Public Health Nutrition: 19(2), 195–203 doi:10.1017/S1368980014003103

Using both principal component analysis and reduced rank regression to study dietary patterns and diabetes in Chinese adults

Carolina Batis[1], Michelle A Mendez[1], Penny Gordon-Larsen[1], Daniela Sotres-Alvarez[2], Linda Adair[1] and Barry Popkin[1,*]

[1]Department of Nutrition and Carolina Population Center, University of North Carolina at Chapel Hill, 137 East Franklin Street, Chapel Hill, NC 27516, USA: [2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Submitted 2 October 2013: Final revision received 14 November 2014: Accepted 28 November 2014: First published online 14 January 2015

Abstract

Objective: We examined the association between dietary patterns and diabetes using the strengths of two methods: principal component analysis (PCA) to identify the eating patterns of the population and reduced rank regression (RRR) to derive a pattern that explains the variation in glycated Hb (HbA1c), homeostasis model assessment of insulin resistance (HOMA-IR) and fasting glucose.

Design: We measured diet over a 3 d period with 24 h recalls and a household food inventory in 2006 and used it to derive PCA and RRR dietary patterns. The outcomes were measured in 2009.

Setting: Adults (n 4316) from the China Health and Nutrition Survey.

Results: The adjusted odds ratio for diabetes prevalence (HbA1c ≥ 6·5 %), comparing the highest dietary pattern score quartile with the lowest, was 1·26 (95 % CI 0·76, 2·08) for a modern high-wheat pattern (PCA; wheat products, fruits, eggs, milk, instant noodles and frozen dumplings), 0·76 (95 % CI 0·49, 1·17) for a traditional southern pattern (PCA; rice, meat, poultry and fish) and 2·37 (95 % CI 1·56, 3·60) for the pattern derived with RRR. By comparing the dietary pattern structures of RRR and PCA, we found that the RRR pattern was also behaviourally meaningful. It combined the deleterious effects of the modern high-wheat pattern (high intakes of wheat buns and breads, deep-fried wheat and soya milk) with the deleterious effects of consuming the opposite of the traditional southern pattern (low intakes of rice, poultry and game, fish and seafood).

Conclusions: Our findings suggest that using both PCA and RRR provided useful insights when studying the association of dietary patterns with diabetes.

Keywords
Dietary patterns
Principal component analysis
Reduced rank regression
Diabetes

The prevalence of type 2 diabetes among adults in China has increased dramatically from 3·0 %[1] in 1994 to 11·6 % in 2010[2]. It is known that diet plays a key role in the prevention of diabetes and the research devoted to better understanding the role of dietary intake ranges from investigating particular nutrients and food groups to examining dietary patterns[3–5].

The study of dietary patterns considers a combination of foods and therefore more closely approximates the population's dietary exposures. Two ways to select the foods for a given dietary pattern are factor or principal component analysis (PCA), which derives linear functions of foods that best explain the variations in intake, and reduced rank regression (RRR), which derives the linear functions of foods that best explain the variations in outcome variables (e.g. disease-related nutrients or biomarkers)[6]. The strength of one method is the limitation of the other. PCA patterns have public health relevance, because they describe the actual dietary patterns of the population, whereas the foods in RRR patterns are not necessarily consumed together and could be behaviourally irrelevant. Conversely, RRR patterns are by definition associated with the outcome or response variables, which might not be the case for PCA patterns. Therefore the two methods can complement each other and provide useful insights when compared. RRR patterns can put PCA findings into perspective by indicating the largest possible strength of association a data-driven dietary pattern can have. PCA patterns can put into perspective how behaviourally meaningful RRR dietary

*Corresponding author: Email popkin@unc.edu

patterns are by indicating which patterns the population follows. In addition, RRR can generate hypotheses about which food components of a dietary pattern are related to diabetes and PCA can indicate whether these foods define the eating patterns of the population.

Few studies have looked at the association between dietary patterns and diabetes or insulin resistance among Chinese adults[7–9] and none have used RRR. Our aim was to use both PCA and RRR to complement each other in comparing the patterns and their strength of association with diabetes and insulin resistance. For RRR we selected response variables that directly represent our outcomes of interest: glycated Hb (HbA1c), fasting glucose and homeostasis model assessment of insulin resistance (HOMA-IR). Typically, to incorporate information about biological pathways, RRR is used on intermediate variables[10]. However, we did not focus on the biological pathways but on the dietary pattern that was most closely related to our outcome of interest. Therefore our RRR dietary pattern should be considered an initial hypothesis and not a pattern with an established association.

Methods

Study design and participants

The China Health and Nutrition Survey (CHNS) is an ongoing longitudinal study of eight waves (1989–2009). The sample was drawn with a multistage, random cluster process in nine provinces. The survey was conducted according to the guidelines in the Declaration of Helsinki and the protocols, instruments and the process for obtaining informed consent were approved by the institutional review committees of the University of North Carolina at Chapel Hill (UNC-CH) and the Chinese Institute of Nutrition and Food Safety (INFS), China Center for Disease Control and Prevention. Details are provided elsewhere[11].

The CHNS collected blood samples for the first time in 2009; therefore the present analysis uses the diabetes-related biomarkers measured in 2009 and the exposures, dietary intakes and covariates measured in 2006. Eligible individuals were 18 to 65 years old in 2006, not previously diagnosed with diabetes (because treatment might have affected dietary intake and/or diabetes-related biomarkers) and not pregnant in 2006 or 2009. Of the 7858 eligible individuals in 2006, we excluded those not followed up in 2009 or with missing biomarkers (*n* 2667) or dietary data (*n* 225) and those not fasting before blood was drawn (*n* 212) or with missing covariates (*n* 438). Our final sample numbered 4316. We conducted several sensitivity analyses to address the potential of selection bias (see 'Statistical analysis' section).

