British Journal of Nutrition (2014), **112**, 61–69
doi:10.1017/S0007114514000221

Principal components analysis of diet and alternatives for identifying the combination of foods that are associated with the risk of disease: a simulation study

Ioannis Bakolis[1]*, Peter Burney[2] and Richard Hooper[3]

[1] *Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK*

[2] *Respiratory Epidemiology and Public Health Group, Imperial College, National Heart and Lung Institute, Emmanuel Kaye Building, Manresa Road, London SW3 6LR, UK*

[3] *Centre for Primary Care and Public Health, Blizard Institute, Barts and The London School of Medicine and Dentistry, Abernethy Building, 2 Newark Street, Whitechapel, London E1 2AT, UK*

(Submitted 5 February 2013 – Final revision received 6 December 2013 – Accepted 13 January 2014 – First published online 11 April 2014)

British Journal of Nutrition

https://doi.org/10.1017/S0007114514000221 Published online by Cambridge University Press

Abstract

Dietary patterns derived empirically using principal components analysis (PCA) are widely employed for investigating diet–disease relationships. In the present study, we investigated whether PCA performed better at identifying such associations than an analysis of each food on a FFQ separately, referred to here as an exhaustive single food analysis (ESFA). Data on diet and disease were simulated using real FFQ data and by assuming a number of food intakes in combination that were associated with the risk of disease. In each simulation, ESFA and PCA were employed to identify the combinations of foods that are associated with the risk of disease using logistic regression, allowing for multiple testing and adjusting for energy intake. ESFA was also separately adjusted for principal components of diet, foods that were significant in the unadjusted ESFA and propensity scores. For each method, we investigated the power with which an association between diet and disease could be identified, and the power and false discovery rate (FDR) for identifying the specific combination of food intakes. In some scenarios, ESFA had greater power to detect a diet–disease association than PCA. ESFA also typically had a greater power and a lower FDR for identifying the combinations of food intakes that are associated with the risk of disease. The FDR of both methods increased with increasing sample size, but when ESFA was adjusted for foods that were significant in the unadjusted ESFA, FDR were controlled at the desired level. These results question the widespread use of PCA in nutritional epidemiology. The adjusted ESFA identifies the combinations of foods that are causally linked to the risk of disease with low FDR and surprisingly good power.

Key words: **Principal components analysis: Dietary patterns, Nutritional epidemiology: Logistic regression: Monte Carlo simulation**

Over the last two decades, there has been an explosion in the use of principal components analysis (PCA) to identify dietary patterns in nutritional epidemiological studies[1–10]. The main reason for the explosion of PCA is that recent results have raised questions about the role of diet in the aetiology of certain chronic diseases. A lack of empirical evidence from randomised controlled trials based on observational findings has been noted in chronic diseases such as asthma[11], cancer[12,13] and CVD[14], and it has been argued that single foods or nutrients may be less important than dietary patterns in causing disease. Since the use of a single variable to explore associations between diet and disease has led to unreliable results, an alternative is to look at a small number of dietary dimensions each made up of a combination of foods with the use of PCA.

PCA of data from FFQ allocates food items according to the degree with which their reported intakes are correlated. The principal components identified are referred to as 'dietary patterns' and can be investigated in relation to health and disease. This approach proved successful, for example, in two large US cohorts, in which the same two patterns of diet – a 'prudent' dietary pattern characterised by intake of vegetables, fruit, legumes, whole grains, fish and poultry and a 'Western' dietary pattern characterised by intake of red

Abbreviations: ESFA, exhaustive single food analysis; FDR, false discovery rate; PCA, principal components analysis.

* **Corresponding author:** I. Bakolis, email i.bakolis@imperial.ac.uk

meat, processed meat, refined grains, sweets and desserts, French fries and high-fat dairy products – were identified and subsequently linked to differences in the occurrence of CHD[15,16], colon cancer[17,18] and chronic obstructive pulmonary disease[19,20].

PCA is often claimed to resolve the issue of confounding between different dietary exposures[21,22], though, even where there is a causal effect from diet, it is not clear whether the foods singled out in a PCA are those that are directly associated with the risk of disease, or are simply confounded with other foods. Some authors have also suggested that by aggregating the effects of different foods, PCA can demonstrate the effects that are too small to be detected by analysing each food separately[2,23]. Slattery[24] claimed that eating patterns derived from PCA characterise the diet-associated disease risk better than any one food or nutrient. However, in order for this statement to be useful in disease prevention, we need PCA to identify all of the foods (and only those foods) that, in combination, increase or decrease the risk of disease.

As far as we are aware, justifications of this kind for using PCA have not been critically evaluated. We set out to investigate, using simulation, whether PCA really performs better in these respects than an analysis of individual food intakes.

