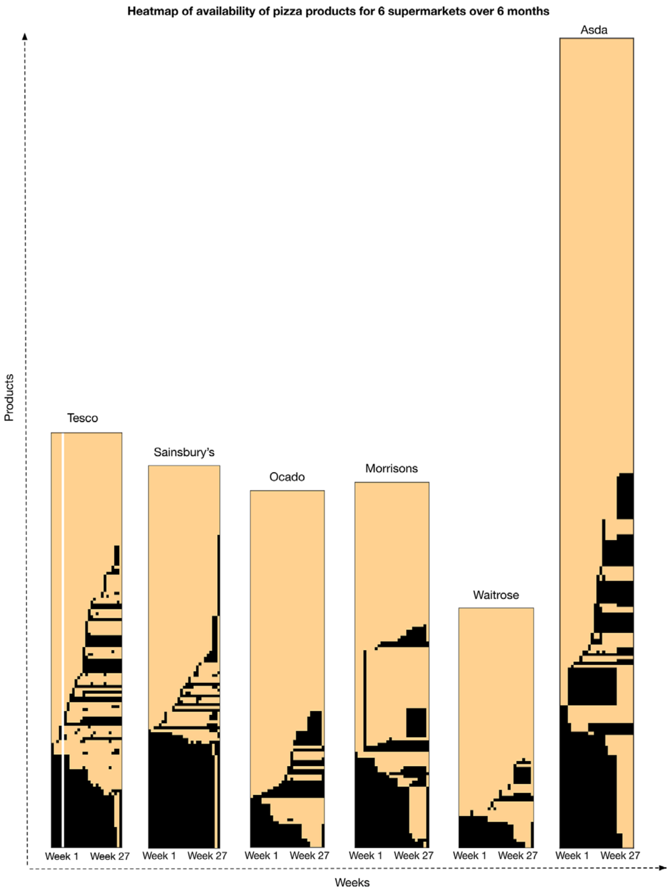

Figure 4 Heatmap of availability of pizza products for six supermarkets over 6 months. Heatmap illustrating changes in pizza product range for six supermarkets over 6 months (27 weeks) of foodDB data collection. Columns represent weekly snapshots, and rows represent a product's availability in that snapshot. Each sand-coloured square shows that the product is available, while each black-coloured square means that product is not available in that snapshot.

webpages from 2011 to present; however, it is far from comprehensive, with HTML pages for many products only collected once, if at all. As a proof of concept, the foodDB codebase has been adapted to collect data on all soft drinks from UK online supermarkets collected in the Internet Archive, using foodDB functionality to process, clean and store individual product details. The usage of such data from commercial websites for research purposes, such as in foodDB, is covered by a document published by the UK Government's Intellectual Property Office in October 2014 entitled 'Exceptions to copyright',[33] which contains a section on 'Text and data mining for non-commercial research'. This section of the document states ' An exception to copyright exists which allows researchers to make copies of any copyright material for the purpose of computational analysis if they already have the right to read the work (that is, they have 'lawful access' to the work)'.

A core strength of foodDB is its flexibility to respond to an ever-changing online supermarket environment. The foodDB software is object-oriented and modular. A core set of classes provides the mechanisms for collection, processing and storage of all data, while subclasses allow for the same code to be easily adapted and modified using inheritance to deal with individual differences between supermarket websites. This allows for minimal interruption to data collection when individual supermarket websites change, and also allows for new data sources/ online providers to be added as required. While only the full datasets for six supermarkets over 6 months are reported here, full or partial data have also been collected for four other online supermarkets/suppliers Iceland, Marks and Spencer, Tesco Ireland and Cook—using this same core codebase and work is in progress to incorporate these and others into foodDB. In addition to intracountry analyses, this expansion of foodDB will allow for in-depth analyses of food and drink marketplaces between countries.[34] foodDB is able to extract ingredient and nutrient data for a high percentage of products, although we currently make the assumption that data provided online is correct. This assumption is supported by the requirement in the Consumer Rights Act 2015[35 36] which requires companies in the UK ensure that goods ordered online must be as described, fit for purpose, and of satisfactory quality. Further, European Union (EU) legislation requires that labelling is consistent between online purchases and buying in-store.[37] While the European Union legislation also requires reporting of certain nutrients, this is, not an exhaustive list, and is reflected by the fact that the foodDB snapshots reported are able to report data extraction of major macronutrients for over 90% of products, but under 10% for other nutrients. In addition to monitoring and analysis of the food and drink marketplace, rather than replacing databases currently used for dietary surveys, we believe foodDB is an important data source for providing up-to-date macronutrient data. We are also investigating the potential for appending averaged or estimated micronutrients in order to enhance foodDB for such research purposes.