Measurement of variables

Diet was assessed with a combination of 24 h recalls on three consecutive days and a food inventory at the household level collected over the same 3 d period. The start day was randomly allocated from Monday to Sunday. For the food inventory, all foods in the household (purchased or produced at home) were measured on a daily basis with digital scales. We estimated the total household food consumption by measuring the changes in the household food inventory and the waste. For the 24 h recall, trained interviewers recorded the types and amounts of all food items consumed, the type of meal and the place of consumption. For foods consumed at home, we estimated the amount of each dish from the household food inventory an individual consumed based on the proportion of the dish he or she reported consuming.

The food groups included in our analysis were based on a food group classification system developed specifically for the CHNS by researchers from the UNC-CH and INFS[12] which classifies foods according to their nutritional and behavioural significance.

We collected blood samples by venepuncture after an overnight fast. We measured glucose in the serum with a glucose oxidase phenol 4-aminoantipyrine peroxidase kit (Randox, Crumlin, UK) in a Hitachi 7600 analyser. We measured HbA1c in whole blood by HPLC with an automated glycohaemoglobin analyser (model HLC-723 G7; Tosoh, Tokyo, Japan). We measured insulin in serum by radioimmunology in a gamma counter XH-6020 analyser (North Institute of Bio-Tech, China). We estimated HOMA-IR (= [fasting insulin (μU/ml) × fasting glucose (mmol/l)]/22·5)[13].

We defined diabetes based on HbA1c ≥ 6·5 %[14]. Despite the controversies surrounding use of HbA1c as a diagnostic tool, its advantage over a single measure of glucose is that it captures long-term glycaemic exposure[15]. In addition, HbA1c correlates well with the risk of long-term diabetes complications and has been shown to be a reliable method for diabetes diagnosis in the Chinese population[16–18].

The demographic and lifestyle covariates we included in the analysis were gender, age, geographic region (north, central or south), urbanicity scale, education level, income, smoking status, physical activity, alcohol intake and BMI. We assessed physical activity with detailed self-reports of time spent and intensity levels for occupational and domestic activities; for both type of activities we estimated metabolic equivalents and added them up into a single variable[19]. We determined the level of urbanization with an urbanicity scale developed for the CHNS that includes such components as population density, economic activity, transportation infrastructure, sanitation and housing types[20]. We estimated BMI from measured weight and height.

Statistical analysis

We analysed dietary patterns with twenty-nine food groups. Most food groups had a high proportion of non-consumers possibly due to the fact that dietary intake was measured over a 3 d period (eighteen food groups

had <30 % of consumers, eight had 30–80 % of consumers and three had >80 % of consumers). We used dichotomous variables because the original continuous variables resembled more a categorical variable due to the high number of zeros. Therefore for the twenty-six food groups with <80 % of consumers, we categorized food group intake as non-consumers or consumers. Otherwise (for rice, fresh non-leafy vegetables and fresh leafy vegetables) we categorized intake as below or above the median. Analyses performed on continuous food groups yielded similar patterns, suggesting that in our data the reported amount consumed did not provide additional information to consumption *v.* non-consumption in the dietary patterns structure. Food groups consumed by fewer than 5 % of our participants were not included in the dietary pattern analysis. For a full list of food groups included and not included and their descriptions, see Supplemental Table 1 (online supplementary material).

We performed the PCA with the procedure PROC PLS with a PCR method option (SAS statistical software package version 9·3). Ten components had an eigenvalue greater than 1, but there was a clear break in the scree plot after the second component (see Supplemental Figure 1, online supplementary material). Based on this, interpretability and previous work[21], we decided to retain two components. RRR was performed with PROC PLS and an RRR option with HbA1c, HOMA-IR and fasting glucose as response variables. Due to non-normality, we performed natural log transformation of all the response variables. Because RRR can potentially derive patterns that are confounded by non-dietary factors[22], we also used adjusted food groups with the residual method, a strategy previously used[23,24]. To estimate the residuals, we ran several logistic regressions with each binary food group as the dependent variable and geographic region, urbanicity index, education and income as the independent variables. Our variable selection for the estimation of residuals is based on previous work[21], where we found in this population that these variables were the most influential on dietary patterns. We used these residuals (difference between the observed and the predicted probability) as intake variables on the RRR procedure. We did not include energy intake in the residual method, because excessive energy intake could be in the causal pathway between the dietary pattern and diabetes. Also because the RRR dietary pattern derived from the residuals was less confounded by non-dietary factors, we do not present the RRR performed with the original dichotomized intake variables in our main results; that is presented in Supplemental Tables 2 and 3 (online supplementary material). We did not use residuals in the PCA because our aim was to describe how the population eats, so there was no interest in adjusting by the factors that may influence these dietary patterns.

For the RRR we retained only the first factor, because this one explained most of the variation in the response variables and was the only one with significant associations. Even though we retained two factors for the PCA, the results are still comparable, because the number of factors retained does not affect the structure of the derived patterns or the explained variation of each. For each participant we calculated a score for each dietary pattern (for each PCA and RRR) as a weighted sum of the food groups based on the factor loadings. The higher the score, the more closely the participant's diet conforms to the dietary pattern.

Although we knew the dietary pattern scores from the RRR would have a stronger association with the outcomes, we ran multiple linear (for HbA1c, HOMA-IR and fasting glucose) and logistic (for diabetes) regressions for each dietary pattern to identify how different the strength of association was between RRR and PCA factors. Because HbA1c, HOMA-IR and fasting glucose were natural-log-transformed, the exponentiated regression coefficients are the ratio of the expected geometric means of the original outcome variable[25]. Therefore we subtracted 1 from the exponentiated coefficients and multiplied by 100 so that these could be interpreted as the percentage change in the outcome due to a 1-unit increase in the independent variable. We present results both by quartiles of the dietary pattern score and by the continuous increment in the score (1 SD unit increase). First we adjusted by gender, smoking (yes/no), alcohol consumption (more than 3 times/week *v.* less than 3 times/week), education (none, primary school, more than lower middle school), region (south, central, north), age, income, urbanicity index and physical activity (continuous). We did not adjust by BMI initially because we hypothesized it was in the causal pathway. But to estimate the dietary pattern association independent of BMI, we additionally adjusted by BMI (continuous) in a second model. We accounted for the clustering at the household level in the estimation of the variance by using a cluster-robust variance estimator (Stata command: vce (cluster *clustvar*)). This estimator relaxes the assumption of independence of the observations and produces the correct standard errors[26]. Results for fasting glucose are presented only in Supplemental Tables 3 and 4 (online supplementary material).

