Estimating the long-term health impacts of changes in alcohol consumption in England during the COVID-19 pandemic

Funded by The National Institute of Health Research (NIHR)

Appendix 3. Technical appendix

HEALTH LUMEN

Contents

Microsimulation framework

The HealthLumen microsimulation consists of two models. The first model is a sophisticated regression model which calculates the predictions of risk factor trends over time based on data from rolling cross-sectional studies. The second model performs the microsimulation of a virtual population, generated with demographic characteristics matching those of the observed data. The health trajectory of each individual from the population is simulated over time allowing them to contract, survive or die from a set of diseases or injuries related to the analysed risk factors. The detailed description of the two modules is presented below.

Model one: Predictions of risk factors over time

Alcohol consumption is modelled as well as a range of other , body mass index (BMI), tobacco consumption, physical activity, salt intake, physical activity, waist circumference, nitrogen dioxide exposure and particulate matter 2.5µm. In general, the risk factors are categorised into three groups based on the cut-offs provided in published guidelines.

For each RF, let *N* be the number of categories for a given risk factor, e.g., *N* = 3 for alcohol. Let k = 1, 2, …, *N* number these categories and $p_k(t)$ denote the prevalence of individuals with RF values that correspond to the category k at time t. We estimate $p_k(t)$ using multinomial logistic regression model with prevalence of RF category k as the outcome, and time *t* as a single explanatory variable. For $k < N$, we have

$$\ln\left(\frac{p_k(t)}{p_1(t)}\right) = \beta_0^k + \beta_1^k t \qquad (0.1)$$

The prevalence of the first category is obtained by using the normalisation constraint $\sum_{k=1}^{N} p_k(t) = 1$. Solving equation (0.1) for $p_k(t)$, we obtain

$$p_k(t) = \frac{\exp\left(\beta_0^k + \beta_1^k t\right)}{1 + \sum_{k'=1}^{N} \exp\left(\beta_0^{k'} + \beta_1^{k'} t\right)}, \qquad (0.2)$$

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

Multinomial logistic regression

Measured data consist of sets of probabilities, with their variances, at specific time values (typically the year of the survey). For any particular time, the sum of these probabilities is unity. Typically, such data might be the probabilities of low-risk, medium-risk, or high-risk as they are extracted from the survey data set. Each data point is treated as a normally distributed[1]

[1] Depending on the circumstances this assumption will be more or less accurate and more or less necessary. In general, it is both extremely useful and accurate. For simple surveys the individual Bayesian prior and posterior

random variable; together they are a set of N groups (number of years) of K probabilities $\{\{t_i, \mu_{ki}, \sigma_{ki} | k \in [0,K\text{-}1]\} | i \in [0,N\text{-}1]\}$. For each year the set of K probabilities form a distribution – their sum is equal to unity.

The regression consists of fitting a set of logistic functions $\{p_k(\mathbf{a}, \mathbf{b}, t) | k \in [0,K\text{-}1]\}$ 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 three states already mentioned, the $k = 0$ regression function represents the probability of being a low-risk drinker over time, $k = 1$ the probability of being an increasing-risk drinker and $k = 2$ the probability of being a high-risk drinker.

The regression equations are most easily derived from a familiar 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 $p_k(\mathbf{a},\mathbf{b};t)$ are chosen to be ratios of sums of exponentials (This is equivalent to modelling the log probability ratios, p_k/p_0, as linear functions of time.)

$$S(\mathbf{a},\mathbf{b}) = \tfrac{1}{2} \sum_{k=0}^{k=K-1} \sum_{i=0}^{i=N-1} \frac{\left(p_k(\mathbf{a},\mathbf{b};t_i) - \mu_{ki}\right)^2}{\sigma_{ki}^2} \tag{0.3}$$

$$p_k(\mathbf{a},\mathbf{b},t) \equiv \frac{e^{A_k}}{1 + e^{A_1} + .. + e^{A_{K-1}}}$$

$$\mathbf{a} \equiv (a_0, a_1, .., a_{K-1}), \quad \mathbf{b} \equiv (b_0, b_1, .., b_{K\text{-}1}) \tag{0.4}$$

$$A_0 \equiv 0, \quad A_k \equiv a_k + b_k t$$

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 $2(K\text{-}1)$ regression parameters to be determined.

