A large-scale study of food purchases and health outcomes: study and implications

Abstract

Background: Conditions associated to the metabolic syndrome have an enormous impact on people’s health. It is estimated that more than 300k premature deaths in Europe are caused by obesity and more than half of European citizens will be obese by 2050. That represents a heavy burden on healthcare, with more than €70B spent every year in Europe. Since healthy eating is one of the most effective intervention to counter the risks of metabolic syndrome, monitoring food consumption at scale is key for effective prevention. Traditional nutrition studies are costly and, most often, of limited scale. To partly fix that, researchers have resorted to digital data to infer what people eat and to estimate how that relates to their health.

Objective: For the entire city of London, we study the association between food purchases in grocery stores, as measured by the digital traces of customer loyalty cards, and consumption of medicines prescribed to treat different aspects of the metabolic syndrome.

Methods: We combine two sets of geo-referenced data. The first contains the record of all the 1.6B food items that 1.6M loyalty card owners bought in 2015 in all the London stores of Tesco, the largest grocery retailer in UK. From it, we extract food nutrients and other indicators of food healthiness. The latter contains information about all the 1.1B medical prescriptions written by every General Practitioner in 2016 in London, from which we infer the incidence of three aspects related to the metabolic syndrome: hypertension, high cholesterol, and diabetes. We study the data at the level of 942 Middle Super Output Areas, geographical areas containing an
average of about 8,000 residents. Using correlations, regression, and classification models, we explore the association between and medicine prescriptions and food purchases.

**Results:** Areas with high incidence of the ailments considered are characterized by high consumption of fat (Spearman rank correlation $\rho - [0.28, 0.36]$), sugar ($\rho - [0.31, 0.49]$), and especially carbohydrates ($\rho - [0.45, 0.57]$) and low consumption of protein ($\rho - [-0.22, -0.50]$) and fibre ($\rho - [-0.19, -0.50]$). A regression model that uses the diversity of nutrients and calorie intake can explain more than half of the variability in diabetes incidence (coefficient of determination $R^2 = 0.598$) and about a third for high cholesterol ($R^2 = 0.345$) and hypertension ($R^2 = 0.388$). A classifier trained over all the available features can identify unhealthy areas from their food consumption with up to 91% accuracy.

**Conclusions:** Our study shows that analytics of digital records of grocery purchase can be used as a cheap and scalable tool for health surveillance: the distribution of the food nutrients is far more predictive of food-related illnesses than socio-economic conditions. Being able to identify areas at higher health risk is just a first step. We detail how different stakeholders could help with prevention strategies.

**Keywords:** nutrition; grocery; loyalty card; GP prescriptions; diabetes; hypertension; cholesterol; metabolic syndrome; Tesco; London

**Introduction**

**Background**

Conditions associated to the metabolic syndrome have an enormous impact on people health. More than 300k premature deaths in Europe are caused by obesity [1]. In the United States, 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 risk of diabetes, heart disease, and even some cancers, which results in increased health-care spending (£70B in Europe every year), and ultimately costs lives.

Healthy eating is one of the most effective intervention to counter the risks of metabolic syndrome [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 metabolic disorders are hard to untangle. Many studies rely on data limited in scale and many others rely on people’s memory of what they have been eating, introducing “recall bias”. 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] to infer what Web users are likely to eat. The presence of recipes or of food-related posts, however, does not translate in a full understanding of what a large sample of the population eats.

To further those studies, we explore finer-grained associations between food purchases and medicine prescriptions at the level of Middle Super Output Areas (MSOA) of the entire city of London. To do that, we analyze for the first time the digital traces of purchases done using customer loyalty cards in all the Tesco grocery stores in London.

**Prior Work**

The incidence of conditions connected to the metabolic syndrome has been related to both lack of physical exercise and unhealthy food consumption.

Physical activity helps to lower the risk of a wide range of ailments, including heart disease, diabetes, breast cancer and depression. By exercising, people boosts their immune system and achieve a 20-50% reduction in sick days [8]. Kids who run around for just six minutes increase their levels of immune cells by nearly 50%, elderly women who take half-hour-a-day walks reduce their risk of getting an upper-respiratory illness down to 20% [9], and conditioned runners have their risk to just 8% [10]. But academics remain skeptical that exercise alone can reverse obesity trends. A change in diet is the most effective way to counter obesity.