As a sensitivity analysis, we derived the PCA patterns again in all included and excluded participants with diet available (*n* 7633) and found that the patterns were very similar to those of the main analysis. (Differences in loadings were less than an absolute level of 0·05. These factor loadings are not shown.) We also estimated the scores for the dietary patterns among the excluded participants and found that their distribution was very similar to those included in the analysis. Selection bias occurs when, by analysing only those included in the sample, we condition on common effects of the exposure and the outcome[27]. Because being included in the sample was not associated with exposure to dietary patterns, selection bias was less likely. The comparison of dietary patterns and covariates distribution between included and excluded

participants is presented in Supplemental Table 5 (online supplementary material).

Results

Compared with non-diabetics, those classified as diabetic were older, had a higher BMI and higher energy intake, and had lower education, income and physical activity levels. In addition, a higher proportion of those classified as diabetic lived in the central region and in more urbanized areas, and a higher proportion of diabetic males consumed alcohol regularly (Table 1).

The first factor from the PCA was inversely associated with the intake of rice and positively associated with the intake of wheat buns and breads; cakes, cookies and pastries; deep-fried wheat; fruits; eggs and egg products; soya milk; cow's milk; and instant noodles and frozen dumplings (Table 2). We previously found a similar dietary pattern in this population that we call 'modern high-wheat'[21]. The second factor from the PCA, which we call 'traditional southern', was positively related to intake of rice, high-fat pork, organ meats, poultry and game, and fish and seafood; and inversely associated with intake of wheat flour, wheat buns and breads, and corn and coarse grains. The factor loadings from the RRR seemed to be close to the modern high-wheat dietary pattern (PCA 1) and the opposite of the traditional southern (PCA 2) at the same time. As in the modern high-wheat, the RRR pattern was also inversely associated with intake of rice and positively associated with intake of wheat buns and breads, deep-fried wheat and soya milk. And in the opposite direction from the traditional southern

Table 1 Baseline characteristics of participants by diabetes status: adults (*n* 4316) from the China Health and Nutrition Survey (diabetes-related biomarkers measured in 2009; exposures, dietary intakes and covariates measured in 2006)

	Diabetes (HbA1c ≥ 6·5 %)	
	No (*n* 4071)	Yes (*n* 245)
Age (years)		
Mean	46·5	51·7
SD	10·5	8·6
Region (%)		
South	45·9	22·0
Central	32·0	57·6
North	22·1	20·4
Male (%)	45·3	49·8
BMI (kg/m^2)		
Mean	23·2	25·7
SD	3·1	3·8
Energy intake (kJ/d)		
Mean	7021	7245
SD	2180	2270
Energy intake (kcal/d)		
Mean	1678·0	1731·5
SD	521·1	542·6
Highest level of education attained (%)		
None	20·5	25·3
Primary school	20·5	25·3
More than lower middle school	59·0	49·4
Income* (%)		
Low	33·3	37·6
Medium	33·6	30·2
High	33·1	32·2
Urbanicity* (%)		
Low	33·8	26·1
Medium	33·7	34·7
High	32·6	39·2
Currently smoking (%)		
Female	3·0	3·3
Male	58·0	58·2
Alcohol intake ≥3 times/week (%)		
Female	2·0	1·6
Male	30·9	37·7
Physical activity* (%)		
Low	32·7	40·8
Medium	33·4	32·7
High	34·0	26·5

HbA1c, glycated Hb.
*Cut-off points for low, medium and high categories are based on tertiles of the entire sample.

Table 2 Factor loadings* and explained variation of dietary patterns from PCA and RRR among adults (*n* 4316) from the China Health and Nutrition Survey (diabetes-related biomarkers measured in 2009; exposures, dietary intakes and covariates measured in 2006)

	PCA		
	Modern high-wheat	Traditional southern	RRR†
Food groups			
Rice	−0·25	0·34	−0·22
Wheat noodles	–	–	0·30
Wheat flour	–	−0·36	–
Wheat buns and breads	0·33	−0·26	0·46
Cakes, cookies and pastries	0·28	–	–
Deep-fried wheat	0·38	–	0·22
Corn and coarse grains	–	−0·30	–
Fresh legumes	–	–	−0·24
Fruits	0·32	–	–
High-fat pork	–	0·26	–
Organ meats	–	0·25	–
Poultry and game	–	0·31	−0·37
Eggs and egg products	0·25	–	−0·23
Fish and seafood	–	0·37	−0·29
Soya milk	0·34	–	0·24
Cow's milk	0·26	–	–
Instant noodles and frozen dumplings	0·23	–	–
Explained variation in food groups (%)	8·47	7·64	4·42
Explained variation in responses (%)			
HbA1c	0·96	2·95	1·40
HOMA-IR	0·08	0·09	0·41
Fasting glucose	0·26	0·18	0·62

PCA, principal component analysis; RRR, reduced rank regression; HbA1c, glycated Hb; HOMA-IR, homeostasis model assessment of insulin resistance.
*Factor loadings < |0·20| are not shown. The following food groups had factor loadings < |0·20| in all patterns and are not shown in the table: starchy roots and tubers; starchy root and tuber products; dried legumes; legume products; nuts and seeds; fresh vegetables; non-leafy fresh vegetables; leafy, pickled, salted or canned vegetables; dried vegetables; high-fat red meat; low-fat pork; processed meats.
†Performed on residuals estimated for each food group with a multiple regression including geographic region, urbanicity, income and education.