For a given dimension K there are K-1 independent functions p_k – the remaining function being determined from the requirement that complete set of K form a distribution and sum to unity.

Note that the parameterization ensures that the necessary requirement that each p_k be interpretable as a probability – a real number lying between 0 and 1.

The minimum of the function S is determined from the equations

probabilities are Beta distributions – the likelihood being binomial. For reasonably large samples, the approximation of the beta distributions by normal distributions is both legitimate and a practical necessity. For complex, multi-PSU, stratified surveys, it is again assumed that these base probabilities are approximately normally distributed and, again, it is an assumption that makes the analysis tractable.

Depending on the nature of the raw data set it may be possible to use non-parametric statistical methods for this analysis. This is possible for the HSE and GHS data sets of this study but when this has been done the authors can report no discernible difference in the results.

$$\frac{\partial S}{\partial a_j} = \frac{\partial S}{\partial b_j} = 0 \qquad \text{for j=1,2,....,k-1} \tag{0.5}$$

noting the relations

$$\frac{\partial p_k}{\partial A_j} = \frac{\partial}{\partial A_j}\left(\frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}}\right) = p_k\delta_{kj} - p_k p_j$$
$$\frac{\partial}{\partial a_j} = \frac{\partial}{\partial A_j} \tag{0.6}$$
$$\frac{\partial}{\partial b_j} = t\frac{\partial}{\partial A_j}$$

The values of the vectors **a**, **b** that satisfy these equations are denoted $\hat{\mathbf{a}}, \hat{\mathbf{b}}$. They provide the trend lines, $p_k\left(\hat{\mathbf{a}}, \hat{\mathbf{b}}; t\right)$, for the separate probabilities. 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 $\hat{\mathbf{a}}, \hat{\mathbf{b}}$. 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 $(\mathbf{a}-\hat{\mathbf{a}}), (\mathbf{b}-\hat{\mathbf{b}})$ about the maximum likelihood estimate $\hat{\mathbf{S}} \equiv S\left(\hat{\mathbf{a}}, \hat{\mathbf{b}}\right)$. Hence

$$\begin{aligned} S(\mathbf{a},\mathbf{b}) &= \tfrac{1}{2}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1}\frac{\left(p_k(\mathbf{a},\mathbf{b};t_i)-\mu_{ki}\right)^2}{\sigma_{ki}^2} \\ &\equiv S\left(\hat{a},\hat{b}\right) + \tfrac{1}{2}\left(a-\hat{a}, b-\hat{b}\right)P^{-1}\left(a-\hat{a}, b-\hat{b}\right)+... \\ &\approx S\left(\hat{a},\hat{b}\right) + \tfrac{1}{2}\sum_{i,j}(a_i-\hat{a}_i)\frac{\partial^2\hat{S}}{\partial\hat{a}_i\partial\hat{a}_j}\left(a_j-\hat{a}_j\right) + \tfrac{1}{2}\sum_{i,j}(a_i-\hat{a}_i)\frac{\partial^2\hat{S}}{\partial\hat{a}_i\partial\hat{b}_j}\left(b_j-\hat{b}_j\right) + \\ &+ \tfrac{1}{2}\sum_{i,j}\left(b_i-\hat{b}_i\right)\frac{\partial^2\hat{S}}{\partial\hat{b}_i\partial\hat{a}_j}\left(a_j-\hat{a}_j\right) + \tfrac{1}{2}\sum_{i,j}\left(b_i-\hat{b}_i\right)\frac{\partial^2\hat{S}}{\partial\hat{b}_i\partial\hat{b}_j}\left(b_j-\hat{b}_j\right) \end{aligned} \tag{0.7}$$

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 $N\left(\hat{p}_k(t),\sigma_k^2(t)\right)$ by expanding p_k about its maximum likelihood estimate (the trend line) $\hat{p}_k(t)=p\left(\hat{\mathbf{a}},\hat{\mathbf{b}},t\right)$

$$\begin{aligned} p_k\left(\mathbf{a},\mathbf{b},t\right) &= p_k\left(\hat{\mathbf{a}}+\mathbf{a}-\hat{\mathbf{a}},\hat{\mathbf{b}}+\mathbf{b}-\hat{\mathbf{b}},t\right) \\ &= \hat{p}_k\left(t\right)+\left(\nabla_{\hat{a}},\nabla_{\hat{b}}\right)\hat{p}_k\left(t\right)\begin{pmatrix}\mathbf{a}-\hat{\mathbf{a}}\\ \mathbf{b}-\hat{\mathbf{b}}\end{pmatrix}+\ldots \end{aligned} \tag{0.8}$$

