REGULAR ARTICLE   **Open Access**

Large-scale and high-resolution analysis of food purchases and health outcomes

Luca Maria Aiello[1]* , Rossano Schifanella[2], Daniele Quercia[1,3] and Lucia Del Prete[4]

*Correspondence:
lajello@gmail.com
[1] Nokia Bell Labs, Cambridge, UK
Full list of author information is
available at the end of the article

Abstract

To complement traditional dietary surveys, which are costly and of limited scale, researchers have resorted to digital data to infer the impact of eating habits on people's health. However, online studies are limited in resolution: they are carried out at country or regional level and do not capture precisely the composition of the food consumed. We study the association between food consumption (derived from the loyalty cards of the main grocery retailer in London) and health outcomes (derived from publicly-available medical prescription records of all general practitioners in the city). The scale and granularity of our analysis is unprecedented: we analyze 1.6B food item purchases and 1.1B medical prescriptions for the entire city of London over the course of one year. By studying food consumption down to the level of nutrients, we show that nutrient diversity and amount of calories are the two strongest predictors of the prevalence of three diseases related to what is called the "metabolic syndrome": hypertension, high cholesterol, and diabetes. This syndrome is a cluster of symptoms generally associated with obesity, is common across the rich world, and affects one in four adults in the UK. Our linear regression models achieve an R^2 of 0.6 when estimating the prevalence of diabetes in nearly 1000 census areas in London, and a classifier can identify (un)healthy areas with up to 91% accuracy. Interestingly, healthy areas are not necessarily well-off (income matters less than what one would expect) and have distinctive features: they tend to systematically eat less carbohydrates and sugar, diversify nutrients, and avoid large quantities. More generally, our study shows that analytics of digital records of grocery purchases can be used as a cheap and scalable tool for health surveillance and, upon these records, different stakeholders from governments to insurance companies to food companies could implement effective prevention strategies.

Keywords: Nutrition; Diabetes; Hypertension; Cholesterol; Metabolic syndrome; Digital purchase records; Grocery; Loyalty card; London

1 Introduction

More than 300k premature deaths in Europe are caused by obesity [1]. In the United Sates, 36% of adults and 17% of children are not just overweight but obese [2]. In UK one in four adults is obese [3] and it is estimated that more than half of European citizens will be obese by 2050. Obesity has long term costs. It raises the risks of diabetes and heart diseases, which result in increased health-care spending (70B of Euros in Europe every year), and ultimately cost lives.

Healthy eating is one of the most effective intervention to counter such risks [4]. In the developing world, people can now afford to eat more food, particularly processed food high in fat and sugar. Monitoring dietary habits of people and persuading them to eat better and exercise more ranks high on the lists of priorities for governments around the world.

Factors associated with food-related disorders are hard to untangle. On top of that, many studies about dietary habits rely on data limited in scale. To partly address the lack of data, computer scientists have recently resorted to the Web. They have analyzed nutrition sites containing food recipes across the world [5], and food images posted on social media [6, 7], and they have done so to infer what Web users are likely to eat. Approaches based on this type of data either suffer from limited spatial resolution (they provide reliable estimates of food consumption but do so at a geographic resolution no lower than city level) or, given their biases, cannot capture reliably what actually people eat.

To further those studies, we explore fine-grained associations between food purchases and disease prevalence at the level of Middle Super Output Areas (MSOA) for the entire city of London. To do that, we analyze, for the first time, the purchases derived from the loyalty cards of the main grocery retailer in the country and match them with the prevalence of the three main medical conditions associated with what is called the "metabolic syndrome". "Syndrome" is the medical term for a collection of symptoms whose common cause is not properly understood, and the "metabolic syndrome" is a cluster of (obesity) symptoms that includes hypertension, cholesterol, and diabetes. This syndrome is common in rich countries and affects one in four adults in the UK. The prevalence of these symptoms is derived from prescription data made publicly available by all the general practitioners in the city. In so doing, we make four main contributions:

- Based on the literature, we formulate seven main research questions that relate food consumption to the three diseases, and operationalize six metrics to answer them (Sect. 3).
- We combine two sets of geo-referenced data in London (Sect. 4). One set contains records of every single food item customers in bought in 2015 at the largest grocery retailer in the country and the second set contains every single medical prescription written in 2016 by all London's General Practitioners (medical doctors or, simply, GPs).[a] From the anonymized and aggregated food purchased data, we extract food nutrients. From the publicly available prescription data, we infer the prevalence of the three ailments: hypertension, cholesterol, and diabetes. All this data will be made available on the project's site.[b]
- We find that, on average, Londoners' diet does not meet the official recommendations of the World Health Organization (Sect. 5), in that, fat and sugar are more prevalent than what the health organization's guidelines would recommend. We also learn that the prevalence of the three medical conditions is related to two features, that is, it is related to food quantities (overall calories) and inversely related to nutrient diversity. These features are not only descriptive but also predictive: the R^2 of a linear regression that predicts diabetes in London neighborhoods reaches 0.6; also, a binary classifier can identify (un)healthy areas from their food consumption with an accuracy as high as 91%.

- We conclude by showing how our methodology and findings might improve evidence-based public outreach initiatives and inform the design of new consumer technologies (Sect. 6).

2 Related work

It is assumed that people are able to freely choose what they eat, yet that does not entirely reflect reality. The amount of food consumed by an individual is also influenced by: (i) the habits and associations around food formed at a young age [8]; (ii) external factors such as portion sizes, food cost and availability [9]; and (iii) biological factors (e.g., upon the consumption of sugary food, the brain releases dopamine, a chemical that signals pleasure and is involved in drug addiction [10]). To produce medical evidence on which factors impact what people eat, the gold standard is represented by randomized controlled clinical trials. Such trials are typically limited in scale and may produce conflicting results (e.g., a meta-review found that most common foods are linked to both a higher and lower risk of cancer [11]).

The availability of large datasets now makes it possible to study health outcomes at unprecedented scales [12]. However, health records of individual patients are not widely available, not least because of privacy concerns. By contrast, a variety of Web sites have been proven to be a good source of data for public health surveillance: search query logs have been used to forecast the spreading of influenza epidemics [13]; microblogging platforms such as Twitter have been used to monitor public health at scale [14, 15] and to estimate the prevalence of a wide range of pathologies, from mental illnesses [16] to obesity [17]; and pictures on social media have been used to estimate values of Body Mass Index (BMI) [18, 19].

On the web, in addition to communities of general interest, there are communities specialized in food. Such communities have allowed researchers to study dietary patterns [20–22], and how these patterns change depending on a country's culture [23]. Databases containing millions of receipts have been used to quantify each receipt's healthiness based on its ingredients and associated images [5, 24], and the resulting proxies for healthiness have been recently incorporated into food recommender systems [25, 26].

Given its scale, web data allows for the study of various societal aspects concerning food consumption. After collecting food-related tweets in 50 US states, researchers found that caloric values of the foods mentioned in the tweets related to state-wide obesity rates [27]. In USA, large areas suffer from obesity: food deserts, for example, are areas that have limited access to nutritious food and typically happen to be of low income. Recently, De Choudhury *et al.* analyzed millions of food-related Instagram images [28] and found that food deserts indeed consume food high in fat, cholesterol, and sugar more than the other locations do [6]. In addition to what social media users might eat, images might reveal perceptions. Indeed, from a food image, one can infer two quantities: first, how healthy the food in the image is (based on what the image depicts); and second, how healthy the social media user perceives it to be (based on the user's comments). Researchers have computed these two quantities for a variety of countries and found that the gap between the two relates to health outcomes [7]. All these works which have relied on Web data, however, invariably suffer from a number of self-presentation and self-selection biases and, as such, the resulting datasets might not reflect actual food consumption [29, 30].

To partly address these biases, a few studies have analyzed data of grocery purchases at scale. Nevalainen *et al.* collected loyalty card data in Finland voluntarily provided by more

than 13K customers, and analyzed consumptions across categories [31]. Guidotti *et al.* mined millions of transactions recorded by two retailers to train algorithms that suggest what online shoppers might like [32]. Mamiya *et al.* collected grocery purchase data from multiple stores in the metropolitan area of Montreal (Canada), performed an extensive analysis of consumption of carbonated drinks [33], and found that education inversely correlates with the consumption of soft drinks [34]. None of these studies, however, have related food consumption to health outcomes.

To sum up, prior work has shown that controlled dietary studies at scale are costly and, as such, are rare, and that social media data—albeit useful to better understand nutrition habits (especially across countries and cultures)—is affected by a number of biases and typically comes at coarse-grained spatial resolutions. What is needed is a new approach to measuring at scale what real people (as opposed to study participants) eat and drink, ideally under naturalistic conditions. Before introducing the datasets and methods with which we tackle this challenge, we introduce our seven research questions.

3 Research questions

To begin with, let us look at one of the simplest food-health relationships: people consume more calories than they use, and the surplus is stored as fat [35]. We test that with our first question:

RQ1: Is calorie consumption positively associated with hypertension, high cholesterol, and diabetes?