It is assumed that people are able to choose what they eat, but that is only partly true. The amount of food consumed by an individual is influenced by: i) the habits and associations around food formed at a young age; ii) external factors (e.g., portion sizes); 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 [11]).

To produce medical evidence on which factors impact what people eat, the gold standard is the randomized, controlled clinical trial. Researchers track the effect of diets over long periods of time by asking participants to fill food diaries. However, such data comes with limitations. Under a research study, participants might be more likely to selectively remember more of the unhealthy foods they ate, and this could artificially inflate the correlation between eating certain foods and the disease under study [12]. At times, such methods produce research with conflicting results. A meta-review found that most common foods are linked with both a higher and lower risk of cancer [13].

The combination of large-scale data analysis and large datasets now makes it possible to study *health outcomes* at unprecedented scales [14]. However, health records of individual patients are not widely available, not least because of privacy
concerns. By contrast, most data coming from social media and the Web has 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 [15]. Microblogging platforms like Twitter have been used to monitor the public health discourse [16, 17] and to estimate the incidence of a wide range of pathologies from mental illnesses [18] to obesity [19].

Profile pictures from social media reveal our body mass [20], and if analyzed with the latest computer vision technologies, they have been found to be predictive of a person's BMI [21]. On the web, we find not only pictures but also countless sites dedicated to food. Digital traces from these online communities have allowed researchers to study dietary patterns of large populations [22, 23, 24], also in relation to culture [25]. Large databases of receipts have been studied to quantify receipt healthiness based on their ingredients and associated images [26, 5]. Information from these sites have been incorporated into food recommender systems [27] with the idea of nudging people towards healthier choices by proposing them healthy dishes that are similar to those they usually eat [28].

Web data allows for the study of segments of an entire population. 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 [29]. In USA, large areas suffer from obesity. Food deserts, for example, are low-income areas that have limited access to nutritious food. Recently, De Choudhury et al. analyzed millions of food-related Instagram images [30] and found that food deserts indeed consume food high in fat, cholesterol, and sugar more than other locations do [6]. In addition to which nutrients social media users objectively consume, images can disclose what social media perceive to have consumed. Indeed, from a food image, one can infer two quantities: first, how healthy a computer vision algorithm labels the image to be (based on what the image objectively 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]. Despite that, social media and Web-based sources suffer from a number of self-presentation and self-selection biases which might yield a distorted picture of the actual food consumption patterns [31, 32].

To sum up, prior work has shown that it is a challenge to run a convincing nutrition study because people do not want to alter their diet patterns for the duration of randomized trials, and because social media users represent specific segments of the population. What is needed is a new approach to measuring what real people (as opposed to study participants) eat and drink, ideally under naturalistic conditions.

**Study objectives and contributions**
For the entire city of London, we study the association between food purchases in grocery stores, measured by the digital traces recorded after the use of customer loyalty cards, and consumption of medicines prescribed to treat the metabolic syndrome. In so doing, we make the following main contributions:

- We combine two sets of geo-referenced data at zipcode level in London. One set contains 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 by all London’s General Practitioners (medical doctors) in 2016. From the anonymized and aggregated food purchased data, we extract food nutrients. From the publicly available prescription data, we infer the incidence of three food-related ailments: hypertension, cholesterol, and diabetes.
- Based on the latest nutrition studies, we formulate seven main research questions that relate food consumption with the three chronic diseases. We then operationalize the metrics that answer these questions.
- Our results revealed some associations with the three diseases that are in line with previous studies, and other associations that have not been explored in the literature. We estimate that the average diet of Londoners fails to comply with recommendations of the World Health Organization by a large margin: people consume way more sugar and fat than the recommended limits and very little fibre. Also, we learned that the two most important predictors of the three chronic diseases have little to do with individual nutrients and more to do both with the composition of a diet (the highest negative correlation of diabetes is with “nutrient diversity”) and with quantities of available food (the highest positive correlation of diabetes is with the weight of the “average” item purchased in a neighborhood). These features are not only descriptive but also predictive: the $R^2$ of a linear regression that predicts diabetes reaches 0.6, which means that the diversity of nutrients and item weight alone explain up to 60% of the variance in prevalence of diabetes in London neighborhoods.
- We conclude by showing how our methodology and findings can improve evidence-based public outreach initiatives and can inform the design of new consumer technologies.