Denoting mean values by angled brackets, the variance of p_k is thereby approximated as

$$\begin{aligned} &\sigma_k^2\left(t\right)\equiv\left\langle\left(p_k\left(\mathbf{a},\mathbf{b},t\right)-\hat{p}_k\left(t\right)\right)^2\right\rangle=\left(\nabla_{\hat{a}}\hat{p}_k\left(t\right),\nabla_{\hat{b}}\hat{p}_k\left(t\right)\right)\left\langle\begin{pmatrix}\mathbf{a}-\hat{\mathbf{a}}\\ \mathbf{b}-\hat{\mathbf{b}}\end{pmatrix}\begin{pmatrix}\mathbf{a}-\hat{\mathbf{a}}\\ \mathbf{b}-\hat{\mathbf{b}}\end{pmatrix}^T\right\rangle\times \\ &\left(\nabla_{\hat{a}}\hat{p}_k\left(t\right),\nabla_{\hat{b}}\hat{p}_k\left(t\right)\right)^T=\left(\nabla_{\hat{a}}\hat{p}_k\left(t\right),\nabla_{\hat{b}}\hat{p}_k\left(t\right)\right)\,P\,\left(\nabla_{\hat{a}}\hat{p}_k\left(t\right),\nabla_{\hat{b}}\hat{p}_k\left(t\right)\right)^T \end{aligned} \tag{0.9}$$

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

$$\sigma_k^2\left(t\right)=\begin{pmatrix}\frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{a}_1} & \frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{a}_2} & \frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{b}_1} & \frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{b}_2}\end{pmatrix}\begin{bmatrix}P_{aa11} & P_{aa12} & P_{ab11} & P_{ab12}\\ P_{aa21} & P_{aa22} & P_{ab21} & P_{ab22}\\ P_{ba11} & P_{ba12} & P_{bb11} & P_{bb12}\\ P_{ba21} & P_{ba22} & P_{bb21} & P_{bb22}\end{bmatrix}\begin{pmatrix}\frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{a}_1}\\ \frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{a}_2}\\ \frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{b}_1}\\ \frac{\partial\hat{p}_k\left(t\right)}{\partial\hat{b}_2}\end{pmatrix} \tag{0.10}$$

where $P_{cdij}\equiv\left\langle\left(c_i-\hat{c}_i\right)\left(d_j-\hat{d}_j\right)\right\rangle$. The 95% confidence interval for $p_k(t)$ is centred given as $\left[\hat{p}_k(t)-1.96\sigma_k(t),\mathrm{p_k}(t)+1.96\sigma_k(t)\right]$.

Model two: Microsimulation Model

Microsimulation model overview

Simulated people are generated with the correct demographic statistics in the simulation's start-year. In this year women are stochastically allocated the number and years of birth of their children – these are generated from known fertility and mother's age at birth statistics (valid in the start-year). If a woman has children then those children are generated as members of the simulation in the appropriate birth year. The microsimulation is provided with a list of relevant diseases. These diseases used the best available incidence, mortality, survival, relative risk and prevalence statistics (by age and sex). The virtual population is initialised with diseases by simulating each individual from birth until the start year of the model simulation. It assumed that a person could die before the model start year. It is assumed that at initialisation the diseases are independent random variables. In the course of their lives, simulated people can die from one of the diseases caused by the particular risk factor that they might have acquired or from some other cause. The probability that a person of a given age and gender dies from a cause other than the disease are calculated in terms of known death and disease statistics valid in the start-year. It is constant over the course of the simulation. The survival rates from the risk factor-related diseases will change as a consequence of the changing distribution of the risk factor in the population.

The microsimulation incorporates a sophisticated economic module. The module employs Markov-type simulation of long-term health benefits, health care costs and non-health care related costs of specified interventions. It synthesises and estimates evidence on and cost-utility analysis. The model is used to project the differences in quality-adjusted life years (QALYs), lifetime health-care costs, premature mortality costs and indirect costs as a consequence of interventions. Over a specified time scale. Outputs can be discounted for any specific discount rate.

