## Supplementary information S1: Predictions of BMI over time

BMI prevalence is categorised as described in Table 1.

Table 1. Description of the categories used for BMI prevalence

|  |  |  |
| --- | --- | --- |
| Risk factor (RF) | Number of categories (K) | Categories |
| BMI | 4 | 1. BMI < 25 kg m-2 (normal weight)
2. BMI from 25 to 29.99 kg m-2 (overweight)
3. BMI ≥ 30 kg m-2 (obesity class I & class II)
4. BMI ≥ 40 kg m-2 (obesity class III)
 |

Letbe the number of categories for BMI, e.g. *K* = 4 in this paper. Let $k $= 1, 2, …,**number these categories and $p\_{k}(t)$ denote the prevalence of individuals with BMI values that correspond to the category $k$ at time *t*. We estimate $p\_{k}(t)$ in two steps. In the first step, we add the third and fourth BMI categories (i.e. obesity class I, II and III)) together and treat them as one category. The total number of categories we need to predict is then reduced to *K* = 3. The new BMI distribution is denoted by {, , }. In the second step, we separate  back into two categories and predict their trends again. The main reason of dividing the projection into two steps is that the proportion of morbid obesity of a particular year can be closed to zero. The regression algorithm cannot handle this directly and this can lead to inaccurate estimation. By dividing the projection into two steps, we effectively scale up the proportion of morbid obesity to avoid the problem and scale the results back after the regression.

We estimate  using multinomial logistic regression model with time *t* as a single explanatory variable. In the first step, for $k<3$, we have

 

The prevalence of the third category,, is obtained by using the normalisation constraint  Solving equation for  $,$we obtain

 

which is subjected to all constraints on the prevalence values, i.e. normalisation and [0, 1] bounds.

In the second step, we repeat the process above to break  into two categories and predict the trend of  and . In this step, *K* = 2. For $k<2$, we have

 

where

 

 

And

 

The same method applied to the first step is repeated for the second step with  and .

### Multinomial logistic regression

Measured data is extracted from the survey data set. They consist of sets of probabilities with their variances. Each set represents the probabilities of individuals of normal weight, overweight, obesity class I & class II and obesity class III at specific time values (i.e., the year of the survey). For any particular time the sum of these probabilities is unity.

Each data point is treated as a normally distributed random variable; together they are a set of *N* groups (number of years) of *K* probabilities {{*t*i, μki, σki |*k*∈[1,*K*]} | *i*∈[1,*N*}, where ,, denote the year of the survey, the mean probability of  BMI category of the year and its variance respectively.

The regression consists of fitting a set of logistic functions {*p*k(**a**, **b***, t*)|*k*∈[1,*K*]} to these data – one function for each *k*-value. At each time value the sum of these functions is unity. Thus, for example, when measuring obesity in the four states, the *k*= 1 regression function represents the probability of being normal weight over time, *k*= 2 the probability of being overweight, *k*= 3 the probability of being of obesity class I & class II and  the probability of being of obesity class III.

The regression equations are derived from a least square minimization. In the following equation set the weighted difference between the measured and predicted probabilities is written as *S*; the logistic regression functions *pk*(**a**,**b***;t*) are chosen to be ratios of sums of exponentials (This is equivalent to modelling the log probability ratios,  as linear functions of time.)

 

 

The parameters *A*0, *a*0 and *b*0 are all zero and are used merely to preserve the symmetry of the expressions and their manipulation. For a *K*-dimensional set of probabilities there will be 2(*K*-1) regression parameters to be determined due to the normalisation constraint.

The minimum of the function *S* is determined from the equations

 

noting the relations

 

The values of the vectors **a**, **b** that satisfy these equations are denoted respectively. They provide the trend lines, , for the probabilities of each BMI category. The confidence intervals for the trend lines are derived most easily from the underlying Bayesian analysis of the problem.

### Bayesian interpretation

The 2*K*-2 regression parameters {**a,b**} are regarded as random variables whose posterior distribution is proportional to the function exp(-*S*(**a**,**b**)). The maximum likelihood estimate of this probability distribution function, the minimum of the function S, is obtained at the values . Other properties of the (2*K*-2)-dimensional probability distribution function are obtained by first approximating it as a (2*K*-2)-dimensional normal distribution whose mean is the maximum likelihood estimate. This amounts to expanding the function *S*(**a**,**b**) in a Taylor series as far as terms quadratic in the differences  about the maximum likelihood estimate . Hence

 

The (2*K*-2)-dimensional covariance matrix *P* is the inverse of the appropriate expansion coefficients. This matrix is central to the construction of the confidence limits for the trend lines.

### Estimation of the confidence intervals

The logistic regression functions *p*k(*t*) can be approximated as a normally distributed time-varying random variable  by expanding *p*k about its maximum likelihood estimate (the trend line) 

 

Denoting mean values by angled brackets, the variance of *p*k is thereby approximated as

 

When *K*=3 this equation can be written as the 4-dimensional inner product

 

where . The 95% confidence interval for *p*k(*t*) is centred given as .