Research questions

Metabolic syndrome might have a simple main cause. People consume more calories than they use, and the surplus is stored as fat. This is therefore our first question:

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

However, it might be less about the number of calories than about their concentration, and that is for two main reasons. The first has to do with the pleasure
centers within people's brains (i.e., dopamine reward system). 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 [11]. 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. The second reason has to do with the stretch receptors in our stomachs. People tend to eat the same amount of food at a meal, regardless of calorie count, mainly because stretch receptors in the stomach are based on food volume. When much of that volume is a zero-calorie component like fiber\textsuperscript{a}, then a person can eat more food and gain less weight. So we ask:

**RQ2: Is calorie concentration positively associated with the three chronic diseases?**

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 fibers, of which at least “a quarter of the American population is not reaching an adequate intake” [12]. 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 an invitation, and that invitation is insulin. Without insulin, the cells cannot accept glucose and, as a result, the glucose builds up in the blood. Inappropriate fat storage may keep cells from responding properly to insulin, causing so-called insulin resistance. Eventually blood-sugar levels rise out of control and the patient develops diabetes. Even among healthy individuals, a high-fat diet impair 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 [33]. When people consume a sugary food the brain releases dopamine [34]. Dopamine signals pleasure and is a chemical involved in drug addiction [11]. This evidence leads to our third research question:

**RQ3: Are fat, carbohydrates and sugar positively associated with the three chronic diseases?**

Not all fats affect the muscle cells in the same way. Palmitate (the saturated fat found mostly in meat, dairy, and eggs) causes insulin resistance. On the other hand, oleate (the mono-unsaturated fat found mostly in nuts, olives, and avocados) protects against the detrimental effects of the saturated fat. Research findings on the impact

\textsuperscript{a} Most plant foods are both high in nutrients and low in calories.
of saturated fats are controversial though. In 2014, a large meta-analysis showed no relationship between saturated fats and heart disease [35].

**RQ4: What is the relationship between saturated fats and the three chronic diseases?**

On a more positive note, consider fibers. Humans evolved over millions of years eating mostly wild plants, likely in excess of one hundred grams daily [36]. That is much more than what the average person eats today. Given their health benefits, we posit the following question:

**RQ5: Are fibers negatively associated with the three chronic diseases?**

Going beyond individual nutrients, one could study their composite impact. More specifically, research has shown that healthy diets is associated with diversity of nutrients [37]. Therefore, our next research question is:

**RQ6: Is nutrient diversity negatively associated with the three chronic diseases?**

Finally, the amount of food people consume is often influenced by external factors, including the size of their plate [38]. Our next research question is then:

**RQ7: Is the overall weight of food consumed positively associated with the three chronic diseases?**

**Methods**

**Datasets**

**Food Purchases**

At all the 411 Tesco 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 are stored in the following anonymized form: customer postcode, store postcode, productID, and timestamp. For each productID, the record further reports its net weight, total energy, fats, saturated fats, carbohydrates, free sugars, proteins, and fibers. The last six elements are expressed as grams of substances contained in the product. Using standard guidelines [39], 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 fibers\(^b\).  

**Chronic diseases**

\(^b\) Fibers have a calorie intake of either 2 or 0 Kcals depending on the type of fiber, which is quite small since they mostly go through the digestive system without being assimilated.
At all the 1,174 general practices (GPs) in Greater London, 1.1B medicine items were prescribed in the entire 2016. Such prescription data\(^c\) has been recently made publicly available [40] in the following form: \(GP\) *postcode*, *medicineID*, and *timestamp*. For each *medicineID*, the record further reports its active ingredients, from which the corresponding diseases can be inferred.

Food purchases were recorded at postcode level. To match food purchases with chronic diseases, medicines should be associated with patients' postcodes. To determine the likely fraction of each postcode's residents served by a given GP, we gather open data that reports, for every postcode, the number of people who were registered to each GP in London as of year 2015 [41]. We associate the GP's prescriptions to a postcode proportionally to the fraction of its residents associated with that GP. As a result, prescriptions are now associated with postcodes. We then map 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 [42]. In conclusion, for each area \(a\), 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 [43]. To capture the incidence of hypertension, we consider prescriptions of 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 [44, 45]. Cholesterol also seems to help some cancers migrate and invade more tissue [46]. To capture the incidence of high cholesterol, we consider lipid-regulating drugs (e.g., Statins).

*Diabetes*. Diabetes is characterized by chronically elevated levels of sugar in the blood [47]. 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 [48]). 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.

