Developing a whole systems obesity classification for the UK Biobank Cohort

Stephen D Clark^{1x}, Mark Birkin, Nik Lomax^{1*}, Michelle A Morris^{2*}

*Equal contribution

^x Corresponding

Email- tra6sdc@leeds.ac.uk

Postal address- Leeds Institute for Data Analytics, University of Leeds, LS2 9JT

Affiliations:

1 – Leeds Institute for Data Analytics & School of Geography, University of Leeds

2 – Leeds Institute for Data Analytics & School of Medicine, University of Leeds

Abstract

The number of people who are obese and overweight presents a global challenge, and the development of effective interventions is hampered by a lack of research which takes a joined up, whole system, approach that considers multiple elements of the complex obesity system together. We need to better understand the collective characteristics and behaviours of the overweight and obese population and how these differ from those who maintain a healthy weight. Using the UK Biobank cohort of 500 000 adults, we develop an obesity classification system using k-means clustering. Variable selection from the UK Biobank cohort is informed by the Foresight obesity system map across key domains (Societal Influences, Individual Psychology, Individual Physiology, Individual Physical Activity, Physical Activity Environment). This paper presents the first study of UK Biobank participants to adopt this obesity system approach to characterising participants. Our classification identifies six groups of people, similar in respect to their exposure to known drivers of obesity: ‘Active workers’, ‘Retirees with healthy lifestyles’, ‘Stressed and not in work’, ‘Deprived with less healthy lifestyles’, ‘Comfortable professionals’ and ‘Comfortable families’. Pen portraits are developed to describe the characteristics of these different groups. Multinomial logistic regression is used to demonstrate that the classification can effectively detect groups of individuals more likely to be overweight or obese. The group identified as ‘Comfortable families’ are observed to have a higher proportion of healthy weight, while three groups have increased relative risk of being overweight or obese: ‘Active workers’, ‘Stressed and not in work’ and ‘Deprived with less healthy lifestyles’. This work presents an innovative new approach to better understand the complex drivers of obesity which has the potential to produce meaningful tools for policy makers to better target interventions across the whole system to reduce overweight and obesity.

Introduction

Obesity presents a global challenge for society, with 650 million people, (13% of the total adult population), estimated as being obese worldwide (World Health Organisation, 2020, World Obesity, 2020) with an additional 39% of adults being classed as overweight. This wicked problem is not limited to adults, with there being over 340 million (18%) overweight or obese children and adolescents (aged 5-19 years) globally in 2016 (World Health Organisation, 2020). Overweight and obesity prevalence continues to increase, and with it related comorbidities, in spite of the fact that it is preventable. In recent years there has been significant investment by United Kingdom (UK) research funders (Medical Research Council, 2020a, Medical Research Council, 2020b) to better understand, and subsequently prevent, weight related health problems, however, overweight and obesity still prevail.

Research to date concludes that the drivers of obesity are complex and multifaceted, and not as simple as consuming less food and drink or moving more (Butland et al., 2007, Lee et al., 2017). Biological, social/cultural, ecological and psychological factors combine to create a tangled web of obesity promoting behaviours and environments (Ulijaszek, 2015). In turn interventions and policy decisions to prevent overweight and obesity are complex. Success will inevitably be limited through tackling individual elements, such as proximity to fast food outlets (Burgoiné et al., 2016) or green spaces (Sarkar, 2017), in isolation. Interventions leading to the greatest overweight and obesity prevention are hypothesised to come from a whole systems approach at a macro level – tackling multiple components of the obesity system at a large geographic scale (Butland et al., 2007, Swinburn et al., 1999). In the UK, ranked the 6th most obese OECD country (Devaux et al., 2017), the obesity system was comprehensively mapped in 2007 as part of the UK government's Foresight initiative (Miles and Keenan, 2003, Government Office for Science, 2007), yet attempts to utilise this, in its entirety, in obesity prevention have been limited (Butland et al., 2007, Local Government Association, 2017, Morris et al., 2018, Bagnall et al.,

2019). In a recent systematic review just thirty qualitative studies were identified as having taken a whole systems approach to obesity or other complex public health challenges (Bagnall et al., 2019). More research took a quantitative approach, with 44 studies identified. These studies followed a range of study designs (Allender et al., 2016, Johnston et al., 2014, Lake et al., 2017) , but in the most part did not report their methods clearly, which given the complexity associated with a whole systems approach, makes interpreting the findings difficult. Linked data on all elements of the whole system are not readily available at an individual level, nor at the population level, further adding to the challenge and complexity associated with taking a whole systems approach to obesity (or other complex public health challenges) (Bagnall et al., 2019, Johnston et al., 2014, Morris et al., 2018). A comprehensive data mapping exercise against the Foresight obesity system map was completed in 2018, concluding that more can be done using traditional and novel data sources to incorporate more aspects of the obesity system into ongoing research (Morris et al., 2018)

Social and environmental interventions present opportunities for change that can be implemented at both a micro and macro level, in communities and whole countries by local and national governments. However, it is not yet clear what the most effective approaches would be and how to synchronise changes across the system. In order to inform such change, we first need to better understand the collective characteristics and behaviours of people who are overweight or obese and how these differ from those who maintain a healthy weight.

Traditionally, insights of this kind are generated from cohort studies that collect a wealth of data spanning individual behaviours, demographic characteristics, anthropometric measures and health related metrics. These encompass multiple areas of the obesity system, however, mostly relate to the individual, rather than wider environmental, societal and food system determinants.

A more holistic approach in obesity research has existed for some time, for example, clusters of lifestyle behaviours have been used in different settings to assist policy makers in targeting

populations for a range of interventions. Often such clusters focus on individual elements of lifestyle, for example, dietary patterns (Greenwood et al., 2000, Kant, 2004). In some cases broader lifestyle clusters are identified (Cameron et al., 2011, Patterson et al., 1994). While insightful, data driven clusters of these kinds are not easily comparable against other cohorts and populations. Variables that drive the clusters are selected through a range of variable selection methods such as principal component or factor analysis, or more simply, by information that is available within the cohort. More generic clustering solutions such as geodemographic classifications, originally generated with marketing in mind, have been utilised by local government organisations and have demonstrated utility in highlighting groups with higher prevalence of obesity and related comorbidities (Nottingham City Council et al., 2011, Petersen et al., 2011, South East Public Health Observatory, 2011, Sunderland City Council, 2017). Some geographic solutions have been tailored to specific application domain areas, for example CACI's ACORN Wellbeing classification (CACI, 2021) which segments the population into four categories: Health Challenges; At Risk; Caution; Healthy - and further segregates these into 25 groups. Input data for such geodemographic classifications are largely derived from census data, open data or aggregated commercial sources. In research cohort studies rich data are available at an individual level, without the need to combine aggregated data sources using computational models to estimate patterns. Therefore, combining insight from geodemographic classifications with rich cohort data using a robust framework such as the Foresight system map presents exciting possibilities for better understanding and leveraging the insight relevant to the whole obesity system. The UK Biobank cohort is an example of an important study that is already contributing to obesity research, albeit focusing on specific areas of the obesity system, rather than the whole system. Using UK Biobank, Burgoine et al. (2018) and Mason et al. (2018) examine the influence of the availability of fast food on obesity. Activity levels are studied as commuting behaviour in two studies reported by Flint and Cummins (2016) and Flint et al. (2016) and by an examination of how the built environment can influence activity patterns

(Sarkar, 2017, Sarkar et al., 2017). Other potential influences for obesity have also been studied using UK Biobank participant data, including socio-economic factors (Tyrrell et al., 2016), smoking (Dare et al., 2015), presence of morbidities (Lyall et al., 2017), work patterns (Wyse et al., 2017) and ethnicity (Ntuk et al., 2014).