Yet that could only be part of the story. It might be less about the number of calories than about their concentration. When one regularly eats calorie-dense animal products and junk foods, what changes is not only the taste buds but also the brain chemistry. Calorie-diluted foods (e.g., green smoothies) do not lead to a dopamine response but calorie-dense foods (e.g., ice creams) with the same amount of calories do. Fatty and sugary foods are energy dense, and their overconsumption has often been compared to drug addiction [10]. To put it simply, given two foods with the same amount of calories but different concentrations, the delivery of pleasure within people's brains is quicker for the calorie-dense food. So our second research question is:

RQ2: Is calorie concentration positively associated with the three medical conditions?

Not all calories are created equal though. The U.S. government's official Dietary Guidelines for Americans recommends the reduction of sugar, calories, saturated fat, sodium, and trans fat and, at the same time, it recommends increasing fibres, of which at least "a quarter of the American population is not reaching an adequate intake" [35]. Therefore, it would make sense to look at individual nutrients, and we do so next. The digestive system breaks down *carbohydrates* into a simple sugar called glucose. To get from the bloodstream into your cells, glucose requires insulin. Without insulin, the glucose builds up in the blood. Inappropriate *fat* storage may keep cells from responding properly to insulin, causing insulin resistance. Eventually blood-sugar levels rise out of control and the patient develops diabetes. Even among healthy individuals, a high-fat diet impairs the body's ability to handle *sugar*. But it is not all to do with fat. Obesity might be caused by diets rich in carbohydrates and sugar in the first place. Interestingly, sugar seems to change the brain's circuitry [36, 37]. When people consume a sugary food, their brains release dopamine [37], which signals pleasure. This evidence leads to our third research question:

RQ3: Are fat, carbohydrates and sugar positively associated with the three medical conditions?

Not all fats affect the muscle cells in the same way. Palmitate (the saturated fat found mostly in meat, diary, and eggs) causes insulin resistance. On the other hand, oleate (the mono-unsaturated fat found mostly in nuts, olives, and avocados) protects us against the detrimental effects of saturated fat. Research findings on the impact of saturated fats are controversial though. In 2014, a large meta-analysis showed no relationship between saturated fats and heart disease [38]. Hence:

RQ4: What is the relationship between saturated fats and the three medical conditions?

On a more positive note, consider fibres. Humans evolved over millions of years eating mostly wild plants, likely in excess of one hundred grams daily [39]. That is much more than what the average person eats today. Given their health benefits, we posit the following question:

RQ5: Are fibres negatively associated with the three medical conditions?

Going beyond individual nutrients, one could study their composite impact. More specifically, research has shown that healthy diets is associated with diversity of foods [40]. Therefore, our next research question is:

RQ6: Is nutrient diversity negatively associated with the three medical conditions?

Finally, the amount of food people consume is often influenced by external factors, including the size of their plate [41]. Our final research question is then:

RQ7: Is the overall weight of food consumed positively associated with the three medical conditions?

4 Methods

4.1 Datasets

4.1.1 Food purchases

At all our retailer's 411 shops in Greater London, 1.6M customers used their loyalty cards and bought 1.6B food products in the entire year of 2015. Given the use of loyalty cards, purchase data is stored in the following anonymized form: customer postcode area/region, store postcode, productID, and timestamp. Each productID is associated with the product's net weight, total energy, fats, saturated fats, carbohydrates, sugars, proteins, and fibres. The last six elements are expressed as grams of substances contained in the product. Using standard guidelines [42], we map grams into corresponding calories by simply multiplying them by fixed factors: 9 Kcals per gram for fats, 4 Kcals for proteins and carbohydrates, and 2 Kcals for fibres.[c]

4.1.2 Chronic diseases

At all the 1174 general practices (GPs) in Greater London, 1.1B medicine items were prescribed in the entire 2016. Such prescription data has been recently made publicly available[d] in the following form: GP identifier, medicineID, and timestamp. Each medicineID is associated with the medicine's active ingredients, from which the corresponding diseases can be inferred. To do so, we map the prescriptions to their medicines' active ingredients and, in turn, to the chronic diseases they are supposed to treat. The mapping of an active ingredient to the most likely disease is done based on the OpenPrescribing taxonomy [43]. As a result, for each GP, we know the number of prescriptions that are meant to treat a given disease.

We are interested not in all prescriptions but only in those related to three main factors that are generally grouped under the heading of "metabolic syndrome": high blood pressure (hypertension), an excess of cholesterol in the blood, and high blood-sugar levels (diabetes).

Hypertension. Hypertension is a long-term medical condition in which the blood pressure in the arteries is persistently elevated. It has been identified as the most important risk factor for death in the Western world [44]. To capture the prevalence of hypertension, we consider the prescriptions in three categories of the OpenPrescribing taxonomy: antihypertensive drugs (e.g., Hydralazine Hydrochloride), alpha-adrenoceptor blocking drugs (e.g., Doxazosin Mesilate), and renin-angiotensin system drugs (e.g., Lisinopril).

Cholesterol. One important risk factor for coronary heart diseases is cholesterol. If cholesterol level is low, an obese and diabetic still does not develop atherosclerosis [45, 46]. Cholesterol also seems to help some cancers migrate and invade more tissue [47]. To capture the prevalence of high cholesterol, we consider one category in the OpenPrescribing taxonomy: lipid-regulating drugs (e.g., Statins).

Diabetes. Diabetes is characterized by chronically elevated levels of sugar in the blood [48]. Insulin is the hormone that keeps the blood sugar in check. The disease is caused by either the pancreas gland not making enough insulin (type 1 diabetes) or by the body becoming resistant to insulin's effects (type 2 diabetes, which accounts for 90-95 percent of diabetes cases [49]). Type 2 diabetes is a consequence of dietary choices (of "high-fat and high-calorie diets") and, as such, is preventable and often treatable. Diabetics are more likely to suffer from strokes and heart failure [50]. To capture the prevalence of diabetes, we consider the prescriptions in four categories of the OpenPrescribing taxonomy: insulin, antidiabetic drugs (e.g., gliclazide), active ingredients for the treatment of hypoglycaemia (e.g., glucagon), and agents for diabetic diagnostic and monitoring (e.g., glucose blood testing reagents).[e]

4.1.3 Mapping food purchases and chronic diseases

To map all our data, we use the postcode area/region as our initial unit of geographic aggregation. We map our food purchases using the customers' area of residence, which are available in our grocery dataset. We then map our prescriptions based on what the anonymized prescription dataset offers—that is, based on the fraction of each GP's patients living across areas. More specifically, for each GP, we perform two steps. First, we consider the fraction of the GP's patients who live in each area since patient counts are publicly available[f] (e.g., 50% of the GP's patients live in area X). Second, we assign the GP's prescriptions to an area, and the assignment is proportional to the fraction of the GP's patients who live in the area (e.g., half of the GP's prescriptions are assigned to area X). We repeat these two steps for all GPs in the city. As a result, we obtain the number of prescriptions containing each medicine in each area, and normalize that number by the population, determining the per-capita prevalence of that medicine in that area.

As we shall see shortly, to support our analyzes, we need to match our data with census data, which is not available at postcode area/region level though. Census data is typically defined at four different spatial resolutions:[g] Lower Super Output Area (LSOA), Medium Super Output Area (MSOA), Ward, and Local Authority (LA, or, more informally, Borough). Among these four aggregations, we opt for MSOA because, at this resolution, our aggregate metrics for food consumption and disease prevalence start to be significant, not

least because they concern a sufficient number of residents: our data covers 937 MSOAs in London, which have an average of 8250 residents. As such, from now on, we refer to MSOAs as areas or neighborhoods.

Given our three diseases, we consider prescriptions related to them and express each of their prevalence at area level in terms of the number of prescriptions for each disease per capita. For example, in the case of diabetes, for any area a we have:

$$\text{prevalence-diabetes}@a = \frac{\text{\#prescriptions for diabetes}@a}{\text{\#residents}@a}. \tag{1}$$

In a similar way, we compute prevalence-hypertension$@a$ and prevalence-cholesterol$@a$.

4.1.4 Socio-economic indicators

The prevalence of chronic diseases is not only impacted by food consumption but also mediated by socio-economic conditions. Higher-income and well-educated people may have better access to doctors, gyms, parks and healthy food. There is an inverse relationship between education levels and the likelihood of getting fat in Australia, Canada and England [51]. The same applies in USA: "obesity rates in children with college-educated parents are less than half the rates of children whose parents lack a high-school degree" [52]. In developed countries, there is a difference between cities and suburban areas: the more affluent urbanites are usually fitter than rural residents [53].

To control for these factors in our study, we collected data on socio-economic conditions from the 2015 UK census and from the Index of Multiple Deprivation (IMD) 2015 that is based on a basket of measures of deprivation for small areas across England.[h] We focus on the socio-economic factors that have been found to be associated with specific food consumption patterns and, ultimately, with chronic diseases. These factors—available at MSOA level—are average income, education level, gender distribution (%female), and average age.

4.2 Estimating eating habits of an area

To estimate the eating habits of people living in an area, we pool together all the food items purchased by its residents and look at the nutritional properties of the average item. We do so by defining six metrics below.

To capture calorie consumption, we compute the average amount of calories contained in the food items purchased in an area:

$$\text{calorie-consumption}@a = \frac{\sum_{p \in P_a} \text{Kcal}(p)}{|P_a|}, \tag{2}$$

where P_a is the set of all food products purchased by residents of area a, p is one of such products, and $\text{Kcal}(p)$ is the value of kilocalories in p.