Materials and methods

Simulation

Simulations were performed in Stata 10 (Stata Corporation). To simulate a dietary dataset with a realistic correlation structure, we sampled with replacement from a reference dataset of real FFQ data. We used two different sets of reference dietary data to allow replication of our findings. These datasets comprised 856 adults aged 16–50 years living in Greenwich who took part in the Food, Lifestyle and Asthma in Greenwich (F.L.A.G.) survey (dataset 1)[25] and 201 adults aged 29–54 years living in Ipswich and Norwich who took part in the UK European Community Respiratory Health Survey (ECHRS) II diet survey (dataset 2)[26]. Quantitative FFQ recorded frequencies of 217 different foods (from never to 6 d/week) in the FLAG survey and of seventy-four different foods (from never to 7 d/week) in the UK ECRHS II diet survey over the previous 12 months. Using both datasets, we created food intake variables by estimating weekly intake (g) of foods and food groups by multiplying frequency of consumption by the weight of standard portion sizes using British food composition tables[27]. The UK ECHRS II study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving human participants were approved by the Ipswich Hospital and Norfolk and Norwich Hospital ethics committees in the UK[26]. The F.L.A.G. survey was approved by the Greenwich Research Ethics Committee[25]. Written informed consent was obtained from all the participants in both studies.

In each simulation, we assumed that disease risk depended on a linear combination of m food intakes derived from the FFQ. Let us suppose that these foods are indexed $i_1, i_2, \ldots, i_m$, and absolute food intakes $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$ are standardised to have zero mean and unit standard deviation. We assumed a logistic model for the disease risk, p, that is:

$$p = 1/(1 + \exp(-(a + b_1 x_{i_1} + b_2 x_{i_2} + \ldots + b_m x_{i_m}))). \qquad (1)$$

This model has been widely used in previous simulated epidemiological studies[28,29]. We chose the constant a so that the baseline risk at the average intake of all foods was 0·15, i.e. $a = \ln(0{\cdot}15/(1 - 0{\cdot}15))$. The constants $b_1, b_2, \ldots, b_m$ were chosen so that the OR per standard deviation of food intake was 1·5 or 1/1·5 depending on whether the food was assumed to increase or decrease the risk of disease, i.e. $b_j = \pm\ln(1{\cdot}5)$. This is comparable to significant OR for dietary patterns that were identified in the systematic review by Newby & Tucker[4].

The value of m was chosen to be about one in seven of the total number of foods on the FFQ, i.e. $m = 30$ (of 217) in dataset 1 and $m = 10$ (of 74) in dataset 2. The m foods were chosen in two different ways. First, they were randomly chosen in each simulation. Results of these simulations inform us about the average performance of different methods when we do not restrict *a priori* the combinations of foods that might be important for disease risk. We considered two models of this kind: in model 1, all m foods were assumed to be protective; in model 2, all m foods were assumed to increase the risk of disease. Second, they were predetermined in each simulation to be foods making up a 'Western' dietary pattern, which was assumed to be positively associated with the risk of disease (model 3). These foods are listed in Table 1, and were chosen as food intakes with the highest positive loadings on a 'Western' dietary pattern obtained using PCA from the original reference dataset (further details are available from the authors). Model 3 might be expected to favour PCA as a means of identifying dietary associations, since the model is based on a principal component in the population.

Having simulated the food intake data for a new individual in the sample, we then calculated the probability of the outcome, p, using equation 1. We determined whether or not the individual had the disease by generating a uniform random number between 0 and 1, and observing whether it was less than p.

The simulated dietary data were then subjected to a PCA (conducted on the correlation matrix of the reported food intakes) with varimax rotation, and the resulting dietary patterns were investigated for their associations with the risk of disease using logistic regression. We considered the results of extracting two, five and ten principal components, since this was the range of components identified in the majority of dietary pattern studies[3,4].

For comparison, we looked at the results of analysing each food on the FFQ separately in relation to the risk of disease, a process we refer to as an exhaustive single food analysis (ESFA). This is a univariate or independent screening approach as we know from, for example, SNP screening in statistical genetics[30]. All regression analyses of dietary patterns and individual food intakes were adjusted for total energy intake. Adjustment for energy intake is strongly advised in the analysis of observational nutritional studies[31,32]. As pointed out by Willet[31], adjustment for

British Journal of Nutrition

https://doi.org/10.1017/S0007114514000221 Published online by Cambridge University Press

Table 1. List of foods comprising a 'Western' dietary pattern for each dataset*

Dataset 1	Dataset 2
Roast potatoes	Sausages
Ham	Doughnuts, pastries and tarts
Ice cream	Beer
Pork (roast, chops)	Corned beef and luncheon meat
Pork stew, casserole	Hard cheeses
Omelette/scrambled egg	Tomato ketchup
Fruit pies, tarts, crumbles	Pizza
Beef stew, casserole, mince, curry	Beef burger
Sponge cakes	Fried egg, scrambled egg, omelette
Fried fish in batter/breadcrumb	Chips
Baked beans	
Chocolate biscuits	
Sandwich/cream biscuits	
Corned beef, spam, luncheon meat	
White bread and rolls	
Fizzy soft drinks (e.g. Coke)	
Bacon	
Fried egg	
Milk chocolate	
Breadcrumbed chicken (e.g. chicken nuggets)	
Crisps	
Sponge puddings	
Tomato ketchup	
Chocolate snack bars	
Meat pizza	
Other fried snacks	
Chips	
Sausages (beef, pork)	
Beef burger, hamburger	
Pies/pasties/sausage rolls	

* Dataset 1 is from the Food, Lifestyle and Asthma in Greenwich (F.L.A.G) survey[25]; Dataset 2 is from the UK European Community Respiratory Health Survey (ECRHS) II diet survey[26].

total energy intake should be considered because the level of intake might be a risk factor, might distort the effect of a food or a nutrient on the potential outcome, and variations in nutrient intake between individuals might reflect variations in individuals' energy intake levels.

