IDENTIFYING SMALL GROUPS OF FOODS THAT CAN PREDICT ACHIEVEMENT OF KEY DIETARY RECOMMENDATIONS: DATA MINING OF THE UK NATIONAL DIET AND NUTRITION SURVEY, 2008-12

Running head: Data mining National Diet & Nutrition Survey

Philippe J. GIABBANELLI[1,2], Jean ADAMS[1]

[1]UKCRC Centre for Diet and Activity Research (CEDAR), MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge CB2 0QQ, United Kingdom

[2]Department of Computer Science, Northern Illinois University, DeKalb IL, USA.

Note. PJG was based at affiliation (a) when this work was performed; he is now based at affiliation (b).

*Correspondence concerning this article should be addressed to: Dr Jean Adams, UKCRC Centre for Diet and Activity Research (CEDAR), MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge CB2 0QQ, United Kingdom; jma79@medschl.cam.ac.uk

RESEARCH ETHICS

Ethical approval for the National Diet and Nutrition Survey (NDNS) has obtained from the Oxfordshire A Research Ethics Committee and all participants provided informed consent to take part in the survey. Further ethical approval was not required for this secondary analysis of anonymised data.

ACKNOWLEDGEMENTS

Thanks to Simon Wheeler at the MRC Epidemiology Unit, Cambridge University for help with interpreting NDNS data.

FINANCIAL SUPPORT

The work was undertaken by the Centre for Diet and Activity Research (CEDAR), a UKCRC Public Health Research Centre of Excellence. Both authors gratefully acknowledge funding from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council (MRC), the National Institute for Health Research (NIHR), and the Wellcome Trust, under the auspices of the UK Clinical Research Collaboration (reference MR/K023187/1).

CONTRIBUTIONS

JA conceived the original idea for this work. PJG designed and conducted the data analysis, and produced all figures and tables. Both authors interpreted the results and drafted the manuscript.

COMPETING INTERESTS

None.

ABSTRACT

Background

Many dietary assessment methods attempt to estimate total food and nutrient intake. If the intention is simply to determine whether participants achieve dietary recommendations, this leads to much redundant data. We used data mining techniques to explore the number of foods that intake information was required on to accurately predict achievement, or not, of key dietary recommendations.

Methods

We built decision trees for achievement of recommendations for fruit & vegetables, sodium, fat, saturated fat, and free sugar using data from the UK National Diet and Nutrition Survey (NDNS, 2008-12). Decision trees describe complex relationships between potential predictor variables (age, sex, and all foods listed in the NDNS database) and outcome variables (achievement of each of the recommendations).

Results

4156 individuals were included in the analysis. Information on consumption of 113 out of 3911 (3%) foods, plus age and sex was required to accurately categorise individuals according to all five recommendations. The best trade-off between decision tree accuracy and number of foods included occurred at between 11 (for fruit and vegetables) and 32 (for fat, plus age) foods, achieving an accuracy of 73% (for fat) to 83% (for fruit and vegetables), with similar values for sensitivity and specificity.

Conclusions

Using information on intake of 113 foods, it is possible to predict with 73-83% accuracy whether individuals achieve key dietary recommendations. Substantial further research is required to make use of these findings for dietary assessment.

Keywords

Data mining; diet; dietary assessment; dietary pattern analysis; nutrition

INTRODUCTION

The intention of many dietary assessment methods is to capture information on all foods consumed, or at least those believed to make the largest contribution to total intake,[1] in order to estimate total nutrient intake. For some purposes, this detailed estimation of total nutrient intake may lead to collection of much redundant data. This is particularly the case when assessing adherence with policy targets and messages such as 'five-a-day' portions of fruit and vegetables.

The collection of substantial redundant information places unnecessary burden on research participants, and unnecessarily uses scarce research resources. To take a first step to overcoming this problem, we applied data mining techniques to explore how many, and which, foods information was required on to accurately predict achievement, or not, of key dietary recommendations.