Table 3 Percentage change in HbA1c and HOMA-IR related to quartiles of dietary pattern score and linear dietary pattern score increase (1 SD) among adults (*n* 4316) from the China Health and Nutrition Survey (diabetes-related biomarkers measured in 2009; exposures, dietary intakes and covariates measured in 2006)

		Quartile 2		Quartile 3		Quartile 4		Dietary pattern score (1 SD increase)	
	Quartile 1	% change	95 % CI	% change	95 % CI	% change	95 % CI	% change	95 % CI
HbA1c*									
PCA, modern high-wheat									
Unadjusted model	0	2·02	0·95, 3·10	3·38	2·28, 4·48	4·39	3·31, 5·47	1·33	0·97, 1·70
Adjusted model 1	0	0·77	−0·25, 1·79	1·51	0·39, 2·65	1·70	0·51, 2·90	0·32	−0·09, 0·74
Adjusted model 2	0	0·53	−0·45, 1·53	1·28	0·19, 2·39	1·44	0·29, 2·60	0·30	−0·10, 0·71
PCA, traditional southern									
Unadjusted model	0	−3·80	−4·86, −2·74	−5·79	−6·75, −4·81	−5·73	−6·75, −4·70	−2·17	−2·57, −1·78
Adjusted model 1	0	−1·55	−2·64, −0·45	−2·30	−3·40, −1·18	−2·18	−3·34, −1·01	−0·79	−1·24, −0·33
Adjusted model 2	0	−1·34	−2·41, −0·26	−1·94	−3·02, −0·84	−1·86	−3·01, −0·70	−0·72	−1·17, −0·26
RRR†									
Unadjusted model	0	1·76	0·72, 2·81	3·35	2·32, 4·38	4·13	3·08, 5·18	1·46	1·10, 1·83
Adjusted model 1	0	1·97	1·02, 2·92	3·28	2·34, 4·22	3·82	2·88, 4·77	1·46	1·13, 1·79
Adjusted model 2	0	1·77	0·84, 2·70	3·01	2·11, 3·93	3·49	2·59, 4·41	1·33	1·02, 1·65
HOMA-IR*									
PCA, modern high-wheat									
Unadjusted model	0	7·42	0·66, 14·63	16·76	9·35, 24·68	20·56	12·86, 28·78	6·24	3·86, 8·68
Adjusted model 1	0	1·81	−4·45, 8·47	5·97	−0·75, 13·14	1·13	−6·08, 8·90	−1·05	−3·68, 1·65
Adjusted model 2	0	−0·36	−6·32, 5·97	3·73	−2·61, 10·48	−1·27	−8·06, 6·01	−1·24	−3·78, 1·38
PCA, traditional southern									
Unadjusted model	0	−5·03	−10·76, 1·07	−10·88	−16·42, −4·98	−5·21	−10·73, 0·65	−1·94	−4·00, 0·17
Adjusted model 1	0	−3·93	−10·03, 2·59	−11·87	−18·12, −5·15	−10·34	−16·92, −3·24	−4·10	−6·71, −1·41
Adjusted model 2	0	−1·97	−7·87, 4·30	−8·80	−15·04, −2·10	−7·57	−14·14, −0·49	−3·50	−6·06, −0·87
RRR†									
Unadjusted model	0	2·71	−3·24, 9·03	11·55	4·79, 18·74	11·41	4·46, 18·83	4·97	2·60, 7·40
Adjusted model 1	0	6·13	0·16, 12·45	13·98	7·34, 21·03	10·15	3·53, 17·19	4·88	2·61, 7·19
Adjusted model 2	0	4·13	−1·53, 10·10	11·24	4·97, 17·88	6·88	0·63, 13·52	3·62	1·45, 5·84

HbA1c, glycated Hb; HOMA-IR, homeostasis model assessment of insulin resistance; PCA, principal component analysis; RRR, reduced rank regression. Model 1 adjusted by gender, smoking, alcohol, education, region, age, income, urbanicity index, physical activity; model 2 adjusted by variables in model 1 plus BMI.

*Regression was performed with logarithms of HbA1c and HOMA-IR; therefore coefficients are interpreted as percentage change·

†Performed on residuals estimated for each food group with a multiple regression including geographic region, urbanicity, income and education.

pattern, it was negatively related to rice, poultry and game, and fish and seafood. In addition, the RRR pattern was positively associated with wheat noodles and negatively associated with fresh legumes, items that were not related to the PCA patterns. Eggs and egg products were the only items that were associated in the opposite direction in the RRR and the modern high-wheat patterns.

As expected, the percentage variation explained in food groups was higher for the PCA factors (8·47 % and 7·64 %, respectively, in PCA 1 and PCA 2 *v.* 4·42 % in the RRR pattern). The percentage variation explained in the responses tended to be higher for the RRR, except for the percentage explained for HbA1c which was higher in PCA 2 (traditional southern) than in RRR (2·95 % *v.* 1·40 %); this could be related to using adjusted food groups in the RRR.

Both PCA factors (modern high-wheat and traditional southern) had very strong associations with HbA1c and diabetes that were greatly weakened after adjustment by covariates. Conversely, the estimates for the RRR were only slightly closer to the null after adjustment, which is also related to the use of residuals in the RRR (Tables 3 and 4). For the adjusted estimates, comparing the fourth quartile with the first, regression coefficients from the three dietary patterns were significantly different from zero. For HbA1c, the association was positive for the modern high-wheat dietary pattern (% change = 1·70 (95 % CI 0·51, 2·90)) and the RRR (% change = 3·82 (95 % CI 2·88, 4·77)) and negative for the traditional southern dietary pattern (% change = −2·18 (95 % CI −3·34, −1·01)). For HOMA-IR only the PCA traditional southern pattern had a negative association (% change = −10·34 (95 % CI −16·92, −3·24)) and the RRR pattern had a positive one (% change = 10·15 (95 % CI 3·53, 17·19)). For diabetes only the dietary pattern from the RRR had a significant positive association (OR = 2·37 (95 % CI 1·56, 3·60)).