Green et al. (2016) conducted a classification exercise, based solely on data from their obese participants, meaning that the purpose of the classification was not to differentiate different weight statuses. Furthermore, this classification did not use the Foresight obesity system map to inform the classification development. We believe that a classification developed from the robust Foresight obesity system map framework, applied to a cohort with a range of weight statuses is novel and will generate substantial transferrable utility. The aim of this study, is to (i) develop an obesity classification system where variable selection is informed by the Foresight system obesity map, applied to a sizeable sub-sample of the UK Biobank cohort of 500 000 adults (Sudlow et al., 2015), and (ii) test this against overweight and obesity outcomes. We hypothesise that distinct classes/clusters will be generated that differentiate different weight statuses.

Material and methods

The UK Biobank is a large prospective cohort study of the 40 to 70 year old population, with baseline data collected between 2006 and 2010 in England, Scotland and Wales (Allen et al., 2012, Sudlow et al., 2015). The data were collected during an initial assessment at 22 regional centres, and provides information on the participants' socio-demographic and economic situation, various physiological measurements, cognitive abilities and a limited number of biomarkers. Participants were asked to consent to the linking of their health episode and death data. Subsequently, sub-samples of the participants were invited back for repeat-assessments, for example to contribute to an imaging study or to measure physical activity levels using accelerometers. The cohort employed a range of data collection methods, included self-reported

dietary surveys and BMI generated from height and weight measured at the initial assessment visit.

In order to produce a classification of participants that aligns to known drivers for obesity, we select UK Biobank variables (Hanscombe et al., 2019) that map onto the Foresight obesity systems map (Jebb et al., 2007, Butland et al., 2007). This involved two ‘sifting’ exercises. The first sift was undertaken by three researchers (a human geographer, a nutritionist and a statistician) and involved the independent consideration of every UK Biobank variable then available and using the UK Biobank ‘Data Showcase’ to evaluate its utility when mapped to the Foresight obesity map. These three considerations were then discussed and a candidate sub-set of UK Biobank variables were requested. Having obtained these variables from UK Biobank a second sift was conducted by the same team. Access to the variables at the participant level allowed a more nuanced consideration of distributions and cross tabulations to further evaluate the utility of a sub-set of variables. Also at this point it became clear that many of the candidate variables had only partial coverage of UK Biobank participants (e.g. measurement from wearable accelerometers (n=103 695) and food intake diaries (n=70 714)). If such variables were to be included, the classification sample would be much reduced. This process is summarised in Figure 1.

For analysis k-means classification analysis is used to derive the obesity classification (Everitt, Landau, & Leese, 2001). What classification methods attempt to do is to group together participants that are most similar on a number of variables that describe their characteristics or nature, so that participants in each class will be more similar to their fellow class members than those in other classes. The advantage of this k-means approach for clustering over others, e.g. hierarchical classification, is that a potentially large distance matrix (of size 379k by 379k) is not required, only the distances from each observation to the k class centres. For k-means to be successfully applied it is necessary that the variables are not skewed, so that symmetric classes

can form, and that each variable used is measured on a similar scale, so that each contributes a similar weight to the classification process, achieved here by converting each variable to a z-score. To correct for skewness Tukey's ladder of power approach is used (Mosteller and Tukey, 1977). Prior to classification, the correlation between variables is measured to assess whether there is any potential redundancy in the variables where two or more variables are essentially capturing the same dimension.

To determine the number of classes, scree plots of the within classes sum of squares and the reductions in this value from 1 to 12 classes are used. For k-means analysis there can be no incomplete cases, therefore where information is either not known or not supplied by the participant, they are not used. Given that k-means clustering can only use continuous or integer count variables, we selected these for inclusion in the classification, and in addition the categorical variables with a match to the obesity systems map are used to profile the new classes.

To gain an understanding of the nature of each class, the centre of each class (which is essentially the mean value of each classification variable (which are given in Table 1) for the observations that are part of the class) is calculated help to characterise the class. This helps to derive 'pen portraits' for each class – some classes will contain participants that, in aggregate, have high (or low) values for certain variables, and what these variables are provides the narrative for the pen portrait. Also, how the classes profile against other variables that have not been used to drive the classification are insightful. These additional variables come from UK Biobank and are also variables that are highlighted within the Foresight Obesity map. Pen portrait names were assigned to the classes following a workshop with 35 multidisciplinary academics in attendance that aimed to generate meaningful names that were not considered to be stigmatising.

In the results presented below, descriptive statistics are tabulated and multinomial regression analysis is used to test whether the new classification can predict weight status. The utility of the multinomial regression is to test the ability of our classification to capture, in a statistically

significant sense, whether membership of each of our classes is linked to a participant obesity status. In the regression, the relative risk ratios indicate risk of being overweight or obese compared to a healthy weight (with underweight omitted) and compared to the reference class ‘Comfortable professionals’. An adjusted model is also estimated that additionally adjusts for gender, ethnicity, health, qualifications and employment status since these are variables that map easily onto the Foresight obesity map but due to their character are not included in the classification. Both models are estimated in Stata using robust sandwich estimators, with statistical clustering on the assessment centre attended by the participant.

Results

Variable selection

Figure 1 presents results of UK Biobank variable mapping to the Foresight map, using a modified PRISMA flow chart where 21 variables are identified for inclusion in the classification and a further 20 variables used to profile the classification (see Appendix S1 for how these variables are constructed). The selection of 21 classification variables derived from UK Biobank variables are shown in Table 1.

Figure 1 : Modified PRISMA flow diagram for variable selection into Foresight informed k-means obesity classification of the UK Biobank cohort

Table 1 : Variables used for classification

UK Biobank Variable	Foresight Obesity Variable	Foresight Theme
Recreational PC use (hours) TV watching (hours) Smoking Duration (years)	1.04 Passive entertainment options 1.12 TV watching 1.16 Smoking cessation	Societal Influences
Household Size (people) Leisure and Social Activities (count) Sleep Duration (hours) Stress (count)	2.02 Face to face social interaction 2.02 Face to face social interaction 2.04 Stress 2.04 Stress	Individual Psychology
Metabolic Equivalent of Task (MET) Mean Hand Grip Strength (Kg)* Peak Expiratory Flow (litres/min)* Time Spent Outdoors in Winter (hours) Time Spent Outdoors in Summer (hours)	3.01 Physical activity 3.02 Functional fitness 3.02 Functional fitness 3.04 Level of recreational activity 3.04 Level of recreational activity	Individual Physical Activity
Vehicles per household member	4.11 Dominance of motorised transport	Physical Activity Environment
Pulse Rate (bpm)	5.02 Resting metabolic rate	Physiology
Townsend Deprivation Index (score) Length of Working Week (hours)	6.01 Purchasing power 6.06 Pressure on job performance	Food Production
Vegetable Consumption (tablespoon/day) Fruit Consumption (pieces/day) Low Fat Meat (% of meat consumption) Alcohol (units of alcohol)	7.08 Food variety 7.08 Food variety 7.08 Food variety 7.09 Alcohol consumption	Food Consumption
Age [†]	Outside Foresight	Outside Foresight

* These variables are bi-modal by gender. To correct for this each observation is standardised by the use of gender specific means and variances.

[†]Age is not included in the Foresight systems map, but something we felt important to include since other studies have highlighted different obesity outcomes by age, through changes in food consumption (Ford et al., 2019) and activity patterns (Jensen and Hsiao, 2010).

NB. Node numbers initially reported in Morris et al. (2018)

After removing those participants that are pregnant, whose weight status is impacted by the pregnancy (371), those aged younger than 40 or older than 69, which was the recruitment criteria for the cohort (2 431), and those that have missing data (120 564), there are 379 140 participants available for classification. The distribution of key demographics for both the larger sample

(excluding just those who are pregnant, younger and older) and the sub-sample used for classification is provided in Table 2 and shows that the sub-sample compares well with the full sample for most measures, with the exception of Townsend deprivation, where there is an indication that the sub-sample is less deprived and also has an under-representation for Scotland. For this classification sample, none of the variables had a pairwise correlation greater than 0.7.

