A Model of Individual BMI Trajectories

Laurens Bogaardt, Anoukh van Giessen, H. Susan J. Picavet and Hendriek C. Boshuizen

National Institute for Public Health and the Environment (RIVM),
Antonie van Leeuwenhoeklaan 9, 3721MA Bilthoven, The Netherlands.

A risk factor model of BMI is an important building block of health simulations aimed at estimating government policy effects with regard to overweight and obesity. We created a model which generates representative population level distributions and which also mimics realistic BMI trajectories at an individual level so that policies aimed at individuals can be simulated. The model is constructed by combining several datasets. First, the population level distribution is extracted from a large, cross-sectional dataset. The trend in this distribution is estimated from historical data. In addition, longitudinal data are used to model how individuals move along typical trajectories over time. The model faithfully describes the population level distribution of BMI, stratified by sex, level of education and age. It is able to generate life course trajectories for individuals which seem plausible, but it does not capture extreme fluctuations, such as rapid weight loss.

Keywords: micro-simulation, risk factor, BMI, life course, individual trajectories

1. Introduction

Overweight and obesity pose significant health risks in many countries (Dai et al. 2020). The link between a high body mass index (BMI) and various health outcomes, including diabetes and cardiovascular disease, has been well established (GBD 2019 Risk Factors Collaborators 2020). Consequently, governments spend much effort to curb the increasing trends (Ministerie van Volksgezondheid, Welzijn en Sport 2018, Van Rinsum et al. 2018). Investigating what course of action is most fruitful is often done using health simulations (Levy et al. 2011). A risk factor model of BMI is the first building block of such analyses.

In a health simulation, the evolution of a population is simulated and the outcome of a baseline scenario is compared to the outcome of an alternative scenario in which some preventive intervention has taken place. It usually starts with a population described by means of a set of characteristics like sex, age and some risk factor, such as BMI. Relevant diseases are also included. The population's characteristics are then evolved forward in time where transitions are often encoded by a Markov process, meaning each transition only depends on the current state of the population (Sonnenberg & Beck 1993).

One strategy to define a population is to first divide each variable into discrete categories. For BMI, the categorisation of underweight, normal weight, overweight and obese is typical. For all combinations of the categorised variables, the initial population's prevalence is determined and this list of values is updated each time step. This is called a macro-simulation. When the population is subdivided on multiple characteristics, such as the status of various diseases, the total number of groups grows exponentially.

At some point, this leads to computational limitations and it pays to switch to a micro-simulation framework. Here, individuals are sampled from an initial distribution and their set of characteristics is updated each time step, describing a unique life course for each individual (Levy et al. 2011).

Unlike in a macro-simulation, the values of these life course trajectories in a micro-simulation can be measured on a continuous scale. One commonly used method to ensure all individuals together adhere to the population level distribution is to assign to each individual a percentile score indicating where they lie within the distribution. It is then assumed that this relative position stays fixed, even though the shape of the distribution, and the corresponding value, may change as the individual ages throughout the simulation (McPherson et al. 2007, OECD 2019). A more realistic model can have the relative positions fluctuate over time as well (Vuik & Cecchini 2021). This is the approach we take to model BMI and figure 1 shows a comparison between the resulting life course trajectories.
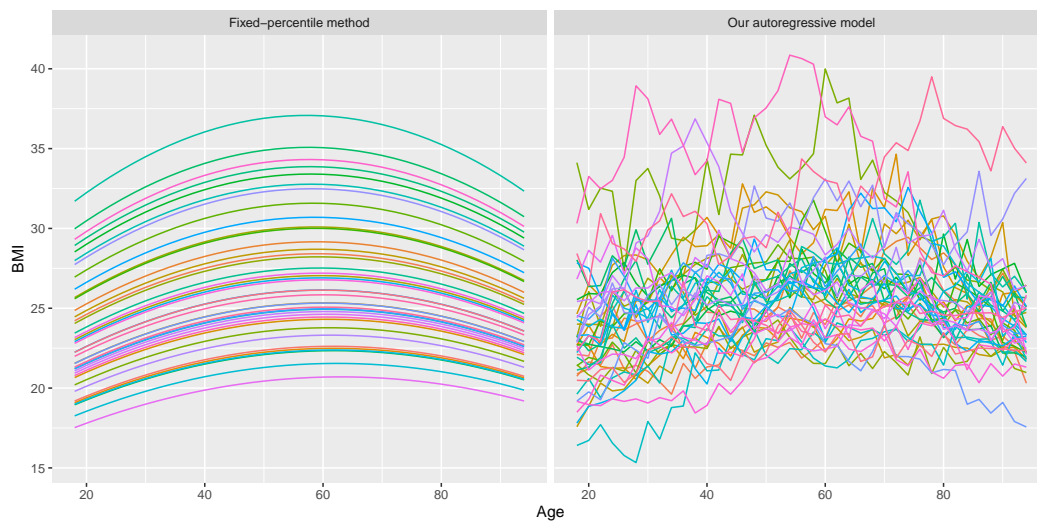

FIG. 1. A sample of trajectories following the fixed-percentile method compared with those generated by our model.

In this article, we create a risk factor model of BMI for micro-simulations which is able to faithfully describe the population level distribution, stratified by sex, level of education and age. It can also predict future trends in obesity as well as produce unique life course trajectories for individuals which seem plausible, but it does not capture extreme fluctuations, such as rapid weight loss. It is fit on the adult population of the Netherlands with the purpose of facilitating simulations concerning government policy with regard to overweight and obesity (Ten Dam et al. in press). Although our model serves as a starting point for such analyses by describing the current dynamics of BMI, it makes no claims about the effect of any particular preventive intervention. Our method is simple, tractable and a natural generalisation of the fixed-percentile method. The generic nature means it can be applied to data from various countries and need not be limited to BMI.

To arrive at a realistic and useful BMI model, we would ideally use a large, representative, longitudinal dataset. Unfortunately, longitudinal panels often only include a small number of individuals and most Dutch cohorts are not representative of the population at large. To overcome these limitations, we can cleverly combine a large, cross-sectional dataset with historical- and longitudinal data. The cross-sectional dataset provides representative information about the population level distribution of BMI, described in section 2. In section 3, we use additional historical data to assess the trend in this distribu-

tion. At an individual level, BMI fluctuates over time, following typical trajectories. These trajectories are modelled using the longitudinal data, described in section 4. Finally, in section 5, we will discuss some limitations of our model and conclude by suggesting potential future improvements in section 6.

2. Population Level Distribution

2.1 *Approach*

We will build up our model of BMI starting with the population level distribution. BMI is defined as body weight divided by the square of height and typical values range between $15\,kg/m^2$ and $45\,kg/m^2$. At the population level, its distribution shows positive skew, meaning that it is best modelled by a probability density function flexible enough to deal with non-normal data (Majer et al. 2013). Although we are modelling the generative process behind BMI, there is no need for a biological interpretation of this process, so any flexible probability density function will do. To fit such a distribution on the adult population of the Netherlands requires a cross-sectional dataset which is representative of the Netherlands. In addition, it must be large enough to accommodate various stratifications. For government policy analyses, stratifying the population by sex, level of education and age is usually adequate. We will develop completely separate models for both sexes, as we believe the differences in biological and environmental influences for males and females warrant distinct models.