To capture calorie concentration rather than simple calorie counts, we compute:

$$\text{calorie-concentration}@a = \frac{\sum_{p \in P_a} \text{Kcal}(p)}{\sum_{p \in P_a} \text{weight}(p)}, \tag{3}$$

which reflects the concentration of calories in the area's "average" product.

For each area, we also compute the average number of calories given by individual nutrients in a product, on average:

$$\text{nutrient-calories}@a = \frac{\sum_{p \in P_a} \text{Kcal}(\text{nutrient}, p)}{|P_a|}, \tag{4}$$

where P_a is the set of all food products purchased at area a; p is one of such products; Kcal(nutrient, p) is the energy intake given by nutrient in p. The nutrients we consider are: fats (fats–calories@a), saturated fats (saturated–calories@a), carbohydrates (carbs–calories@a), sugars (sugar–calories@a), proteins (proteins–calories@a), and fibres (fibres–calories@a).

We also capture the diversity of nutrients consumed in the area. This is computed as the Shannon entropy of the distribution of the calories given by all the nutrients:

$$H(\text{nutrient})@a = -\sum_{\text{nutrient}} \text{prob}(\text{nutrient}, a) \cdot \log \text{prob}(\text{nutrient}, a), \tag{5}$$

where prob(nutrient, a) can be thought as the fraction of area a's total calories coming from nutrient, which, in turn, can be written as:

$$f_{\text{nutrient-calories}}@a = \frac{\text{nutrient-calories}@a}{\text{calorie-consumption}@a}. \tag{6}$$

For example, $f_{\text{fat-calories}}@a$ is the fraction of a's total calories coming from fat.

Finally, we also compute the average item weight:

$$\text{item-weight}@a = \frac{\sum_{p \in P_a} \text{weight}(p)}{|P_a|}. \tag{7}$$

For the sake of reproducibility of our analysis, we publicly share our data aggregated at MSOA level.
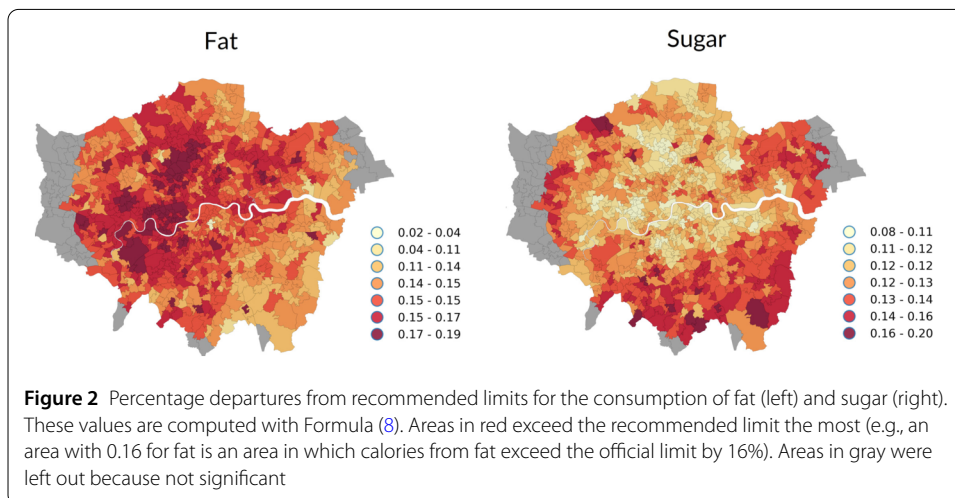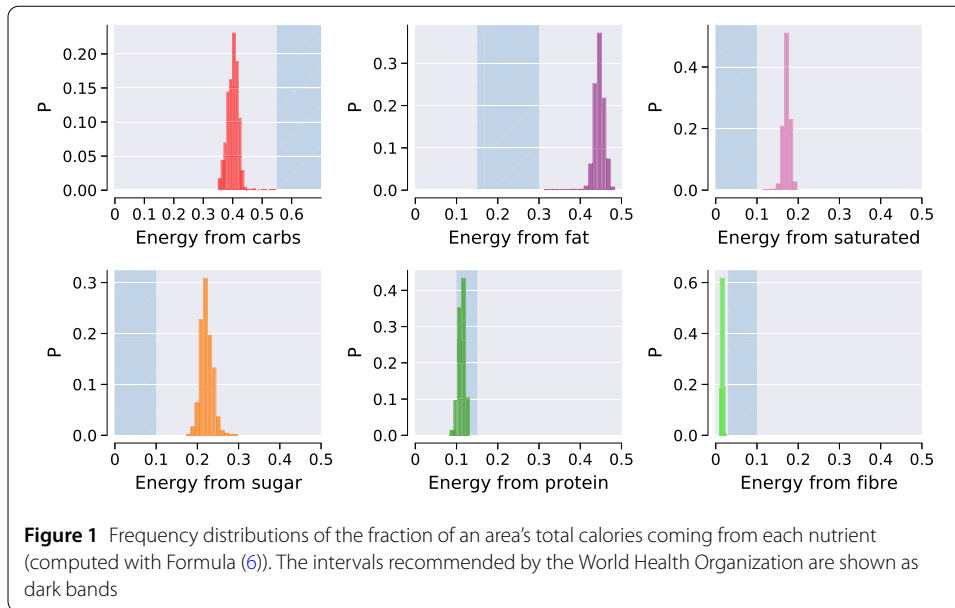
5 Results

5.1 Relative abundance of nutrients

The nutrition guidelines by the World Health Organization (WHO) recommend to limit the relative energy supply derived from each nutrient within specific ranges [54]; for example, fats should contribute no more than 30% to the total intake. By plotting the distributions of the $f_{\text{nutrient-calories}}@a$ values across neighborhoods (Fig. 1), one sees that Londoners, on average, buy a healthy share of protein, yet they buy unhealthy nutrients (e.g., sugar, fat, saturated fat) more than the recommended limits, and carbohydrates and fibres less than the recommended amounts. The extent to which residents collectively depart from recommended limits changes across the city and is defined as departure-nutrient@a:

$$\begin{cases} |f_{\text{nutrient-calories}}@a - \max_{\text{nutrient}}| & \text{if } f_{\text{nutrient-calories}}@a > \max_{\text{nutrient}}, \\ 0 & \text{if } \min_{\text{nutrient}} \leq f_{\text{nutrient-calories}}@a \leq \max_{\text{nutrient}}, \\ |\min_{\text{nutrient}} - f_{\text{nutrient-calories}}@a| & \text{if } f_{\text{nutrient-calories}}@a < \min_{\text{nutrient}}, \end{cases} \tag{8}$$

That is, an area's departure from the nutrient's recommended level is zero, if the fraction of area a's total calories coming from nutrient is within the recommended min-max band for

Figure 1 Frequency distributions of the fraction of an area's total calories coming from each nutrient (computed with Formula (6)). The intervals recommended by the World Health Organization are shown as dark bands

Figure 2 Percentage departures from recommended limits for the consumption of fat (left) and sugar (right). These values are computed with Formula (8). Areas in red exceed the recommended limit the most (e.g., an area with 0.16 for fat is an area in which calories from fat exceed the official limit by 16%). Areas in gray were left out because not significant

that nutrient (i.e., within $[\min_{\text{nutrient}}, \max_{\text{nutrient}}]$). The departure from the recommended levels of, for example, fats and sugars are mapped in Fig. 2.

To dwell on the health impact of such departure, we now match food purchases with health outcomes and start to answer our seven research questions.