Data mining, an overview

Unlike traditional statistical approaches such as multiple regression, data mining allows multiple, non-linear, relationships and interaction effects to be efficiently captured.[2; 3] Several data mining tools exist. In this study, we use 'classifiers'. A classifier is a function that labels individuals on an outcome (e.g. achieving a dietary recommendation or not) based on a group of predictor variables (e.g. how much of each individual food was consumed). The analysis package is first provided with a 'training set' of individual-level data in which both the outcome and the predictor variables are known, and uses this to learn how the predictor variables are related to the outcome. This produces the classifier function, which can then be used to infer the outcome in a new case based on just the predictor variables. Finally, the accuracy of the classifier is evaluated on a new 'testing set' of data.

There are numerous ways to build classifiers. We used 'decision trees'.[2; 4; 5] Decision trees provide a graphical illustration of a classifier composed of a number of predictor variables. A decision tree involves repeated 'cuts' of the data according to the level of included predictor variables to identify groups of individuals who are similar in terms of the outcome variable of interest. This produces a decision tree where the path from the root to the outcome corresponds to successive 'cuts', or divisions, of the population.

Figure 1 provides a simplified, hypothetical example of a decision tree where the intention is to identify whether or not individuals achieve the recommended intake of fruit and vegetables (the outcome) using information on consumption of carrots and white bread (the two predictor variables). Figure 1a shows the decision tree based on the 'cuts' represented in Figure 1b. Figure 1b is a simple graphical plot of consumption of both carrots and white bread with all individuals labelled according to whether or not they achieve the recommended intake of fruit and vegetables. There appear to be five 'clusters' of participants in Figure 1b in terms of meeting fruit and vegetable recommendations. A series of 'cuts' can isolate these clusters. The first cut (labelled 'A' in both Figure 1a and Figure 1b) divides the population according to consumption of carrots. The next two cuts (labelled 'B' and 'C') then divide the resulting two groups according to consumption of white bread. Finally, a fourth cut (labelled 'D') divides those with a medium carrot and medium white bread intake according to a more fine-grained assessment of carrot intake.

To build decision trees with different numbers of predictor variables, the minimum number of individual cases that can be further divided by a subsequent 'cut' is varied. If a small group of individuals can be further sub-divided, a sizable tree including many predictor variables can result. However, if limits are placed on the minimum size of group that can be further sub-divided, a smaller decision tree, including fewer predictor variables, results. In the current study, we make use of this feature to explore the effect of including more or fewer predictor variables on the accuracy of decision trees.

A small number of studies have applied data mining techniques to nutritional data. These have primarily focused on dietary pattern analysis, exploring which dietary components are predictive of a range of health outcomes.[6] [7] [8] [9] However, we are not aware of any other uses of data mining to identify which foods are predictive of achievement, or not, of key dietary recommendations.

Aims

Our aim was: to use data mining techniques to determine the number of foods that intake information was required on to accurately predict achievement, or not, of dietary recommendations for intake of fruits & vegetables, free sugars, sodium, fat, and saturated fat.

METHODS

We built decision trees for achievement of key dietary recommendations using data from the first four years of the rolling programme of the UK's national dietary surveillance dataset: the National Diet and Nutrition Survey (NDNS).

Data source

The NDNS is an annual cross-sectional survey assessing the diet, nutrient intake and nutritional status of the general population aged 18 months and upwards living in private households in the UK.[10] Since 2008, an annual 'rolling programme' has been in place, allowing data to be combined over years. We used data from years 1-4 of this programme, collected in 2008-12.

The NDNS aims to collect data from a sample of 1,000 respondents per year: at least 500 adults (aged 19 years and older) and at least 500 children (aged 1.5 to 18 years). Households across the UK are selected to take part in the NDNS using a multi-stage probability design. In each wave, a random sample of primary sampling units is selected for inclusion. These are small geographical areas that allow more efficient data collection by enabling it to be geographically focused. Within these primary sampling units, private addresses are randomly selected for inclusion. If, on visiting, it is found that more than one household lives at a particular address, one is randomly selected for inclusion. Within participating households, up to one adult and one child are randomly selected to take part as 'respondents'. Data collection includes completion of four-day estimated food diary – where participants estimate the weight of foods consumed using food labels and household measures.[11]

NDNS data were obtained from the UK Data Archive – an online resource that makes research data available to the UK research community.