2.2 *Data*

The Public Health Monitor dataset is a Dutch cross-sectional dataset based on a large, health-related questionnaire administered by the Community Health Services, Statistics Netherlands and the National Institute for Public Health and the Environment (GGD'en et al. 2012). The dataset contains self-reported weight and height measurements, rounded to nearest kilogram and centimetre. The included survey weights allow for representative analyses of the Netherlands in 2012. The questionnaire was repeated in 2016 and in 2020. However, the data for 2020 are likely to be affected by the COVID pandemic, which is not an effect we would like to include in our model. Furthermore, the data for 2016 will be reserved for external validation, described in the online supplementary information (Bogaardt et al. 2023). This leads to our choice of fitting the model on the 2012 data.

Table 1 provides some descriptive statistics of the variables used which excludes some of the original data, removed in the cleaning process. This is also described in the online supplementary information. Education was measured as the highest level reached and categorised into three levels. The lowest level applies to people with intermediate secondary education or less, the medium level aggregates higher secondary- and intermediate vocational education and highest level includes higher vocational education and university.

TABLE 1 *Descriptive statistics of the Public Health Monitor dataset.*

	Number	Percentage		Value
Total	350166	100%	Minimum age	19.0
Male	160225	45.8%	Mean age	56.8
Female	189941	54.2%	Maximum age	107.0
Low education	156041	44.6%	Minimum BMI	14.0
Medium education	99302	28.4%	Mean BMI	25.7
High education	94823	27.1%	Maximum BMI	45.0

2.3 *Methods*

We describe the population level distribution of BMI using the sinh-arcsinh normal distribution (Jones & Pewsey 2009, 2019). It is a variation on the normal distribution following the transformation shown in equation 2.1, where z is a normally distributed random variable with zero mean and unit standard deviation. This transformation will also be relevant in section 4, when we apply it to our longitudinal data to turn the BMI values into z-scores. Whenever $\nu = 0$ and $\tau = 1$, a regular normal distribution is obtained where μ and σ control location and scale. Increased values of ν and τ result in positive skew and negative kurtosis, respectively.

$$\text{BMI} = \mu + \sigma \times \tau \times \sinh\left(\frac{\sinh^{-1}(z) + \nu}{\tau}\right) \tag{2.1}$$

We can incorporate explanatory variables into our analysis by making use of the *GAMLSS* package in R which allows us to model each of the four distribution parameters as functions of the predictors (Rigby & Stasinopoulos 2005, Stasinopoulos & Rigby 2007, R Core Team 2021). Initially, we fitted age-dependent splines to all four of the parameters, by sex and level of education. This indicated an approximately quadratic relationship with age for the μ and σ parameters, whereas the level of education mostly impacted their intercept. The ν and τ parameters barely differed by age or education. So we repeated the analysis for a restricted, parametric model, given by equation 2.2. The limited number of parameters results in a parsimonious model and subsequently makes the application in health simulations easier. The two historical trend coefficients, μ_{year} and σ_{year}, are added here for completeness but may be ignored for now; how their values are estimated will be described in section 3.

$$
\begin{aligned}
\mu &= \mu_{education} + \mu_{age} \times age + \mu_{age^2} \times age^2 + \mu_{year} \times year \\
\sigma &= \sigma_{education} + \sigma_{age} \times age + \sigma_{age^2} \times age^2 + \sigma_{year} \times year \\
\nu &= \nu_{intercept} \\
\tau &= \tau_{intercept}
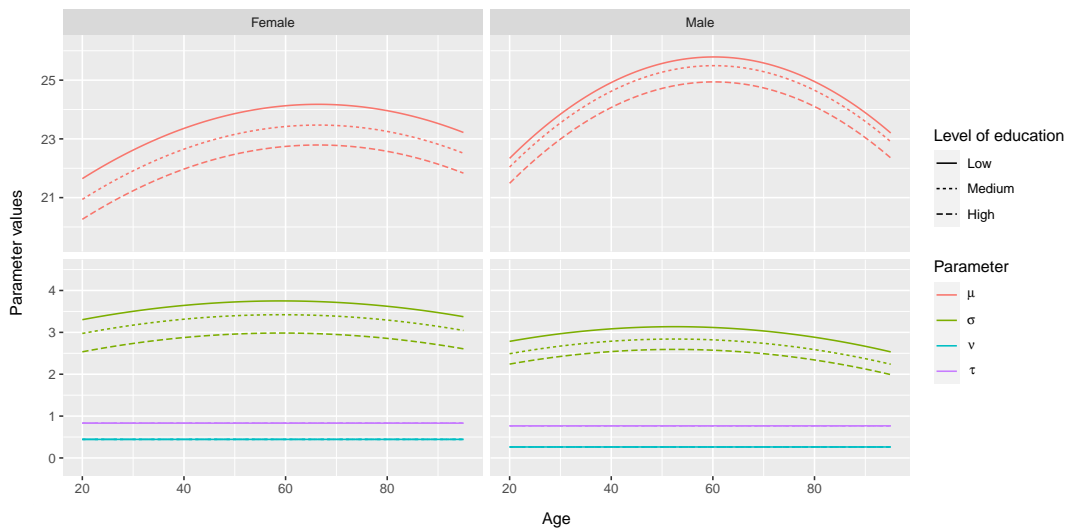\end{aligned}
\tag{2.2}
$$

FIG. 2. The population level distribution parameters for BMI by sex, level of education and age.

2.4 *Results*

The resulting parameter values are summarised in figure 2 while the underlying coefficients are listed in the first columns of table 3. Note again that we are creating separate models for males and females.

2.5 *Internal Validation*

To compare the model's predictions and the underlying data, we can categorise BMI into underweight (values below $18.5\,kg/m^2$), normal weight ($18.5 - 25\,kg/m^2$), overweight ($25 - 30\,kg/m^2$) or obese (values of $30\,kg/m^2$ and above) and determine the observed and modelled prevalences of each of these categories. Figure 3 shows these prevalences, stratified by sex, level of education and age. The model's values are necessarily more smooth due to the low degree of the age polynomial in equation 2.2, but the prediction shows good correspondence with the observed data.
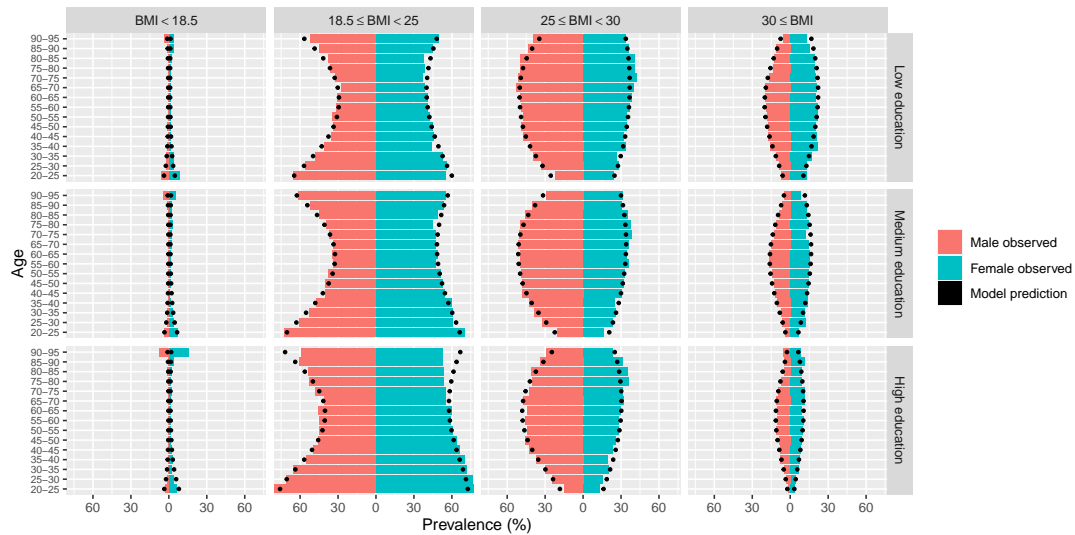

FIG. 3. The observed and modelled prevalences of BMI categories in 2012 by sex, level of education and age.