ESFA was, in the first instance, unadjusted for the effects of other foods, but in order to deal with confounding, we also carried out ESFA adjusting for other foods using three different methods:

(a) Adjusting for the first five principal components of diet. Because net confounding by correlated foods and patterns could result in biased associations between diet and disease, adjustment for dietary patterns has been suggested in the literature[33]. We chose five components of diet because that was the number of dietary patterns of diet that were identified in the original populations (data not shown).

(b) Adjusting for all foods that were significant in the unadjusted ESFA. In this case, as a first step, we ran an ESFA procedure adjusting for total energy intake keeping the food variables (covariates) with those regression coefficient estimates that had a P value lower than a specific threshold. The threshold was determined by the procedure of Benjamini & Hochberg[34], controlling the rate of our false discoveries at 20 %. Then, we re-ran an ESFA procedure adjusting for energy intake and for all these foods (covariates) that were significant in the first round of analysis (unadjusted ESFA). This method is conceptually similar to the iterative sure independence screening method proposed by Fan & Lv[35]. However, herein, we chose our covariates based on a multiple test procedure and not on a penalised likelihood method. Furthermore, we aimed to control the false discovery rate (FDR), whereas the iterative sure independence screening method focused on missed discoveries.

(c) Adjusting for a propensity score for predicting the amount of each index food intake consumed from other food intakes[36]. Once the propensity score was estimated, it was used as a confounder in our multivariate model[37].

The process of simulation and testing was repeated a large number of times (10 000) in order to determine the long-run performance of PCA and ESFA in detecting the associations between diet and disease.

Using the 'powerlog' sample size calculation routine in Stata[38], we determined that a sample size of 330 would achieve 80 % power at the 5 % significance level to detect an OR of 1·5 per standard deviation, using an unadjusted logistic regression with no allowance for multiple testing. We present results herein for sample sizes of 300, 1200 and 4800.

Evaluating the performance of exhaustive single food analysis and principal components analysis

First, we investigated the statistical power with which ESFA and PCA could detect whether there was any association between diet and disease. For the ESFA, we considered that an association had been found if any of the food intakes were significantly associated with the risk of disease after applying a Bonferroni correction for the number of foods tested (family-wise $P<0·05$)[39]. For the PCA, we considered an association had been found if any of the dietary patterns were significantly associated with the risk of disease after applying a Bonferroni correction for the number of patterns identified.

We also wanted to see how well the two procedures identified the specific combination of food intakes that were causally associated with the risk of disease. We compared the power and the FDR of ESFA and PCA for detecting these associations. In this context, we extended the concept of 'power' to mean the proportion of foods included in the model that were correctly identified as significantly associated with the disease outcome (Fig. 1). The FDR is the proportion of discoveries, or significant findings, that are false (Fig. 1).

For the ESFA, we considered that there was a 'significant' effect of a food if it was identified as such using the multiple testing procedure of Benjamini & Hochberg[34], with a nominal FDR set to 20 %. For the PCA, we considered that there was a 'significant' effect of a food if it had a correlation $>0·3$ or $<-0·3$ with a dietary pattern that was significantly associated with the risk of disease ($P<0·05$) – this being the way in which individual foods tend to be highlighted

	Foods not causally linked with disease	Foods causally linked with disease
Foods declared non-significant	TN	FN
Foods declared significant	FP	TP

Fig. 1. How the results of dietary analyses can be broken down. TN, true negatives; FN, false negatives; FP, false positives; TP, true positives. Power = TP/(FN + TP). False discovery rate (FDR) = FP/(FP + TP) if FP + TP > 0. FDR = 0 if FP + TP = 0.

in a PCA. It should be noted that the procedure of Benjamini & Hochberg[34] is designed to control the FDR at no more than the nominal level, but here false discoveries (of foods) occur not just as random errors, but also because of confounding with other foods, so the nominal rate may be exceeded.

Results

Table 2 displays the estimates of the power of PCA and ESFA for detecting an association between diet and disease, for different sample sizes and numbers of principal component scenarios. In model 3, the estimates of power for a sample size of 300 were all close to 100 %, so the estimates for a sample size of 100 are also given. Neither method consistently outperformed the other in this respect.

Table 3 shows the estimates of the power and FDR of PCA and ESFA for identifying the combinations of food intakes that were causally associated with the risk of disease. In most parts of the table, ESFA ‘dominates’ PCA in the sense of having both a higher power and a lower FDR. In the remainder of the table, neither one dominates the other.

Adjusting for other foods that were significant in an unadjusted analysis successfully controlled the FDR at about the 20 % nominal level, though with some loss of power, particularly with low sample sizes (Table 4). Attempting to control the FDR of ESFA by adjusting for principal components of diet, or by adjusting for a propensity score, was not successful (Table 4).