Capturing local geographical variability of food and drink availability within individual online supermarkets is a current limitation of foodDB. Initial analyses have shown that some products are only available for purchase in specific regions of the UK, as well as showing regional differences in price and promotions of some products. foodDB currently collects data on the core set of products only, and future work will address these regional variations.

At present, the size of foodDB means that conceptually straightforward tasks (eg, linking products sold in different sizes, or different brands of the same product) require considerable resources to be completed manually, which restricts the scope of possible analyses. For example, while it was possible to classify and verify a single snapshot of ready meals in foodDB for the cross-sectional analyses reported here through a combination of computational and manual methods, the resources required to process a large number of snapshots for the longitudinal study was prohibitive. The pizza categories, like

Table 4 Changes to nutritional composition of pizzas between 30 November 2017 and 1 June 2018

Supermarket	Pizza products with nutritional composition information*, N (%)†	Products that changed nutritional content during data collection period, % (95% CIs)‡					
		Total fat	Saturated fat	Sugar	Salt	Any traffic light nutrient	Any change in traffic light colour§
Tesco	143 (100.0)	4.9 (1.4% to 8.4%)	4.9 (1.4% to 8.4%)	6.3 (2.3% to 10.3%)	5.6 (1.8% to 9.4%)	7.7 (3.3% to 12.1%)	3.5 (0.5% to 6.5%)
Sainsbury's	126 (95.5)	15.1 (8.8% to 21.3%)	14.3 (8.2% to 20.4%)	14.3 (8.2% to 20.4%)	15.9 (9.5% to 22.3%)	15.9 (9.5% to 22.3%)	7.9 (3.2% to 12.7%)
Ocado	123 (100.0)	10.6 (5.1% to 16%)	10.6 (5.1% to 16.0%)	12.2 (6.4% to 18.0%)	12.2 (6.4% to 18.0%)	13.0 (7.1% to 19.0%)	3.3 (0.1% to 6.4%)
Morrisons	126 (88.1)	7.9 (3.2% to 12.7%)	8.7 (3.8% to 13.7%)	7.9 (3.2% to 12.7%)	6.3 (2.1% to 10.6%)	10.3 (5.0% to 15.6%)	3.2 (0.1% to 6.2%)
Waitrose	83 (100.0)	7.2 (1.7% to 12.8%)	7.2 (1.7% to 12.8%)	9.6 (3.3% to 16.0%)	10.8 (4.2% to 17.5%)	10.8 (4.2% to 17.5%)	7.2 (1.7% to 12.8%)
Asda	165 (59.1)	7.3 (3.3% to 11.2%)	6.1 (2.4% to 9.7%)	7.9 (3.8% to 12.0%)	6.7 (2.9% to 10.5%)	8.5 (4.2% to 12.7%)	1.2 (–0.5% to 2.9%)
All	766 (84.8)	8.7 (6.7% to 10.7%)	8.5 (6.5% to 10.5%)	9.5 (7.5% to 11.6%)	9.3 (7.2% to 11.3%)	10.8 (8.6% to 13.0%)	4.0 (2.7% to 5.4%)

*Product contained nutritional composition information on all of total fat, saturated fat, total sugars and salt on either the front of pack label, or nutrition table.
†Base for the percentage is distinct pizza products (see table 3).
‡Base for percentage is pizza products with nutritional info.
§Nutritional composition change resulted in change of at least one traffic light colour on front-of-pack label.

others such as cereals and soft drinks, are well defined, and require little manual processing, which allowed us to conduct longitudinal analyses on this category. Work is ongoing to applying machine learning techniques to the processing stage of foodDB, which would allow for automatic mapping of categories and subcategories in a validated foodDB hierarchy, allowing for easier comparison across ranges and time. This work will also allow for the automatic identification of duplicate products across supermarkets, for example, the same branded soft drink available for purchase in multiple stores. Including these features will require a degree of manual classification in order to train a classification model, but will then run automatically, removing the need for human involvement in all but edge cases. Further improvements to foodDB also includes increasing the number of derived variables such as the Food Standards Agency (FSA)/Ofcom nutrient profiling score used to distinguish between foods that can and cannot be marketed to children.[38] Creating scores with the FSA model requires data from the nutrient composition table, and a measure for the fruit, nut and vegetable (FNV) content for each product, extracted from the ingredient list. The formatting of ingredient lists is highly inconsistent, and often the FNV data are not included, thus making calculation of the FSA score problematic. Current work on foodDB will improve the ingredient parsing to extract FNV data wherever possible, and where there are missing data, use an appropriate measure, for example, the k-nearest neighbour algorithm, to estimate the FNV value and thus calculate the product's FSA score. Incorporating a method for systematically calculating these scores will allow us to more accurately assess nutritional quality, across and between large numbers of products over time. foodDB is a methodological step forward for food composition databases, which are the bedrock of nutritional epidemiology. The first 6 months of data collection have demonstrated that automatically scraping data from online supermarkets can produce food composition databases with sufficient accuracy, transparency, granularity, flexibility and regularity to monitor a highly dynamic food and drink marketplace, to reveal important relationships between food marketing and nutrition and to support measurements of dietary quality over time that incorporate changes in both food consumption and the nutritional composition of commonly consumed branded foods.