Inclusion and exclusion criteria

NDNS participants were included in the analysis if they completed three or four days of the estimated food diary. As recommendations for fruit and vegetable intake only apply to those aged 11 years or older, children aged less than 11 years were excluded from this component of the analysis.

Outcomes of interest – achievement of dietary recommendations

Information on which foods were consumed, and how much participants estimated was consumed, was combined with nutritional information to determine mean daily intake of fruit and vegetables (80g portions), and sodium (mg); and mean daily percentage of energy derived from fat, saturated fat, and free sugars for each individual. This information was then used to determine whether or not each individual met international, or UK, recommendations for these variables.

We used UK recommendations or fruit and vegetable and sodium intake, as these have been graded according to age. It is recommended that individuals aged 11 years and older consume at least five 80g portions of fruit and vegetables per day. This includes a maximum of one portion of juice, with additional juice portions not counted. For sodium, current UK recommendations are that those aged 11 years and older consume no more than 2400mg per day; children aged 7-10 years, no more than 2000mg; children aged 4-6 year, no more than 1200mg; and children aged 1-3 years, no more than 800mg.[12]

The World Health Organization recommends population food and nutrient intake goals for the avoidance of diet related diseases. These state that no more than 30% of energy should be derived from fat, no more than 10% from saturated fatty acids, and no more than 10% from free sugars.[13]

Predictor variables of interest – foods consumed

In total, 3911 different foods (including drinks) have been recorded in NDNS food diaries. We used total estimated weight (in grams) of each individual food eaten by each individual as potential predictor variables. Age and sex were also included as potential predictor variables. The use of including markers of socio-economic position (education, income, and social class) as potential predictor variables was explored but these were found to add no additional increase in accuracy over and above age, sex and individual foods. Decision trees reported here do not include any socio-economic predictor variables.

Data analysis

Our analysis scripts and detailed decision trees are available at https://osf.io/znv82. In all cases except sodium, the proportion of individuals achieving the recommendations was substantially less than 50%; for sodium substantially more than 50% of individuals achieved the recommendations (Table 1). As detailed in Supplementary File 1, this imbalance in outcome variables can lead to low-quality classifiers. To correct this, we pre-processed the data using the Synthetic Minority Over-sampling TEchnique (SMOTE),[14] which creates new cases for the group which accounted for less than 50% of participants by interpolating between existing cases that lie together. WEKA software[15] was then used to build decision trees using the J48 algorithm and error pruning.

For each outcome of interest we built a series of decision trees with different numbers of predictor variables by varying the minimum number of individual cases that could be further divided. For each of the decision trees built, we calculated the number of predictor variables used and overall accuracy in correctly classifying individuals. We used the standard 10-fold cross-validation procedure[16] in which the entire eligible NDNS dataset was split into 10 approximately equally sized parts. Nine parts were used in turn as training sets, and the remaining 10th part was used as testing set. The ability of decision trees to correctly identify those who achieved the recommendations (sensitivity) and those who did not (specificity) was also calculated. Adaptive sampling was used to identify the maximum overall accuracy that could be achieved, as well as the optimum trade-off between minimising number of predictor variables and maximising overall accuracy.

RESULTS

Overall, 91% of households eligible for inclusion agreed to take part in the first four waves of NDNS. Within these, 56% (2083 adults and 2073 children; 4156 participants in total) of individuals selected to take part completed three or four days of the estimated food diary and were included in the analysis for sodium, free sugars, fat and saturated fat. Of these 4156 participants, 2967 (71.4%) were aged 11 years or older and included in the analysis for fruit and vegetables. There were no missing data on sex or age.

The distributions of age and sex in the analytical sample compared to the UK population as a whole are shown in Table 1. As the NDNS sample contains relatively equal numbers of children aged 18 years or younger, and adults, distributions are provided separately for adults and children in this table. The main differences between the age and sex distributions in the analytical sample and UK population were that the analytical sample had a higher proportion of adult women and a lower proportion of young adults (aged 19-29 years) than the UK population.