---

\(^c\) Prescription data is available for year 2015 too. We have replicated all the results using 2015 data: the results are all very similar.
To capture the incidence of diabetes, we consider prescriptions of 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).

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

$$\text{diabetes}_a = \frac{\text{prescriptions for diabetes}_a}{\text{residents}_a}.$$

In a similar way, we compute $\text{hypertension}_a$, $\text{cholesterol}_a$, $\text{smoking}_a$, and $\text{obesity}_a$.

To prevent data sparsity, we conduct the analysis at for geographical areas coarser than postcodes. We aggregate the data at postcode level into 942 Medium Super Output Areas (MSOAs) which contain an average of 8,250 residents. To ease presentation, we refer to MSOAs as areas or neighborhoods.

**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 [50]. 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” [51]. In developed countries, there is a difference between cities and suburban areas: the more affluent urbanites are usually fitter than rural residents [52].

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 [53]. We focus on the factors that have been found associated with specific food consumption patterns and, ultimately, with chronic diseases. These are average income, education level, gender distribution (%female), and average age at MSOA level.

**Estimating eating habits of an area**

From the purchase records, there is no way to know how many people the food purchased with a loyalty card will feed. Therefore, 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 using a set of metrics defined below.

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

\[
\text{calorie consumption} @ a = \frac{\sum_{p \in P_a} Kcal(p)}{|P_a|},
\]

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

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

\[
\text{calorie concentration} @ a = \frac{\sum_{p \in P_a} Kcal(p)}{\sum_{p \in P_a} \text{weight}(p)},
\]

which reflects the concentration of calories in the area’s “average” product.

For each area, we also compute the average number\(^4\) of calories given by individual nutrients in a product, on average:

\[
\text{nutrient} @ a = \frac{\sum_{p \in P_a} Kcal(\text{nutrient}, p)}{|P_a|},
\]

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

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:

\[
\text{nutrient diversity} @ a = -\sum_{j} p(\text{nutrient}_j, a) \cdot \log p(\text{nutrient}_j, a),
\]

\(^4\) We conducted the experiments also using the relative fraction of calories given by a nutrient. For the sake of brevity, we do not report the results here, but all outcomes were very similar to those reported in this paper.
\[
p_{\text{nutrient}, a} = \frac{\sum_{p \in P_a} Kcal(nutrient, p)}{\sum_{p \in P_a} Kcal(p)}.
\]

Last, we also compute the average item weight:

\[
\text{item weight } @ a = \frac{\sum_{p \in P_a} \text{weight}(p)}{|P_a|}.
\]

Reproducibility

For the sake of reproducibility of our analysis, we publicly share all the data used in work, aggregated at MSOA level\(^e\).

Results

Relative abundance of nutrients

Figure 1. Distribution of the fraction of energy coming from different nutrients in the different areas in London. The intervals recommended by the guidelines for a healthy diet by the World Health Organization are highlighted.

The nutrition guidelines by the World Health Organization (WHO) recommend to limit the relative energy supply derived from each nutrient within specific ranges; for example, fats should contribute no more than 30\% [54] to the total intake. By plotting the distributions of \( p_{\text{nutrient}, a} \) values across neighborhoods (Figure 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 fibers less than the recommended

\(^e\) The URL with the link to the data will be provided upon submission of the camera-ready version.
amounts. The extent to which residents collectively depart from recommended limits changes across the city (see Figure 2 for fats and sugars), with the area of Chelsea being closest to the recommended limits and Sutton being the farthest.

These overall results reflect food purchases for an entire year. It would be therefore likely to observe long-term consequences when dwelling on the food data. To this end, we match it with health outcomes (derived from our medical prescriptions) and start to answer our research questions.

Figure 2. Extent to which residents collectively depart from recommended limits in consumption of fat (left) and free sugar (right). Areas in red exceed the recommended limit the most, areas in gray were left out because not significant.

**Correlation between nutrients and medical prescriptions**

Figure 3. Spearman rank correlations between incidence of chronic diseases, nutrients and calories in purchased foods. All correlations are significant with p<0.001.