We compute the Spearman rank correlation between disease prevalence (as per Formula (1)) and all the food-related metrics (as per Formulae (2)–(7)). As shown in Fig. 3, calorie consumption is strongly correlated with cholesterol and hypertension, while calorie concentration is strongly correlated with diabetes (RQs1+2). To check whether the relationships between calories and chronic diseases are linear or not, we produce a set of *x-y* plots arranged in three columns and nine rows in Fig. 4. Each column corresponds to one of the three chronic diseases, and each row corresponds to one of the features derived from our food purchases. For example, in the first row, we have plots that relate hypertension, cholesterol, and diabetes to calorie concentration; in the second row, instead, we have
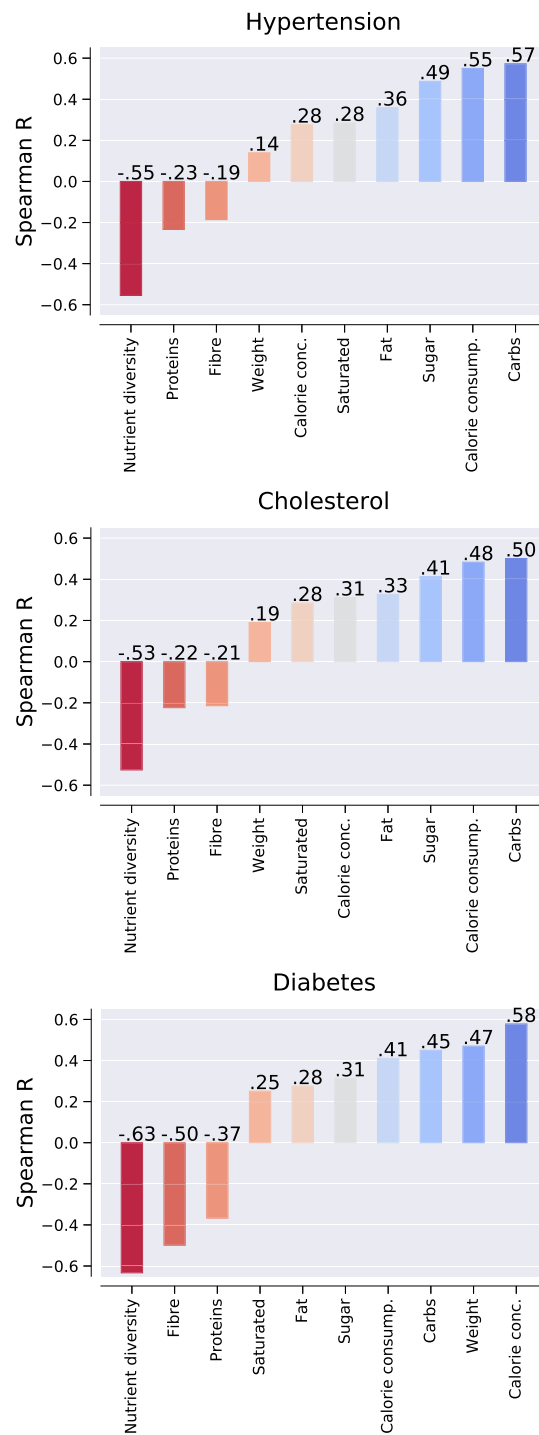
Figure 3 Spearman rank correlations between disease prevalence (as per Formula (1)) and food-related metrics (as per Formulae (2)–(7)). All correlations are significant with $p < 0.001$

plots that relate these three diseases to calorie consumption. In both cases, we see that, as calories increase, the prevalence of any of the three diseases increases, as one would expect. To ease interpretation of these x-y plots, we rescaled the x-axis. For example, we
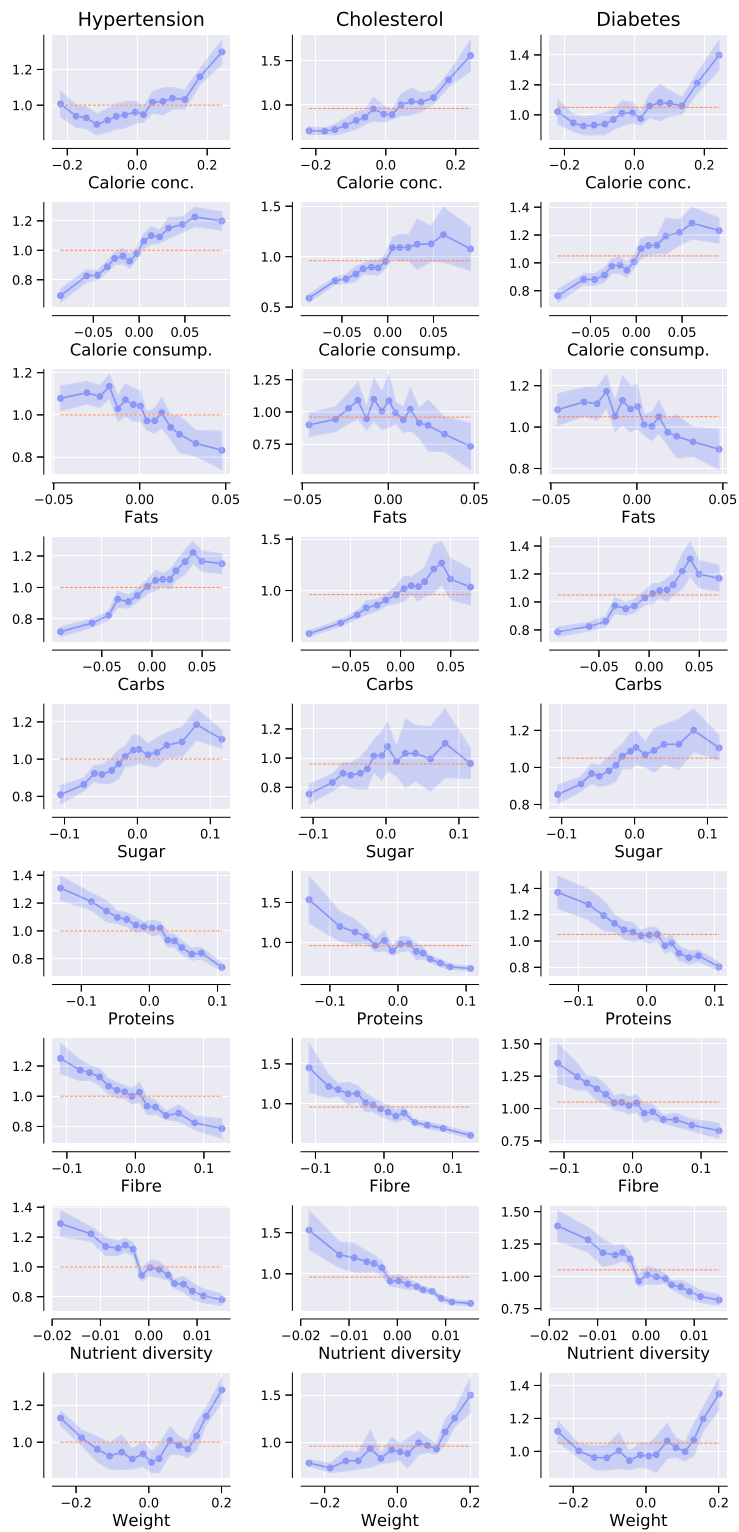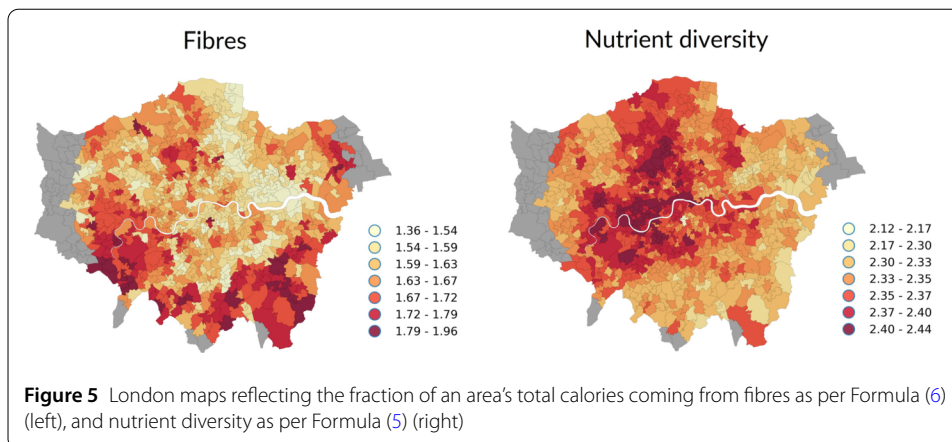
Figure 4 The relationship between hypertension, cholesterol, and diabetes (columns) and food-related metrics (rows). On the *x*-axis, we represent the relative food-consumption values compared to the average (set at 0), which are computed with Formula (9). On the *y*-axis, we represent the per-capita disease prevalence, which is computed with Formula (1): a value of 1 for a disease means that each resident takes, on average, 1 medication for that disease in the year. The dotted red line indicates the average disease prevalence across all areas. The shaded areas show 95% confidence intervals
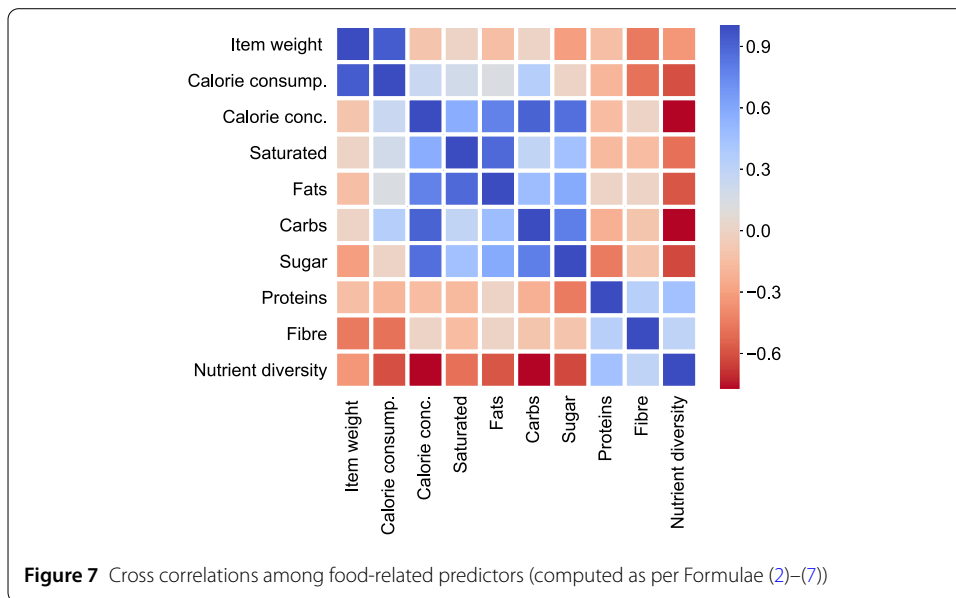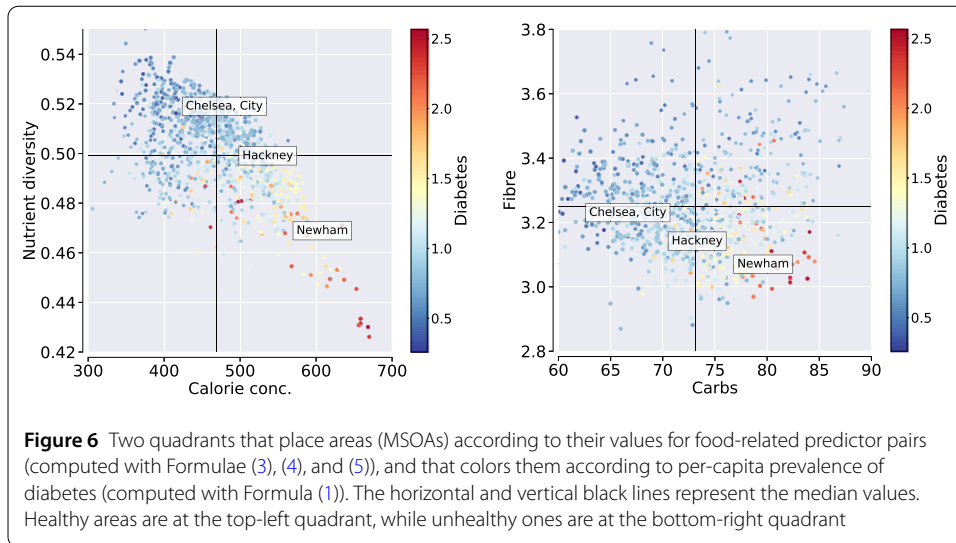
Figure 5 London maps reflecting the fraction of an area's total calories coming from fibres as per Formula (6) (left), and nutrient diversity as per Formula (5) (right)