Figure 2 shows the overall accuracy of decision trees for each of the five outcomes plotted against the number of predictor variables in decision trees. Overall accuracy ranged from 69% (fat; 10 predictor variables) to 84% (fruit and vegetables; 50 predictor variables) depending on the outcome of interest and number of predictor variables included. For all guidelines but sodium, the relationship between the number of predictor variables and the accuracy was best described using a logarithmic trend model ($p<0.01$ in all cases). Thus, increasing the number of predictor variables from around 10 to 30 improved the accuracy by a maximum of around five percentage points, but beyond this adding even a large number of additional predictor variables yielded only a very small additional improvement. We were unable to fit any function to the relationship between accuracy and number of predictor variables for sodium.

Table 2 provides information on the decision tree for each outcome that represented the best trade-off between accuracy and number of predictor variables. Information on the most accurate possible tree for each outcome is also shown in Table 2. Between 11 (for fruit and vegetables) and 33 (for fat) predictor variables provided the best trade-off to identify whether individuals achieved each of the recommendations, achieving overall accuracy of 73% (for fat) to 83% (for fruit and vegetables). Adding further predictor variables beyond this improved accuracy by a maximum of 2% (for saturated fat) and less than 1% (for all other outcomes). Sensitivity and specificity were similar to overall accuracy for fruit and vegetables and free sugars (and saturated fat when the maximum number of predictor variables were included). However, specificity was higher than sensitivity for fat (and saturated fat), but the reverse was seen for sodium. Predictor variables in decision trees with the best trade-off between accuracy and number of predictor variables accounted for between 13% (for fat) and 31% (for free sugars) of total intake of relevant outcome variables.

Predictor variables used in decision trees with the best trade-off between accuracy and number of predictor variables are shown in Table 3. In total, 113 foods (out of a total 3911 [3%] recorded as consumed), age and sex were included in the decision trees for all five outcomes. Overall, there was little overlap in predictor variables across outcomes. Age and two foods were included as predictor variables in the decision trees for three outcomes. A further six foods were included as predictor variables in the decision trees for two outcomes. The remaining 104 foods were included as predictor variables in only one decision tree.

DISCUSSION

Summary of results

This is the first work we are aware of using data mining techniques to explore the number of foods that information is required on to predict achievement of dietary recommendations. In total, information on consumption of 113 of 3911 foods (3%), plus age and sex was required to accurately categorise individuals according to all five dietary recommendations (fruit & vegetables, free sugars, sodium, fat, and saturated fat). The best trade-off between decision tree accuracy and number of foods included was achieved at between 11 (for fruit and vegetables) and 32 (for fat, plus age) foods. These decision trees had an overall accuracy of 73% (for fat) to 83% (for fruit and vegetables), with similar values for sensitivity and specificity. Few individual foods were present in the decision tree for more than one dietary recommendation, although age was present in three.

Strengths and limitations of methods

We used data from a population-based sample meaning our findings are likely to be generalizable across the UK and to other countries with similar dietary profiles. However, diets vary internationally[17] and our results may not be more widely generalizable. The analytical sample had a slightly higher proportion of adult women and lower proportion of younger adults (aged 19-29 years) than the UK population as a whole.

The data used were collected using ‘estimated’ food diaries – where portion sizes were estimated but not weighed. These are considered to be one of the more accurate methods of measuring dietary intake,[18] meaning that both the predictor and outcome variables are likely to be valid. However, even estimated food diaries have their limitations, particularly in terms of participant burden and under-reporting of energy intake.[19; 20] Doubly labelled water has been used to estimate total energy expenditure in a subsample of NDNS participants and compare this to reported energy intake from food diaries. This reveals that reported energy intake is 12-34% lower than estimated total energy expenditure, depending on the age of participants.[11] This mismatch may be due to intentional or unintentional misreporting; participants changing their food intake in response to recording it; or a variety of other reasons. However, misreporting is unlikely to affect all foods and nutrients equally. For example, participants may be more likely to misreport confectionary than vegetable intake. For this reason, misreporting is not adjusted for in NDNS and we have not adjusted for misreporting here.