Table 2 : Comparison of the characteristics of the full sample and the sub-sample used for classification

Variable	Category	Full sample	Classification sample
N		499,704	379,140
Gender	Male	45.6%	46.0%
	Female	54.4%	54.0%
Age	40-44	10.3%	10.7%
	45-49	13.2%	13.6%
	50-54	15.3%	15.5%
	55-59	18.2%	18.2%
	60-64	24.3%	24.1%
	65-69	18.7%	17.9%
Ethnicity	White	94.1%	95.0%
	Mixed	0.6%	0.6%
	Asian	2.3%	2.0%
	Black	1.6%	1.3%
	Other	0.9%	0.8%
	Not available	0.6%	0.3%
Townsend Deprivation	Lowest 20%	20.0%	20.8%
	20% to 40%	20.0%	20.8%
	40% to 60%	20.0%	20.4%
	60% to 80%	20.0%	19.9%
	Highest 20%	20.0%	18.1%
	Not available	0.1%	Not available
Assessment Region	Stockport (pilot)	0.8%	Not available
	North	41.4%	42.5%
	Midlands	15.7%	16.2%
	South	17.2%	19.0%
	London	13.7%	14.3%
	Wales	4.2%	4.1%
	Scotland	7.1%	3.8%
BMI	Underweight	0.5%	0.5%
	Healthy	32.3%	33.4%
	Overweight	42.2%	42.7%
	Obese	24.3%	23.2%
	Not available	0.6%	0.2%

The profiling variables are shown in Table 3 and cover aspects such as the socio-demographic composition (e.g. gender and ethnicity), health status and illnesses (including specific morbidities), socio-economic composition (e.g. employment and occupation), the nature of their work (tasks involved and satisfaction) and geography.

Table 3 : Variables used for profiling of the classification

UK Biobank Variable	Foresight Obesity Variable	Foresight Theme
Education	1.1 Education	Societal Influences
Ethnicity	1.5 Sociocultural valuation of food	
Job involves shift work	2.4 Stress	Individual Psychology
Work satisfaction	2.4 Stress	
Diagnosed diabetes	2.10 Use of medicines and	Individual Psychology & Physiology
Diagnosed cancer	5.12 Reliance on pharma remedies	
Diagnosed other illness		
Overall health	3.2 Functional fitness	Individual Physical Activity
Limiting illnesses	3.2 Functional fitness	
Breathless walking	3.2 Functional fitness	
Job involves heavy lifting/physical tasks	3.6 Level of occupational activity	
Job involves walking/standing	4.12 Dominance of sedentary employment	Physical Activity Environment
BMI	5.24 Level of fat free mass	Physiology
Waist/hip circumference	5.24 Level of fat free mass	
Body fat percentage	5.24 Level of fat free mass	
Gender	6.10 Female Employment	Food Production
Employment	6.15 Level of Employment	
Occupation	6.15 Level of Employment	
Assessment Centre visited	Outside Foresight : Geographical location	Outside Foresight

Cluster analysis

The scree plots in Figure 2 show that after 6 classes the reduction in the within squares sum of squares in the left hand plot plateaus out, and after 6 classes the first differences in the right hand side figure shows similar reductions as k increases, suggesting 6 as the number of useful classes in these data.

Figure 2: Scree plots of the within class sum of squares and its first difference for a range of classes

The classification centres and the number of participants in each class are shown in Table 4.

Radial plots for the classifications are available in supplementary figure S1.

Table 4 : Class centres on classification variables

	Classification Group					
	Active worker s	Retiree s with healthy lifestyle s	Stresse d and not in work	Deprive d with less healthy lifestyle s	Comfortabl e professional s	Comfortabl e families
Age (years)	52.4	62.5	60.7	60.6	53.6	47.9
Recreational PC use (hours)	1.0	1.0	0.7	1.1	1.1	1.1
TV watching (hours)	2.7	2.7	3.6	3.7	2.2	1.9
Smoking Duration (year)	8.8	3.5	1.1	33.4	5.0	3.6
Household Size (people)	2.6	2.0	2.1	2.0	1.9	3.9
Leisure and Social Activities (count)	0.9	1.5	0.7	0.8	0.9	1.0
Sleep Duration (hours)	6.9	7.4	7.2	7.3	6.9	7.1
Stress (count)	2.4	2.1	3.1	2.9	2.6	2.3
Metabolic Equivalent of task (MET)	21.4	12.7	5.9	7.7	6.5	9.0
Mean Hand Grip Strength (Kg) ¹	0.37	-0.05	-0.63	-0.27	0.20	0.51
Peak Expiratory Flow (litres/min) ¹	0.19	-0.02	-0.56	-0.48	0.31	0.59
Time Spent Outdoors in Winter (hours)	4.9	2.2	1.6	1.9	0.7	1.1
Time Spent Outdoors in Summer (hours)	7.1	4.6	3.5	4.2	2.0	2.7
Vehicles per household member	0.74	0.88	0.55	0.64	0.96	0.48
Pulse Rate (bpm)	67.6	66.7	73.0	71.3	69.1	67.6
Townsend Deprivation Index (score)	5.7	3.7	6.0	6.2	4.7	4.9
Length of Working Week (hours)	37.0	5.4	7.7	10.4	37.2	32.5
Vegetable Consumption (tablespoon/day)	5.1	5.9	4.6	4.4	4.5	4.8
Fruit Consumption (pieces/day)	2.7	4.2	3.1	2.2	2.8	2.9
Low Fat Meat (proportion meat consumption)	0.5	0.6	0.6	0.5	0.6	0.6
Alcohol (units of alcohol)	22.3	14.2	6.3	21.6	15.3	13.0
Size (n)	36405	81931	60847	50850	68024	81083

¹ These variables are bi-modal by gender. To correct for this each observation is standardised by the use of gender specific means and variances

Using the characteristics of each classification group as described in Table 4 alongside the distribution of counts in various categories of profiling variables (which are not used in the classification) as shown in supplementary tables S1 to S18 (with indicators of particularly high and low percentages relative to the percent's in the whole classification sample) a pen-portrait is constructed for each classification and presented in Box 1.

Box 1: Pen portraits for Classification groups

Active Workers

This is the youngest group, comprised largely of men in larger households, with a trend towards the more deprived Townsend decile areas. More people are in the Black ethnic group than average. The majority are employed, many spend a large portion of their working day walking, standing and involved in heavy physical activity. There is strong representation in skilled trade, managerial and process/plant and machine operative occupations and this group undertake more shift work than other groups. Time spent outdoors in both summer and winter is much higher than for other groups. They sleep less and have a longer working week than other groups. Diets are generally about average although this group has higher alcohol consumption than other groups. They are also more likely to smoke than many other groups. The proportion in this group who are overweight is high relative to other groups, and more people are obese than the average.

Retirees with healthy lifestyles

This is the oldest group, a large proportion are retired. They are predominantly White, living in the least deprived areas and their educational profile is close to the sample average. They spend a relatively large proportion of their time socialising and enjoy a good diet. Smoking and drinking are low. This group sleep well and report lower levels of stress than all other groups. Generally this group reports good or excellent health. A large proportion are overweight but prevalence of healthy weight is relatively high when compared with other groups, and obesity rates are relatively low.

Stressed and not in work

This is a relatively older group, comprised predominantly of women with higher than average representation in the Asian and Black ethnic groups. A large proportion of people are retired, looking after family or unable to work due to sickness and educational attainment is lower than most other groups. They tend to live in more deprived areas. They spend a relatively large amount of time watching TV and comparatively little time outside in both summer and winter. Diet is about average but smoking and drinking are much lower than other groups. This group report being stressed and are less satisfied with their health than most other groups, with relatively high prevalence of limiting long term illness. The proportion of people of healthy weight is relatively low compared to other groups and obesity levels are relatively high.

Deprived with less healthy lifestyles

This is a relatively older group with people living in smaller households in the more deprived areas. There are more men than women in this group. Many are retired, but also a relatively large proportion are unable to work due to sickness. Of those who do work, the proportion in managerial and professional occupations is relatively low. They report being more stressed than many other groups, and while a proportion are moderately happy with their health, more report being moderately or very unhappy when compared with other groups. This group has the highest prevalence of limiting long term illness. They spend a relatively large amount of time watching TV as well as recreational computer use. Smoking duration is very high and alcohol consumption is high relative to other groups, while diet is relatively poor with lower consumption of fruit, vegetables and lean meat. Relative to other groups, the proportion of people with a healthy weight is low while obesity rates are high.