We compute the Spearman rank correlation between prescriptions and all the measures of food consumption. As shown in Figure 3, calorie consumption is strongly correlated with cholesterol and hypertension and, 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 Figure 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 with calorie consumption; in the second row, instead, we have plots that related the diseases with calorie concentration. In both cases, we see that, as calories increase, the prevalence of any of the three diseases increases. To ease interpretation of these x-y plots, we rescaled the x-axis in terms of premium. For example, we transform the average item’s weight in area a as:

\[ \text{premium item weight} @ a = \frac{\text{item weight} @ a - \mu(\text{item weight})}{\mu(\text{item weight})} \]

where \( \mu(\text{item weight}) \) is the average weight across all areas. If the premium is zero, then the area’s weight is equal to the average value in London. If the premium is 0.1, then the area’s weight is 10% higher than the average value. In the plot related to calorie consumption (first row in Figure 4), we see that as the consumption exceeds the average value (i.e., \( x > 0 \) ) the likelihood of any of the three diseases stops increasing linearly and starts to exponentially increase.

All these three nutrients are associated with chronic illness in expected ways: fat and sugar are positively associated with the three diseases, while fibers are negatively associated (RQs3+5) (Figure 3). Indeed, the prevalence of each of the diseases increases as carbohydrates (fourth row in Figure 4) increases. The relationships between the diseases and the nutrients of fat (third row in Figure 4) is not as clear cut as one would have expected: they come with high variability and, as such, the corresponding relationships are not conclusive. The correlation between diseases and consumption of saturated fat is comparable to that with fat – a little weaker, in fact (RQ4). By contrast, it is pretty clear that both more proteins (sixth row) and more fibers (seventh row) are associated with a lower prevalence of the three chronic diseases.

Both nutrient diversity and average item weight show high correlations, of opposite sign. Item weight resembles calorie consumption, as the two quantities are not orthogonal. Nutrient diversity has the highest negative correlation with incidence of all three chronic diseases (RQs6+7). In fact, in Figure 4, one sees that the presence of any of the three diseases rapidly decreases with nutrient diversity, while it exponentially increases with calorie concentration and average weight after a certain threshold. West and central London tend to consume fibers and diversify their nutrients the most (Figure 5). This is further confirmed by the presence of the City (central London) and Chelsea (West London) as consumers of high-fiber and highly-diverse foods in the quadrants of Figure 6, in which we place neighborhoods
in a bi-dimensional nutrient space and relate their position with prevalence of diabetes-related prescriptions.
Figure 4. The relationship between hypertension, cholesterol, and diabetes (columns) and food-related features (rows). Straight lines represent the average disease incidence.

Figure 5. London maps reflecting purchases of fibers (left) and nutrient diversity (right).

Figure 6. London areas placed quadrants with (top) energy from fibers per item on the x, and energy from carbs on the y, and (bottom) average energy —calories— per item on the x, and nutrient diversity on the y. Each point represents a single MSOA. The horizontal and vertical black lines represent the median values. Colors reflect the prevalence of diabetes-related prescriptions.
So far we have considered nutrients individually. However, nutrients 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 Figure 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 fibers. These correlations meet expectation, so we can safely presume that our food purchases, in aggregate, reflect typical food consumption and, for example, are not skewed towards specific niches.

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. Each of the three chronic diseases is the outcome variable of an ordinary least squares regression, with nutrient measures and a set of control variables as the predictor variables. Where necessary, predictor variables undergo a logarithmic
transformation, and in addition, we apply a min-max rescaling of each variable, which allows us to judge the relative influence of each factor: the larger the absolute value of the coefficient associated to a feature is, the higher the relative importance of feature in predicting the outcome.

We first try a regression that considers individual nutrients (carbs, fats, sugar, proteins, fibre) as independent variables. Table 1 shows the results. To sum up, carbohydrates, fats and to less extent sugar are associated with the three chronic diseases, and the presence of proteins and fibers counter 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 features. In particular for diabetes we observe that it seems to be more prevalent among males, as expected from previous studies [56]. Overall, nutrients and demographic features jointly explain more than one third of the variability of hypertension ($R^2 = 0.388$) and cholesterol ($R^2 = 0.345$) and almost 60% of the variance in prevalence of diabetes in London neighborhoods. The Durbin-Watson statistic reported in the table checks for the presence of autocorrelation of residuals in the regression: when values are close to 2 indicate no autocorrelation.

During our previous correlation analysis, we have found that nutrients are not the variables with the highest correlations. Features with the highest correlations included nutrient diversity for all ailments, calorie consumption for cholesterol and hypertension and calorie concentration for diabetes. One might now wonder whether a linear regression analysis solely based on a combination of diversity and calorie intake would be informative. Indeed, 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 measures alone explain up to 38% of the variance in the prevalence of hypertension, up to 34% for cholesterol, and up to 59% for diabetes.