Data mining using decision trees is computationally and statistically efficient. For example, inclusion of all 3911 foods consumed by NDNS participants in regression models with achievement of dietary recommendations as outcomes would be computationally, and statistically, demanding and unlikely to produce satisfactory results. Decision trees also produce transparent, and intuitively understandable, outputs (ours are provided at https://osf.io/znv82).(21)

Many of food included in the analysis had very skewed distributions. Indeed, the vast majority of foods in the database (3618) were eaten by less than 150 people. Decision trees seek to maximize information gain at each step, rather than working with the distribution as a whole as in traditional regression analysis. If an item is very discriminatory and helps differentiate between those who do and do not meet a particular guideline then it will be included, even if it is only consumed by a small number of people. Conversely, if an item is eaten by almost everyone but is not discriminatory, then it would be unlikely to be included. There was no overall trend between the proportion of participants who ate a food and the chance that that food was included in a decision tree (data not shown).

We used adaptive sampling to identify decision trees that achieved the best trade-off between accuracy and number of predictor variables included. Thus, instead of systematically calculating the accuracy of all decision trees including all possible number of predictor variables, we focused on identifying the relationship between accuracy and number of predictor variables (logarithmic in most cases), where the optimum trade-off between accuracy and number of predictor variables occurred (i.e. where the logarithmic curve flattened out). This means we cannot be absolutely sure that we have identified the decision trees with the best trade-off between accuracy and number of predictor variables in all cases. However, given the very small additional improvements in accuracy achieved by the most accurate, versus best trade-off, decision trees, we are certainly likely to have identified the near-best trade-off decision trees.

We used estimated dietary records as our ‘gold standard’ tool for determining whether or not individuals achieved recommendations. Further work will be required to compare the accuracy of our decision trees to other methods of estimating who achieves dietary recommendations, such as food frequency questionnaires.

Interpretation and implications of findings and areas for future work

Our findings indicate that information on only a small number of foods is required to determine whether individuals achieve five important dietary recommendations. If such binary outcomes are the key outcome of interest, then more detailed dietary assessment methods, may inappropriately use scarce research resources and be unnecessarily burdensome to participants.

Whilst our results suggest that information on only a limited number of foods needs to be captured when assessing whether guidelines are met, substantial further research will be needed before these findings could be applied in the form of a new dietary assessment instrument. Firstly, it would be helpful to replicate our analyses in a different, but comparable, sample. We have not done is as we are not aware of a comparable UK population-representative sample in whom diet diaries have been collected. Our decision trees used information on exact intake of 113 foods over 3-4 days. Assessing exact intake of a small number of foods may be no less burdensome for participants than assessing estimated intake of all foods using a food diary. Future work could compare the accuracy of decision trees based on exact intake of 113 foods, approximate intake of these foods (e.g. using the ordinal categories often used in food frequency questionnaires), and exact and approximate intake of foods at the food group, rather than individual food, level. Acceptability to research participants and resource implications of collecting the data required in all cases should also be compared.

Our analysis focused on which foods can be used to predict whether or not individuals achieve dietary recommendations. But it is not necessarily the case that it is the foods included in the decision tress which cause people to achieve the recommendations or not. Only a maximum of 32% of total intake of relevant nutrients or foods were accounted for by predictor variables in decision trees with the best trade-off between accuracy and number of predictor variables. Thus, decision trees did not particularly include foods that account for the majority of intake of nutrients and foods of interest – as might be expected in a food frequency questionnaire. The complex relationships between individual foods included in our decision trees and the dietary recommendations they are associated with may offer further useful insights and could be studied further.

CONCLUSION

We used data mining techniques to explore the number of foods that consumption information was required on to accurately predict achievement, or not, of five key dietary recommendations. Information on consumption of 11-32 foods (plus age and sex) was sufficient to identify with 73-83% accuracy whether individuals achieved individual dietary recommendations. In total, information on 113 foods was required to predict achievement of all five recommendations studied. This method could be used to develop a new dietary assessment questionnaire.

REFERENCES

1. Willett WC, Sampson L, Stampfer MJ *et al.* (1985) Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol* **122**, 51-65.

2. Crutzen R, Giabbanelli P (2013) Using Classifiers to Identify Binge Drinkers Based on Drinking Motives. *Subst Use Misuse*.

3. Dierker L, Rose J, Tan X *et al.* (2010) Uncovering multiple pathways to substance use: a comparison of methods for identifying population subgroups. *J Prim Prev* **31**, 333-348.

4. McKenzie DP, McFarlane AC, Creamer M *et al.* (2006) Hazardous or harmful alcohol use in Royal Australian Navy veterans of the 1991 Gulf War: identification of high risk subgroups. *Addict Behav* **31**, 1683-1694.