Comfortable professionals

A relatively young group, these are people predominantly in employment within managerial, professional, associate professional and administrative occupations. They live in smaller households in less deprived areas. Educational attainment is high, with a large proportion educated to A-level or above. This group report the longest working week relative to other groups as well as lower sleep. They spend less time outside than other groups and are more likely to use a computer for recreational use. In terms of BMI, this group has a profile close to the sample average. They are also very close to average across a range of other characteristics including health, ethnicity and satisfaction with their work and health. This group is the most average in comparison to the whole UK Biobank cohort.

Comfortable families

This is the youngest group, living in larger households in less deprived areas. There are more females than males in this group and the proportion of Asian and Black ethnicity is higher than in many other groups. The majority are in paid employment, with a large proportion in managerial, professional, and associate professional occupations. Working weeks are relatively long but prevalence of stress is relatively low. Diet is fairly average compared with other groups and smoking prevalence is relatively low. Time spent watching TV is relatively low, although recreational computer use is higher than in many other groups. Prevalence of long-standing illness is low compared with other groups. Prevalence of healthy BMI is relatively high and the proportion who are obese is relatively low compared with other groups.

Table 5 identifies a distinct profile for our classes by BMI category. Recall that BMI was not used during the classification exercise, yet the use of proxies for the drivers of obesity identified in the Foresight Obesity System Map have revealed differing weight status outcomes. The classes with the greatest proportion of people with a healthy weight are the ‘Comfortable families’ and the ‘Retirees with healthy lifestyles’ classes. Conversely the class with the lowest proportion in healthy weight are the ‘Deprived with less healthy lifestyles’. The class of ‘Active workers’ have high levels in the overweight category, and ‘Stressed and not in work’ and ‘Deprived with less healthy lifestyles’ have similar high proportions in the obese category.

Table 5 : Distribution of classification by BMI category and mean

Weight status	Active workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
BMI category							
Underweight	0.2%	0.4%	0.7%	0.6%	0.5%	0.5%	0.5%
Healthy weight	26.0%	37.1%	28.9%	25.1%	34.6%	40.6%	33.4%
Overweight	47.2%	45.5%	39.9%	43.7%	41.4%	40.6%	42.7%
Obese	26.5%	16.9%	30.1%	30.3%	23.5%	18.2%	23.2%
NA	0.1%	0.1%	0.4%	0.3%	0.1%	0.1%	0.2%
BMI							
Median BMI	27.4	26.1	27.3	27.6	26.5	25.9	26.6
Mean BMI	27.8	26.6	28.1	28.2	27.3	26.5	27.3

The estimates from the multinomial regression models shown in Figure 3 identify that relative risk ratio for being overweight or obese compared to a healthy weight and being in the most typical UK Biobank class of ‘Comfortable professionals’. These relative risk ratios are different from 1.0 (excepting ‘Retirees with healthy lifestyles’) indicating that there is significant differentiation between our classes. This relative risk ratio is higher for classes: ‘Active workers’, ‘Stressed and not in work’ and ‘Deprived with less healthy lifestyles’. The ‘Comfortable families class’ have lower risk of being overweight or obese and there are no significant differences observed for the ‘Retirees with healthy lifestyles’ with respect to risk of overweight, but significantly less risk of being obese. A second model is estimated using some of the aforementioned profiling variables that are not used in the classification: gender, ethnicity, self-rated overall health, qualifications, and current employment status. In general, further adjustment with these additional profiling variables attenuates the relative risk ratios.

Figure 3 - Multinomial logistic regression models to investigate the association between our new whole system classes and weight status. The relative risk ratios indicate risk of being overweight or obese compared to a healthy weight (with underweight omitted) and compared to the reference class ‘Younger busy professionals’. Adjusted models are adjusted for gender, ethnicity, health, qualifications and employment status

Discussion

We present the development of a whole system obesity classification which can differentiate weight status in a cohort of adults in the UK. The approach to variable selection is, to the best of our knowledge, novel and has the potential to be replicated across other cohorts worldwide.

The classes we identified in this study highlight six groups of people, similar in respect to their exposure to known drivers of obesity. The ‘Comfortable professionals’ class was most typical of the UK Biobank participant characteristics, and this extended to their BMI (almost exactly to 1 decimal place), despite this not being used in the clustering process. The classification highlights three classes with increased relative risk of being overweight or obese : ‘Active workers’, ‘Stressed and not in work’ and ‘Deprived with less healthy lifestyles’ and one with significantly lower relative risk of both overweight and obesity: ‘Comfortable families’. For the ‘Retirees with healthy lifestyles’ no significant difference in relative risk of being overweight is observed, however a significantly lower risk of being obese is detected. This is reassuring that the classification appears to be sensitive to the different weight statuses.

On completion of the research the classification will be deposited back into the UK Biobank and made available to other researchers.

Strengths and limitations

This work is an important development that recognises the need for a whole systems approach to obesity research and policy making and incorporates this into the development of a classification to identify groups of individuals at highest risk of obesity - according to elements identified by the Foresight obesity systems map (Butland et al., 2007). Cohort studies like the UK Biobank study capture a rich array of information on a large number of people, encompassing many of the seven Foresight themes: Societal influences, individual psychology, individual

physical activity, physical activity environment, physiology, food production and food consumption. We found that Foresight themes relating to individual behaviours were better captured within the cohort data, with the environmental, societal and food production areas more difficult to populate. Given that the UK Biobank cohort was first established prior to the 2007 Foresight report, it is not surprising that they didn't consider collecting information from participants about their exposures to aspects of the wider obesity system, especially so since the aims of the UK Biobank cohort study are broad in respect to helping to gain a better understanding of determinants of disease, not just obesity. The Foresight obesity systems map incorporates some upstream determinants of behaviours, for example, elements of food production in addition to food consumption. (Morris et al., 2018) With regards to food consumption, the quantity of this information was challenging to incorporate into the classification. This is not so much of an issue since we were not trying to generate dietary pattern clusters, instead we have generated whole systems obesity clusters so therefore to collapse the dietary information into a small group of food consumption variables is acceptable, a process explained in supplementary appendix S1. While some dietary information is lost using this approach, it achieved our goal of condensing the dietary information such that it did not dominate the classification. It is also important to acknowledge that dietary data collected in the UK Biobank was self-reported and therefore subject to bias. However, a recent validation study of the Oxford WebQ dietary questionnaire suggests reasonable agreement against biomarkers when compared with dietary recall interviews (Greenwood et al., 2019). Other self-reported measures, such as TV watching were also used in our classification and we are not aware that these have undergone such scrutiny.

It was originally envisaged that more variables would be available to map onto the Physical Activity Environment Foresight theme from the UK Biobank Urban Morphometric Platform (Sarkar et al., 2015) bulk data, in particular data on the greenness of the participants' home environment. However, on receipt of these data we found that there were issues around the

completeness and comprehension. For the example of greenness - data were not available for participants recruited through Edinburgh, Glasgow, Oxford and Reading assessment centre and for the combined Nottingham and Birmingham centre, only available for the Nottingham based participants we included. We additionally observed some inconsistencies whereby Newcastle and Middlesbrough were intensely green compared to other urban assessment centres. Local knowledge led us to query this, but we were unable to resolve this discrepancy. We found inconsistent data within density metrics, with regards to both (i) the format of the density units where some cells present densities and others count of outlets and (ii) where counts are presented the sum of outlets within a hierarchal structure do not add correctly. We also observed missingness within the health specific destination accessibility measure. The limitations in these data diminished their utility for our purposes, where we wish to use them for all UK Biobank participants, rather than for specific assessment centres with complete data. Given the important governance restrictions around acquiring fine grained geographic data in combination with rich demographic information, we were not able to request small area geographic identifiers in our study in order to attach open data on greenness and other points of interest. Additionally, given that other UK Biobank projects were already approved to develop indicators of physical activity and food environments, we did not attempt to re-generate these for this study. This outcome does weight our consideration more towards the participant's behavioural and individual status rather than their environmental situation. A future next step could be to improve upon these indicators so that they can be better applied in whole systems obesity research in this cohort and subsequently enhance the utility of the classification presented here.