3. Historical Trend

3.1 *Approach*

In this section, we will expand our model by adding a historical trend. In many western countries, including the Netherlands, overweight and obesity have become more prevalent over time. The main drivers are changes in food habits and a more sedentary lifestyle (Martínez-González et al. 1999, Swinburn et al. 2011, López-Valenciano et al. 2020). Previous work has suggested that this transition increased the spread and skewness of the BMI distribution, rather than solely influencing the mean (Penman & Johnson 2006, Majer et al. 2013, Yamada et al. 2020). Therefore, a linear trend is added to the μ and σ parameters of the population level distribution, as seen in equation 2.2. We will use historical data from the Netherlands to estimate the values of the corresponding μ_{year} and σ_{year} coefficients.

3.2   Data

Historical data about BMI were assessed by Statistics Netherlands and the National Institute for Public Health and the Environment (CBS & RIVM 2021). We use the prevalences of overweight and obesity for adults between 1990 and 2021 which were standardised to the demography of the Netherlands in 2021 (VZinfo.nl 2021). Due to this standardisation step, any trend we extract from these data will mostly be indicative of changes in lifestyle and its influence on BMI rather than also including trends in demography. Table 2 shows some of the figures from this dataset.

TABLE 2 *Historical prevalences of overweight and obesity in the Netherlands.*

	Overweight males	Obese males	Overweight females	Obese females
1990	31.3%	5.4%	26.7%	6.8%
2000	35.4%	8.3%	32.2%	9.8%
2010	42.3%	9.8%	30.4%	12.5%
2020	40.9%	12.3%	31.5%	15.2%

3.3   Methods

To estimate the μ_{year} and σ_{year} coefficients in equation 2.2 which determine the historical trend, we first realise that if we keep ν and τ constant, particular values for μ and σ imply prevalences of overweight and of obesity. Likewise, these prevalences imply values for μ and σ. So from the historical prevalences of overweight and obesity in the Netherlands, we can determine how the μ and σ parameter must have changed over time. Subsequently, we perform a linear regression through these values to estimate the μ_{year} and σ_{year} coefficients which best fit the historical data. These coefficients are estimated for males and females separately and are expressed as the change in μ and σ per year relative to 2012, the calibration year of our model. No dependency on the level of education nor on age is included.

3.4   Results

The first part of table 3 lists the coefficients of the population level distribution, previously visualised in figure 2. The trend coefficients, μ_{year} and σ_{year}, are added as the final two columns of this table. Taken together with equations 2.1 and 2.2, it describes the BMI distribution of the Netherlands stratified by sex, level of education, age and calendar time, valid for some time range around 2012.

TABLE 3 *The coefficients for the population level BMI distribution of the Netherlands.*

	$\mu_{education}^{low}$	$\mu_{education}^{med}$	$\mu_{education}^{high}$	μ_{age}	μ_{age^2}
Male	18.05	17.75	17.20	0.2573	-0.002137
Female	19.00	18.29	17.61	0.1560	-0.001174

	$\sigma_{education}^{low}$	$\sigma_{education}^{med}$	$\sigma_{education}^{high}$	σ_{age}	σ_{age^2}
Male	2.210	1.924	1.677	0.03491	-0.0003327
Female	2.728	2.398	1.960	0.03460	-0.0002926

	$\nu_{intercept}$	$\tau_{intercept}$	μ_{year}	σ_{year}
Male	0.2626	0.7680	0.03624	0.01969
Female	0.4436	0.8302	0.01440	0.03594

3.5 *Internal Validation*

A comparison between the model's predictions and the underlying data can be made. Figure 4 shows the predicted prevalences of overweight and obesity by year, including extrapolations between 1980 and 2040. Here, we average over demography, which is kept constant at the 2021 level. Only the μ and σ parameter are changed as a result of the historical trend coefficients. The predicted trends show good correspondence with the observed data. Note that, although a linear trend is added to the μ and σ parameters, this need not imply that the prevalences of overweight and obesity follow a straight line.
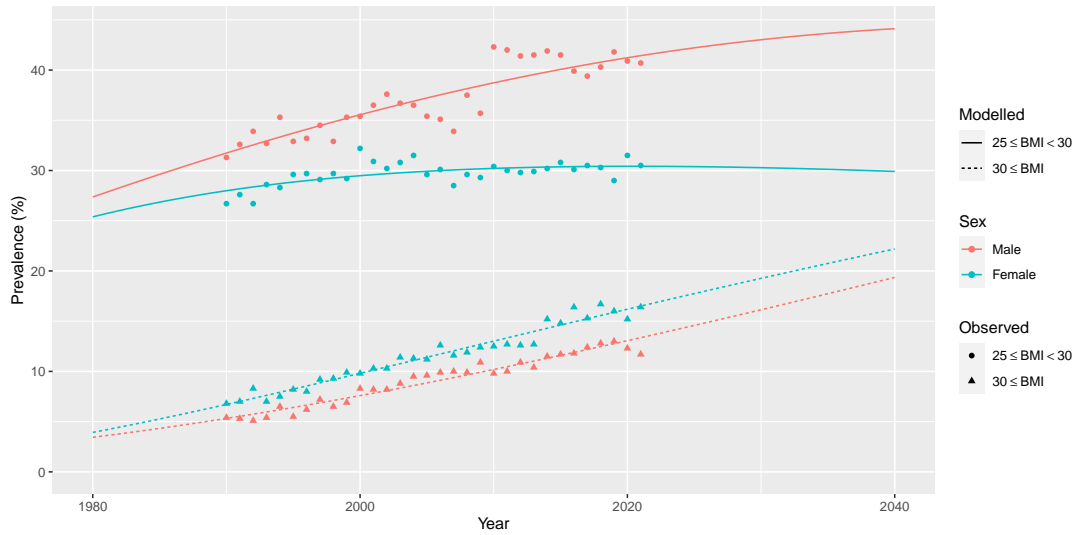

FIG. 4. The observed and modelled prevalences of overweight and obesity by sex and year.

4. Individual Trajectories

4.1 *Approach*

At this stage, our risk factor model can only describe BMI at the population level. For many micro-simulations, such a model is already a useful building block. Then, simulated individuals can be assigned a percentile score indicating where they lie within the distribution which, in combination with their sex, level of education and age, produces a BMI value (McPherson et al. 2007, OECD 2019). As the individual ages throughout the simulation, this value can change while its relative position remains fixed. This is visualised as the fixed-percentile method in figure 1.

We go one step further and allow for individuals to move within the population level distribution. Instead of looking at their percentile scores, we will examine and model their z-scores. Although the information contained in these two types of scores is equivalent, z-scores follow a normal distribution which make them easier to model. Our approach is to reduce the BMI values in our longitudinal data to z-scores following the transformation of equation 2.1 and using values for μ, σ, ν and τ which correspond to the individual's sex, level of education and age and to the calendar time of the measurement. Subsequently, we assume that individuals' z-scores follow stochastic trajectories over their lifetimes and that their BMI trajectories are those z-scores back-transformed to the original scale. Then, the union of all trajectories adheres to the shape of the population level distribution.