Discussion

In some scenarios, ESFA had greater power than PCA to detect an association of diet with the risk of disease. Allowing for multiple testing using the procedure of Benjamini & Hochberg[34], ESFA also typically had a higher power and a lower FDR for identifying the combinations of foods that were causally linked with the risk of disease than PCA in

Table 2. Power (%) of exhaustive single food analysis (ESFA) and principal components analysis (PCA) for detecting any association between diet and disease*

			PCA		
			No. of components		
Dataset and model†	Sample size	ESFA	2	5	10
Dataset 1: F.L.A.G. survey[25]					
Model 1	300	77·2	98·2‡	98·2‡	97·2‡
	1200	100·0	100·0	100·0	100·0
	4800	100·0	100·0	100·0	100·0
Model 2	300	94·2	98·1‡	99·9‡	98·4‡
	1200	100·0	100·0	100·0	100·0
	4800	100·0	100·0	100·0	100·0
Model 3	100	70·1	99·5‡	99·5‡	99·3‡
	300	100·0	100·0	100·0	100·0
	1200	100·0	100·0	100·0	100·0
	4800	100·0	100·0	100·0	100·0
Dataset 2: UK ECRHS II diet survey[26]					
Model 1	300	30·2	44·5‡	49·5‡	49·4‡
	1200	99·7	82·8	92·5	96·6
	4800	100·0	96·1	99·5	99·9
Model 2	300	81·2	61·6	71·4	76·3
	1200	100·0	90·0	97·6	99·5
	4800	100·0	97·8	99·8	99·9
Model 3	100	43·1‡	72·2‡	73·5‡	65·0‡
	300	99·9	99·9	99·9	100·0‡
	1200	100·0	100·0	100·0	100·0
	4800	100·0	100·0	100·0	100·0

F.L.A.G., Food, Lifestyle and Asthma in Greenwich; ECRHS, European Community Respiratory Health Survey.

* All estimates of power have a standard error < 0·5 %.

† In models 1 and 2, one in seven foods (thirty in dataset 1 and ten in dataset 2) are selected at random in each replication from the foods on the FFQ. In model 1, all selected food intakes are negatively associated with the risk of disease; in model 2, all selected food intakes are positively associated with the risk of disease. In model 3, foods comprising a ‘Western’ dietary pattern (thirty in dataset 1 and ten in dataset 2; see Table 1) are used in each replication, with all these food intakes being positively associated with the risk of disease.

‡ Power of PCA exceeds that of ESFA.

Table 3. Power and false discovery rate (FDR) (%) of exhaustive single food analysis (ESFA) and principal components analysis (PCA) for detecting the foods that are causally linked to the risk of disease*

				PCA					
				No. of components					
		ESFA		2		5		10	
Dataset and model†	Sample size	Power	FDR	Power	FDR	Power	FDR	Power	FDR
Dataset 1: F.L.A.G. survey[25]									
Model 1	300	49·1	70·3	35·5	85·7	46·1	85·3	55·2‡	85·5‡
	1200	89·7	80·3	36·2	86·1	49·5	86·1	58·6	85·8
	4800	97·3	83·6	38·2	86·5	51·5	86·1	60·9	86·0
Model 2	300	55·3	71·8	35·0	85·2	47·8	85·8	55·6	85·4
	1200	88·6	80·5	36·8	86·2	49·6	86·2	59·0	85·9
	4800	98·7	83·5	39·0	86·4	50·1	86·4	60·2	86·1
Model 3	300	88·6	76·7	55·3	82·4	74·8	80·7	79·8	81·7
	1200	99·7	82·8	78·2§	74·3§	86·9§	77·4§	86·4§	80·7§
	4800	100·0	84·9	87·3§	71·5§	93·1§	75·2§	87·7§	80·4§
Dataset 2: UK ECRHS II diet survey[26]									
Model 1	300	21·7	37·8	21·0	46·5	29·3‡	64·9‡	34·4‡	75·4‡
	1200	90·1	66·1	35·8	73·6	49·1	83·1	60·5	84·1
	4800	99·5	79·6	45·3	83·1	61·4	85·7	75·2	85·4
Model 2	300	53·3	48·0	27·2	58·6	39·5	75·3	47·9	80·3
	1200	92·9	73·2	39·5	78·3	54·4	84·5	68·0	84·6
	4800	98·4	81·7	46·8	84·3	63·9	85·9	78·2	85·9
Model 3	300	92·4	63·0	54·7	83·8	67·6	81·4	80·0	81·3
	1200	99·9	80·6	61·8	83·9	77·3	82·1	89·1	84·5
	4800	100·0	84·2	66·1	83·5§	87·9	81·8§	91·8	84·5

F.L.A.G., Food, Lifestyle and Asthma in Greenwich; ECRHS, European Community Respiratory Health Survey.

* All estimates of power and FDR have a standard error <0·5 %.

† In models 1 and 2, one in seven foods (thirty in dataset 1 and ten in dataset 2) are selected at random in each replication from the foods on the FFQ. In model 1, all selected food intakes are negatively associated with the risk of disease; in model 2, all selected food intakes are positively associated with the risk of disease. In model 3, foods comprising a 'Western' dietary pattern (thirty in dataset 1 and ten in dataset 2; see Table 1) are used in each replication, with all these food intakes being positively associated with the risk of disease.

‡ Power of PCA exceeds that of ESFA, but FDR is also higher.

§ FDR of PCA is lower than that of ESFA, but power is also lower.

which foods were singled out if they correlated highly ($>0{\cdot}3$ or $<-0{\cdot}3$) with a significant dietary pattern. Even when a simplified 'Western' dietary pattern was the real culprit, PCA could not outperform ESFA in reconstructing the foods that were linked with the risk of disease. These findings were replicated in two different FFQ datasets.