Compared with the RRR dietary pattern, the strength of association of the PCA modern high-wheat pattern was 56 %, 89 % and 73 % weaker for HbA1c, HOMA-IR and diabetes, respectively, whereas for the traditional southern it was 43 % and 68 % weaker for HbA1c and diabetes and 2 % stronger for HOMA-IR (for estimates from adjusted model 1, comparing fourth *v.* first quartiles). Additionally adjusting by BMI brought all the estimates closer to the null.

Discussion

In the present study we used both PCA and RRR to study the association between dietary patterns and diabetes in China. From the PCA, a modern high-wheat dietary

Table 4 Association between diabetes (HbA1c ≥ 6·5 %) and quartiles of dietary pattern score and linear dietary pattern score increase (1 SD) among adults (*n* 4316) from the China Health and Nutrition Survey (diabetes-related biomarkers measured in 2009; exposures, dietary intakes and covariates measured in 2006)

	Quartile 1	Quartile 2		Quartile 3		Quartile 4		Dietary pattern score (1 SD increase)	
Diabetes prevalence (%)									
PCA, modern high-wheat	3·80	4·91		6·59		7·41		–	
PCA, traditional southern	8·43	5·84		4·07		4·36		–	
RRR*	3·15	5·47		6·67		7·41		–	
		OR	95 % CI	OR	95 % CI	OR	95 % CI	OR	95 % CI
PCA, modern high-wheat									
Unadjusted model	1	1·31	0·86, 1·98	1·79	1·20, 2·65	2·03	1·38, 2·98	1·20	1·06, 1·35
Adjusted model 1	1	1·06	0·67, 1·68	1·30	0·82, 2·07	1·26	0·76, 2·08	1·01	0·86, 1·18
Adjusted model 2	1	1·00	0·63, 1·59	1·22	0·77, 1·95	1·13	0·68, 1·86	0·99	0·84, 1·16
PCA, traditional southern									
Unadjusted model	1	0·67	0·48, 0·94	0·46	0·32, 0·67	0·50	0·34, 0·71	0·74	0·65, 0·85
Adjusted model 1	1	0·90	0·62, 1·32	0·74	0·48, 1·15	0·76	0·49, 1·17	0·88	0·75, 1·05
Adjusted model 2	1	0·93	0·64, 1·35	0·82	0·52, 1·29	0·86	0·54, 1·35	0·91	0·76, 1·09
RRR*									
Unadjusted model	1	1·78	1·16, 2·73	2·20	1·45, 3·33	2·46	1·63, 3·71	1·36	1·19, 1·55
Adjusted model 1	1	1·86	1·20, 2·88	2·21	1·44, 3·39	2·37	1·56, 3·60	1·35	1·19, 1·53
Adjusted model 2	1	1·74	1·12, 2·71	2·08	1·35, 3·21	2·15	1·41, 3·29	1·31	1·15, 1·49

HbA1c, glycated Hb; PCA, principal component analysis; RRR, reduced rank regression.
Model 1 adjusted by gender, smoking, alcohol, education, region, age, income, urbanicity index and physical activity; model 2 adjusted by variables in model 1 plus BMI.
*Performed on residuals estimated for each food group with a multiple regression including geographical region, urbanicity, income and education.

pattern was positively associated with HbA1c, whereas a traditional southern dietary pattern was negatively associated with HbA1c and HOMA-IR. Compared with the RRR, the association of the PCA patterns was about 50 % and about 70 % weaker for HbA1c and diabetes, respectively. But the negative association of the traditional southern dietary pattern with HOMA-IR was comparable to the RRR one. Moreover, the RRR pattern was closely related to the structure of both PCA dietary patterns. It combined the deleterious effects of following the modern high-wheat pattern (high intakes of wheat buns and breads, deep-fried wheat and soya milk) with the deleterious effects of following a diet that is the opposite of the traditional southern one (low intakes of rice, poultry and game, and fish and seafood). This gives public health relevance to the RRR pattern, because it was not only associated with markers of diabetes but was also related to dietary patterns actually followed by this population.

It is also useful to identify the food groups that differ between the PCA and the RRR. For example, we can hypothesize that consumption of wheat noodles and non-consumption of fresh legumes are important in a pattern associated with diabetes even if irrelevant for defining a behavioural pattern. Conversely, intakes of wheat flour; cakes, cookies and pastries; corn and coarse grains; fruits; high-fat pork; organ meats; cow's milk; and instant noodles and frozen dumplings were not key parts of a diabetes-related dietary pattern in this population, even if they defined behavioural dietary patterns. To confirm this we looked at the independent association between each food group and the diabetes-related markers (see Supplemental Table 6, online supplementary material). Indeed, we found that most food groups associated with the RRR including those not in the PCA patterns (wheat noodles and fresh legumes) had independent associations that were in the same direction as their loading in the RRR pattern; whereas most food groups that were associated with the PCA patterns but not with the RRR pattern had a null or even inverse association to the one they had in the PCA pattern (e.g. high-fat pork). This is expected as the PCA identifies the food groups that are distinctive to a behaviour, but the food groups in the pattern could have null or even contradictive associations among them in relation to the outcome, as individuals do not always select foods based on health reasons. On the other hand, the RRR derives the pattern that best explains the variation in the outcome and therefore most food groups within this pattern had a consistent association with the outcome. The only exception we found was soya milk which had a negative association with HOMA-IR but a positive loading in the RRR pattern; this could be due to the fact that soya milk is frequently consumed with deep-fried wheat products.