Individuals with the same sex, level of education and age can still have vastly different trajectories of BMI. Such heterogeneity is the result of various stochastic components which determine a life course and our model needs to be able to capture this. It may even be possible to identify subgroups with fundamentally different behaviour. For example, it has been suggested that fussy eaters follow different BMI trajectories from people who are less choosy (Herle et al. 2020). This leads to growth mixture models where each subgroup's behaviour is determined by a separate model. However, we will assume all individual life course trajectories can be captured by the same mechanism, where three stochastic components form the source of individual variations.

The first component of these trajectories is the long-term effect. Heritability studies indicate a fairly strong genetic component to BMI, suggesting that individuals have a predisposition to belong to the upper- or lower percentiles of the distribution (Mathias et al. 2003, Goode et al. 2007, Ordoñana et al. 2007). Upbringing and other environmental causes may, too, have long-term effects, influencing someone's habits over their entire lifetime.

Fluctuations do occur however. Previous research has examined the temporal stability of BMI values (Wilsgaard et al. 2001, Ulmer et al. 2003, Juhola et al. 2011, Bayer et al. 2011). These indicate slow changes over time, meaning values close in time correlate more than values measured further apart. Life events and new habits may influence eating- and exercise patterns, thereby affecting BMI, but it seems reasonable to assume that such effects eventually wane, making place for new habits. Modelling these dynamics correctly can be important when the relationship between a risk factor and a disease is simulated. Many such relationships are non-linear and an individual whose BMI fluctuates wildly can experience a different cumulative risk than one who follows a stable trajectory, even when their mean is similar (GBD 2019 Risk Factors Collaborators 2020). That's why, in addition to the long-term effects, we also allow for medium-term effects.

Finally, BMI is subject to short-term fluctuations. Eating or drinking right before weighing yourself influences body weight and, therefore, BMI. Instrumental error originating from the measurement device also contributes. Whether these short-term fluctuations should be modelled as part of the generative process of BMI is up for debate. Instrumental error shows up in both cross-sectional and longitudinal data and increases the population variance (Biehl et al. 2013). Such noise is best removed. Short-term effects with a physical interpretation, such as eating or drinking, or simply a short-lived new habit, might reasonably be considered a part of someone's BMI. Yet, these minor fluctuations have little causal influence on the onset of BMI-related diseases and are, for health modelling purposes, less relevant. Regardless, short-term fluctuations are estimated in our statistical analysis.

4.2 *Data*

To understand how individuals move along typical trajectories and to distinguish between long-term, medium-term and short-term effects, we require longitudinal data with multiple measurements for each participant over a relatively long time. The Doetinchem Cohort Study provides such data. This study has followed a sex- and age stratified random sample from the population registers of the municipality of Doetinchem in the Netherlands for the past 30 years (Verschuren et al. 2008). Its aim is to study lifestyle factors and biological risk factors on aspects of health. The participants underwent a health examination about every 5 years since 1987. A key feature of this panel is that characteristics such as weight and height were measured by research assistants instead of being self-reported. Table 4 provides some descriptive statistics of its latest round, focussing on the variables needed for our model which excludes some of the original data, removed in the cleaning process described in the online supplementary information (Bogaardt et al. 2023).

TABLE 4 *Descriptive statistics of the Doetinchem dataset.*

	Number	Percentage		Value
Total	3426	100%	Minimum age	45.9
Male	1606	46.9%	Mean age	64.0
Female	1820	53.1%	Maximum age	85.8
Low education	1491	43.5%	Minimum BMI	16.0
Medium education	1055	30.8%	Mean BMI	26.8
High education	880	25.7%	Maximum BMI	49.8

The Doetinchem dataset is not completely representative of the Netherlands. For example, it lacks representation of ethnic subgroups and the study was not able to sample new cohorts from younger age categories (Picavet et al. 2017). So, in addition to the Doetinchem data, we also make use of the Longitudinal Internet studies for the Social Sciences (LISS) panel (Scherpenzeel & Das 2010). The LISS panel is a representative sample of Dutch individuals who participate in monthly Internet surveys. The panel is based on a true probability sample of households drawn from the population register. Households that could not otherwise participate are provided with a computer and Internet connection. A longitudinal survey is fielded in the panel every year, covering a large variety of domains including health, work, education, income, housing, time use, political views, values and personality. Table 5 provides some descriptive statistics of its latest round which excludes some of the original data, removed in the cleaning process described in the online supplementary information (Bogaardt et al. 2023).

TABLE 5 *Descriptive statistics of the LISS dataset.*

	Number	Percentage		Value
Total	4076	100%	Minimum age	19.0
Male	1926	47.3%	Mean age	55.6
Female	2150	52.7%	Maximum age	104.4
Low education	1135	27.8%	Minimum BMI	12.9
Medium education	1395	34.2%	Mean BMI	26.1
High education	1546	37.9%	Maximum BMI	58.5

The LISS data were collected between 2007 and 2021, so they do not stretch over as long a period as the Doetinchem data. It also includes self-reported rather than measured values. On the other hand, it does contain a large group of young adults and is sampled with shorter time intervals, meaning both datasets complement each other well. The datasets were simply joined together before the analysis.

4.3 *Methods*

We model an individual's trajectory as a sequence of z-scores which indicate the positions relative to the population level distribution. Following our analysis in section 2, we have a description of the BMI distribution using the sinh-arcsinh transformation and the four parameters μ, σ, ν and τ. These parameters depend on the individual's sex, level of education and age. From the analysis in section 3, we also know how they depend on calendar time. Consequently, all BMI measurements from our longitudinal studies can be transformed using the specific parameters associated with the individuals' characteristics and with the dates of the measurements. This removes any dependency on these variables from the z-scores.

The individual trajectories are assumed to contain long-term, medium-term and short-term effects. This can be described by a generalised autoregressive model. The long-term effects are then opera-

tionalised as a random intercept, which indicates the tendency to belong to the upper- or lower per-
centiles of the BMI distribution. In the fixed-percentile method, this random intercept is the only effect
which determines the trajectories. Our model extends this method by including medium-term effects
which are represented by a standard autoregressive process (AR1). This process assumes that, at each
time period, a random shock occurs which either pushes the BMI z-score up or down. The effect of a
shock decays exponentially over time, similar to how habits wax and wane. The overall effect is the sum
of all previous shocks, which results in a meandering BMI value with temporal autocorrelation, as vi-
sualised in figure 1. Short-term effects are modelled as additional, uncorrelated error representing daily
fluctuations in weight. Note that no random slope is included. Models with these three components are
described in detail in Diggle (1988), Diggle et al. (1994) and Verbeke & Molenberghs (2000).