In principal components regression, there is no standard procedure for identifying foods as 'significantly associated' with the disease outcome. We used a definition that matches what researchers typically do when they single out a combination of foods to mention in their abstract[40] or when they devise an intervention based on the results[10]; that is, to list the foods that are highly correlated with a dietary pattern that is itself strongly associated with the disease outcome. We wanted to know whether this method was better at correctly identifying the combinations of foods that are associated with the risk of disease than an ESFA. 'Significant' foods are defined differently in each case (they are different methods), but the performance of each method is assessed using the same criterion: how often does it get it right? In fact, we use two criteria: FDR and power. The two methods can differ in both these respects, making comparison a little harder (think of two different diagnostic tests that might differ both in their sensitivity and specificity); nevertheless, if we observe one method to have both a lower FDR and a higher power, then we can conclude that it has superior performance.

It is common to try to control the FDR at a low level[34]. We have used a nominal FDR of 20 %; in genetic studies, where the use of FDR is well established, FDR between 5 and 20 % are recommended depending on the circumstances[41]. It should be noted that 20 % is still well below the FDR of individual hypothesis testing using $P<0{\cdot}05$ as a cut-off[42]. However, it is concerning that the observed FDR of both ESFA (nominally controlled at 20 %) and PCA increase in an uncontrolled fashion as the sample size and power increase (Table 3). It is worth noting that when one in seven foods are causally linked with the risk of disease, as here, an FDR of about 86 % would be achieved by selecting 'significant' foods entirely at random. The uncontrolled FDR occurs because all food intakes are correlated to some extent with the causal foods, leading to false positive findings (more so as power increases). We tried a variety of approaches to control for other foods in the ESFA, and found that the FDR could be successfully controlled at the nominal level by adjusting for foods that were significant in a univariate analysis. Achieving a given power requires about twice the sample size of an unadjusted ESFA (with its inflated FDR) and four times the sample size from our original calculation, i.e. for an unadjusted analysis with the criterion $P<0{\cdot}05$.

Our models included one in seven of the foods on the FFQ. We repeated our simulations with a smaller number of foods in the models, and obtained qualitatively similar findings

Table 4. Power and false discovery rate (FDR) (%) of exhaustive single food analysis with different methods of adjustment for other foods*

		Adjusted for five principal components		Adjusted for foods that are significant in the unadjusted analysis		Adjusted for propensity scores	
Dataset and model†	Sample size	Power	FDR	Power	FDR	Power	FDR
Dataset 1: F.L.A.G. survey[25]							
Model 1	300	1·7	27·5	4·8	29·4	0·9	8·5
	1200	59·6	53·7	50·9	35·3	66·3	47·1
	4800	95·1	72·6	91·4	24·3	98·6	69·7
Model 2	300	7·3	26·8	5·2	17·2	0·2	0·8
	1200	67·5	57·3	64·5	41·9	24·9	19·4
	4800	95·6	75·6	95·3	27·4	94·2	54·9
Model 3	300	4·3	52·8	2·0	18·5	0·5	21·3
	1200	32·7	75·6	57·5	38·8	19·7	77·7
	4800	72·5	80·6	94·4	22·9	76·4	76·8
Dataset 2: UK ECRHS II diet survey[26]							
Model 1	300	9·6	30·2	4·2	15·9	5·7	7·7
	1200	77·5	55·4	67·5	19·7	73·6	16·0
	4800	98·6	74·8	99·2	18·1	99·7	35·5
Model 2	300	32·3	35·1	23·0	16·8	3·0	5·5
	1200	88·0	66·0	90·7	21·4	66·8	22·5
	4800	98·2	79·5	99·9	21·4	98·6	58·0
Model 3	300	17·8	48·8	12·2	19·1	1·7	5·1
	1200	65·3	72·0	79·3	20·2	52·8	32·6
	4800	88·7	80·4	99·9	16·9	85·2	70·5

F.L.A.G., Food, Lifestyle and Asthma in Greenwich; ECRHS, European Community Respiratory Health Survey.

* All estimates of power and FDR have a standard error <0·5 %.

† In models 1–3, one in seven foods (thirty in dataset 1 and ten in dataset 2) are selected at random in each replication from the foods on the FFQ. In model 1, all selected food intakes are negatively associated with the risk of disease; in model 2, all selected food intakes are positively associated with the risk of disease. In model 3, foods comprising a ‘Western’ dietary pattern (thirty in dataset 1 and ten in dataset 2; see Table 1) are used in each replication, with all these food intakes being positively associated with the risk of disease.

(see online supplementary Tables S1–S3). We did not consider interactive effects of foods in our models; this requires further investigation. Studies have claimed that dietary constituents interact with each other in complex ways[43,44], and these interactions have an effect, for example, on the risk of CVD[45] and breast cancer[46]. PCA is often recommended as a way of dealing with interactions between foods[2,20]. Furthermore, foods could be associated with the risk of disease in a non-linear way. However, it is questionable whether linear combinations of food intakes produced by the PCA adequately address the issues of modelling interactions or capture potential non-linear effects between diet and disease.