5. Hillemacher T, Frieling H, Wilhelm J *et al.* (2012) Indicators for elevated risk factors for alcohol-withdrawal seizures: an analysis using a random forest algorithm. *J Neural Transm* **119**, 1449-1453.

6. Lazarou C, Karaolis M, Matalas A-L *et al.* (2012) Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. *Comput Methods Programs Biomed* **108**, 706-714.

7. Kastorini C-M, Papadakis G, Milionis HJ *et al.* (2013) Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: A case/case-control study. *Artif Intell Med* **59**, 175-183.

8. Thangamani D, Sudha P (2014) Identification Of Malnutrition With Use Of Supervised Datamining Techniques -Decision Trees And Artificial Neural Networks. *International Journal Of Engineering And Computer Science* **3**, 8236-8241.

9. Einsele F, Sadeghi L, Ingold R *et al.* (2015) A Study about Discovery of Critical Food Consumption Patterns Linked with Lifestyle Diseases using Data Mining Methods. *Proceedings of the International Conference on Health Informatics*, 239-245.

10. Bates B, Lennox A, Swan G (editors) (2010) *National Diet and Nutrition Survey: Headline results from Year 1 of the Rolling Programme (2008/2009).* London: Foods Standards Agency and Department of Health.

11. Bates B, Lennox A, Prentice A *et al.* (editors) (2014) *National Diet and Nutrition Survey Results from Years 1, 2, 3 and 4 (combined) of the Rolling Programme (2008/2009 – 2011/2012).* London: Public Health England.

12. Scientific Advisory Committee on Nutrition (2003) *Salt and Health*. London: The Stationary Office.

13. World Health Organisation (2003) Diet, nutrition and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation. *WHO Technical Report Series* **916**.

14. Chawla N, Bowyer K, Hall L *et al.* (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321-357.

15. Bouckaert RR, Frank E, Hall MA *et al.* (2010) WEKA - Experiences with a Java Open-Source Project. *Journal of Machine Learning Research* **11**, 2533-2541.

16. Kuncheva L (2004) *Fundamentals of pattern recognition Combining pattern classifiers: Methods and algorithms.* Hoboken, New Jersey: John Wiley & Sons.

17. Imamura F, Micha R, Khatibzadeh S *et al.* Dietary quality among men and women in 187 countries in 1990 and 2010: a systematic assessment. *The Lancet Global Health* **3**, e132-e142.

18. Bingham S, Gill C, Welch A *et al.* (1994) Comparison of dietary assessment methods in nutritional epidemiology: weighed records v. 24 h recalls, food-frequency questionnaires and estimated-diet records. *Br J Nutr* **72**, 619-643.

19. Poslusna K, Ruprich J, de Vries JH *et al.* (2009) Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. *Br J Nutr* **101 Suppl 2**, S73-85.

20. Burrows TL, Martin RJ, Collins CE (2010) A systematic review of the validity of dietary assessment methods in children when compared with the method of doubly labeled water. *J Am Diet Assoc* **110**, 1501-1510.

21. Crutzen R, Giabbanelli PJ, Jander A *et al.* (2015) Identifying binge drinkers based on parenting dimensions and alcohol-specific parenting practices: building classifiers on adolescent-parent paired data. *BMC Public Health* **15**, 747.

FIGURE TITLES AND LEGENDS

Figure 1. Schematic illustration of a decision tree (left, Figure 1a.) and how this is formed through repeated ‘cuts’ of the data (right, Figure 1b)

Figure 1a. Schematic illustration of a decision tree

Figure 1b. Schematic illustration of how a decision tree is formed through repeated ‘cuts’ of the data

Figure 2. Overall accuracy (with 95% confidence margins) of decision trees against number of predictor variables included