In developing the new classification we found that both age and gender were important with respect to the outcome BMI and also as a confounding factor with some of the classification variables – e.g. hand grip strength. Supplementary material in appendix S1 explains in detail how gender was used to standardise these variables. While these characteristics do not appear in the Foresight maps as features of the obesity system, they are still important to consider when

developing methods to better understand obesity. This is suggestive that perhaps the Foresight obesity system map should in fact encompass more nodes within its complex system map.

This study employed a comprehensive variable selection process for the development of the obesity classification, driven by the Foresight obesity system map. This process was rigorous and completed by three independent researchers from a range of backgrounds and agreement on variable selection reached by informed discussions. However, there are limitations to this approach, especially with respect to the omission of categorical variables, which cannot easily be used in the k-means clustering algorithm. In spite of this, k-means is a recognised and robust method of identifying clusters within a cohort like this. The quality of the key variable in this study offers advantages over other studies, for example height and weight were measured, rather than self-reported, thus reducing social desirability bias in our outcome of interest, overweight and obesity. That said, we only considered baseline BMI, as the follow up measurements in the cohort were only for subsamples. We did investigate agreement in measurement between the multiple time points and found that the baseline information was a strong indicative measure of BMI for us to test our classification against. In the same vein, we did not make use of other candidate variables which were only available for a subset of UK Biobank participants, for example the accelerometer data.

The UK Biobank cohort study offers rich and varied data with respect to numerous aspects of lifestyle and sociodemographic characteristics, making it a strong option to use for development of this novel classification. Some caution must be taken when interpreting the results in this paper given that the UK Biobank participants are not generalisable to the wider UK population (Fry et al., 2017). That said, this bias in the data does not impact upon the methods and process we present for selecting variables and development of such a classification.

Policy recommendations and future applications

The recommendation to use the Foresight obesity systems map as a framework to inform variable selection for clustering solutions can be applied to data sources available to the researchers and policy makers. Therefore, while classes derived from the clustering remain relevant to the population the data relate too, the methods for variable selection are consistent, meaning that methods can be reproduced for different data and more insightfully compared. However, in order for this to be most effectively applied we would recommend a broader representation of the data from each of the foresight themes, which may be achieved through a more targeted data collection in such cohort studies, or through collating other sources of information on the system, such as those provided by consumer data. A combination of existing data sources would present a powerful alternative to new primary data collection.

A further development of our classification, which would enrich understanding of areas and groups of people most in need of positive change, would be to incorporate a geographic identifier to the whole system classification, akin to geodemographic classifications such as the Output Area Classification from the Office for National Statistics (Gale et al., 2016) or commercial classifications like Cameo (TransUnionUK, 2017). With geographic identifiers incorporated the potential use for this type of classification would extend to better targeting of resource and support, for example they could be used by national policy makers to allocate funds to areas most in need of making system wide changes and in turn by public health directorates in local authorities to allocate resources to neighbourhoods most requiring support.

Overweight and obesity have been substantial public health challenges for some time, but in light of the recent ‘call to action’ within the UK’s National Obesity Strategy, citing COVID-19 as a wakeup call, methods to better target resources to improve health through reducing prevalence overweight and obesity, is more important than ever (Department of Health and Social Care, 2020).

Conclusion

This work presents an innovative new approach to better understanding the whole systems drivers of obesity which has the potential to produce meaningful tools for policy makers to better target interventions across the whole system to reduce overweight and obesity.

References

- ALLEN, N., SUDLOW, C., DOWNEY, P., PEAKMAN, T., DANESH, J., ELLIOTT, P., GALLACHER, J., GREEN, J., MATTHEWS, P., PELL, J., SPROSEN, T. & COLLINS, R. 2012. UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1, 123-126.
- ALLENDER, S., MILLAR, L., HOVMAND, P., BELL, C., MOODIE, M., CARTER, R., SWINBURN, B., STRUGNELL, C., LOWE, J., DE LA HAYE, K., ORELLANA, L. & MORGAN, S. 2016. Whole of Systems Trial of Prevention Strategies for Childhood Obesity: WHO STOPS Childhood Obesity. *Int J Environ Res Public Health*, 13, 1143.
- BAGNALL, A. M., RADLEY, D., JONES, R., GATELY, P., NOBLES, J., VAN DIJK, M., BLACKSHAW, J., MONTEL, S. & SAHOTA, P. 2019. Whole systems approaches to obesity and other complex public health challenges: a systematic review. *BMC Public Health*, 19, 8.
- BURGOINE, T., FOROUHI, N. G., GRIFFIN, S. J., BRAGE, S., WAREHAM, N. J. & MONSIVAIS, P. 2016. Does neighborhood fast-food outlet exposure amplify inequalities in diet and obesity? A cross-sectional study. *Am J Clin Nutr*, 103, 1540-7.
- BURGOINE, T., SARKAR, C., WEBSTER, C. J. & MONSIVAIS, P. 2018. Examining the interaction of fast-food outlet exposure and income on diet and obesity: evidence from 51,361 UK Biobank participants. *Int J Behav Nutr Phys Act*, 15, 71.
- BUTLAND, B., JEBB, S., KOPELMAN, P., MCPHERSON, K., THOMAS, S., MARDELL, J. & PARRY, V. 2007. *Tackling obesity: future choices-project report*, Department of Innovation, Universities and Skills London.
- CACI. 2021. *The Wellbeing acorn user guide* [Online]. Available: http://www.caci.co.uk/sites/default/files/resources/Wellbeing_Acorn_User_Guide.pdf [Accessed].
- CAMERON, A. J., CRAWFORD, D. A., SALMON, J., CAMPBELL, K., MCNAUGHTON, S. A., MISHRA, G. D. & BALL, K. 2011. Clustering of obesity-related risk behaviors in children and their mothers. *Annals of epidemiology*, 21, 95-102.
- DARE, S., MACKAY, D. F. & PELL, J. P. 2015. Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the UK general population. *PLoS One*, 10, e0123579.

- DEPARTMENT OF HEALTH AND SOCIALCARE. 2020. *Tackling obesity: government strategy* [Online]. Available: <https://www.gov.uk/government/publications/tackling-obesity-government-strategy> [Accessed].
- DEVAUX, M., GORYAKIN, Y., CECCHINI, M., HUBER, H. & COLOMBO, F. 2017. *OECD Obesity Update 2017* [Online]. OECD: Paris, France. Available: <https://www.oecd.org/els/health-systems/Obesity-Update-2017.pdf> [Accessed].
- FLINT, E. & CUMMINS, S. 2016. Active commuting and obesity in mid-life: cross-sectional, observational evidence from UK Biobank. *Lancet Diabetes Endocrinol*, 4, 420-35.
- FLINT, E., WEBB, E. & CUMMINS, S. 2016. Change in commute mode and body-mass index: prospective, longitudinal evidence from UK Biobank. *Lancet Public Health*, 1, e46-e55.
- FORD, N., TROTT, P. & SIMMS, C. 2019. Food portions and consumer vulnerability: qualitative insights from older consumers. *Qualitative Market Research: An International Journal*.
- FRY, A., LITTLEJOHNS, T. J., SUDLOW, C., DOHERTY, N., ADAMSKA, L., SPROSEN, T., COLLINS, R. & ALLEN, N. E. 2017. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*, 186, 1026-1034.
- GALE, C. G., SINGLETON, A. D., BATES, A. G. & LONGLEY, P. A. 2016. Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*, 1-27.
- GOVERNMENT OFFICE FOR SCIENCE. 2007. *Reducing obesity: future choices* [Online]. Available: <https://www.gov.uk/government/publications/reducing-obesity-future-choices> [Accessed].
- GREEN, M. A., STRONG, M., RAZAK, F., SUBRAMANIAN, S. V., RELTON, C. & BISSELL, P. 2016. Who are the obese? A cluster analysis exploring subgroups of the obese. *J Public Health (Oxf)*, 38, 258-64.
- GREENWOOD, D., CADE, J., DRAPER, A., BARRETT, J., CALVERT, C. & GREENHALGH, A. 2000. Seven unique food consumption patterns identified among women in the UK Women's Cohort Study. *European journal of clinical nutrition*, 54, 314-320.
- GREENWOOD, D. C., HARDIE, L. J., FROST, G. S., ALWAN, N. A., BRADBURY, K. E., CARTER, M., ELLIOTT, P., EVANS, C. E., FORD, H. E. & HANCOCK, N. 2019. Validation of the Oxford WebQ online 24-hour dietary questionnaire using biomarkers. *American journal of epidemiology*, 188, 1858-1867.
- HANSCOMBE, K. B., COLEMAN, J. R. I., TRAYLOR, M. & LEWIS, C. M. 2019. ukbtools: An R package to manage and query UK Biobank data. *PLoS One*, 14, e0214311.
- JEBB, S., KOPELMAN, P. & BUTLAND, B. 2007. Executive summary: foresight 'tackling obesities: future choices' project. *Obesity Reviews*, 8, vi-ix.