Our model is described in equation 4.1. Here, Z_i is the vector of z-scores of individual i measured at
k time periods. This includes the three stochastic components of our model. RI_i is the random intercept
of individual i, which has the same value for all measurements, and J_k is the k-by-k matrix with only
ones. $AR1_i$ is the vector of k correlated values, where the level of correlation between measurements
is a function of the time between these measurements. Finally, ε_i is a k-vector of independent, random
errors and I_k is the identity matrix. This model was fit to the transformed longitudinal BMI data using
the *nlme* package in R (Pinheiro et al. 2021, R Core Team 2021).

$$
\begin{aligned}
Z_i &= \beta_0 + \beta_1 \times age + RI_i + AR1_i + \varepsilon_i \\
RI_i &\sim \mathcal{N}(0, \sigma^2_{intercept} \times J_k) \\
AR1_i &\sim \mathcal{N}(0, \Sigma) \text{ where } \Sigma_{tt'} = \sigma^2_{correlated} \times e^{-\frac{|t-t'|}{\tau_{AR1}}} \\
\varepsilon_i &\sim \mathcal{N}(0, \sigma^2_{uncorrelated} \times I_k)
\end{aligned}
\tag{4.1}
$$

Our longitudinal data could, in principle, be used to estimate both a population level distribution
and a historical trend of BMI. But this would only be useful if these data are representative of the
Dutch population. We do not want to make this assumption and prefer to rely on the Public Health
Monitor and the VZinfo datasets to model the population level distribution and historical trend of BMI.
Whatever trend is contained in the longitudinal data, this should largely be removed when transforming
the BMI values to z-scores because the transformation itself depends on calendar time via equation 2.2.
Nonetheless, if these data are not representative, this procedure need not result in values with zero mean
and unit standard deviation; some bias may remain. Ultimately, we want a model which suits the entire
population, so which produces true z-scores. A simple solution is to model the bias by including a fixed
intercept and fixed slope in our statistical analysis, as shown in equation 4.1. We are not interested in the
values of this intercept and slope, but by including the terms, the other parameters are not compromised.

4.4 *Results*

The estimated values, excluding the fixed intercept and slope, are listed in table 6.

TABLE 6 *The parameters of the individual BMI z-score trajectories.*

	$\sigma^2_{intercept}$	τ_{AR1}	$\sigma^2_{correlated}$	$\sigma^2_{uncorrelated}$
Male	0.4744	34.53	0.4761	0.04951
Female	0.6287	19.21	0.3323	0.03905

The resulting covariance structure is also visualised in figure 5, which shows the exponential decay
and the asymptotes of the covariance between two BMI z-scores given their difference in time. Note

that the three variances add up to one. Presented this was, the values for the three variances indicate the relative contributions of the long-term, medium-term and short-term effects.
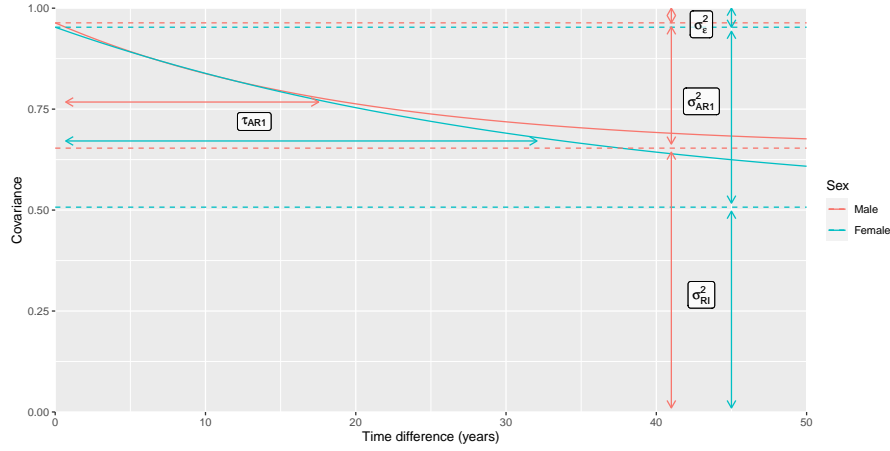

FIG. 5. The exponential decay in the covariance between two BMI z-scores given their difference in time.

4.5 *Internal Validation*

To provide some feeling for the model, we can generate z-score trajectories and visually compare these to the data of a few individuals in the Doetinchem Cohort Study and LISS panel. Although it depends on the precise random samples which are drawn, figure 6 shows that, on the face of it, the generated z-scores and the observed data are comparable. This give credence to our idea of modelling the transformed BMI values using a generalised autoregressive process with long-term, medium-term and short-term effects. By extension, we can conclude that the trajectories in figure 1 mimic realistic BMI trajectories.
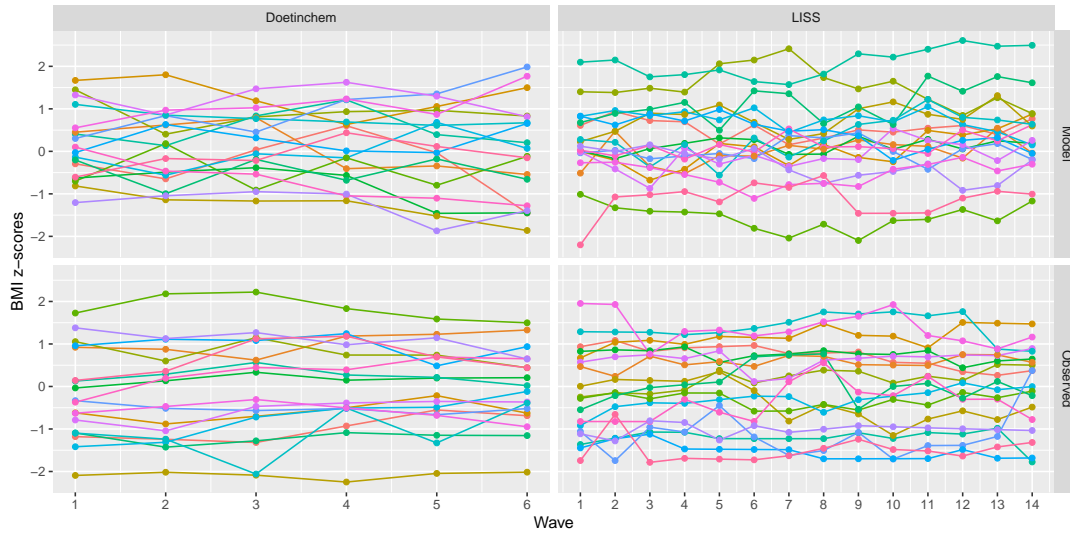

FIG. 6. A sample of observed and model-generated z-scores, coloured by individual, for Doetinchem and LISS.

The model we have created is meant as a building block for micro-simulations. In a macro-simulation framework, a different modelling strategy is required, usually in the form of transition matrices. Estimating such matrices is not trivial and often simulations include no transitions at all (Hendriksen et al. 2015). Others may include the smallest number of transitions between categories necessary to track changes in the population level distribution with age (Van de Kassteele et al. 2012). Our model can be converted to a transition matrix for application in a macro-simulation. The code to do so is available in the online supplementary materials (Bogaardt et al. 2023). This retains much of the richness of our model, including the stratifications as well as the historical trends and additional fluctuations.