normalize the average item's weight in area a as follows:

$$\text{relative-item-weight}@a = \frac{\text{item-weight}@a - \mu(\text{item-weight})}{\mu(\text{item-weight})}, \tag{9}$$

where $\mu(\text{item-weight})$ is the average weight across all areas. If the rescaled value is zero, then the area's weight is equal to the average value in London. If the value is 0.1, then the area's weight is 10% higher than the average value. By observing, for example, the resulting plot related to calorie concentration (first row in Fig. 4), we see that, as the consumption exceeds the average value (i.e., $x > 0$), the per-capita prevalence of any of the three diseases considerably increases.

We also see that the four main nutrients are associated with the three diseases in expected ways: carbohydrates, fat and sugar are positively associated with the three diseases, while fibres are negatively associated (RQs3+5) (Fig. 3). Indeed, the prevalence of each of the diseases increases as carbohydrates (fourth row in Fig. 4) increases. The relationships between the diseases and fat (third row in Fig. 4) is not as clear cut as one would have expected though: the relationships with fat and saturated fat come with high variability (RQ4). By contrast, more proteins (sixth row) and fibres (seventh row) are associated with lower disease prevalence.

To go beyond individual nutrients, we also find that both nutrient diversity (mapped in Fig. 5) and average item weight show high correlations, which of course have opposite signs: item weight (which correlates with calorie consumption) is positively correlated with disease prevalence, while nutrient diversity is negatively correlated (RQs6+7). Indeed, in Fig. 4, one sees that the prevalence of any of the three diseases rapidly decreases with nutrient diversity, while it considerably increases with calorie concentration and average item weight. This is further confirmed by the quadrant in Fig. 6, which places areas according to the prevalence of pairs of nutrients (computed with Formulae (3), (4), and (5)), and colors them according to the prevalence of diabetes (computed with Formula (1)): residents of the City (central London) and Chelsea (West London), who tend to be highly educated and well-off, consume fibres and diversify nutrients; those of Newham, which is a deprived yet rapidly developing neighborhood in East London, consume considerable quantities of calories, do not diversify their nutrients, and end up with a high prevalence of diabetes; interestingly, the residents of Hackney, which is a deprived yet

Figure 6 Two quadrants that place areas (MSOAs) according to their values for food-related predictor pairs (computed with Formulae (3), (4), and (5)), and that colors them according to per-capita prevalence of diabetes (computed with Formula (1)). The horizontal and vertical black lines represent the median values. Healthy areas are at the top-left quadrant, while unhealthy ones are at the bottom-right quadrant

Figure 7 Cross correlations among food-related predictors (computed as per Formulae (2)–(7))

highly-educated neighborhood in East London, enjoy healthier eating habits (i.e., low consumption of carbohydrates and high nutrient diversity) and do not suffer from diabetes as much as Newham's residents do.

So far we have considered our food-related metrics individually. However, these metrics are not orthogonal, and the presence of one is generally associated with the presence of another. In fact, by correlating the presence of a nutrient with the presence of another (cross-correlation matrix in Fig. 7), we see that an item's average weight (on the first row of the correlation matrix) is generally not related to any nutrient, as one would expect; carbohydrates (second row) are associated with calorie concentration and sugar (sugar is indeed one type of carbohydrate); high calorie concentration (third row), in turn, comes with food high in carbohydrates, fat, and sugar; and nutrient diversity (last row) is generally found in food high in proteins and fibres.

5.2 Predicting medical prescriptions from nutrients

As a next step, we go beyond studying correlations and aim at predicting the number of prescriptions from the food data. To do that, we first need to account for the dependencies between nutrients. Also, we should account for any factor other than nutrients that impacts health outcomes. The literature typically controls for socio-economic conditions, which have been shown to be a proxy for access to knowledge and capabilities, including access to nutritional information and physical exercising [55]. To account for all these aspects, we use linear regression analysis. The prevalence of each of the three chronic diseases is the outcome variable of an ordinary least squares regression, while our food-related metrics and a set of control variables are the predictor variables. Where necessary, predictor variables undergo a logarithmic transformation, and in addition, we apply a min-max rescaling of each variable in the range $[0, 1]$, which allows us to assess the relative influence of each factor: the larger the absolute value of the coefficient associated to a feature, the higher the relative importance of feature in predicting the outcome.

We first try a regression that considers individual nutrients (carbohydrates, fats, sugar, proteins, fibres) as independent variables. Table 1 shows the results and suggests that carbohydrates, fats and sugar are associated with the three chronic diseases, while the presence of proteins and fibres counters that association. Among the control variables, income has very little predictive power when combined with the other factors (it has always either a low coefficient or a high p-value). Education and gender are more informative predictors among the socio-demographic variables. In particular, the prevalence of diabetes seems to be more prevalent among males, which is in line with previous findings [56]. Overall, nutrients and demographic features jointly explain more than one third of the variability in the linear regressions for hypertension ($R^2 = 0.388$) and cholesterol ($R^2 = 0.345$), and almost 60% in the regressions for diabetes ($R^2 = 0.598$), and such an explanatory power is not impacted by any autocorrelation. That is because the Durbin–Watson statistic values reported at the bottom of all the regression tables reflect the impact of autocorrelations on the residuals, are defined in the range $[0, 4]$, and, in our case, take values close to 2, which indicate no autocorrelation.

Previously, we have seen that nutrient diversity and calorie consumption tend to be highly correlated with disease prevalence. As such, one might now wonder whether a linear regression analysis solely based on the combination of nutrient diversity and calorie intake would be informative. Indeed, we find that it is (Table 2), and in two main ways. First, based on the regression coefficients, both indicators seem to matter, with nutrient diversity being the most powerful of the two. Second, after controlling for socio-economic variables, these two metrics alone explain up to 38% of the variance in the prevalence of hypertension, up to 34% for cholesterol, and up to 59% for diabetes.

To better make sense of the predictive power of our features, we run a sensitivity analysis where we measure the R^2 values for regressions run with different feature sets (Fig. 8). First, we note that demographic features alone are considerably less predictive than nutrients alone, although they improve the overall accuracy when combined with nutrients. Second, and more interestingly, the prediction performance of only the combination of nutrient diversity and calorie intake is comparable to the more complex combination of all the individual nutrients.