Population module

The population module contains several datasets which can be edited by the end user through a user interface. The population is created in the start-year and propagated forwards in time by allowing females to give birth and also has the ability to incorporate population projections (i.e. migration through minimum arrivals and departures). People within the model can die from specific diseases or from other causes. The <deaths by year by sex by age> file is a necessary input to the model when population projections are being used valid in the start year and usually referred to as the deaths from all causes file. This module is flexible and allows the user to run open and closed populations with no births.

Distributions

Distribution name	symbol	note
MalesByAgeByYear	$p_m(a)$	Input in year$_0$ – probability of a male having age a
FemalesByAgeByYear	$p_f(a)$	Input in year$_0$ – probability of a female having age a
BirthsByAgeofMother	$p_b(a)$	Input in year$_0$ – conditional probability of a birth at age a| the mother gives birth.
NumberOfBirths	$p_l(n)$	λ≡TFR, Poisson distribution, probability of giving birth to n children
MaleDeathByAge	$p_{Wm}(a)$	Input in year$_0$, probability of a male dying at age a
FemaleDeathByAge	$p_{Wf}(a)$	Input in year$_0$, probability of a female dying at age a

Birth model

Any female in the child bearing years {*AgeAtChild.lo, AgeAtChild.hi*} is deemed capable of giving birth. The number of children, n, that she has in her life is dictated by the Poisson distribution $p_l(n)$ where the mean of the Poisson distribution is the Total Fertility Rate (TFR) parameter[2].

The probability that a mother (who does give birth) gives birth to a child at age a is determined from the BirthsByAgeOfMother distribution as $p_b(a)$. For any particular mother the births of multiple children are treated as independent events, so that the probability that a mother who produces N children produces n of them at age a is given as the Binomially distributed variable,

$$p_b\left(n\,at\,a \mid N\right) = \frac{N!}{n!\left(N-n\right)!}\left(p_b\left(a\right)\right)^n\left(1-p_{bm}\left(a\right)\right)^{N-n} \tag{0.11}$$

The probability that the mother gives birth to n children at age a is

$$p_b\left(n\,at\,a\right) = e^{-\lambda}\sum_{N=n}^{\infty}\frac{\lambda^N}{N!}p_b\left(n\,at\,a \mid N\right) = e^{-\lambda}\sum_{N=n}^{\infty}\frac{\lambda^N}{n!\left(N-n\right)!}\left(p_b\left(a\right)\right)^n\left(1-p_b\left(a\right)\right)^{N-n} \tag{0.12}$$

Performing the summation in this equation gives the simplifying result that the probability p_b(*n at a*) is itself Poisson distributed with mean parameter $\lambda p_b(a)$,

$$p_b\left(n\,at\,a\right) = e^{-\lambda p_b(a)}\frac{\left(\lambda p_b\left(a\right)\right)^n}{n!} = p_{\lambda p_b(a)}\left(n\right) \tag{0.13}$$

Thus, on average, a mother at age a will produce $\lambda p_b(a)$ children in that year.

The gender of the children[3] is determined by the probability p_{male}=1-p_{female}. In the baseline model this is taken to be the probability $N_m/(N_m+N_f)$.

[2] This could be made to be time dependent; in the baseline model it is constant.

[3] The probability of child gender can be made time dependent.

Time dependent birth rates

The total fertility rate (TFR) parameter for future years can be input from file if known – or otherwise modelled. In general, the TFR parameter is kept constant overtime. In each year of their simulated life (y at age a), mothers of child-bearing age can use the appropriate Poisson parameter $\lambda(a)p_b(a)$ to generate the number of children in that year. Each child is recorded in the mother's Life Event list and processed as part of the current family at the end of the mother's life.

Population dynamics

In some year, Y, the population will consist of N_m males and N_f females with their respective age distributions. In the next year, Y', the numbers will have been depleted by deaths and augmented by the $N_{newborn}$ births. The new, primed, population is determined from the old by the following equation set

$$N_{newborn} = \lambda N_f \sum_{a=AgeAtChild.lo}^{a=AgeAtChild.hi} p_f(a)\left(1-p_f(a)\right)p_b(a) \tag{0.14}$$

$$N'_m = N_m \sum_{a=1}^{a=Age.hi} p_m(a)\left(1-p_m(a)\right) + p_{male}N_{newborn} \tag{0.15}$$