The ability to convert our model to a transition matrix naturally leads to another method to validate the results. We can first categorise all BMI values in the Doetinchem and LISS datasets into underweight, normal weight, overweight and obese. Then, every transition from one category to the next is counted and compared with the predicted transitions, shown in figure 7. The model approaches the observed transition rates between the four BMI categories but overestimates the fluctuations which occur for higher BMI values. This indicates that not all aspects of the life course trajectories are being reproduced, but that a large part of the relevant dynamics is indeed captured by our model. The differences seen between Doetinchem and LISS are mostly due to the different time intervals between measurements, where the Doetinchem data were recorded about every 5 years and the LISS data approximately every year.
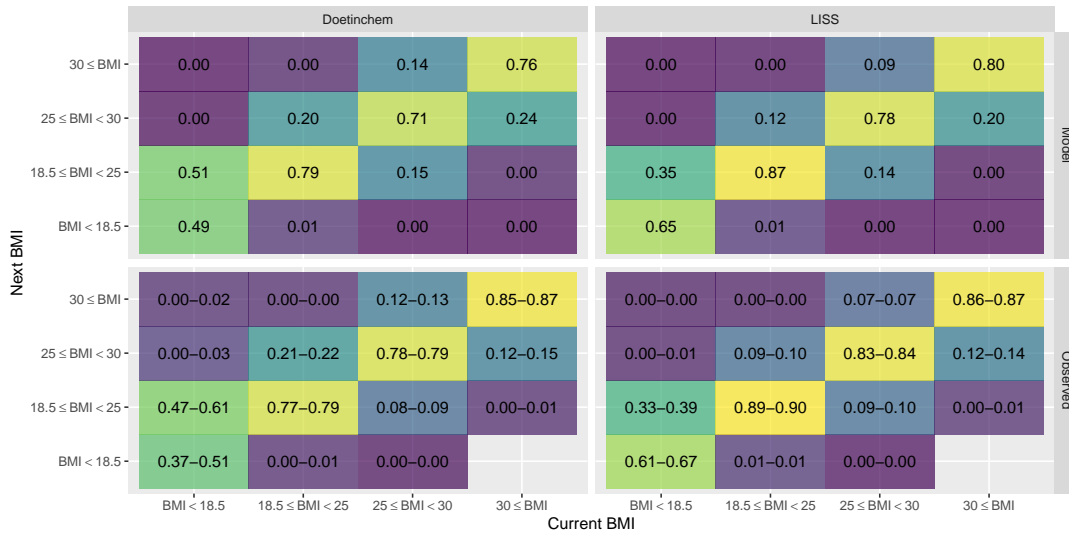

FIG. 7. The observed and modelled transition matrix for the Doetinchem Cohort Study and the LISS panel.

5. Discussion

To summarise, we fitted a population level distribution of BMI, stratified by sex, level of education and age, to a cross-sectional dataset of the Netherlands in 2012. We added a trend using historical prevalences of overweight and obesity. Finally, we modelled individual trajectories as the z-scores of this distribution by making use of longitudinal data and a generalised autoregressive model with long-term, medium-term and short-term effects.

In section 2.5, we discussed the goodness of fit of the population level distribution and showed that our simple model produces representative population level distributions. Some limitations do remain.

For one, the BMI values used to fit the model are self-reported. Previous research has shown that there can be a discrepancy between self-reported and measured BMI values and that individuals with a high BMI tend to underreport their weight (Olfert et al. 2018). Although it may be possible to compensate for such discrepancies, we accept this as a weakness of our work and simply note that our model describes the population level distribution of self-reported BMI values.

Another limitation is that the cross-sectional data may be influenced by missingness due to BMI-related mortality. If a higher BMI leads to a higher mortality rate, the trajectories of the population's heaviest people will be censored. The resulting distribution then shows a lower mean BMI, especially at older ages. Whether this is an issue depends on the type of micro-simulation the model will be used in. If the simulation does not include a separate mortality process, describing the current BMI distribution of a country is best. If, however, mortality due to risk factors is added on top, then the BMI model should aim to describe the generative process. By obtaining information about the link between BMI and mortality, it becomes possible to compensate for this type of missingness. An easy method is to weight each participant in the cross-sectional data with the inverse of their BMI-related mortality rate before fitting the distribution.

In section 3.5, we discussed the goodness of fit of our historical trend and showed that by adding a linear trend to the μ and the σ parameter, we can captures much of the recent change in the prevalences of overweight and obesity. Nonetheless, such a linear trend has limitations. For one, we are ignoring any interaction with education or age. The analysis in the supplementary materials indicated that this is a reasonable simplification (Bogaardt et al. 2023). But more importantly, a linear trend predicts unrealistically low BMI values in the distant past and an ever increasing prevalence of overweight and obesity in the future. The ideal way to operationalise such a time effect would be to fit a sigmoidal function which asymptotically levels off to fixed values in both the distant past and the distant future. Due to limited data and our inherent ignorance of the future, however, a linear trend was fit. This entails that our model can only be valid for some time range around 2012, say, between 1980 and 2040.

A further limitation is that we are ignoring recent government actions to reverse the overweight and obesity trend in the Netherlands (Ministerie van Volksgezondheid, Welzijn en Sport 2018, Van Rinsum et al. 2018). Perhaps it is reasonable to assume that the increase in overweight and obesity has already started to slow down. If that is the case, our trend will not adequately predict the future and should have been fit on a more recent subset of the historical data. Alternatively, we could attempt to quantify the effects of these government actions and take those into account when extrapolating the trend.

In section 4.5, we discussed the goodness of fit of the generalised autoregressive model applied to the individual trajectories of z-scores. Although we have shown that our model fits the data reasonably well, these data contain intermittent missing values and dropouts. Our analysis assumes data are missing-at-random (Diggle et al. 1994, Verbeke & Molenberghs 2000). This is probably a strong assumption as it is not unlikely that a higher BMI leads to a higher probability to miss a wave. In fact, BMI affects mortality rates and some of the dropout in our longitudinal data may be due to obesity-related death (Hoogenveen 2000). The fact that the Doetinchem and LISS datasets are not completely representative of the Netherlands is also a form of missing-not-at-random. In the ideal case, we would model the missing values using some mechanism which predicts intermittent missingness and dropout, but our solution to add a fixed intercept and slope may go a long way.

Another point to note is the linear dependence between the three stochastic components of the generalised autoregressive model. The random intercept, AR1 process and the uncorrelated error are not distinct mechanisms; an AR1 process with perfect temporal correlation leads to a random intercept, whereas a process with zero correlation leads to random noise. So, in a way, our model contains three distinct AR1 processes. As a result, small changes in the data may lead to relatively large changes in the

parameters even though the resulting trajectories are nearly indistinguishable. The three stochastic components are, therefore, best interpreted together as part of a single mechanism, illustrated in figure 5. In light of this, it may be concluded that the individual trajectories of males and females are quite similar.

Figure 6 shows a few individuals who drastically gain- and lose weight over time. Such outlying transitions are not captured by our model. Contrarily, figure 7 suggests our model slightly overestimates the amount of fluctuation which occurs for higher BMI values. Both observations can be true if there is a part of the population who have a particularly stable BMI and another whose BMI changes more rapidly. We had assumed the parameters listed in table 6, which produce mildly fluctuating trajectories, hold true for the entire population. Future work could possibly extend the model using a growth mixture model where one subgroup follows fairly stable trajectories and another subgroup produces larger variations over time (Herle et al. 2020). Regardless, our risk factor model already captures the majority of the heterogeneity within and between individual trajectories, and certainly more than the fixed-percentile method does. Modelling these dynamics correctly can be important when the relationship between a risk factor and a disease is simulated. Many such relationships are non-linear and an individual whose BMI fluctuates wildly can experience a different cumulative risk than one who follows a stable trajectory, even when their mean is similar (GBD 2019 Risk Factors Collaborators 2020).

The risk of disease due to overweight is what drives governments to develop policies to reduce BMI (Ministerie van Volksgezondheid, Welzijn en Sport 2018). Health simulations help to investigate what course of action is most fruitful by comparing a baseline scenario to an alternative scenario in which some preventive intervention has taken place. Interventions usually only apply to a subsection of the population where some criteria determine who is included. These criteria must be encoded into the health simulation. Our model allows for additional selection criteria based on the variations in an individual's BMI over time. Furthermore, it adds realism to the simulated selection procedure because, given the fluctuating BMI, an individual may satisfy the criteria one moment while fail to meet them the next, as occurs in real life.