- JENSEN, G. L. & HSIAO, P. Y. 2010. Obesity in older adults: relationship to functional limitation. *Curr Opin Clin Nutr Metab Care*, 13, 46-51.
- JOHNSTON, L. M., MATTESON, C. L. & FINEGOOD, D. T. 2014. Systems science and obesity policy: a novel framework for analyzing and rethinking population-level planning. *Am J Public Health*, 104, 1270-8.
- KANT, A. K. 2004. Dietary patterns and health outcomes. *J Am Diet Assoc*, 104, 615-35.
- LAKE, A. A., HENDERSON, E. J. & TOWNSHEND, T. G. 2017. Exploring planners' and public health practitioners' views on addressing obesity: lessons from local government in England. *Cities & health*, 1, 185-193.
- LEE, B. Y., BARTSCH, S. M., MUI, Y., HAIDARI, L. A., SPIKER, M. L. & GITTELSON, J. 2017. A systems approach to obesity. *Nutr Rev*, 75, 94-106.
- LOCAL GOVERNMENT ASSOCIATION 2017. Making obesity everybody's business: a whole systems approach to obesity. *Local Government Association*.
- LYALL, D. M., CELIS-MORALES, C., WARD, J., ILIOTROMITI, S., ANDERSON, J. J., GILL, J. M. R., SMITH, D. J., NTUK, U. E., MACKAY, D. F., HOLMES, M. V., SATTAR, N. & PELL, J. P. 2017. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. *JAMA Cardiol*, 2, 882-889.
- MASON, K. E., PALLA, L., PEARCE, N., PHELAN, J. & CUMMINS, S. 2020. Genetic risk of obesity as a modifier of associations between neighbourhood environment and body mass index: an observational study of 335 046 UK Biobank participants. *BMJ Nutrition, Prevention & Health*.
- MASON, K. E., PEARCE, N. & CUMMINS, S. 2018. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *The Lancet Public Health*, 3, e24-e33.
- MEDICAL RESEARCH COUNCIL. 2020a. *Nutrition and Obesity* [Online]. Available: <https://mrc.ukri.org/funding/science-areas/population-systems-medicine/our-science-and-contacts-psmb/nutrition-and-obesity/> [Accessed].
- MEDICAL RESEARCH COUNCIL. 2020b. *Obesity research priorities* [Online]. Available: <https://mrc.ukri.org/research/initiatives/obesity-research/obesity-research-priorities/> [Accessed].
- MILES, I. & KEENAN, M. Ten Years of Foresight in the UK. documento presentado en la segunda conferencia internacional sobre prospectiva tecnológica, Tokio, 2003.
- MORRIS, M. A., WILKINS, E., TIMMINS, K. A., BRYANT, M., BIRKIN, M. & GRIFFITHS, C. 2018. Can big data solve a big problem? Reporting the obesity data landscape in line with the Foresight obesity system map. *International Journal of Obesity*, 42, 1963-1976.

- MOSTELLER, F. & TUKEY, J. W. 1977. *Data analysis and regression: a second course in statistics*.
- NOTTINGHAM CITY COUCL, NHS NOTTINGHAM CITY & NOTTINGHAM1. 2011. *Healthy Weight Strategy for Nottingham City 2011-2020* [Online]. Available: https://www.nottinghamcity.nhs.uk/media/3850/child_obesity_strategy.pdf [Accessed].
- NTUK, U. E., GILL, J. M., MACKAY, D. F., SATTAR, N. & PELL, J. P. 2014. Ethnic-specific obesity cutoffs for diabetes risk: cross-sectional study of 490,288 UK biobank participants. *Diabetes Care*, 37, 2500-7.
- PATTERSON, R. E., HAINES, P. S. & POPKIN, B. M. 1994. Health lifestyle patterns of US adults. *Preventive medicine*, 23, 453-460.
- PETERSEN, J., GIBIN, M., LONGLEY, P., MATEOS, P., ATKINSON, P. & ASHBY, D. 2011. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems*, 13, 173-192.
- SARKAR, C. 2017. Residential greenness and adiposity: Findings from the UK Biobank. *Environ Int*, 106, 1-10.
- SARKAR, C., WEBSTER, C. & GALLACHER, J. 2017. Association between adiposity outcomes and residential density: a full-data, cross-sectional analysis of 419 562 UK Biobank adult participants. *The Lancet Planetary Health*, 1, e277-e288.
- SOUTH EAST PUBLIC HEALTH OBSERVATORY. 2011. *Cardiovascular disease PCT health profile: Wakefield District* [Online]. Available: <http://www.wakefieldjsna.co.uk/site/wp-content/uploads/CVD-Profile-2011.pdf> [Accessed].
- SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J., LANDRAY, M., LIU, B., MATTHEWS, P., ONG, G., PELL, J., SILMAN, A., YOUNG, A., SPROSEN, T., PEAKMAN, T. & COLLINS, R. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, 12, e1001779.
- SUNDERLAND CITY COUNCIL. 2017. *Health Weight and Excess Weight* [Online]. Available: https://www.sunderland.gov.uk/media/20675/ALS-2017-Profile-Healthy-Weight/pdf/ALS_2017_Profile_-_Healthy_Weight.pdf?m=636746790855830000 [Accessed].
- SWINBURN, B., EGGER, G. & RAZA, F. 1999. Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Prev Med*, 29, 563-70.
- TRANSUNIONUK. 2017. *CAMEO UK* [Online]. Available: <https://www.transunion.co.uk/product/cameo> [Accessed].
- TYRRELL, J., JONES, S. E., BEAUMONT, R., ASTLEY, C. M., LOVELL, R., YAGHOOTKAR, H., TUKE, M., RUTH, K. S., FREATHY, R. M., HIRSCHHORN, J. N., WOOD, A. R., MURRAY, A., WEEDON, M. N.

- & FRAYLING, T. M. 2016. Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *BMJ*, 352, i582.
- ULIJASZEK, S. 2015. With the benefit of Foresight: Obesity, complexity and joined-up government. *BioSocieties*, 10, 213-228.
- WORLD HEALTH ORGANISATION. 2020. *Obesity and Overweight* [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> [Accessed].
- WORLD OBESITY. 2020. *Prevelence of Obesity* [Online]. Available: <https://www.worldobesity.org/about/about-obesity/prevalence-of-obesity> [Accessed].
- WYSE, C. A., CELIS MORALES, C. A., GRAHAM, N., FAN, Y., WARD, J., CURTIS, A. M., MACKAY, D., SMITH, D. J., BAILEY, M. E. S., BIELLO, S., GILL, J. M. R. & PELL, J. P. 2017. Adverse metabolic and mental health outcomes associated with shiftwork in a population-based study of 277,168 workers in UK biobank[.] *Ann Med*, 49, 411-420.