Finally, we build classification models that can identify areas that are healthy or unhealthy in terms of the prevalence of the three chronic diseases. We first formulate this

Table 1 Linear regressions that predict the three chronic diseases from individual nutrients and socio-demographic control variables (income, gender, age, education level)

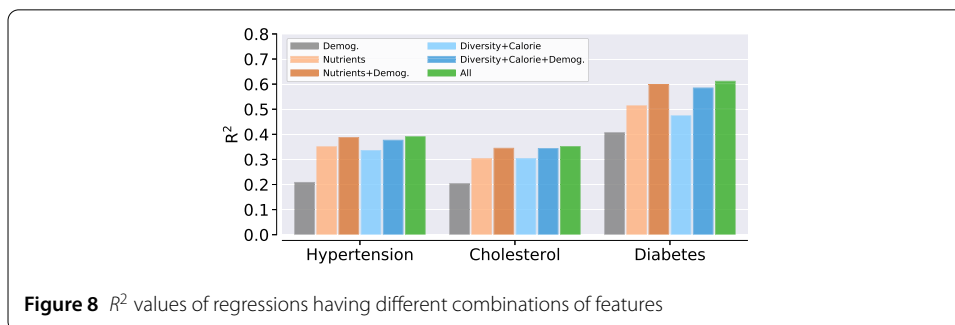| Feature	Coefficient	Std. error	p-value
Hypertension			
α (intercept)	0.3241	0.050	<0.001
Carbs	0.7652	0.075	<0.001
Fats	0.3225	0.066	<0.001
Sugar	0.2402	0.072	0.001
Proteins	−0.2868	0.050	<0.001
Fibre	−0.0627	0.051	0.216
Income	0.0477	0.042	0.259
%Females	−0.2124	0.038	<0.001
Average age	0.1664	0.037	<0.001
Education	0.2451	0.042	<0.001
Durbin–Watson stat. = 2.048		Adj R^2 = 0.388	
Cholesterol			
α (intercept)	0.2645	0.047	<0.001
Carbs	0.5877	0.070	<0.001
Fats	0.3382	0.062	<0.001
Sugar	0.2441	0.067	<0.001
Proteins	−0.2745	0.046	<0.001
Fibre	−0.0268	0.047	0.569
Income	−0.0184	0.039	0.640
%Females	−0.2322	0.036	<0.001
Average age	0.1272	0.035	<0.001
Education	0.1751	0.039	<0.001
Durbin–Watson stat. = 2.001		Adj R^2 = 0.345	
Diabetes			
α (intercept)	0.5073	0.041	<0.001
Carbs	0.57659	0.061	<0.001
Fats	0.5002	0.054	<0.001
Sugar	0.4992	0.059	<0.001
Proteins	−0.5137	0.041	<0.001
Fibre	−0.1312	0.041	0.002
Income	−0.1222	0.034	<0.001
%Females	−0.3536	0.031	<0.001
Average age	−0.0290	0.030	0.342
Education	−0.0947	0.035	<0.006
Durbin–Watson stat. = 2.000		Adj R^2 = 0.598	

Figure 8 R^2 values of regressions having different combinations of features

classification problem as a binary classification: the goal is to identify areas that fall into the top and bottom quartiles of each of the three diseases (higher scores corresponded to higher presence of a disease). As such, we define the response variable y_i as 0, if area i is in the first quartile of the disease distribution; and as 1, if it is in the last quartile.

Table 2 Linear regressions that predict the three chronic diseases from item weight and nutrient diversity (plus control variables such as income, gender, age, education level)

Feature	Coefficient	Std. error	*p*-value
Hypertension			
α (intercept)	0.5582	0.064	<0.001
Calorie consumption	0.30228	0.070	<0.001
Nutrient diversity	−0.5182	0.069	<0.001
Income	0.0615	0.041	0.131
%Females	−0.2210	0.038	<0.001
Average age	0.1627	0.037	<0.001
Education	−0.2309	0.041	<0.001
Durbin–Watson stat. = 2.033		Adj R^2 = 0.377	
Cholesterol			
α (intercept)	0.5465	0.059	<0.001
Calorie consumption	0.1395	0.064	0.03
Nutrient diversity	−0.4943	0.064	<0.001
Income	0.0017	0.037	0.96
%Females	−0.2364	0.035	<0.001
Average age	0.11790	0.034	0.001
Education	−0.1785	0.037	<0.001
Durbin–Watson stat. = 1.981		Adj R^2 = 0.344	
Diabetes			
α (intercept)	0.7582	0.038	<0.001
Calorie concentration	0.1301	0.028	<0.001
Nutrient diversity	−0.6353	0.043	<0.001
Income	−0.0790	0.034	0.019
%Females	−0.3693	0.031	<0.001
Average age	−0.0606	0.030	0.042
Education	0.1047	0.032	0.001
Durbin–Watson stat. = 1.964		Adj R^2 = 0.585	

This formulation effectively prunes the middle quartiles and makes it possible to focus on the classification of extreme samples. In so doing, we also ensure a roughly 1:1 ratio of positive to negative examples in each class. In a second experiment, we define a 3-class classification problem by keeping the points in the two previous classes (top and bottom quartiles) and defining a third class. This class has the same number of points as the two other classes, and these points are randomly selected from the mid quartiles. We compute the mean accuracy of 10 iterations using a Random Forest classifier in both experiments. Results are reported in Table 3. The performance of each model can be interpreted relative to a baseline random classifier, which after a sufficient number of iterations averages out with an accuracy of 0.5 for the binary case, and 0.33 for the ternary case. We then test three combinations of predictors, one at a time: the demographic features (gender, age, income, education), the two most predictive features from the food data (i.e., calorie concentration and nutrient diversity), and those two combinations jointly. As expected from the previous analysis, demographic factors have the lowest predictive power yet are orthogonal to food-related predictors. The binary classifier that uses all features correctly identifies (un)healthy areas 91% of the times for diabetes, 82% of the times for hypertension, and 81% of the times for cholesterol.

6 Discussion
6.1 Main results
All the previous results suggest that, in London, socio-economic conditions matter far less than what people eat. As opposed to having high levels of education or of median income,

Table 3 Accuracy of Random Forest classifier in predicting the prevalence of diseases in London areas. The results of two classifiers are reported: (i) binary classification of areas in the top and bottom quartiles of the three diseases' prevalence; (ii) ternary classification where an equally-sized class containing training instances randomly sampled from the two central quartiles is added. The predictive features are six: gender, average age, education level, item weight, nutrient diversity, and calorie concentration. The accuracy of a random baseline classifier is 0.5 for the binary case, and 0.33 for the ternary case. Numbers in parenthesis represent the standard deviation on the 10-fold cross validation

	Medicine	Accuracy			
		Random	Demographic	Diversity + Calorie	All
Binary	Hypertension	0.50	0.60 (0.06)	0.80 (0.05)	0.82 (0.05)
	Cholesterol	0.50	0.59 (0.06)	0.81 (0.05)	0.81 (0.05)
	Diabetes	0.50	0.79 (0.06)	0.86 (0.05)	0.91 (0.04)
Ternary	Hypertension	0.33	0.40 (0.05)	0.54 (0.05)	0.57 (0.04)
	Cholesterol	0.33	0.41 (0.05)	0.53 (0.06)	0.54 (0.07)
	Diabetes	0.33	0.53 (0.04)	0.63 (0.04)	0.68 (0.03)

eating less calories and opting for a diet with diverse nutrients are both strongly associated with healthy areas. Indeed, one of the surprising results from the regression analysis is that income is not a significant predictor. Previous studies have shown that income is correlated with general health conditions including mental disorders, self-reported bad health, and lower chances of long-term survival [57]. Yet not all diseases equally interact with socio-economic variables. In contrast to the prevalence of cancer, the prevalence of illnesses connected to the metabolic syndrome such as circulatory diseases or high cholesterol does not significantly change across socio-economic groups [58]. Indeed, previous population surveys point out that, after controlling by education, the link between metabolic syndrome and other socio-economic factors such as racial background [59] or income [60] weakens markedly, which is in line with our results.

In terms of which nutrients matter the most, as opposed to unhealthy areas, healthy ones tend to buy more fibres and far less carbohydrates (including sugar). Also, it is less about calorie consumption and more about calorie concentration, which have been previously found to lead to forms of addiction [10]. By combining all the predictors together, one obtains a model that is not only *descriptive* of how health outcomes are associated with food purchases but also *predictive* of such outcomes: it turns out that, from food purchases, we can accurately predict whether an urban area will suffer from diabetes, for example.

6.2 Theoretical implications
This study has two main theoretical implications. The first is a call for enlightened nutrition research. The question for food companies is how to continue to make money even as they cut calories. The answer might come from a shift in how food companies have approached the formulation of their products so far. Food product research has focused its attention on taste, not nutrition. That needs to change. The combination of nutritional research with recent advances in biomedical research promises to create foods that are not only delicious but also provide concrete medicinal benefits.

The second theoretical implication has to do with studying how entire neighborhoods eat. Past research has explored two relationships. The first is between "where people live" and their health: income inequality, unemployment rates, and education have all been shown to relate to people's health. The second is between "what people eat" and their health: nutrition research has long tested associations between eating patterns and health

outcomes with survey data. A third relationship transitively follows which has never been tested before: the relationships between "where people live" and "what they eat". We have now tested it and found that, indeed, healthy neighborhoods eat less and diversify nutrients more than what neighborhoods suffering from chronic diseases tend to do.

6.3 Practical implications

This study suggests practical implications for a variety of stakeholders. We have shown that unhealthy products have a negative impact on community health. The bad news is that many unhealthy products are very popular. The good news is that as many as five stakeholders have incentives for change: food companies do not wish to be seen as the cause of people's obesity; insurance companies (especially those in life insurance) have our health at heart; technology companies are entering the digital health market; governments want to be seen to act; and local communities increasingly want to be empowered to tackle their own needs.

Food Companies. By simply cutting out bad ingredients, adding good ones or introducing new products, the food industry could reformulate their offering and elaborate plans to improve nutrition.