With RRR we found that a dietary pattern high in wheat products and low in legumes, poultry and fish was positively associated with diabetes, which is consistent with the literature. Evidence suggests that the glycaemic index and staples like noodles and bread are associated with greater risk of diabetes, whereas higher intakes of dietary fibre and legumes have a protective effect against the disease[28–32]. In meta-analyses stratified by region, it has been reported that fish intake has a protective effect

against diabetes in Asian countries[33–35]. There is also evidence in the Chinese population that the intake of poultry is associated with decreased risk of type 2 diabetes[36]. The inverse association of eggs and the positive association of soya milk in our RRR pattern are inconsistent with what previous studies have reported[29,37]. One possible explanation is that the Chinese consume eggs as a replacement for red meat. In the case of soya milk, as explained above, it is possible that because deep-fried dough is commonly accompanied with soya milk for breakfast in China, both foods remained in the same pattern in our analysis.

Moreover, in our analysis rice was inversely associated with the RRR pattern. Yet rice intake has been associated with risk of diabetes in the USA, China and Japan[38]. In Shanghai rice has been shown to be the top contributor to the glycaemic load in the diet[28]. However, a randomized trial substituting brown rice for white rice had no effect on metabolic risk factors[39]. In addition, studies assessing dietary patterns in China have found that individuals with high intakes of rice and vegetables; moderate intakes of fish, poultry and pork; and low intakes of wheat and other cereals had the lowest prevalence of glucose tolerance abnormalities[7] or that a dietary pattern low in rice and high in wheat was positively associated with insulin resistance[9].

A drawback of the RRR is that the derived patterns have the potential to be confounded by other non-dietary factors. For example, it is possible that rice was inversely associated with the RRR pattern partially because in the South rice intake is high and diabetes prevalence is low. The concept of using residuals is to first adjust the food group before including these in the dietary pattern analysis. If confounding factors are not strongly related to both the foods and the responses, then using or not using residuals is irrelevant, as previous studies have found[23]. In our analysis it made a difference to use the residual method, as the structure of the dietary pattern differed (i.e. the loading for rice became weaker and other food groups emerged) and the change in estimates when adjusting by covariates was smaller (Supplemental Tables 2 and 3). Even if the adjusted estimates are relatively similar when using or not using residuals, it is preferable to have a dietary pattern that is already less biased, and therefore using the residuals was a useful approach.

Four studies have used RRR on biomarkers and dietary intake data to derive dietary patterns that predict incident diabetes in American and European populations[23,24,40,41]. In all of the studies the dietary patterns associated with incident diabetes were characterized by refined grains and caloric soft drinks, and some were associated with processed meat, red meat and low-caloric soft drinks and negatively associated with vegetables and wine. Even when the patterns had items in common, Imamura *et al.* found that the patterns from the European studies[23,41] were not generalizable to the US population[24]. Therefore, the RRR pattern we found in China might be even less comparable. Nevertheless, we also found that refined carbohydrates were a very important part of this dietary pattern.

Several studies have compared dietary patterns derived from RRR and PCA with different health outcomes and have mainly focused on comparing which method yields more significant associations[6,42–46]. All except one[42] concluded that RRR derived stronger or more statistically significant patterns. The first RRR and PCA factors of the majority of the studies that present the factor loadings for both methods were relatively similar[42,44,45]. In our analysis, we also found that the RRR pattern was closely related to the PCA patterns. When using RRR, it is important to compare those patterns with the PCA patterns to determine whether the RRR pattern has behavioural significance in the population under study.

To best of our knowledge, the present study is the first that compares RRR and PCA in relation to diabetes in the Chinese population. A strength of our analysis is that we were not limited to a specific urban area or province (the surveyed provinces represent 56 % of the Chinese population). Also due to the longitudinal nature of the study, the temporal sequence is unambiguous, because the diet was measured in 2006 and the outcome was measured in 2009. A limitation is that biomarkers of glucose homeostasis were measured for the first time in 2009, and we could not distinguish between incident and prevalent diabetes. To avoid reverse causality (i.e. participants improving their diets because of diabetes diagnoses), we excluded all the participants who reported being previously diagnosed with diabetes.

The modern high-wheat pattern was only associated with HbA1c and not with HOMA-IR, unlike the other two patterns. Because the modern high-wheat dietary pattern had the weakest associations overall, it is possible that the association with HOMA-IR was not even identified, as it is partially based on a short-term measure (fasting glucose) and therefore is more subject to random error.

The dietary assessment in the CHNS is very detailed and precise[11,12,47]; however, because it covers only 3 d of intake, it is not the best measure of usual intake. Yet the variation explained in the response variables and food groups was comparable to that of other studies using an FFQ[6,23]. Another limitation is that the proportion of consumers during the 3 d was very low ($\leq$5 %) for many key food groups, such as processed meats, Western-style fast foods, salty snacks, ready-to-eat cereals and porridge, and calorically sweetened beverages. It is possible that these foods are important for the development of diabetes in this population, but we were not able to include these in our dietary pattern analysis.

The results of the association between dietary patterns and diabetes-related outcomes might be biased because of residual confounding, particularly from physical activities

that are hard to measure precisely with a self-report questionnaire and that are closely related to both diabetes and diet in terms of energy intake and type of dietary pattern. Our analysis only included occupational and domestic physical activities, because 20 % of the sample lacked information on active leisure and transportation physical activities. However, this does not seem a concern because occupational and domestic physical activities represent about 95 % of all physical activity[48]. Redoing the analysis using all of the participants with complete information on active leisure and transportation physical activities (n 3377) and adjusting accordingly did not change the results. Domestic and occupational physical activities were well correlated with transportation and active leisure physical activities, and hence adjusting by domestic and occupational was enough. We also found that occupational, domestic and transportation physical activities were negatively associated with the modern high-wheat dietary pattern and that active leisure was positively associated with the traditional southern dietary pattern (data not shown). This association with diet in the expected direction and also with diabetes (Table 1), along with results from previous studies that have documented associations between changes in physical activity over time and overweight and other cardiometabolic risk factors[48,49], suggest that our measurement of physical activity was valid and that the residual confounding might be minimal.