$$N'_f = N_f \sum_{a=1}^{a=Age.hi} p_f(a)\left(1-p_f(a)\right) + p_{female}N_{newborn} \tag{0.16}$$

$$p'_m(a+1) = \frac{N_m}{N'_m} p_m(a)\left(1-p_m(a)\right) \tag{0.17}$$

$$p'_m(a+1) = \frac{N_m}{N'_m} p_m(a)\left(1-p_{\Omega m}(a)\right) \tag{0.18}$$

$$p'_f(a+1) = \frac{N_f}{N'_f} p_f(a)\left(1-p_{\Omega f}(a)\right) \tag{0.19}$$

$$p'_m(0) = \frac{1}{N'_m} p_{male}N_{newborn} \tag{0.20}$$

$$p'_f(0) = \frac{1}{N'_f} p_{female}N_{newborn} \tag{0.21}$$

The Population editor' menu item Population Editor\View\Population dynamics\male implements these equations and draws projected populations year by year.

Deaths from modelled diseases

The simulation models any number of specified diseases some of which may be fatal. In the start year the simulation's death model uses the diseases' own mortality statistics to adjust the probabilities of death by age and gender. In the start year the net effect is to maintain the same probability of death by age and gender as before; in subsequent years, however, the rates at which people die from modelled diseases will change as modelled risk factors change. This the population dynamics sketched above will be only an approximation to the simulated population's dynamics. The latter will be known only on completion of the simulation.

Multiple population processing

Multiple populations can be used in a simulation provided they are non-overlapping (people cannot belong to both).

In a simulation, Monte Carlo trials are allocated between current different populations in proportion to their total person count (malesCount+femalesCount). The idea being to provide a representative sample of the combined population.

In a simulation, a population (pop) is current if the simulated year Y satisfies

$$pop \to startYear \le Y \le pop \to stopYear \tag{0.22}$$

Open populations

This model is an *open* population model which allows people to enter and to depart from the population (departure probability $p_\delta(t)$).

Open population, births and deaths

In the year y the number of males and females in the population are denoted as $\{N_m(a,y), N_f(a,y)\}$,

And we suppose that they have departure probabilities $\{p_{m\delta}(a,y), p_{f\delta}(a,y)\}$. The number of new arrivals into each age in the year Y are denoted $\{N_{mArr}(a,y), N_{fArr}(a,y)\}$.

The following analysis applies equally to males and females and we drop the gender suffix. The male and female populations grow according to the recursion relations

$$N(a+1,y+1) = N(a,y)\left(1-p_\Omega(a)\right)\left(1-p_\delta(a,y)\right) + N_{Arr}(a,y) \quad (a>1) \tag{0.23}$$

$$N(1,y+1) = N_{Newborn}(y)\left(1-p_\Omega(0)\right)\left(1-p_\delta(0,y)\right) + N_{Arr}(0,y) \quad (a=0) \tag{0.24}$$

The longitudinal modelling of populations having known cross sectional data

Given a set of X-sectional population projections $\{K_m(a,y), K_f(a,y)|0<=a<=100;\ Y_0<=y<=Y_1\}$ (the K- population) the question arises of how to model the lives of individuals within the population (the N-population). In the absence of precise arrival (immigration) and departure (emigration) statistics, many solutions exist. The population is constructed iteratively: given the population in year Y the next year' population is calculated from the known birth and death rates; the departure probabilities and arrival numbers are found by matching with the projected K-population.

Minimum arrival and departure model

The minimum arrival and departure model fixes the modelled N-population in the start year and compensates in subsequent years either by having non-zero departure statistics (if N>K) or by importing new people (K>N).

From equation (0.23):

$$if\ N(a,y)\left(1-p_{\Omega}(a)\right)>K(a+1,y+1)$$

$$\left(1-p_{\delta}(a,y)\right)=\frac{K(a+1,y+1)}{N(a,y)\left(1-p_{\Omega}(a)\right)}\quad (a>1)$$

$$\Rightarrow$$

$$N(a+1,y+1)=N(a,y)\left(1-p_{\Omega}(a)\right)\left(1-p_{\delta}(a,y)\right)=K(a+1,y+1)\quad (a>1) \qquad (0.25)$$

$$if\ N(a,y)\left(1-p_{\Omega}(a)\right)<K(a+1,y+1)$$

$$N_{Arr}(a,y)=K(a+1,y+1)-N(a,y)\left(1-p_{\Omega}(a)\right)\quad (a>1)$$

$$\Rightarrow$$

$$N(a+1,y+1)=N(a,y)\left(1-p_{\Omega}(a)\right)+N_{Arr}(a,y)=K(a+1,y+1) \qquad (0.26)$$