Acknowledgements Tesco data were sourced from the Tesco supermarket website at https://www.tesco.com/groceries/. Sainsbury's data were sourced from the Sainsbury's supermarket website at https://www.sainsburys.co.uk/shop/gb/groceries. Asda data were sourced from the Asda supermarket website at https://groceries.asda.com/. Waitrose data were sourced from the Waitrose supermarket website at https://www.waitrose.com/ecom/shop/browse/groceries. Morrisons data were sourced from the Morrisons supermarket website at https://groceries.morrisons.com/. Ocado data were sourced from the Ocado online supermarket website at https://www.ocado.com/.

Contributors RAH contributed to conception and design of the project and research questions, architecture and development of the software and database, statistical analysis and interpretation of data, drafting the manuscript, obtaining funding. PS contributed to design of the project and research questions, statistical analysis and interpretation of data, drafting of the manuscript, obtaining funding. VA contributed to data processing and analysis, interpretation of data, critical revision of the manuscript. MR contributed to interpretation of data, critical revision of the manuscript, obtaining funding, supervision. All authors have approved the submitted manuscript.

Funding This work was supported by a pump priming research grant from the Nuffield Department of Population Health at the University of Oxford, and is currently supported by the NIHR Biomedical Research Centre at Oxford. PS is supported by a British Heart Foundation fellowship (FS/15/34/31656). MR is supported by the British Heart Foundation (006/PSS/CORE/2016/OXFORD).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Anonymised datasets used in the analyses are available upon request. If you believe foodDB data would be of value to your research, please get in touch with the authors via email. Use of foodDB data is only possible for non-commercial purposes.