Subsequently, the effect size of the proposed intervention must be analysed. Our model is a prediction model which describes the associations between sex, level of education, age, calendar time and BMI. We have not looked into the determinants of BMI nor claim any causal connection between variables. Nonetheless, in health simulations, causal connections are essential. When a lifestyle intervention is simulated, its effect on BMI must be modelled to be able to determine any health impact. In addition, assumptions must be made about if and how this effect wanes over time. It is tempting to assume the temporal autocorrelation found in our prediction model implies something about the speed at which intervention effects decay (Bayer et al. 2011). We believe this to be unfounded; the immediate and lasting effects are best researched for every intervention separately (Van Rinsum et al. 2018).

Assuming it is clear what effect a preventive intervention will have, and on whom, this effect can be coded into the health simulation. Effects are frequently modelled by adjusting the parameters of the risk factor model. Although our model of BMI has many parameters which can be tweaked, we do not think all possible changes can be justified. Likely, altering the historical trend when predicting the future or changing an individual's random intercept can be defended. A better alternatively is to model the effect of interventions independent of the baseline risk factor model. For example, a hypothetical intervention may decrease BMI by one point initially and have a lasting effect of half a point. Such deviations can simply be added on top of a realistic BMI model such as ours. Policy effects aimed at an individual's life course may also be modelled as a decrease in the z-score. At the population level, such effects are visible because the union of all z-score trajectories would no longer produce a normal distribution with zero mean, but one with a slightly negative mean. Following the transformation shown in equation 2.1, the shape of the population level distribution of BMI would then be altered as well.

6. Conclusion

We have create a risk factor model of BMI for micro-simulations which is able to faithfully describe the population level distribution, stratified by sex, level of education and age. It can also predict future trends in obesity as well as produce unique life course trajectories for individuals which seem plausible, but it does not capture extreme fluctuations, such as rapid weight loss. Our method is simple, tractable and a natural generalisation of the fixed-percentile method. The generic nature means it can be applied to data from various countries. Even when only cross-sectional data are available, it is probably valid to use the longitudinal parameters from our analysis to construct a country-specific model of individual BMI trajectories. Likewise, the method of fitting a flexible distribution at the population level and modelling longitudinal z-scores by a generalised autoregressive model could be applied to other continuous variables such as daily sugar intake, blood pressure or cholesterol levels. Finally, the risk factor model presented here can be simplified to a transition matrix for use in macro-simulations.

In future work, the model may be extended and generalised in multiple ways. First of all, a growth mixture model could be used to capture the variation between trajectories better, as previously discussed in section 5. Secondly, the set of covariates can be expanded by any other predictor which has some association with BMI. The sole requirement is that this predictor is found in both the cross-sectional and the longitudinal data. Adding immigration-status may be helpful in the development of an accurate health simulation model to guide government policy (Ten Dam et al. in press). Additionally, the BMI trajectories of children and adolescents could be modelled. And finally, our generalised autoregressive model can be extended to multivariate outcomes to model multiple lifestyle characteristics. An example of previous work includes a model similar to ours that was used to simultaneously analyse CD4 T-cell and β-2-microglobulin measurements, both markers of AIDS (Sy et al. 1997). One may also think of alcohol consumption and smoking, known to correlate with BMI (Chiolero et al. 2008, Traversy & Chaput 2015). A model which predicts various risk factors simultaneously can be a great addition to the analysis of preventive interventions on lifestyle choices.

References

Azzalini, A. & Genz, A. (2022), *The R package mnormt: The multivariate normal and t distributions*.
URL: *http://azzalini.stat.unipd.it/SW/Pkg-mnormt/*

Bayer, O., Krüger, H., Von Kries, R. & Toschke, A. M. (2011), 'Factors associated with tracking of BMI: A meta-regression analysis on BMI tracking', *Obesity* **19**(5), 1069–1076.

Biehl, A., Hovengen, R., Meyer, H. E., Hjelmesæth, J., Meisfjord, J., Grøholt, E.-K., Roelants, M. & Strand, B. H. (2013), 'Impact of instrument error on the estimated prevalence of overweight and obesity in population-based surveys', *BMC Public Health* **13**(1), 1–6.

Bogaardt, L., Van Giessen, A., Picavet, H. S. J. & Boshuizen, H. C. (2023), 'A model of individual BMI trajectories', `https://github.com/rivm-syso/A-Model-of-Individual-BMI-Trajectories`.

CBS & RIVM (2021), 'Gezondheidsenquête/Leefstijlmonitor', `https://www.rivm.nl/leefstijlmonitor`.

Chiolero, A., Faeh, D., Paccaud, F. & Cornuz, J. (2008), 'Consequences of smoking for body weight, body fat distribution, and insulin resistance', *Am. J. Clin. Nutr.* **87**(4), 801–809.

Dai, H., Alsalhe, T. A., Chalghaf, N., Riccò, M., Bragazzi, N. L. & Wu, J. (2020), 'The global burden of disease attributable to high body mass index in 195 countries and territories, 1990–2017: An analysis of the global burden of disease study', *PLOS Med.* **17**(7), 1–19.

Diggle, P. J. (1988), 'An approach to the analysis of repeated measurements', *Biometrics* **44**(4), 959–971.

Diggle, P. J., Heagerty, P., Liang, K. Y. & Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford Statistical Science Series, Oxford University Press.

GBD 2019 Risk Factors Collaborators (2020), 'Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019', *Lancet* **396**(10258), 1223–1249.

GGD'en, CBS & RIVM (2012), 'Gezondheidsmonitor volwassenen 2012', `https://www.monitorgezondheid.nl/`.

Goode, E. L., Cherny, S. S., Christian, J. C., Jarvik, G. P. & De Andrade, M. (2007), 'Heritability of longitudinal measures of body mass index and lipid and lipoprotein levels in aging twins', *Twin Res. Hum. Genet.* **10**(5), 703–711.

Hendriksen, M. A. H., Van Raaij, J. M. A., Geleijnse, J. M., Breda, J. & Boshuizen, H. C. (2015), 'Health gain by salt reduction in Europe: A modelling study', *PLOS One* **10**(3), 1–12.

Herle, M., Micali, N., Abdulkadir, M., Loos, R., Bryant-Waugh, R., Hübel, C., Bulik, C. M. & De Stavola, B. L. (2020), 'Identifying typical trajectories in longitudinal data: modelling strategies and interpretations', *Eur. J. Epidemiol.* **35**(3), 205–222.

Hoogenveen, R. T. (2000), Ageing and mortality, Technical report, RIVM.

Huntington-Klein, N. (2021), *vtable: Variable Table for Variable Documentation*. R package version 1.3.3.
URL: *https://CRAN.R-project.org/package=vtable*

Jones, M. C. & Pewsey, A. (2009), 'Sinh-arcsinh distributions', *Biometrika* **96**(4), 761–780.

Jones, M. C. & Pewsey, A. (2019), 'The sinh-arcsinh normal distribution', *Significance* **16**(2), 6–7.

Juhola, J., Magnussen, C. G., Viikari, J. S. A., Kähönen, M., Hutri-Kähönen, N., Jula, A., Lehtimäki, T., Åkerblom, H. K., Pietikäinen, M., Laitinen, T., Jokinen, E., Taittonen, L., Raitakari, O. T. & Juonala, M. (2011), 'Tracking of serum lipid levels, blood pressure, and body mass index from childhood to adulthood: The cardiovascular risk in young Finns study', *J. Pediatr.* **159**(4), 584–590.

Levy, D. T., Mabry, P. L., Wang, Y. C., Gortmaker, S., Huang, T. T. K., Marsh, T., Moodie, M. & Swinburn, B. (2011), 'Simulation models of obesity: a review of the literature and implications for research and policy', *Obes. Rev.* **12**(5), 378–394.