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension</td>
<td>Adjusted R2 = 0.388</td>
<td></td>
<td></td>
</tr>
<tr>
<td>α (intercept)</td>
<td>0.3241</td>
<td>0.050</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Carbs</td>
<td>0.7652</td>
<td>0.075</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fats</td>
<td>0.3225</td>
<td>0.066</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sugar</td>
<td>0.2402</td>
<td>0.072</td>
<td>0.001</td>
</tr>
<tr>
<td>Proteins</td>
<td>-0.2868</td>
<td>0.050</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fibre</td>
<td>-0.0627</td>
<td>0.051</td>
<td>0.216</td>
</tr>
</tbody>
</table>
### Cholesterol

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>α (intercept)</td>
<td>0.2645</td>
<td>0.047</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Carbs</td>
<td>0.5877</td>
<td>0.070</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fats</td>
<td>0.3382</td>
<td>0.062</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sugar</td>
<td>0.2441</td>
<td>0.067</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Proteins</td>
<td>-0.2745</td>
<td>0.046</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fibre</td>
<td>-0.0268</td>
<td>0.047</td>
<td>0.569</td>
</tr>
<tr>
<td>Income</td>
<td>-0.0184</td>
<td>0.039</td>
<td>0.640</td>
</tr>
<tr>
<td>%Females</td>
<td>-0.2322</td>
<td>0.036</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average age</td>
<td>0.1272</td>
<td>0.035</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Education</td>
<td>0.1751</td>
<td>0.039</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Durbin-Watson stat.</td>
<td>2.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Diabetes

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>α (intercept)</td>
<td>0.5073</td>
<td>0.041</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Carbs</td>
<td>0.57659</td>
<td>0.061</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fats</td>
<td>0.5002</td>
<td>0.054</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sugar</td>
<td>0.4992</td>
<td>0.059</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Proteins</td>
<td>-0.5137</td>
<td>0.041</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fibre</td>
<td>-0.1312</td>
<td>0.041</td>
<td>0.002</td>
</tr>
<tr>
<td>Income</td>
<td>-0.1222</td>
<td>0.034</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>%Females</td>
<td>-0.3536</td>
<td>0.031</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average age</td>
<td>-0.0290</td>
<td>0.030</td>
<td>0.342</td>
</tr>
<tr>
<td>Education</td>
<td>-0.0947</td>
<td>0.035</td>
<td>&lt;0.006</td>
</tr>
<tr>
<td>Durbin-Watson stat.</td>
<td>2.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Hypertension

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>α (intercept)</td>
<td>0.5582</td>
<td>0.064</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Calorie consumption</td>
<td>0.30228</td>
<td>0.070</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 2. Linear regressions that predict the three chronic diseases from item weight, nutrient diversity and control variables (income, gender, age, education level).
<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nutrient diversity</td>
<td>-0.5182</td>
<td>0.069</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Income</td>
<td>0.0615</td>
<td>0.041</td>
<td>0.131</td>
</tr>
<tr>
<td>%Females</td>
<td>-0.2210</td>
<td>0.038</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average age</td>
<td>0.1627</td>
<td>0.037</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Education</td>
<td>-0.2309</td>
<td>0.041</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Durbin-Watson stat.</td>
<td>2.033</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Cholesterol**

Feature | Coefficient | Std. error | p-value |
---------|-------------|------------|---------|
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      |            |         |

**Diabetes**

Feature | Coefficient | Std. error | p-value |
---------|-------------|------------|---------|
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      |            |         |

Figure 8. Sensitivity analysis of the regression. $R^2$ for different combinations of features are reported.
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 (Figure 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, the prediction performance of nutrient diversity and calorie intake combined is comparable to the one of all the individual nutrients combined.

Finally, we describe how we use our features of food purchases to build classification models which can identify areas that are healthy or unhealthy in terms of the prevalence of the three chronic diseases. We formulate the problem as one of binary classification, in which the goal is to identify areas which fall into the top quartile of each of the three diseases (higher scores corresponded to higher presence of a disease). We define the response variable $y_i$ as 0 if area $i$ is in the 1$^{st}$ quartile of the disease distribution and 1 if it is in the first quartile.