REFERENCES

1. Food and Agriculture Organisation (FAO). International Network of Food Data Systems (INFOODS): International food composition table / database directory. http://www.fao.org/infoods/infoods/tables-and-databases/en/ (Accessed 22ndJun 2018).
2. International Agency for Research on Cancer (IARC)/World Health Organization (WHO). EPIC study. http://epic.iarc.fr/about/background.php (Accessed 22ndJun 2018).
3. Roberts C, Steer T, Maplethorpe N, *et al*. *National diet and nutrition survey results from years 7 and 8 (combined) of the rolling programme (2014/2015 to 2015/2016)*. London: NatCen, Public Health England and Food Standards Agency, 2018.
4. Black R. Begin to imagine: Thoughts and considerations following the 39th NNDC. *Journal of Food Composition and Analysis* 2017;64:143–4.
5. Dunford EK, Neal B. FoodSwitch and use of crowdsourcing to inform nutrient databases. *Journal of Food Composition and Analysis* 2017;64:13–17.
6. Open Food Facts. Methodology. https://world.openfoodfacts.org/ (Accessed 19thJan 2019).
7. Access to Nutrition Index (ATNI). Harnessing the power of the private sector to tackle the world's biggest nutrition challenges. https://www.accesstonutrition.org/ (Accessed 17thJul 2018).
8. Kantar Worldpanel. Consumer panel insights in a wide range of sectors. https://www.kantarworldpanel.com/en/Consumer-Panels-(Accessed 22ndJun 2018).
9. Kretser A, Murphy D, Starke-Reed P. A partnership for public health: USDA branded food products database. *Journal of Food Composition and Analysis* 2017;64:10–12.
10. Perrin C, Battisti C, Chambefort A, *et al*. Range of processed foods available in France and nutrition labelling according to the type of brand. *Journal of Food Composition and Analysis* 2017;64:156–62.
11. Spiteri M, Soler LG. Food reformulation and nutritional quality of food consumption: an analysis based on households panel data in France. *Eur J Clin Nutr* 2018;72:228–35.
12. Goglia R, Spiteri M, Ménard C, *et al*. Nutritional quality and labelling of ready-to-eat breakfast cereals: the contribution of the French observatory of food quality. *Eur J Clin Nutr* 2010;64(Suppl 3):S20–5.
13. Seeuws C. Belgian branded food products database: inform consumers on a healthy lifestyle in a public-private partnership. *Journal of Food Composition and Analysis* 2017;64:39–42.
14. Fuller D, Buote R, Stanley K. A glossary for big data in population and public health: discussion and commentary on terminology and research methods. *J Epidemiol Community Health* 2017;71:1113–7.
15. National Institute of Health Research. Evaluation of the health impacts of the UK Treasury Soft Drinks Industry Levy (SDIL). https://www.journalslibrary.nihr.ac.uk/programmes/phr/1613001/#/ (Accessed 7thJun 2018).
16. Python Software Foundation. Python programming language. https://www.python.org/ (Accessed 19thJune 2018).
17. Python Software Foundation. Classes. https://docs.python.org/2/tutorial/classes.html (Accessed 19th Jun 2018).
18. Reitz K, Cordasco I, Prewitt N. Requests: HTTP for humans. http://docs.python-requests.org/en/master/ (Accessed 19thJun 2018).
19. Muthukadan B. Selenium with python. https://selenium-python.readthedocs.io/ (Accessed 19thJun 2018).
20. Krekel H. pytest documentation. Release 3.6. https://media.readthedocs.org/pdf/pytest/latest/pytest.pdf (accessed 19thJune 2016).
21. MySQL. The world's most popular open source database. https://www.mysql.com/ (Accessed 19thJun 2018).
22. Oracle. MySQL 8.0 reference manual: chapter 15 the InnoDB storage engine. https://dev.mysql.com/doc/refman/8.0/en/innodb-storage-engine.html (Accessed 19thJun 2018).
23. Worldpanel K. Grocery market share (12 weeks ending 20th May 2018). https://www.kantarworldpanel.com/en/grocery-market-share/great-britain (Accessed 19th Jun 2018).
24. Scarborough P, Matthews A, Eyles H, *et al*. Reds are more important than greens: how UK supermarket shoppers use the different information on a traffic light nutrition label in a choice experiment. *Int J Behav Nutr Phys Act* 2015;12:151.
25. NumPy. The fundamental package for scientific computing with Python. http://www.numpy.org/ (Accessed 14thJul 2018).
26. Pandas. An open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. https://pandas.pydata.org/ (Accessed 14thJul 2018).
27. Matplotlib. A Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. https://matplotlib.org/ (Accessed 14thJul 2018).
28. Rayner M, Wood A, Lawrence M, *et al*. Monitoring the health-related labelling of foods and non-alcoholic beverages in retail settings. *Obes Rev* 2013;14 Suppl 1:70–81.
29. Department of Health. Guide to creating a front of pack (FoP) nutrition label for pre-packed products sold through retail outlets. 2013 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207588/FINAL_VERSION_OF_THE_2013_FOP_GUIDANCE_-_WEB.pdf (Accessed 22nd Jun 2018).
30. MySupermarket. Compare supermarkets and save money with MySupermarket. Shop for groceries, household, health & beauty and more. http://www.mysupermarket.co.uk/ (Accessed 16thJul 2018).
31. BrandView. The leading provider of real-time price and promotion tracking. Measure and manage your price position and communicate value to shoppers. http://www.brandview.com/ (Accessed 16thJul 2018).
32. Archive. About the internet archive. https://archive.org/about/ (Accessed 25thJun 2018).
33. GOV.UK. Exceptions to copyright. https://www.gov.uk/guidance/exceptions-to-copyright (Accessed 19thJan 2019).
34. Coyne KJ, Baldridge AS, Huffman MD, *et al*. Differences in the sodium content of bread products in the USA and UK: implications for policy. *Public Health Nutr* 2018;21:632–6.
35. Citizens Advice. The consumer rights Act 2015. https://www.citizensadvice.org.uk/about-us/how-citizens-advice-works/citizens-advice-consumer-work/the-consumer-rights-act-2015/ (Accessed 19thJan 2019).
36. GOV.UK. Consumer Rights Act 2015. https://www.legislation.gov.uk/ukpga/2015/15/contents/enacted (Accessed 19th Jan 2019).
37. European Commission. Food information to consumers - legislation - Food Safety. https://ec.europa.eu/food/safety/labelling_nutrition/labelling_legislation_en (Accessed 19thJan 2019).
38. Department of Health. Nutrient profiling technical guidance. Department of Health: London, 2011. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216094/dh_123492.pdf (Accessed 25thJun 2018).