Insurance Companies. There is at least one corporate sector that benefits from keeping people healthy: insurance companies. Our study encourages new partnerships between insurance firms and large grocery retailers on, for example, data sharing initiatives. Also, retailers could make anonymized purchased data publicly available and launch "hackathons". These are meetings in which participants are asked to come up with a solution to a problem within a day or two, and some of the teams generally offer effective solutions at little cost (the winning team is typically awarded a prize).

Technology Companies. Predictive analytics and wearable sensors will transform how people manage their health. A smartphone app might be able to warn users that, based on which foods they share on social media and what their wearable sensors measure, they will exacerbate a heart condition. The app could even suggest which foods to eat—foods that are both pleasurable and nutritious.

Public authorities. In the past, governments have focused on treating diseases rather than preventing them. Yet state-based prevention strategies might be justified, not least because of externalities. Unhealthy eating harms not only oneself but also others, in that, it results in additional costs for health care. Public authorities could intervene in three main ways.

- *Taxing.* Taxing fat and subsidizing healthy eating is one way of tackling the obesity problem. However, a recent study showed that taxing fat might not help [61]. In the study, subsidies were given to encourage all income groups to buy more fruit and vegetables. Women on higher incomes bought more fruit and vegetables than usual, while those on lower incomes changed their habits less. As a result, women on lower incomes paid much more for food (as taxes were on the food they ate most), and the inequality between the two groups widened. Taxes and subsidies might not change people's habits, and other strategies are needed—notably education.
- *Educating.* One simple state intervention is the launch of new educational programs that inform people about the dangers of not eating well. It has been shown that a short-lived change in diet have long-term consequences. A three-week change of diet aimed at reducing cravings for salt, sugar, and fat has been shown to change participants' taste buds [62].

- *Nudging.* A more viable alternative would be to nudge citizens into healthy behavior. The idea is to provide small impulses so that healthy becomes the obvious choice.

Local Communities. We have shown that, by mining publicly available prescription data, we are able to identify (un)healthy areas. Mining digital health data reflects concrete health outcomes and might well benefit local communities by enabling residents to hold local authorities to account. Additionally, a city health monitor could help assess the benefits of implementing different policies.

6.4 Limitations

Sample bias. Our data comes from one grocery retailer and concerns only those customers who have opted for the use of a loyalty card. Furthermore, the data is anonymized and does not contain personal information such as age or gender. Future work should use additional geographic data to quantify sample biases.

Limited explanatory power. Our study does not fully explain health outcomes, and rightly so. Our food data does fully cover Londoners' food consumption, and our study does not include any data on another important predictor of health outcomes: physical activity.

Average product. From our data, we cannot reconstruct the dietary habits of individual customers and, as such, our results reflect the dietary habits of an area in terms of the area's "average product".

Causality. Our results do not speak to causality. Though the causal direction is difficult to determine from observational data, one could consider different temporal snapshots of both sets of data (food purchases and medical prescriptions), and perform a cross-lag analysis.

6.5 Conclusion

It was healthy and adaptive for our primate brains to drive us to eat carbohydrates and sugar when only wild grass was at hand. However, carbohydrates (including sugar) are now readily available at every corner. People living in areas of London with higher prevalence of medical conditions linked to the metabolic syndrome seem to surrender to their human instincts and end up buying carbohydrates and sugar to a considerable extent. By contrast, people living in healthy neighborhoods seem to counter their evolutionary adaptation and buy considerable quantities of fibre. This difference in purchases is not explained by socio-economic conditions: income does not matter as much as one expects. By transcending conventional class boundaries, human biases, instead, seem to be the main obstacle to healthy eating. Our study suggests that the "trick" to not being associated with chronic diseases is eating less what we instinctively like (by not listening to the dopamine rushes in our brains), balancing all the nutrients, and avoiding the (big) quantities that are readily available.

In the future, we will explore the impact of two additional factors on health outcomes. The first is the city itself: certain city's forms are more appealing to pedestrians than others and, as such, one might wonder which forms are "healthier". The second factor is exercising. We are exploring the possibility of capturing exercising levels across an entire city with wearable devices. Too see why this is important, consider that, by exercising (even a little), an individual boosts his/her immune system, achieves a 20-50 percent reduction in sick days in the short term, and reduces the risk of chronic diseases in the long term [63].

In our cities, food is cheap and exercise discretionary, and health takes its toll. Technology could change that. With modern data analytics, the availability of new open data, recent advances in persuasive computing, and ever increasingly miniaturized health wearables, modern technologies are now best positioned to help people counter the dopamine rushes coming from sugar and fat, eat better, and exercise more.

Abbreviations
GP, General Practitioner; ONS, Office for National Statistics; OA, Output Areas; LSOA, Lower Super Output Areas; MSOA, Medium Super Output Areas; LA, Local Authority; IMD, Index of Multiple Deprivation.

Availability of data and materials
Our food dataset is available upon request. To preserve customer privacy, the data has been aggregated at MSOA level. The dataset is available on the project's site http://www.goodcitylife.org/food.

Competing interests
The authors declare that they have no competing interests. This work was done while LMA and DQ were employees of Nokia Bell Labs and LDP was employee of Tesco Labs. The authors' employers provided support in the form of salaries but did not have any additional role in the study design, data collection and analysis, or preparation of the manuscript. All work was done as part of the respective authors' research, with no additional or external funding.

Authors' contributions
LMA and RS worked on the prescription data collection and performed the analysis. DQ set the hypotheses and research questions. LDP worked on the food data collection. All authors contributed to establish the methodology and to write the manuscript. All authors read and approved the final manuscript.

Author details
[1]Nokia Bell Labs, Cambridge, UK. [2]University of Turin, Turin, Italy. [3]Centre for Urban Science and Progress (CUSP), King's College, London, UK. [4]Tesco Labs, London, UK.

Endnotes
[a] We also conducted all the analyses using prescription data from 2015, and obtained similar results.
[b] http://goodcitylife.org/food
[c] Fibres have a calorie intake of either 2 or 0 Kcals depending on the type of fibre, which is quite small since they mostly go through the digestive system without being assimilated.
[d] GP practice prescribing data—Presentation level: https://data.gov.uk/dataset/prescribing-by-gp-practice-presentation-level
[e] Some people using glucose testing equipment will have type 1 diabetes which is not obesity nor diet-related, but they account for less than 10% of the cases.
[f] Numbers of Patients Registered at a GP Practice January 2015: http://digital.nhs.uk/catalogue/PUB16357
[g] https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography
[h] English indices of deprivation 2015: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

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

References
1. Ekelund U, Ward HA, Norat T, Luan J, May AM, Weiderpass E, Sharp SJ, Overvad K, Østergaard JN, Tjønneland A et al (2015) Physical activity and all-cause mortality across levels of overall and abdominal adiposity in European men and women: the European prospective investigation into cancer and nutrition study (epic). Am J Clin Nutr 101(3):613–621
2. Flegal K, Kruszon-Moran D, Carroll M, Fryar C, Ogden C (2016) Trends in obesity among adults in the United States, 2005 to 2014. 315:2284
3. van Vliet-Ostaptchouk JV, Nuotio M-L, Slagter SN, Doiron D, Fischer K, Foco L, Gaye A, Gögele M, Heier M, Hiekkalinna T et al (2014) The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. BMC Endocr Disord 14(1):9
4. Kaur J (2014) A comprehensive review on metabolic syndrome. Cardiology research and practice 2014
5. Trattner C, Elsweiler D (2017) Investigating the healthiness of Internet-sourced recipes: implications for meal planning and recommender systems. In: Proceedings of the 26th international conference on World Wide Web. WWW. ACM, Geneva, pp 489–498

6.   De Choudhury M, Sharma S, Kiciman E (2016) Characterizing dietary choices, nutrition, and language in food deserts via social media. In: Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing. CSCW. ACM, New York, pp 1157–1170

7.   Ofli F, Aytar Y, Weber I, al Hammour R, Torralba A (2017) Is saki #delicious?: the food perception gap on Instagram and its relation to health. In: Proceedings of the 26th international conference on World Wide Web. WWW. ACM, Geneva, pp 509–518

8.   Boumtje PI, Huang CL, Lee J-Y, Lin B-H (2005) Dietary habits, demographics, and the development of overweight and obesity among children in the United States. Food Policy 30(2):115–128

9.   Lee H (2012) The role of local food availability in explaining obesity risk among young school-aged children. Soc Sci Med 74(8):1193–1203

10. Volkow ND, Wang GJ, Fowler JS, Tomasi D, Baler R Carter CS, Dalley JW (eds) (2012) Food and drug reward: overlapping circuits in human obesity and addiction. Curr Top Behav Neurosci 22:1–24

11. Schoenfeld JD, Ioannidis JP (2013) Is everything we eat associated with cancer? A systematic cookbook review. Am J Clin Nutr 97(1):127–134

12. Ahnert SE (2013) Network analysis and data mining in food science: the emergence of computational gastronomy. Flavour 2(1):4

13. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012

14. Prier KW, Smith MS, Giraud-Carrier C, Hanson CL (2011) Identifying health-related topics on Twitter. In: International conference on social computing, behavioral-cultural modeling, and prediction. Springer, Berlin, pp 18–25