The implementation of this model can be arranged using multiple populations – one population for each year of the simulation. The first population consists of the base line model that matches the N and K populations in the start year; subsequent populations contain the corrections (the arrivals, if any in that year). When arrivals enter the simulated population they have a start year corresponding to this population's start year. They usually will have been modelled from birth in the appropriate risk and disease environment. Arrivals are ordinary members of the modelled population – they simply enter the population at times after the simulation-start time. Arrivals carry with them a population identifier.

The numbers of males and females and their ages are known for all populations. Within the micro simulation multiple populations are sampled at a rate proportional to their population size.

Risk factor module

The distribution of risk factors (RF) in the population is estimated using regression analysis stratified by both sex *S* = {male, female} and age group *A* = e.g. {0-9, 10-19, ..., 70-79, 80+}. The fitted trends are extrapolated to forecast the distribution of each RF category in the future. For each sex-and-age-group stratum, the set of cross-sectional, time-dependent, discrete distributions $D = \{p_k(t)|k = 1, \dots N;\ t > 0\}$, is used to manufacture RF trends for individual members of the population.

Continuous risk factors

In the case of a continuous RF, for each discrete distribution D there is a continuous counterpart. Let β denote the RF value in the continuous scale and let $f(\beta|A,S,t)$ be the probability density function of β for age group A and sex S at time t. Then

$$p_k(t|A,S) = \int_{\beta \in k} f(\beta|A,S,t)\, d\beta. \qquad (0.27)$$

Equations (0.2) and (0.27) both refer to the same quantity. However, equation (0.27) uses the definition of the probability density function to express the age-and-sex-specific percentage of individuals in RF category *k* at time *t*. Equation (0.2) gives an estimate of this quantity using equation (0.1) for all *k* = 0, …, N. The cumulative distribution function of β is

At time *t*, a person with sex S belonging to the age group A is said to be on the p–th percentile of this distribution if $F(\beta|A,S,t) = p/100$. Given the cross-sectional information from the set of distributions D, it is possible to simulate longitudinal trajectories by forming pseudo-cohorts within the population. A key requirement for these sets of longitudinal trajectories is that they reproduce the cross-sectional distribution of RF categories for any year with available data. The method adopted here and in the earlier Foresight report (1) is based on the assumption that person's RF value changes throughout their lives in such a way that they always have the same associated percentile rank. As they age, individuals move from one age group to another and their RF value changes so that they have the same percentile rank but of a different RF distribution. In a nutshell, we assume (in accordance with research on the long-term success rate in dieting) that relatively fat people will remain relatively fat and relatively thin people will remain relatively thin. Crucially it meets the important condition that the cross-sectional RF distributions obtained by simulation match the RF distributions of the observed data.

The above procedure can be explained using the example of the alcohol consumption distribution. The alcohol consumption distributions are known for the population stratified by sex and age for all years of the simulation (by extrapolation of fitted model, see equation (0.1)). A person who is in age group A and who grows ten year older will at some time move into the next age group A' and will have an alcohol consumption that was described first by the distribution $f(\beta|A,S,t)$ and then at the later time t' by the distribution $f(\beta|A',S,t')$. If the

alcohol consumption of that individual is on the p-th percentile of the alcohol consumption distribution, their alcohol consumption will change from β to β' so that

$$\beta = F^{-1}\left(\frac{p}{100}|A,S,t\right) \tag{0.28}$$

$$\beta' = F^{-1}\left(\frac{p}{100}|A',S,t'\right) \Rightarrow \beta' = F^{-1}\left(F(\beta \mid A,S,t)|A',S,t'\right) \tag{0.29}$$

Where F^{-1} is the inverse of the cumulative distribution function of β, which we model with a continuous uniform, normal or lognormal distribution (depending on the risk factor) within the RF categories. Equation (0.29) guarantees that the transformation taking the random variable β to β' ensures the correct cross-sectional distribution at time t'.

The microsimulation first generates individuals from the RF distributions of the set D and, once generated, grows the individual's RF in a way that is also determined by the set D. It is possible to implement equation (0.29) as a suitably fast algorithm.