Although we considered a model based on a ‘Western’ dietary pattern, there is no reason why foods with truly causal effects should be foods that are highly correlated with each other in order to be associated with the risk of disease. Hence, in the present simulation study, foods associated with the risk of disease were not necessarily highly correlated. Note it is not obvious that ESFA will outperform PCA in this case, since ESFA must pay a much higher penalty for multiple testing of all the individual foods, while each principal component may include – even if only by chance – a number of foods that have causal effects in the same direction. Where disease risk is explained by other factors that are confounded with diet, however, this confounding is more likely to be at the level of a dietary pattern than with an individual food. A ‘prudent’ dietary pattern, for example, is associated with older age[47], female sex[48], non-smoking[49], higher income[50], higher educational level[51], exercise[52] and supplement use[53]. These factors are likely to be associated with a number of food intakes contributing to a ‘prudent’ dietary pattern rather than with any one of these foods in particular. This is another reason for adjusting each food effect for others found to be significant, as we suggest this should help control for both measured and some unmeasured confounding. We did not include non-dietary confounders in our simulations because there are just too many different potential confounders and models for their effects to be considered. We suspect, however, that as long as there are foods with truly causal effects, the findings of the present study will generalise to situations where other confounders have been explicitly adjusted for.

PCA does not aim to obtain a clear picture of disease variation but to summarise the overall dietary intake variation in the population. As Jolliffe[54] points out, the mistake when PCA is employed for distinguishing between healthy and unhealthy individuals is to assume that the separation between these groups is along the axis of greatest variation in diet. More appropriate statistical approaches in this respect are linear discriminant analysis[54] and discriminant analysis of principal components[55].

Other multivariate approaches such as cluster analysis[56], and data reduction techniques such as reduced-rank regression[57] provide an alternative way for identifying dietary patterns. The main advantage of cluster analysis over PCA is that cluster analysis creates mutually exclusive groups that can easily be used in the analysis. Reduced-rank regression constructs dietary patterns according to the covariance matrix of specific biomarkers (taken by blood samples, for example) that are associated with dietary intake variables and assumed to be linked with the investigated disease outcome. Thus, the dietary pattern that is identified to be associated with the risk of disease for a specific biological reason and the combinations of foods that characterise the pattern describe a specific biomarker in the causal pathway between diet and disease.

Another approach is factor analysis, an ambiguous term which in certain textbooks include both PCA and common factor analysis. Even when investigators say that they are employing factor analysis in nutritional studies, they may be employing PCA. This misconception depends on the statistical package that these studies have used, because certain statistical packages (e.g. SAS) treat PCA as a special category of factor analysis. Factor analysis, which is a different statistical technique from PCA[58], is not recommended for the analysis of nutritional data[2], although factor analysis based on the principal factor method gives generally similar results to PCA[59].

Since ESFA outperforms PCA in the present simulation study, dealing with high-dimensional multivariate dietary exposures could be treated as a problem of variable model selection, that is, finding the non-zero regression coefficients in an unknown regression model. Our adjusted ESFA is similar to the iterative sure independence screening method for ultra-high-dimensional data[35]. Other forms of penalised likelihood estimation methods have been developed in the last decade to cope with high-dimensional data and have been lately reviewed[60]. These methods could be potentially useful in nutritional epidemiological studies, and further research is needed.

There is a growing interest in designing dietary interventions around foods rather than nutrients[61] and around particular foods rather than dietary patterns[61,62]. Specifically, Jacobs *et al.*[61] suggests that the evidence for beneficial effects of a ‘prudent’ diet comes from interventions that only modified the intake of one or two foods. Furthermore, McCann *et al.*[63] suggests that fruits and vegetables alone provided the highest discrimination among endometrial cancer cases and controls compared with PCA and other methods of characterisation. Mann & Aune[62], evaluating the evidence that fruits and vegetables can prevent diabetes, have called for more studies looking at the effects of specific fruits and vegetables.

In conclusion, an FFQ-wide study of associations between food intakes and disease risk outperforms an analysis of dietary patterns derived from PCA. Analysing each food adjusting for others allows truly causal effects to be identified with a low rate of false discoveries and surprisingly good power. Although PCA has proved extremely popular in nutritional epidemiology to date, we question its routine use in this context.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0007114514000221

Acknowledgements

The present study is part of a programme of research funded by the Department of Health Policy Research Programme (DHTBX P16621).

All authors were responsible for the reported research and made substantial contributions to the conception and design of the simulation study, analysis and interpretation of the data, drafting of the manuscript or revising it critically for important intellectual content. I. B. prepared the manuscript, with R. H. and P. B. contributing to it. I. B. conducted the computer programming and simulation and data analysis. All authors read and approved the final version of the manuscript.

None of the authors has any personal, commercial, political, academic or financial conflicts of interest to declare.