López-Valenciano, A., Mayo, X., Liguori, G., Copeland, R. J., Lamb, M. & Jimenez, A. (2020), 'Changes in sedentary behaviour in European Union adults between 2002 and 2017', *BMC Public Health* **20**(1), 1–10.

Majer, I. M., Mackenbach, J. P. & Van Baal, P. H. M. (2013), 'Time trends and forecasts of body mass index from repeated cross-sectional data: a different approach', *Stat. Med.* **32**(9), 1561–1571.

Martínez-González, M. A., Martínez, J. A., Hu, F. B., Gibney, M. J. & Kearney, J. (1999), 'Physical inactivity, sedentary lifestyle and obesity in the European Union', *Int. J. Obesity* **23**(11), 1192–1201.

Mathias, R. A., Roy-Gagnon, M. H., Justice, C. M., Papanicolaou, G. J., Fan, Y. T., Pugh, E. W., Wilson, A. F. & Framingham Heart Study (2003), 'Comparison of year-of-exam- and age-matched estimates of heritability in the framingham heart study data', *BMC Genetics* **4**, 1–4.

McPherson, K., Marsh, T. & Brown, M. (2007), Tackling obesities: Future choices - modelling future trends in obesity and the impact on health, Technical report.

Ministerie van Volksgezondheid, Welzijn en Sport (2018), Nationaal preventieakkoord, Technical report.
URL: *https://www.rijksoverheid.nl/documenten/convenanten/2018/11/23/nationaal-preventieakkoord*

OECD (2019), SPHeP-NCDs documentation, Technical report.
URL: *http://oecdpublichealthexplorer.org/ncd-doc/index.html*

Olfert, M. D., Barr, M. L., Charlier, C. M., Famodu, O. A., Zhou, W., Mathews, A. E., Byrd-Bredbenner, C. & Colby, S. E. (2018), 'Self-reported vs. measured height, weight, and BMI in young adults', *Int. J. Environ. Res. Public Health* **15**(10), 2–9.

Ooms, J. (2022), *gifski: Highest Quality GIF Encoder*. R package version 1.6.6-1.
URL: *https://CRAN.R-project.org/package=gifski*

Ordoñana, J. R., Rebollo-Mesa, I., González-Javier, F., Pérez-Riquelme, F., Martínez-Selva, J. M., Willemsen, G. & Boomsma, D. I. (2007), 'Heritability of body mass index: A comparison between the netherlands and spain', *Twin Res. Hum. Genet.* **10**(5), 749–756.

Pedersen, T. L. (2021), *ggforce: Accelerating 'ggplot2'*. R package version 0.3.3.
URL: *https://CRAN.R-project.org/package=ggforce*

Pedersen, T. L. & Robinson, D. (2020), *gganimate: A Grammar of Animated Graphics*. R package version 1.0.7.
URL: *https://CRAN.R-project.org/package=gganimate*

Penman, A. D. & Johnson, W. D. (2006), 'The changing shape of the body mass index distribution curve in the population: implications for public health policy to reduce the prevalence of adult obesity', *Prev. Chronic Dis.* **3**(3), 1–4.

Picavet, H. S. J., Blokstra, A., Spijkerman, A. M. W. & Verschuren, W. M. M. (2017), 'Cohort profile update: The Doetinchem cohort study 1987–2017: lifestyle, health and chronic diseases in a life course and ageing perspective', *Int. J. Epidemiol.* **46**(6), 1751–1751g.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2021), *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-152.
URL: *https://CRAN.R-project.org/package=nlme*

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. R version 4.1.1.
URL: *https://www.R-project.org/*

Rigby, R. A. & Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape', *Appl. Stat.* **54**(3), 507–554.

Scherpenzeel, A. & Das, J. W. M. (2010), *'True' longitudinal and probability-based internet panels: Evidence from the Netherlands*, Taylor & Francis, pp. 77–103.

Sonnenberg, F. A. & Beck, J. R. (1993), 'Markov models in medical decision making: A practical guide', *Med. Decis. Mak.* **13**(4), 322–338.

Stasinopoulos, D. M. & Rigby, R. A. (2007), 'Generalized additive models for location scale and shape (GAMLSS) in r', *J. Stat. Softw.* **23**(7), 1–46.

Swinburn, B. A., Sacks, G., Hall, K. D., McPherson, K., Finegood, D. T., Moodie, M. L. & Gortmaker, S. L. (2011), 'The global obesity pandemic: shaped by global drivers and local environments', *Lancet* **378**(9793), 804–814.

Sy, J. P., Taylor, J. M. G. & Cumberland, W. G. (1997), 'A stochastic model for the analysis of bivariate longitudinal AIDS data', *Biometrics* **53**(2), 542–555.

Ten Dam, J., Bogaardt, L., Hoekstra, J., Rodenburg, A. J., Koffijberg, H., Boshuizen, H. C., Hoogeveen, R. T. & Van Giessen, A. (in press), 'Development of a microsimulation model predicting BMI and type 2 diabetes in The Netherlands'.

Traversy, G. & Chaput, J.-P. (2015), 'Alcohol consumption and obesity: An update', *Curr. Obes. Rep.* **4**(1), 122–130.

Ulmer, H., Kelleher, C., Diem, G. & Concin, H. (2003), 'Long-term tracking of cardiovascular risk factors among men and women in a large population-based health system', *Eur. Heart J.* **24**(11), 1004–1013.

Van de Kassteele, J., Hoogenveen, R. T., Engelfriet, P. M., van Baal, P. H. M. & Boshuizen, H. C. (2012), 'Estimating net transition probabilities from cross-sectional data with application to risk factors in chronic disease modeling', *Stat. Med.* **31**(6), 533–543.

Van Rinsum, C., Gerards, S., Rutten, G., Philippens, N., Janssen, E., Winkens, B., Van de Goor, I. & Kremers, S. (2018), 'The coaching on lifestyle (CooL) intervention for overweight and obesity: A longitudinal study into participants' lifestyle changes', *Int. J. Environ. Res. Public Health* **15**(4), 1–27.

Verbeke, G. & Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer.

Verschuren, W. M. M., Blokstra, A., Picavet, H. S. J. & Smit, H. A. (2008), 'Cohort profile: The Doetinchem cohort study', *Int. J. Epidemiol.* **37**(6), 1236–1241.

Vuik, S. & Cecchini, M. (2021), Modelling life trajectories of body-mass index, Technical Report 132.

VZinfo.nl (2021), 'Overgewicht', `https://www.vzinfo.nl/overgewicht`.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
   URL: *https://ggplot2.tidyverse.org*

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686.

Wickham, H. & Miller, E. (2021), *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.4.3.
   URL: *https://CRAN.R-project.org/package=haven*

Wilsgaard, T., Jacobsen, B. K., Schirmer, H., Thune, I., Løchen, M. L., Njølstad, I. & Arnesen, E. (2001), *Am. J. Epidemiol.* **154**(5), 418–426.

Yamada, G., Castillo-Salgado, C., Jones-Smith, J. C. & Moulton, L. H. (2020), 'Obesity prediction by modelling BMI distributions: application to national survey data from Mexico, Colombia and Peru, 1988–2014', *Int. J. Epidemiol.* **49**(3), 824–833.

Yu, G. & Xu, S. (2021), *ggbreak: Set Axis Break for 'ggplot2'*. R package version 0.0.6.
   URL: *https://CRAN.R-project.org/package=ggbreak*