15. Paul MJ, Dredze M (2011) You are what you tweet: analyzing Twitter for public health. In: Proceedings of the 5th international AAAI conference on weblogs and social media. ICWSM. AAAI Press, Menlo Park, pp 265–272

16. De Choudhury M, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. In: Proceedings of the 7th international AAAI conference on web and social media. ICWSM. AAAI Press, Menlo Park, pp 1–10

17. Mejova Y, Haddadi H, Noulas A, Weber I (2015) #foodporn: obesity patterns in culinary interactions. In: Proceedings of the 5th international conference on digital health. DH. ACM, New York, pp 51–58

18. Weber I, Mejova Y (2016) Crowdsourcing health labels: inferring body weight from profile pictures. In: Proceedings of the 6th international conference on digital health conference. DH. ACM, New York, pp 105–109

19. Kocabey E, Camurcu M, Ofli F, Aytar Y, Marin J, Torralba A, Weber I (2017) Face-to-bmi: using computer vision to infer body mass index on social media, 1–10. AAAI Press

20. West R, White RW, Horvitz E (2013) From cookies to cooks: insights on dietary patterns via analysis of web usage logs. In: Proceedings of the 22nd international conference on World Wide Web. WWW. ACM, New York, pp 1399–1410

21. Wagner C, Singer P, Strohmaier M (2014) The nature and evolution of online food preferences. EPJ Data Sci 3(1):38

22. Sajadmanesh S, Jafarzadeh S, Ossia SA, Rabiee HR, Haddadi H, Mejova Y, Musolesi M, Cristofaro ED, Stringhini G (2017) Kissing cuisines: exploring worldwide culinary habits on the web. In: Proceedings of the 26th international conference on World Wide Web. WWW. ACM, Geneva, pp 1013–1021

23. Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L (2011) Flavor network and the principles of food pairing. Sci Rep 1

24. Said A, Bellogín A (2014) You are what you eat! Tracking health through recipe interactions. In: RSWEB workshop at ACM recsys

25. Ge M, Ricci F, Massimo D (2015) Health-aware food recommender system. In: Proceedings of the 9th ACM conference on recommender systems. RecSys. ACM, New York, pp 333–334

26. Elsweiler D, Trattner C, Harvey M (2017) Exploiting food choice biases for healthier recipe recommendation. In: Proceedings of the 40th international conference on research and development in information retrieval. SIGIR. ACM, New York, pp 575–584

27. Abbar S, Mejova Y, Weber I (2015) You tweet what you eat: studying food consumption through Twitter. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems. CHI. ACM, New York, pp 3197–3206

28. Sharma SS, De Choudhury M (2015) Measuring and characterizing nutritional information of food and ingestion content in Instagram. In: Proceedings of the 24th international conference on World Wide Web. WWW. ACM, New York, pp 115–116

29. Mejova Y, Abbar S, Haddadi H (2016) Fetishizing food in digital age:# foodporn around the world. In: Proceedings of the 10th international AAAI conference on web and social media. ICWSM. AAAI Press, Menlo Park, pp 250–258

30. Wagner C, Aiello LM (2015) Men eat on mars, women on venus?: an empirical study of food-images. In: Proceedings of the ACM web science conference. WebSci. ACM, New York, pp 63–1633

31. Nevalainen J, Erkkola M, Saarijärvi H, Näppilä T, Fogelholm M (2018) Large-scale loyalty card data in health research. Digit Health 4:2055207618816898

32. Guidotti R, Rossetti G, Pappalardo L, Giannotti F, Pedreschi D (2018) Personalized market basket prediction with temporal annotated recurring sequences. IEEE Trans Knowl Data Eng

33. Mamiya H, Moodie EE, Buckeridge DL (2017) A novel application of point-of-sales grocery transaction data to enhance community nutrition monitoring. In: AMIA annual symposium proceedings. American Medical Informatics Association, pp 1253

34. Mamiya H, Moodie EE, Ma Y, Buckeridge DL (2018) Susceptibility to price discounting of soda by neighbourhood educational status: an ecological analysis of disparities in soda consumption using point-of-purchase transaction data in Montreal, Canada. Int J Epidemiol 47(6):1877–1886

35. Greger M, Stone G (2016) How not to die: discover the foods scientifically proven to prevent and reverse disease. Pan Macmillan, London

36. Avena NM, Rada P, Hoebel BG (2008) Evidence for sugar addiction: behavioral and neurochemical effects of intermittent, excessive sugar intake. Neurosci Biobehav Rev 32:20–39

37. Iozzo P, Guiducci L, Guzzardi MA, Pagotto U (2012) Brain pet imaging in obesity and food addiction: current evidence and hypothesis. Obes Facts 5:155–164

38. Chowdhury R, Warnakula S, Kunutsor S, Crowe F, Ward HA, Johnson L, Franco OH, Butterworth AS, Forouhi NG, Thompson SG et al (2014) Association of dietary, circulating, and supplement fatty acids with coronary risk: a systematic review and meta-analysis. Ann Intern Med 160(6):398–406

39. Clemens R, Kranz S, Mobley A, Nicklas T, Raimondi MP, Rodriguez J, Slavin J, Warshaw H (2012) Filling America's fibre intake gap: summary of a roundtable to probe realistic solutions with a focus on grain-based foods. J Nutr 142:1390–1401
40. Kant A, Schatzkin A, Harris TB, Ziegler RG, Block G (1993) Dietary diversity and subsequent mortality in the first national health and nutrition examination survey epidemiologic follow-up study. Am J Clin Nutr 57:434–440
41. Wansink B, Van Ittersum K, Painter J (2006) Ice cream illusions: bowls, spoons, and self-served portion sizes. Am J Prev Med 31:240–243
42. Whitney E, Rolfes SR (2007) Understanding nutrition. Cengage Learning, Boston
43. Curtis HJ, Goldacre B (2018) Openprescribing: normalised data and software tool to research trends in English nhs primary care prescribing 1998–2016. BMJ Open 8(2):019921
44. Das P, Samarasekera U (2013) The story of gbd 2010: a "super-human" effort. Lancet 380:2067–2070
45. Roberts W (2010) It's the cholesterol, stupid! Am J Cardiol 106:1364–1366
46. Esselstyn JC (2000) In cholesterol lowering, moderation kills. Clevel Clin J Med 67:560–564
47. Danilo C, Frank PG (2012) Cholesterol and breast cancer development. Curr Opin Pharmacol 12(6): 677–682
48. Neeland I, Turer AT, Ayers CR, Powell-Wiley T, Vega GL, Farzaneh-Far R, Grundy SM, Khera A, McGuire DK de Lemos JA (2012) Dysfunctional adiposity and the risk of prediabetes and type 2 diabetes in obese adults. 308:1150–1159
49. National Center for Chronic Disease Prevention and Health Promotion (2017) National diabetes statistics report. Atlanta, GA: centers for Disease Control and Prevention
50. Pratley RE (2013) The early treatment of type 2 diabetes. Am J Med 126(9):2–9
51. Sassi F, Devaux M, Church J, Cecchini M, Borgonovi F (2009) Education and obesity in four oecd countries. OECD, Directorate for Employment, Labour and Social Affairs, OECD Health Working Papers
52. The Economist (2012) The caveman's curse. why it is easy to get fat and hard to slim down. Special Report
53. Monteiro CA, Conde WL, Popkin BM (2007) Income-specific trends in obesity in Brazil: 1975–2003. Am J Publ Health 97(10):1808–1812
54. Amine E, Baba N, Belhadj M, Deurenberg-Yap M, Djazayery A, Forrester T, Galuska D, Herman S, James W, Mbuyamba J et al (1990) Diet, nutrition, and the prevention of chronic diseases. Report of a Joint WHO/FAO Expert Consultation 797
55. Gidlow C, Johnston LH, Crone D, Ellis N, James D (2006) A systematic review of the relationship between socio-economic position and physical activity. Health Educ J 65(4):338–367
56. Logue J, Walker J, Colhoun H, Leese G, Lindsay R, McKnight J, Morris A, Pearson D, Petrie J, Philip S et al (2011) Do men develop type 2 diabetes at lower body mass indices than women? Diabetologia 54(12):3003–3006
57. Graham H (2004) Socioeconomic inequalities in health in the uk: evidence on patterns and determinants. A Short Report for the Disability Rights Commission
58. Cookson R, Propper C, Asaria M, Raine R (2016) Socio-economic inequalities in health care in England. Fisc Stud 37(3–4):371–403
59. Williams DR, Yu Y, Jackson JS, Anderson NB (1997) Racial differences in physical and mental health: socio-economic status, stress and discrimination. J Health Psychol 2(3):335–351
60. Loucks EB, Rehkopf DH, Thurston RC, Kawachi I (2007) Socioeconomic disparities in metabolic syndrome differ by gender: evidence from nhanes iii. Ann Epidemiol 17(1):19–26
61. Muller L, Lacroix A, Lusk JL, Ruffieux B (2017) Distributional impacts of fat taxes and thin subsidies. Econ J 127(604):2066–2092
62. Grieve FG, Vander Weg MW (2003) Desire to eat high-and low-fat foods following a low-fat dietary intervention. J Nutr Educ Behav 35(2):98–104
63. Nieman DC (2011) Moderate exercise improves immunity and decreases illness rates. Am J Lifestyle Med 5(4):338–345