Conclusion

In sum, we found that using both PCA and RRR provided important insights. The aims of the two methods are different and their results complement each other. According to our findings, we can hypothesize that in the modern high-wheat dietary pattern the key combination of foods associated with diabetes is wheat buns and breads, deep-fried wheat and soya milk and that in the traditional southern dietary pattern the key protective combination of foods is rice, poultry and fish. Because these sets of food groups are typically consumed together in the Chinese population, it could be possible to identify people at higher risk of developing diabetes and to intervene accordingly if further evidence warrants it.

Acknowledgements

Acknowledgements: The authors thank Ms Frances L. Dancy for administrative assistance and Mr Daniel Blanchette who gave exceptional programming assistance. They also thank the institutional review committees of the University of North Carolina at Chapel Hill (UNC-CH) and the Chinese Institute of Nutrition and Food Safety (INFS), China Center for Disease Control and Prevention. *Financial support:* This research received funding from the National Institutes of Health (NIH; grant numbers R01-HD30880, DK056350, R24 HD050924, R01-HD38700 and R01-HL108427); and the Fogarty International Center, NIH gave financial support for the CHNS data collection and analysis files from 1989 to 2011. C.B. was supported by a scholarship from the Mexican council Consejo Nacional para la Ciencia y Tecnologia. NIH, Fogarty International Center and Consejo Nacional para la Ciencia y Tecnologia had no role in the design, analysis or writing of this article. *Conflict of interest:* None. *Authorship:* C.B. designed and conducted the analysis and wrote the manuscript. M.A.M., P.G.-L., D.S.-A., L.A. and B.P. contributed to the interpretation of the data analysis and reviewed the manuscript. C.B. and B.P. had primary responsibility for the final content. *Ethics of human subject participation:* All protocols, instruments and the process for obtaining informed consent for this study were approved by the institutional review committees of the UCH-CH and the INFS, China Center for Disease Control and Prevention.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1368980014003103

References

1. Pan X-R, Yang W-Y, Li G-W *et al.* (1997) Prevalence of diabetes and its risk factors in China, 1994. *Diabetes Care* **20**, 1664–1669.
2. Xu Y, Wang L, He J *et al.* (2013) Prevalence and control of diabetes in Chinese adults. *JAMA* **310**, 948–959.
3. Salas-Salvadó J, Martinez-Gonzalez MA, Bulló M *et al.* (2011) The role of diet in the prevention of type 2 diabetes. *Nutr Metab Cardiovas Dis* **21**, B32–B48.
4. Thomas T & Pfeiffer AFH (2012) Foods for the prevention of diabetes: how do they work? *Diabetes Metab Res Rev* **28**, 25–49.
5. Hu FB (2011) Globalization of diabetes the role of diet, lifestyle, and genes. *Diabetes Care* **34**, 1249–1257.
6. 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.
7. He Y, Ma G, Zhai F *et al.* (2009) Dietary patterns and glucose tolerance abnormalities in Chinese adults. *Diabetes Care* **32**, 1972–1976.
8. Villegas R, Yang G, Gao YT *et al.* (2010) Dietary patterns are associated with lower incidence of type 2 diabetes in middle-aged women: the Shanghai Women's Health Study. *Int J Epidemiol* **39**, 889–899.
9. Zuo H, Shi Z, Yuan B *et al.* (2013) Dietary patterns are associated with insulin resistance in Chinese adults without known diabetes. *Br J Nutr* **109**, 1662–1669.
10. Michels KB & Schulze MB (2005) Can dietary patterns help us detect diet–disease associations? *Nutr Res Rev* **18**, 241–248.
11. Popkin BM, Du S, Zhai F *et al.* (2010) Cohort profile: the China Health and Nutrition Survey – monitoring and understanding socio-economic and health change in China, 1989–2011. *Int J Epidemiol* **39**, 1435–1440.
12. Popkin BM, Lu B & Zhai F (2002) Understanding the nutrition transition: measuring rapid dietary changes in transitional countries. *Public Health Nutr* **5**, 947–953.