Table 1. Comparison of analytical sample to UK population

	Adults aged 19y or older		Children aged <19y	
Variable	Analytical sample (n=2083)	UK population	Analytical sample (n=2073)	UK population
Female, n(%)	1182 (56.8)	25,198,773 (51.5)	1007 (48.6)	6,955,262 (48.8)
Age (adults)				
19-29y, n(%)	296 (14.2)	9,447,071 (19.3)	--	--
30-39y, n(%)	390 (18.7)	8,319,926 (17.0)	--	--
40-49y, n(%)	425 (20.4)	9,268,735 (18.9)	--	--
50-59y, n(%)	363 (17.4)	7,708,532 (15.8)	--	--
60-64y, n(%)	181 (8.7)	3,807,975 (7.8)	--	--
65y+, n(%)	428 (20.6)	10,377,127 (21.2)	--	--
Age (children)				
0-4y, n(%)	--	--	499 (24.1)	3,913,953 (27.5)
5-9y, n(%)	--	--	583 (26.4)	3,516,615 (24.7)
10-14y, n(%)	--	--	547 (26.4)	3,669,326 (25.7)
15-18y, n(%)	--	--	444 (21.4)	3,152,919 (22.1)

Table 2. Prevalence of achieving and not achieving dietary recommendations and accuracy of decision trees to predict this

	Fruit& vegetables	Free sugars	Sodium	Fat	Saturated fat
N (%) achieving recommendation without over-sampling	656 (22.1%)	1472 (35.4%)	2524 (60.7%)	1045 (25.1%)	795 (19.1%)
SMOTE over-sampling %*	252% (YES)	85% (YES)	54% (NO)	197% (YES)	322% (YES)
N achieving recommendation after over-sampling	2309*	2679	2524	3103	3354
N not achieving recommendation after over-sampling	2311*	2684	2513	3111	3361
Decision tree with the best trade-off between accuracy and number of predictor variables					
Overall accuracy	83.1%	76.5%	75.9%	72.4%	79.7%
Sensitivity	82.5%	76.1%	81.9%	66.3%	75.8%
Specificity	83.8%	76.9%	69.8%	78.4%	83.6%
Npredictor variables	11	28	28	33	28
% of all relevant food/nutrient (g) accounted for by predictor variables	21.0%**	31.2%	13.4%	13.0%	27.4%
Most accurate decision tree					
Overall accuracy	83.6%	77.0%	76.1%	72.9%	81.7%
Sensitivity	83.9%	75.7%	80.7%	69.3%	81.4%
Specificity	83.3%	78.3%	71.5%	76.4%	81.9%
N predictor variables	50	64	49	123	156
% of all relevant food/nutrient accounted for by predictor variables	30.8%**	38.6%	25.4%	29.5%	42.7%

*After over-sampling using the SMOTE method (see Appendix); the prevalence affected by over-sampling is underlined

**Percent of all fruit and vegetables (g) recorded, not just those contributing to 5-a-day portions (specifically, fruit juice can only contribute a maximum of one 5-a-day portion)

Table 3. Predictor variables (individual foods, age and sex) included in decision trees for predicting achievement of five dietary recommendations