Disease module

Disease modelling relies heavily on the sets of incidence, mortality, survival, relative risk and prevalence statistics. The microsimulation uses risk dependent incidence statistics and these are inferred from the relative risk statistics and the distribution of the risk factor within the population. In the simulation, individuals are assigned a risk factor trajectory giving their personal risk factor history for each year of their lives. Their probability of getting a particular risk factor related disease in a particular year will depend on their risk factor state in that year. The necessary equations are given below. The microsimulation model has the ability to model discrete multiple stages of a disease.

Once a person has a fatal disease (or diseases) their probability of survival will be controlled by a combination of the disease-survival statistics and the probabilities of dying from other causes. Disease survival statistics are modelled as age and gender dependent exponential distributions.

Relative risks

The reported incidence risks for any disease do not make reference to any underlying risk factor. The microsimulation requires this dependence to be made manifest.

The risk factor dependence of disease incidence has to be inferred from the distribution of the risk factor in the population (here denoted as π); it is a disaggregation process:

Suppose that α is a risk factor state of some risk factor A and denote by $p_A(d|\alpha,a,s)$ the incidence probability for the disease d given the risk state, α, the person's age, a, and gender, s. The relative risk ρ_A is defined by equation (0.30).

$$p_{\mathrm{A}}\left(d\left|\alpha,a,s\right.\right)=\rho_{\mathrm{A}|d}\left(\alpha\left|a,s\right.\right)p_{\mathrm{A}}\left(d\left|\alpha_0,a,s\right.\right)$$
$$\rho_{\mathrm{A}|d}\left(\alpha_0\left|a,s\right.\right)\equiv 1 \qquad (0.30)$$

Where α_0 is the zero risk state (for example, the moderate state for alcohol consumption).

The incidence probabilities, as reported, can be expressed in terms of the equation,

$$\begin{aligned} p\left(d\left|a,s\right.\right) &= \sum_{\alpha} p_{\mathrm{A}}\left(d\left|\alpha,a,s\right.\right)\pi_A\left(\alpha\left|a,s\right.\right) \\ &= p_A\left(d\left|\alpha_0,a,s\right.\right)\sum_{\alpha}\rho_{\mathrm{A}|d}\left(\alpha\left|a,s\right.\right)\pi_A\left(\alpha\left|a,s\right.\right) \end{aligned} \qquad (0.31)$$

Combining these equations allows the conditional incidence probabilities to be written in terms of known quantities

$$p\left(d\left|\alpha,a,s\right.\right)=\rho_{A|d}\left(\alpha\left|a,s\right.\right)\frac{p\left(d\left|a,s\right.\right)}{\sum_{\beta}\rho_{A|d}\left(\beta\left|a,s\right.\right)\pi_A\left(\alpha\left|a,s\right.\right)} \qquad (0.32)$$

Previous to any series of Monte Carlo trials the microsimulation program pre-processes the set of diseases and stores the *calibrated* incidence statistics $p_A(d|a_0, a, s)$.

For each scenario the incidence statistics are calibrated against the baseline trends.

Approximating missing data points

Published disease statistics are frequently incomplete and occasionally inconsistent. The microsimulation program makes use of a number of supporting methods to check and, as necessary, to estimate missing disease statistics.

Public health intervention module

The HL model is flexible and a number of different public health interventions can be simulated in the model from population to individual level interventions. The intervention can be implemented to impact on one or more of the modules such as the risk factor and disease modules.

Model output module

Cross-sectional outputs (epidemiological and economic) per 100,000 of the population are computed for each year of the simulation.

Epidemiological outputs

A range of different epidemiological outputs are produced by the model including:

- Incidence rates
- Cumulative incidence rates
- Prevalence rates
- Mortality rates
- Premature mortality rates
- Quality Adjusted Life Years (QALY)
- Potential Years of Life Lost (PYLL)

Some of these outputs are standard and do not require further explanation. The QALY and PYLL outputs can be discounted if required and this can be defined by the user at the start of a modelling project. The discounting rate each year (Discount(year)) was calculated as shown in equation (0.33).

$$Discount(year) = \frac{1}{(1+R)^{year - year_{start}}} \tag{0.33}$$

Where, $year_{start}$ refers to the start year of the modelling which is 2018 in this study and R is the annual discount rate. The following sections describe some of these outputs in more detail.