Classification radar plots

The following charts show how distinctive each of the derived classification is.

There is one chart for each classification of participants (making 6 in total). Each classification variable is aggregated to a mean for all participants that are allocated to the classification. Where this mean is positioned on a normalised scale (with a mean of zero and unit variance) defined by all participant's data is shown on the chart. Thus a value greater than 0.0 means that the participants in that class tend to have a higher value for this variable than in the classification sample as a whole. The greater the value, the more distinctive. A value less than 0.0 means that the participants tend to have a lower value for this variable. The variables are sorted, clockwise, from the greatest positive standardised difference to the greatest negative.

Active workers

Retirees with healthy lifestyles

Stressed and not in work

Deprived with less healthy lifestyles

Comfortable professionals

Comfortable families

These tables provide percentage counts on how participants in each of the derived classes profiles against a number of key UK Biobank categorical variables. Where the percentage is appreciably more than that for the overall Classification sample it is highlighted in **bold** and where it is lower it is highlighted in **red**.

Table S1 : Distribution of classification by gender (N = 379140)

Gender	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Female	26.1%	57.3%	70.2%	44.1%	54.9%	56.5%	54.0%
Male	73.9%	42.7%	29.8%	55.9%	45.1%	43.5%	46.0%

Table S2 : Distribution of classification by ethnicity (N = 379140)

Ethnicity	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
White (incl. British!)	93.7%	97.9%	91.9%	97.2%	96.8%	92.2%	95.0%
Mixed	0.8%	0.2%	0.5%	0.6%	0.5%	0.9%	0.6%
Asian (incl. Chinese)	1.7%	0.8%	3.9%	0.7%	1.1%	3.4%	2.0%
Black	2.6%	0.4%	2.1%	0.7%	0.8%	2.0%	1.3%
Other	0.9%	0.4%	1.3%	0.6%	0.5%	1.2%	0.8%
PNA/DK	0.3%	0.3%	0.3%	0.3%	0.2%	0.3%	0.3%

Table S3: Distribution of classification by BMI category and mean (N = 379140)

BMI	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Underweight	0.2%	0.4%	0.7%	0.6%	0.5%	0.5%	0.5%
Healthy	26.0%	37.1%	28.9%	25.1%	34.6%	40.6%	33.4%
Overweight	47.2%	45.5%	39.9%	43.7%	41.4%	40.6%	42.7%
Obese	26.5%	16.9%	30.1%	30.3%	23.5%	18.2%	23.2%
NA	0.1%	0.1%	0.4%	0.3%	0.1%	0.1%	0.2%
Median BMI	27.4	26.1	27.3	27.6	26.5	25.9	26.6
Mean BMI	27.8	26.6	28.1	28.2	27.3	26.5	27.3

Table S4 : Distribution of classification by self-rated health (N = 379140)

[illegible]

Table S8 : Distribution of classification by doctor diagnosed other illness (N = 379140)

Other illness	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
None	85.5%	79.6%	70.9%	71.9%	82.0%	86.2%	79.6%
Some	13.4%	19.2%	26.6%	25.8%	16.8%	12.7%	18.9%
PNA/DK	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
NA	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table S9 : Distribution of classification by experience of shortness of breath (N = 379140)

Shortness in breath on walking on level ground	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
No	33.5%	35.3%	31.7%	28.7%	30.5%	33.1%	32.3%
Yes	2.3%	1.8%	6.7%	6.6%	2.6%	2.0%	3.5%
PNA/DK	0.6%	0.2%	1.2%	0.9%	0.5%	0.5%	0.6%
NA	63.7%	62.6%	60.5%	63.8%	66.4%	64.3%	63.6%

Table S10: Distribution of classification by employment status (N = 379140)

[illegible]

Table S11 : Distribution of classification by occupation (N = 379140)

Job/Occupation code	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Managers and senior officials	13.3%	5.6%	3.9%	6.6%	20.8%	17.6%	11.5%
Professional occupations	9.8%	11.3%	7.6%	6.9%	26.5%	24.7%	15.6%
Associate professional and technical occupations	12.2%	6.9%	6.0%	6.4%	16.1%	17.0%	11.0%
Administrative and secretarial occupations	5.6%	6.8%	10.7%	7.7%	15.0%	12.4%	10.1%
Skilled trades occupations	19.3%	2.3%	1.9%	4.0%	3.9%	3.8%	4.7%
Personal service occupations	5.7%	2.2%	3.9%	3.3%	2.9%	5.1%	3.7%
Sales and customer service occupations	2.6%	1.3%	3.1%	2.4%	1.9%	2.2%	2.2%
Process, plant and machine operatives	11.1%	1.1%	1.6%	3.1%	2.1%	1.8%	2.7%
Elementary occupations	8.1%	1.3%	3.6%	3.7%	1.6%	2.3%	2.9%
Other	4.1%	2.7%	2.5%	2.6%	4.4%	3.8%	3.3%
NA	8.1%	58.6%	55.2%	53.3%	4.9%	9.2%	32.3%

Table S12 : Distribution of classification by job involves walking or standing (N = 224599)

Walking or standing	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Never/rarely	13.8%	30.4%	32.1%	30.3%	46.5%	41.6%	36.4%
Sometimes	26.1%	32.4%	29.0%	29.2%	32.3%	31.8%	30.7%
Usually	22.0%	17.3%	14.6%	15.4%	10.9%	13.2%	14.5%
Always	38.0%	19.8%	24.2%	25.1%	10.2%	13.4%	18.4%
PNA/DK	0.1%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%
NA							

Table S13 : Distribution of classification by job involves heavy lifting or physical tasks (N = 224599)

	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Heavy or physical							
Never/rarely	32.0%	67.5%	64.6%	58.7%	77.9%	73.4%	66.3%
Sometimes	30.8%	21.7%	22.9%	25.7%	16.3%	18.6%	20.9%
Usually	17.2%	6.1%	6.1%	8.0%	3.5%	4.4%	6.6%
Always	19.9%	4.6%	6.2%	7.6%	2.4%	3.6%	6.2%
PNA/DK	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%
NA							

Table S14 : Distribution of classification by job involves shift work (N = 224599)

	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Shift work							
Never/rarely	70.7%	88.1%	83.1%	79.7%	87.7%	85.4%	83.5%
Sometimes	12.0%	5.7%	6.1%	7.8%	6.0%	6.5%	7.2%
Usually	3.3%	1.8%	2.6%	2.6%	1.4%	1.7%	2.0%
Always	14.0%	4.4%	8.0%	9.8%	4.8%	6.3%	7.2%
PNA/DK	0.1%	0.1%	0.2%	0.1%	0.1%	0.1%	0.1%
NA							

Table S15 : Distribution of classification by educational qualification (N = 379140)

	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Education							
College or University degree	19.9%	34.7%	22.0%	18.4%	45.9%	48.7%	34.1%
A levels/ AS levels or equivalent	8.9%	11.2%	9.6%	9.0%	14.0%	14.6%	11.7%
O levels/ GCSEs or equivalent	24.3%	22.2%	23.0%	22.5%	20.4%	20.0%	21.8%
CSEs or equivalent	12.5%	2.7%	4.9%	4.8%	4.7%	6.6%	5.5%
NVQ or HND or HNC or equivalent	11.4%	6.5%	5.8%	8.6%	5.6%	4.1%	6.5%
Other professional qualifications eg: nursing, teaching	4.6%	7.3%	5.9%	5.5%	4.3%	2.9%	5.1%
PNA/DK	18.5%	15.4%	28.7%	31.1%	5.1%	3.1%	15.5%
NA	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table S16 : Distribution of classification by work satisfaction (N = 138078)

	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Work satisfaction							
Extremely happy	9.6%	11.5%	7.5%	9.0%	7.1%	7.0%	8.3%
Very happy	34.1%	48.2%	35.7%	34.9%	32.2%	33.4%	35.9%
Moderately happy	45.5%	36.8%	47.0%	45.4%	47.5%	47.9%	45.5%
Moderately unhappy	7.2%	2.0%	5.8%	6.7%	9.7%	8.6%	7.1%
Very unhappy	2.1%	0.5%	1.5%	1.8%	2.2%	2.1%	1.8%
Extremely unhappy	1.1%	0.2%	0.8%	1.0%	1.1%	0.8%	0.8%
I am not employed							
PNA/DK	0.3%	0.7%	1.8%	1.2%	0.2%	0.3%	0.6%
NA							