References

1. Slattery ML, Boucher KM, Caan BJ, *et al.* (1998) Eating patterns and risk of colon cancer. *Am J Epidemiol* **148**, 4–16.
2. Hu FB (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* **13**, 3–9.
3. Kant AK (2004) Dietary patterns and health outcomes. *J Am Diet Assoc* **104**, 615–635.
4. Newby PK & Tucker KL (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev* **62**, 177–203.
5. Naska A, Fouskakis D, Oikonomou E, *et al.* (2006) Dietary patterns and their socio-demographic determinants in 10 European countries: data from the DAFNE databank. *Eur J Clin Nutr* **60**, 181–190.
6. Knudsen VK, Orozova-Bekkevold IM, Mikkelsen TB, *et al.* (2008) Major dietary patterns in pregnancy and fetal growth. *Eur J Clin Nutr* **62**, 463–470.
7. Northstone K & Emmett P (2005) Multivariate analysis of diet in children at four and seven years of age and associations with socio-demographic characteristics. *Eur J Clin Nutr* **59**, 751–760.
8. Iqbal R, Anand S, Ounpuu S, *et al.* (2008) Dietary patterns and the risk of acute myocardial infarction in 52 countries: results of the INTERHEART study. *Circulation* **118**, 1929–1937.
9. Cottet V, Touvier M, Fournier A, *et al.* (2009) Postmenopausal breast cancer risk and dietary patterns in the E3N-EPIC prospective cohort study. *Am J Epidemiol* **170**, 1257–1267.
10. Bertuccio P, Rosato V, Andreano A, *et al.* (2013) Dietary patterns and gastric cancer risk: a systematic review and meta-analysis. *Ann Oncol* **24**, 1450–1458.
11. Shaheen SO, Newson RB, Rayman MP, *et al.* (2007) Randomised, double blind, placebo-controlled trial of selenium supplementation in adult asthma. *Thorax* **62**, 483–490.
12. Greenberg ER, Baron JA, Tosteson TD, *et al.* (1994) A clinical trial of antioxidant vitamins to prevent colorectal adenoma. Polyp Prevention Study Group. *N Engl J Med* **331**, 141–147.

13. Schatzkin A, Lanza E, Corle D, *et al.* (2000) Lack of effect of a low-fat, high-fiber diet on the recurrence of colorectal adenomas. Polyp Prevention Trial Study Group. *N Engl J Med* **342**, 1149–1155.
14. Hennekens CH, Buring JE, Manson JE, *et al.* (1996) Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *N Engl J Med* **334**, 1145–1149.
15. Hu FB, Rimm EB, Stampfer MJ, *et al.* (2000) Prospective study of major dietary patterns and risk of coronary heart disease in men. *Am J Clin Nutr* **72**, 912–921.
16. Fung TT, Willett WC, Stampfer MJ, *et al.* (2001) Dietary patterns and the risk of coronary heart disease in women. *Arch Intern Med* **161**, 1857–1862.
17. Wu K, Hu FB, Fuchs C, *et al.* (2004) Dietary patterns and risk of colon cancer and adenoma in a cohort of men (United States). *Cancer Causes Control* **15**, 853–862.
18. Fung T, Hu FB, Fuchs C, *et al.* (2003) Major dietary patterns and the risk of colorectal cancer in women. *Arch Intern Med* **163**, 309–314.
19. Varraso R, Fung TT, Barr RG, *et al.* (2007) Prospective study of dietary patterns and chronic obstructive pulmonary disease among US women. *Am J Clin Nutr* **86**, 488–495.
20. Varraso R, Fung TT, Hu FB, *et al.* (2007) Prospective study of dietary patterns and chronic obstructive pulmonary disease among US men. *Thorax* **62**, 786–791.
21. Randall E, Marshall JR, Graham S, *et al.* (1990) Patterns in food use and their associations with nutrient intakes. *Am J Clin Nutr* **52**, 739–745.
22. Jacques PF & Tucker KL (2001) Are dietary patterns useful for understanding the role of diet in chronic disease? *Am J Clin Nutr* **73**, 1–2.
23. Newby PK, Weismayer C, Akesson A, *et al.* (2006) Longitudinal changes in food patterns predict changes in weight and body mass index and the effects are greatest in obese women. *J Nutr* **136**, 2580–2587.
24. Slattery ML (2008) Defining dietary consumption: is the sum greater than its parts? *Am J Clin Nutr* **88**, 14–15.
25. Shaheen SO, Sterne JA, Thompson RL, *et al.* (2001) Dietary antioxidants and asthma in adults: population-based case–control study. *Am J Respir Crit Care Med* **164**, 1823–1828.
26. Hooper R, Heinrich J, Omenaas E, *et al.* (2010) Dietary patterns and risk of asthma: results from three countries in European Community Respiratory Health Survey-H. *Br J Nutr* **103**, 1354–1365.
27. Paul AA, Southgate DA & Buss DH (1986) McCance and Widdowson's 'The composition of foods': supplementary information and review of new compositional data. *Hum Nutr Appl Nutr* **40**, 287–299.
28. Fewell Z, Davey Smith G & Sterne JA (2007) The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol* **166**, 646–655.
29. Peduzzi P, Concato J, Kemper E, *et al.* (1996) A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* **49**, 1373–1379.
30. Laird NM & SpringerLink (2011) *The Fundamentals of Modern Statistical Genetics*. New York, NY: Springer Science + Business Media, LLC. http://dx.doi.org/10.1007/978-1-4419-7338-2
31. Willett W (1998) *Nutritional Epidemiology*, 2nd ed. Oxford: Oxford University Press.
32. Jakes RW, Day NE, Luben R, *et al.* (2004) Adjusting for energy intake – what measure to use in nutritional epidemiological studies? *Int J Epidemiol* **33**, 1382–1386.
33. Imamura F, Lichtenstein AH, Dallal GE, *et al.* (2009) Confounding by dietary patterns of the inverse association between alcohol consumption and type 2 diabetes risk. *Am J Epidemiol* **170**, 37–45.
34. Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc: Ser B (Methodol)* **57**, 289–300.
35. Fan J & Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc: Ser B (Methodol)* **70**, 849–911.
36. Imai Kosuke & van Dyk David A (2004) Causal inference with general treatment regime: generalizing the propensity score. **99**, 854–866.
37. Sturmer T, Joshi M, Glynn RJ, *et al.* (2006) A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* **59**, 437–447.
38. Ender PB (2002) powerlog: command to perform logistic regression power analysis. http://www.ats.ucla.edu/stat/stata/ado/analysis/
39. Miller RG (1981) *Simultaneous Statistical Inference*, 2nd ed. Springer: Verlag.
40. Shaheen SO, Jameson KA, Syddall HE, *et al.* (2010) The relationship of dietary patterns with adult lung function and COPD. *Eur Respir J* **36**, 277–284.
41. Benjamini Y & Yekutieli D (2005) Quantitative trait loci analysis using the false discovery rate. *Genetics* **171**, 783–790.
42. Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* **2**, e124.
43. Messina M, Lampe JW, Birt DF, *et al.* (2001) Reductionism and the narrowing nutrition perspective: time for reevaluation and emphasis on food synergy. *J Am Diet Assoc* **101**, 1416–1419.
44. Sacks FM, Obarzanek E, Windhauser MM, *et al.* (1995) Rationale and design of the Dietary Approaches to Stop Hypertension trial (DASH). A multicenter controlled-feeding study of dietary patterns to lower blood pressure. *Ann Epidemiol* **5**, 108–118.
45. Halliwell B (1996) Antioxidants in human health and disease. *Annu Rev Nutr* **16**, 33–50.
46. Edefonti V, Randi G, La Vecchia C, *et al.* (2009) Dietary patterns and breast cancer: a review with focus on methodological issues. *Nutr Rev* **67**, 297–314.
47. Agurs-Collins T, Rosenberg L, Makambi K, *et al.* (2009) Dietary patterns and breast cancer risk in women participating in the Black Women's Health Study. *Am J Clin Nutr* **90**, 621–628.
48. Robinson S, Syddall H, Jameson K, *et al.* (2009) Current patterns of diet in community-dwelling older men and women: results from the Hertfordshire Cohort Study. *Age Ageing* **38**, 594–599.
49. Fung TT, Rimm EB, Spiegelman D, *et al.* (2001) Association between dietary patterns and plasma biomarkers of obesity and cardiovascular disease risk. *Am J Clin Nutr* **73**, 61–67.
50. Perrin AE, Dallongeville J, Ducimetiere P, *et al.* (2005) Interactions between traditional regional determinants and socio-economic status on dietary patterns in a sample of French men. *Br J Nutr* **93**, 109–114.
51. Raberg Kjollesdal MK, Holmboe-Ottesen G & Wandel M (2010) Associations between food patterns, socioeconomic position and working situation among adult, working women and men in Oslo. *Eur J Clin Nutr* **64**, 1150–1157.
52. Lopez-Garcia E, Schulze MB, Fung TT, *et al.* (2004) Major dietary patterns are related to plasma concentrations of markers of inflammation and endothelial dysfunction. *Am J Clin Nutr* **80**, 1029–1035.