13. Matthews DR, Hosker JP, Rudenski AS *et al.* (1985) Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **28**, 412–419.
14. International Expert Committee (2009) International Expert Committee report on the role of the A1c assay in the diagnosis of diabetes. *Diabetes Care* **32**, 1327–1334.
15. Hare MJ, Shaw JE & Zimmet PZ (2012) Current controversies in the use of haemoglobin A1c. *J Intern Med* **271**, 227–236.
16. Xin Z, Yuan M-X, Li H-X *et al.* (2012) Evaluation for fasting and 2-hour glucose and HbA1c for diagnosing diabetes based on prevalence of retinopathy in a Chinese population. *PLoS One* **7**, e40610.
17. Yang C, Liu Y, Li X *et al.* (2012) Utility of hemoglobin A1c for the identification of individuals with diabetes and prediabetes in a Chinese high risk population. *Scand J Clin Lab Invest* **72**, 403–409.
18. Yu Y, Ouyang X-J, Lou Q-L *et al.* (2012) Validity of glycated hemoglobin in screening and diagnosing type 2 diabetes mellitus in Chinese subjects. *Korean J Intern Med* **27**, 41–46.
19. Ng SW, Norton EC & Popkin BM (2009) Why have physical activity levels declined among Chinese adults? Findings from the 1991–2006 China Health and Nutrition Surveys. *Soc Sci Med* **68**, 1305–1314.
20. Jones-Smith JC & Popkin BM (2010) Understanding community context and adult health changes in China: development of an urbanicity scale. *Soc Sci Med* **71**, 1436–1446.
21. Batis C, Sotres-Alvarez D, Gordon-Larsen P *et al.* (2014) Longitudinal analysis of dietary patterns in Chinese adults from 1991 to 2009. *Br J Nutr* **111**, 1441–1451.
22. Tucker KL (2010) Dietary patterns, approaches, and multicultural perspective. *Appl Physiol Nutr Metab* **35**, 211–218.
23. McNaughton SA, Mishra GD & Brunner EJ (2008) Dietary patterns, insulin resistance, and incidence of type 2 diabetes in the Whitehall II Study. *Diabetes Care* **31**, 1343–1348.
24. Imamura F, Lichtenstein AH, Dallal GE *et al.* (2009) Generalizability of dietary patterns associated with incidence of type 2 diabetes mellitus. *Am J Clin Nutr* **90**, 1075–1083.
25. UCLA Statistical Consulting Group (2006) How do I interpret a regression model when some variables are log transformed? http://www.ats.ucla.edu/stat/mult_pkg/faq/general/log_transformed_regression.htm (accessed April 2014).
26. StataCorp LP (2013) *STATA User's Guide Release 13*. College Station Texas: Stata Press.
27. Hernan MA, Hernandez-Diaz S & Robins JM (2004) A structural approach to selection bias. *Epidemiology* **15**, 615–625.
28. Villegas R, Liu S, Gao Y-T *et al.* (2007) Prospective study of dietary carbohydrates, glycemic index, glycemic load, and incidence of type 2 diabetes mellitus in middle-aged Chinese women. *Arch Intern Med* **167**, 2310–2316.
29. Villegas R, Gao YT, Yang G *et al.* (2008) Legume and soy food intake and the incidence of type 2 diabetes in the Shanghai Women's Health Study. *Am J Clin Nutr* **87**, 162–167.
30. Willett W, Manson J & Liu S (2002) Glycemic index, glycemic load, and risk of type 2 diabetes. *Am J Clin Nutr* **76**, issue 1, 274S–280S.
31. Sluijs I, van der Schouw YT, Spijkerman AM *et al.* (2010) Carbohydrate quantity and quality and risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition–Netherlands (EPIC-NL) study. *Am J Clin Nutr* **92**, 905–911.
32. Weickert MO & Pfeiffer AFH (2008) Metabolic effects of dietary fiber consumption and prevention of diabetes. *J Nutr* **138**, 439–442.
33. Xun P & He K (2012) Fish consumption and incidence of diabetes meta-analysis of data from 438,000 individuals in 12 independent prospective cohorts with an average 11-year follow-up. *Diabetes Care* **35**, 930–938.
34. Zheng J-S, Huang T, Yang J *et al.* (2012) Marine *n*-3 polyunsaturated fatty acids are inversely associated with risk of type 2 diabetes in Asians: a systematic review and meta-analysis. *PLoS One* **7**, e44525.
35. Cai H, Zheng W, Xiang YB *et al.* (2007) Dietary patterns and their correlates among middle-aged and elderly Chinese men: a report from the Shanghai Men's Health Study. *Br J Nutr* **98**, 1006–1013.
36. Villegas R, Shu XO, Gao Y-T *et al.* (2006) The association of meat intake and the risk of type 2 diabetes may be modified by body weight. *Int J Med Sci* **3**, 152–159.
37. Shi Z, Yuan B, Zhang C *et al.* (2011) Egg consumption and the risk of diabetes in adults, Jiangsu, China. *Nutrition* **27**, 194–198.
38. Hu EA, Pan A, Malik V *et al.* (2012) White rice consumption and risk of type 2 diabetes: meta-analysis and systematic review. *BMJ* **344**, e1454.
39. Zhang G, Pan A, Zong G *et al.* (2011) Substituting white rice with brown rice for 16 weeks does not substantially affect metabolic risk factors in middle-aged Chinese men and women with diabetes or a high risk for diabetes. *J Nutr* **141**, 1685–1690.
40. Schulze MB, Hoffmann K, Manson JE *et al.* (2005) Dietary pattern, inflammation, and incidence of type 2 diabetes in women. *Am J Clin Nutr* **82**, 675–684.
41. Heidemann C, Hoffmann K, Spranger J *et al.* (2005) A dietary pattern protective against type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)–Potsdam Study cohort. *Diabetologia* **48**, 1126–1134.
42. DiBello JR, Kraft P, McGarvey ST *et al.* (2008) Comparison of 3 methods for identifying dietary patterns associated with risk of disease. *Am J Epidemiol* **168**, 1433–1443.
43. Nettleton JA, Steffen LM, Schulze MB *et al.* (2007) Associations between markers of subclinical atherosclerosis and dietary patterns derived by principal components analysis and reduced rank regression in the Multi-Ethnic Study of Atherosclerosis (MESA). *Am J Clin Nutr* **85**, 1615–1625.
44. Manios Y, Kourlaba G, Grammatikaki E *et al.* (2010) Comparison of two methods for identifying dietary patterns associated with obesity in preschool children: the GENESIS study. *Eur J Clin Nutr* **64**, 1407–1414.
45. Hoffmann K, Boeing H, Boffetta P *et al.* (2005) Comparison of two statistical approaches to predict all-cause mortality by dietary patterns in German elderly subjects. *Br J Nutr* **93**, 709–716.
46. Vujkovic M, Steegers EA, Looman CW *et al.* (2009) The maternal Mediterranean dietary pattern is associated with a reduced risk of spina bifida in the offspring. *BJOG* **116**, 408–415.
47. Zhai F, Guo X, Popkin BM *et al.* (1996) Evaluation of the 24-hour individual recall method in China. *Food Nutr Bull* **17**, 154–161.
48. Ng SW & Popkin B (2012) Time use and physical activity: a shift away from movement across the globe. *Obes Rev* **13**, 659–680.
49. Adair L, Gordon-Larsen P, Du S *et al.* (2014) The emergence of cardiometabolic disease risk in Chinese children and adults: consequences of changes in diet, physical activity and obesity. *Obes Rev* **15**, 49–59.