Table S17 : Distribution of classification by health satisfaction (N = 138078)

Health satisfaction	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Extremely happy	6.9%	7.0%	3.2%	3.2%	4.7%	5.8%	5.2%
Very happy	35.3%	41.3%	24.5%	23.1%	33.3%	36.4%	33.0%
Moderately happy	48.1%	45.7%	52.4%	54.4%	49.2%	46.5%	49.0%
Moderately unhappy	7.4%	4.8%	12.4%	13.0%	10.0%	8.6%	9.1%
Very unhappy	1.5%	0.8%	4.6%	4.1%	2.1%	1.8%	2.4%
Extremely unhappy	0.6%	0.3%	2.1%	1.7%	0.6%	0.6%	0.9%
PNA/DK	0.2%	0.2%	0.8%	0.5%	0.2%	0.3%	0.4%
NA							

Table S18 : Distribution of classification by assessment centre (N = 379140)

Location of assessment centre	Active Workers	Retirees with healthy lifestyles	Stressed and not in work	Deprived with less healthy lifestyles	Comfortable professionals	Comfortable families	Classification sample
Newcastle	7.7%	6.4%	8.3%	8.0%	6.5%	6.0%	7.0%
Middlesbrough	5.6%	5.1%	4.6%	5.1%	4.0%	3.7%	4.5%
Manchester	2.9%	2.5%	2.1%	2.9%	3.2%	3.4%	2.8%
Liverpool	7.2%	7.2%	7.8%	8.3%	5.8%	5.9%	6.9%
Bury	6.1%	5.8%	5.6%	6.7%	5.8%	4.8%	5.7%
Leeds	9.0%	9.2%	8.8%	9.5%	9.5%	8.8%	9.1%
Sheffield	6.9%	7.2%	6.4%	6.5%	6.0%	5.6%	6.4%
Nottingham	7.5%	8.3%	7.4%	7.4%	7.4%	6.5%	7.4%
Stoke	4.5%	3.7%	3.5%	3.5%	3.7%	3.0%	3.6%
Birmingham	4.6%	4.6%	6.7%	5.8%	5.0%	5.1%	5.3%
Oxford	2.6%	3.0%	2.0%	1.9%	3.6%	3.7%	2.9%
Bristol	9.4%	10.4%	7.9%	7.6%	10.1%	10.6%	9.5%
Reading	4.9%	8.1%	4.4%	4.3%	8.4%	7.3%	6.6%
London (Barts)	2.8%	0.9%	3.2%	3.3%	2.5%	3.5%	2.6%
Hounslow	5.5%	5.1%	7.4%	5.6%	5.7%	7.1%	6.1%
Croydon	5.1%	5.0%	6.3%	5.3%	5.0%	6.5%	5.6%
Cardiff	3.9%	3.6%	3.4%	3.4%	3.8%	3.7%	3.6%
Swansea	0.4%	0.4%	0.5%	0.5%	0.3%	0.2%	0.4%
Wrexham	0.2%	0.1%	0.1%	0.2%	0.1%	0.1%	0.1%
Glasgow	2.1%	1.8%	2.1%	2.8%	2.1%	2.3%	2.2%
Edinburgh	1.2%	1.8%	1.2%	1.6%	1.7%	2.1%	1.7%

This note provides pseudo code to illustrate how each classification variable was constructed from a range of UK Biobank variables. The number after the character 'F' denotes the Data-Field number used in the UK Biobank Data showcase (<https://biobank.ndph.ox.ac.uk/ukb/>). The following _X_Y code denotes the baseline data (X=0) and the instance (Y).

Age

```
date_of_attending_assessment_centre_f53_0_0-  
function(year_of_birth_f34_0_0,month_of_birth_f52_0_0)
```

Sedentary activity PC

```
time_spent_using_computer_f1080_0_0
```

Sedentary activity TV

```
time_spent_watching_television_tv_f1070_0_0
```

Smoking years

Never smokers

0

Former smokers

```
age_stopped_smoking_f2897_0_0-age_started_smoking_in_former_smokers_f2867_0_0
```

Current smokers

```
age_0_0-age_started_smoking_in_current_smokers_f3436_0_0
```

Household size

When type_of_accommodation_lived_in_f670_0_0 is NOT Sheltered accommodation or Care home then this is the number_in_household_f709_0_0

Social activities

Count of:

```
leisuresocial_activities_f6160_0_?
```

Sports club or gym

Pub or social club

Religious group

Adult education class

Other group activity

Sleep duration

```
sleep_duration_f1160_0_0
```

Stresses

Count of:

```
sleeplessness_insomnia_f1200_0_0 +
```

```
worrier_anxious_feelings_f1980_0_0 +
```

```
loneliness_isolation_f2020_0_0 +
```

```
seen_doctor_gp_for_nerves_anxiety_tension_or_depression_f2090_0_0 +
```

```
seen_a_psychiatrist_for_nerves_anxiety_tension_or_depression_f2100_0_0 +
```

```
illness_injury_bereavement_stress_in_last_2_years_f6145_0_?
```

Serious illness, injury or assault to yourself

Serious illness, injury or assault of a close relative

Death of a close relative

Death of a spouse or partner

Marital separation/divorce

Financial difficulties

MET activity levels

```
(duration_of_walks_f874_0_0*3.3+
```

```
duration_of_moderate_activity_f894_0_0*4.0+
```

```
duration_of_vigorous_activity_f914_0_0*8.0)/60
```

Hand grip

```
(my_ukb_data$hand_grip_left_0_0+my_ukb_data$hand_grip_right_0_0)/2
```

Peak expiratory flow

```
Ignore peak_expiratory_flow_pef_f3064_0_? less than 25
```

```
(peak_expiratory_flow_pef_f3064_0_1+  
peak_expiratory_flow_pef_f3064_0_2+  
peak_expiratory_flow_pef_f3064_0_3)/3
```

Outdoor activity in winter

```
time_spent_outdoors_in_winter_f1060_0_0
```

Outdoor activity in summer

```
time_spend_outdoors_in_summer_f1050_0_0
```

Vehicles per household member

```
number_of_vehicles_in_household_f728_0_0/number_in_household_f709_0_0
```

Pulse

```
(pulse_rate_automated_reading_f102_0_0+pulse_rate_automated_reading_f102_0_1)/2
```

Townsend

```
# Make wholly positive
```

```
townsend_deprivation_index_at_recruitment_f189_0_0+6.5
```

Working week

```
length_of_working_week_for_main_job_f767_0_0
```

Healthy foods vegetables

```
cooked_vegetable_intake_f1289_0_0+salad_raw_vegetable_intake_f1299_0_0
```

Healthy foods fruit

```
fresh_fruit_intake_f1309_0_0+dried_fruit_intake_f1319_0_0
```

Lean meats

```
Convert categories of consumption to daily
```

Category	Annual amount
Never	0
Less than once per week	18
Once a week	52
2-4 times a week	156
5-6 times a week	286
Daily	365

```
Lean_meat_0_0<-(oily_fish_0_0+fish_0_0+poultry0_0)/  
(oily_fish_0_0+my_ukb_data$fish_LIDA_0_0+poultry_0_0+  
process_meat_0_0+beef_0_0+lambmutton_0_0+pork_0_0)
```

Alcohol

```
average_weekly_red_wine_intake_f1568_0_0*1.5+
```

```
average_weekly_champagne_plus_white_wine_intake_f1578_0_0*1.5+
```

```
average_weekly_beer_plus_cider_intake_f1588_0_0*2.9+
```

```
average_weekly_spirits_intake_f1598_0_0*1.0+
```

```
average_weekly_fortified_wine_intake_f1608_0_0*1.25
```