Dietary recommendation outcome					Food name
Fat	Free sugars	Fruit & veg	Sodium	Saturated fat	
Yes			Yes	Yes	Age
Yes					Alcoholic soft drinks spirit based
		Yes			Almonds kernel only: ground almonds
	Yes				Apple juice unsweetened cartons pasteurised
	Yes				Apple juice unsweetened UHT
		Yes			Apples eating raw flesh & skin only
Yes					Avocado pear flesh only
			Yes		Bacon rashers back grilled lean and fat
			Yes		Bacon rashers back not smoked grilled extra trim
			Yes		Baked beans in tomato sauce with pork sausages
Yes		Yes			Bananas raw flesh only
				Yes	Beefburger and onion grilled
Yes					Black pudding fried
	Yes				Blackcurrant juice drink ready to drink not low calorie
	Yes				Boiled sweets barley sugar butterscotch glacier mints hard candy
			Yes		Bread white crusty
			Yes	Yes	Bread white toasted
Yes					Bread, 50% white and 50% wholemeal flours
			Yes		Bread, white sliced, not fortified
			Yes		Brown sauce bottled
			Yes		Brussels sprouts-fresh boiled
Yes					Butter beans dried boiled
Yes				Yes	Butter salted
				Yes	Butter unsalted
	Yes				Carbonated beverages no juice not low calorie canned
Yes	Yes			Yes	Carbonated beverages no juice not low calorie not canned
		Yes			Celery, fresh raw
Yes					Chapati brown no fat
Yes				Yes	Cheese cheddar any other or for recipes
				Yes	Cheese cheddar English
			Yes		Cheese soft full fat. Philadelphia type
Yes					Chicken fried in olive oil
				Yes	Children’s fromagefrais fruit with added vitamin D
				Yes	Chocolate brownie no nuts purchased
				Yes	Chocolate covered caramels Cadburys caramel
	Yes				Chocolate Swiss roll with buttercream purchased
	Yes				Cola cherry cola canned not low calorie
	Yes				Cola not canned not low calorie not caffeine free
Yes					Coleslaw purchased not low calorie
Yes					Cookies and biscuits with chocolate
				Yes	Cornetto type ice cream chocolate or nut based
	Yes				Cranberry fruit juice drink e.g. Ocean Spray
				Yes	Cream double
	Yes				Cream egg
				Yes	Croissants plain not filled
	Yes				Drinking chocolate instant dry weight
			Yes		Fat spread (62-72% fat) not polyunsaturated
	Yes				Fruit gums winegums
	Yes				Fruit juice drink carbonated not low calorie not canned
	Yes				Fruit juice drink with 5% fruit juice ready to drink
				Yes	Fully coated chocolate biscuits with biscuit filling
Yes					Garlic bread. Lower fat
			Yes		Ham unspecified not smoked not canned
			Yes		Hamburger Big Mac McDonalds
	Yes				High juice ready to drink not blackcurrant or low calorie
	Yes				Ice lollies
	Yes				Jaffa Cakes
				Yes	Kit Kat

Yes					Lager not canned e.g. Heineken
Yes					Lager not canned e.g. Skol
Yes					Lamb scrag and neck stewed lean only
	Yes				Lemonade not low calorie not canned
				Yes	Light spreadable butter (60% fat)
	Yes				Lucozade sport isotonic drink not carbonated
Yes				Yes	Mayonnaise (retail)
			Yes	Yes	Milk chocolate bar
	Yes				Milk shake thick style takeaway
Yes					Milk skimmed after boiling
				Yes	Milk whole pasteurised winter
				Yes	Milk whole summer pasteurised
Yes					Mushrooms fried in olive oil
			Yes		Naan bread plain
		Yes			Oatcakes
	Yes				Olive oil
		Yes			Onions boiled
	Yes				Orange juice unsweetened UHT
Yes					Oven ready chips
			Yes		Papadums/poppadoms fried in vegetable ghee
Yes					Pasta noodles boiled
			Yes		Pasta noodles egg boiled
Yes					Pasta spaghetti boiled white
			Yes		Peanut butter crunchy not wholenut
		Yes			Pears eating raw flesh & skin only no core
Yes					Pepperami
				Yes	Petit Filousfromagefrais
			Yes		Potato cakes (scones) purchased
Yes					Potatoes new boiled skins eaten
			Yes		Potatoes old baked flesh & skin
				Yes	Potatoes old mashed & butter
			Yes		Prawns boiled flesh only
			Yes		Reduced fat spread (41-62%) not polyunsaturated
			Yes		Ribena original blackcurrant drink concentrate
	Yes				Robinsons fruit shoot
			Yes		Rolls white crusty
Yes		Yes		Yes	Sausage roll flaky pastry purchased
Yes					Sausages, pork, grilled
			Yes		Sausages, premium pork, grilled
Yes					Scrambled eggs with skimmed milk and no fat
Yes					Semi-sweet biscuit
			Yes		Sex
	Yes				Soya alternative to milk sweetened plain
		Yes			Spinach fresh raw
				Yes	Spreadable butter (75-80% fat)
	Yes				Sugar white
				Yes	SupernoodlesBatchelorsas served
				Yes	Swiss roll individual chocolate coated purchased
		Yes			Tomatoes raw
			Yes		Turkey slices unsmoked prepack or deli
	Yes				Water for concentrated soft drinks not diet
Yes					White chocolate buttons mice
Yes					Whole milk after boiling
Yes					Wine white dry not canned
	Yes			Yes	Yogurt twinpot with cereal/crumble
		Yes			Yogurt, Greek style, cows, natural, whole milk
			Yes		Yorkshire pudding frozen