This formulation effectively prunes the middle quartiles, since we are interested in classifying the extreme examples. In so doing, we also ensure a roughly 1:1 ratio of positive to negative examples in each class, as opposed to a 1:3 ratio had we retained the 2$^{nd}$ and 3$^{rd}$ quartiles. We compute the mean accuracy of 10 iterations using a Random Forest classifier. 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. We test three groups of features: the demographic features (gender, age, income, education), the two most predictive features from the food data (calorie concentration and nutrient diversity), and those two sets jointly. As expected from previous analysis, demographic factors have lower predictive power but help when combined with the food features. The classifier that uses all features correctly identifies healthy areas 91% of the times for diabetes, 82% of the times for hypertension, and 81% of the times for cholesterol.

Table 3. Accuracy of Random Forest classifier of London areas in the top and bottom quartiles of the three diseases’ prevalence. The predictive features are six: gender, average age, education level, item weight, nutrient diversity, and calorie concentration. The accuracy of a random guess is 0.5.
Discussion

Principal Results

In London, socio-economic conditions matter far less than what people eat. Eating less calories and opting for a diet with diverse nutrients are the most important predictors associated with healthy areas. In terms of which nutrients matter, as opposed to areas suffering from the three diseases, healthy areas tend to buy more fibers and far less carbohydrates (including sugar). Also, it is not only about calorie consumption. It is more about concentration of calories, which previous studies have been found to lead to forms of addiction [11] and which this study related to areas suffering from the three chronic diseases, especially diabetes. By combining these 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.

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, for the first time, 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 (most of past research has looked at that relationship). 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 the chronic diseases tend to do.

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 concentrated mainly on treating diseases rather than preventing them. State-based prevention strategies might be justified, not least because of externalities. Unhealthy eating harms not only oneself but also others, in that, it adds pressure to food supplies and 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 [57]. 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 [58].
**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 healthy areas and areas suffering from chronic diseases. Mining digital health data could 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.

**Limitations**

This study has three main limitations. The first is sample bias: the sample comes from one grocery retailer. To partly tackle biases of self-selection by consumers, we have teamed up with the largest grocery retailer in the country.

The second limitation is that 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.

The third limitation is that our study does not fully explain health outcomes. That was not the intention though as key sets of data are missing on, for example, restaurant consumption and exercising levels across neighborhoods. Without them, health outcomes cannot be fully explained.

**Conclusions**

It was healthy and adaptive for our primate brains to drive us to eat sugar and carbohydrates when only wild grass was at hand. However, carbohydrates (including sugar) are now readily available at every corner (shop). Londoners in neighborhoods suffering from chronic diseases seem to surrender to their human instincts and end up buying carbohydrates and sugar to a considerable extent. By contrast, Londoners in healthy neighborhoods seem to counter their evolutionary adaptation and buy considerable quantities of fibers. 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 for 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. To 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 [8].

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.

Acknowledgements
We would like to thank Tesco for sharing the anonymized data of customer purchase for the purpose of this study.

Conflicts of Interest
This work was done while Luca Maria Aiello and Daniele Quercia were employees of Nokia Bell Labs and Lucia Del Prete 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, decision to publish, or preparation of the manuscript. All work was done as part of the respective authors’ research, with no additional or external funding.

Abbreviations
MSOA: Medium super output area
WHO: World health organization

References
3. Van Vliet-Ostaptchouk JV, Nuotio ML, Slagter SN, Doiron D, Fischer K, Foco L and Joensuu A. The prevalence of metabolic syndrome and metabolically

4. Kaur J. A comprehensive review on metabolic syndrome. Cardiology research and practice; 2014. PMID: 24711954


16. Prier KW, Smith MS, Giraud-Carrier C, and Hanson CL. Identifying health-related topics on twitter. Proceedings of the 4th international conference on social computing, behavioral-cultural modeling, and prediction (SBP); Springer; College Park, MD, USA.

18. De Choudhury M, Gamon M, Counts S, and Horvitz E. Predicting depression via social media. Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM); 2013; Boston, MA, USA.


30. Sharma SS, De Choudhury M. Measuring and characterizing nutritional information of food and ingestion content in Instagram. Proceedings of the 24th ACM International Conference on World Wide Web (WWW); 2015; Firenze, Italy.


51. The caveman’s curse. Why it is easy to get fat and hard to slim down. The economist; 2012.


56. Logue J, Walker JJ, Colhoun HM, Leese GP, Lindsay RS, McKnight JA and Wild SH. Do men develop type 2 diabetes at lower body mass indices than women? Diabetologia; 2011; 54(12):3003-3006. PMID: 21959958