Potential Years of Life Lost (PYLL)

The PYLL ($PYLL_i(year)$) for an individual *i* in the year "year" will be calculated from equation (0.34). Where $yearD_i$ refers to the year in which an individual *i* dies and $ageD_i$ refers to the age at which an individual dies.

$$PYLL_i(year) = \begin{cases} \sum_{a=ageD_i}^{a=LE_i-1} 1 * HealthDiscount(year + a - ageD_i) & \text{if } \mathrm{AgeD}_i < LE_i \\ 0 & \text{if } \mathrm{AgeD}_i \geq LE_i \end{cases} \tag{0.34}$$

For each individual (*i*) the number of years between the age of death and the life expectancy at birth ($LE_i(atbirth)$) will be calculated. $PYLL_i(year)$ is the total number of years of life lost (due to premature death) by individual *i* (if they died in year "year").

$$\overline{annualPYLL}(year) = \begin{cases} \frac{\sum_{i=0}^{N(year)} PYLL_i(year)}{N} \times 100{,}000 & \text{if } \; Age_{death} < LE \\ 0 & \text{if } \; Age_{death} \geq LE \end{cases} \tag{0.35}$$

The average annual PYLL will be calculated each year in the microsimulation. This metric will consider individuals who are alive in a given year (N(year)). For each individual the difference between the reference age (life expectancy) and the age of death will be calculated.

Potential years of life lost error

This error will be calculated in a similar way to the premature mortality costs. The costs are calculated by considering the error generated from each age in each year.

For each age *a*,

p_a = Rate of death at age a in a given year

d_a = Number of inviduals at age a that die in a given year

N_a = Total number of individuals who are alive in a given year

$$p_a = \frac{d_a}{N_a}$$

W_a = The weight per case at age a

$$W_a(y) = \sum_{a=ageD_i}^{a=LE_i-1} HealthDiscount(y + (a - ageD_i))$$

R = Rate of death at age a in the whole population in a given year

D = Total number of individuals who die in a given year

$$R = \frac{d_a}{D}$$

$$95\%\ \text{CI per 100,000 costs at age a} = 1.96\left(\sqrt{\frac{p_a(1-p_a)}{N_a}}\right) * W_a * 100000 \qquad (0.36)$$

Where the standard deviation (σ) is calculated as shown in equation

$$\sigma_a = \sqrt{\frac{p_a(1-p_a)}{N_a}} \qquad (0.37)$$

For each year the standard deviation for all age groups is calculated as shown in equation. A weighted average of the variance from each age group is calculated.

$$\sigma = \sqrt{\sum_{a=0}^{a=LE} RW_a{}^2 \sigma_a{}^2} \qquad (0.38)$$

The 95% CI for the premature mortality costs in a given year will be calculated from

$$\text{PMC 95\%CI per 100,000} = 1.96\sigma * 100000 \qquad (0.39)$$

Health Economic outputs

The following costs are produced by the model during the simulation:

- Direct healthcare costs
- Premature mortality costs
- Potential years of life lost
- Indirect costs
- NHS social care costs

Premature mortality costs (PMC)

$$PMC = \begin{cases} \sum_{i=age_{death}}^{i=LE-1} Income(i)*\text{Discount}(y+(i-age_{death}) & \text{if } age_{death} < LE \\ 0 & \text{if } age_{death} \geq LE \end{cases} \quad (0.40)$$

The premature mortality costs for each individual are calculated by summing over the income costs from the age of death until the LE.

$$\overline{annualPMC}(year) = \begin{cases} \frac{\sum_{i=0}^{N(year)} PMC(i)}{N} \times 100{,}000 & \text{if } age_{death} < LE \\ 0 & \text{if } age_{death} \geq LE \end{cases} \quad (0.41)$$

The average annual PYLL will be calculated each year in the microsimulation. This metric will consider individuals who are alive in a given year (N(year)). For each individual the difference between the reference age (life expectancy) and the age of death will be calculated.

Bibliography

1. Butland B, Jebb S, Kopelman P, McPherson K, Thomas S, Mardell J, et al. Foresight. Tackling obesities: future choices. Project report. Foresight Tackling obesities: future choices Project report. 2007.
2. Cancer Research UK, UK Health Forum. Short and sweet: why the Government should introduce a sugary drinks tax. 2016.