53. Heidemann C, Schulze MB, Franco OH, *et al.* (2008) Dietary patterns and risk of mortality from cardiovascular disease, cancer, and all causes in a prospective cohort of women. *Circulation* **118**, 230–237.
54. Jolliffe IT (2010) *Principal Component Analysis*, 2nd ed. New York/London: Springer.
55. Jombart T, Devillard S & Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.
56. Reedy J, Wirfalt E, Flood A, *et al.* (2010) Comparing 3 dietary pattern methods – cluster analysis, factor analysis, and index analysis – with colorectal cancer risk: The NIH-AARP Diet and Health Study. *Am J Epidemiol* **171**, 479–487.
57. Hoffmann K, Schulze MB, Schienkiewitz A, *et al.* (2004) Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol* **159**, 935–944.
58. Bartholomew DJ, Steele F, Moustaki I, *et al.* (2002) *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Boca Raton, FL: Chapman & Hall/CRC Press.
59. Schulze MB & Hoffmann K (2006) Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. *Br J Nutr* **95**, 860–869.
60. Fan J & Lv J (2010) A selective overview of variable selection in high dimensional feature space. *Stat Sin* **20**, 101–148.
61. Jacobs DR Jr, Gross MD, Tapseli LC, *et al.* (2009) Food synergy: an operational concept for understanding nutrition. *Am J Clin Nutr* **89**, 1543S–1548S.
62. Mann J & Aune D (2010) Can specific fruits and vegetables prevent diabetes? *BMJ* **341**, c4395.
63. McCann SE, Weiner J, Graham S, *et al.* (2001) Is principal components analysis necessary to characterise dietary behaviour in studies of diet and disease? *Public Health Nutr* **4**, 903–908.

British Journal of Nutrition

https://doi.org/10.1017/S0007114514000221 Published online by Cambridge University Press