bhf.org.uk

British Heart Foundation

Modelling the potential impact of a reduction in salt consumption on hypertension, coronary heart disease and stroke in the population of the United Kingdom from 2021 to 2035

Appendix 2

Technical details of the microsimulation

HEALTH LUMEN

Prepared by

HEALTH LUMEN

Microsimulation framework

Our simulation consists of two modules. The first module calculates the predictions of risk factor trends over time based on data from rolling cross-sectional studies. The second module 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.

Microsimulation module one: predictions of risk factors over time

Salt consumption and systolic blood pressure are analysed within the model as risk factors (RF), as described in Table 1.

Table 1. Description of the categories used for the salt and blood pressure risk factors in England

Risk factor (RF)	Number of categories (N)	Categories
Grams of salt consumed per day	3	<6g 6-12g >12g
Systolic blood pressure	3	<120 mmHg 120-140 mmHg >140 mmHg

For the RF, let N be the number of categories for a given risk factor, e.g. $N = 3$ for SAL. Let $k = 1, 2, \ldots, N$ 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 (0.2) 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 for each risk factor

Measured data consist of sets of probabilities, with their variances, at specific time values (typically the year of the data were collected). For any particular time the sum of these probabilities is unity. Typically such data might be the probabilities of low, medium and high SAL intake, as they are extracted from the data set. Each data point is treated as a normally distributed[1] random variable; together they are a set of N groups (number of years: 18 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 SAL in the three states already mentioned, the $k = 0$ regression function represents the probability of low SAL exposure over time, $k = 1$ the probability of medium SAL exposure and $k = 2$ the probability of high SAL exposure.

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).

---

[1]In general, the assumption of a normal distribution is both extremely useful and accurate for both simple and complex surveys: indeed, for simple surveys the individual Bayesian prior and posterior 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.

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

$$p_k(\mathbf{a},\mathbf{b},t) \equiv \frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}}$$
$$\mathbf{a} \equiv (a_0,a_1,..,a_{K-1}), \quad \mathbf{b} \equiv (b_0,b_1,..,b_{K-1})$$
$$A_0 \equiv 0, \quad A_k \equiv a_k + b_k t \qquad (0.4)$$

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.

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

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

noting the relations

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

The values of the vectors **a**, **b** that satisfy these equations are denoted $\hat{\mathbf{a}}, \hat{\mathbf{b}}$. They provide the trend lines $p_k(\hat{\mathbf{a}},\hat{\mathbf{b}};t)$, 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

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

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) $\hat{p}_k(t)=p\left(\hat{\mathbf{a}},\hat{\mathbf{b}},t\right)$

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

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

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

$$\left(\nabla_{\hat{a}}\hat{p}_k(t), \nabla_{\hat{b}}\hat{p}_k(t)\right)^T = \left(\nabla_{\hat{a}}\hat{p}_k(t), \nabla_{\hat{b}}\hat{p}_k(t)\right) P \left(\nabla_{\hat{a}}\hat{p}_k(t), \nabla_{\hat{b}}\hat{p}_k(t)\right)^T$$

(0.9)

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

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

(0.10)

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

Module two: Microsimulation model

Microsimulation initialisation: birth, disease and death models

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) (3). 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 salt-related diseases. These diseases used the best available incidence, mortality, survival, relative risk and prevalence statistics (by age and gender). Individuals in the model are simulated from their year of birth (which may be before the start year of the simulation). In the course of their lives, simulated people can die from one of the diseases caused by excess salt consumption or from some other cause(s). 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 microsimulation incorporates a sophisticated economic module. The module employs a Markov-type simulation of long-term health benefits and health care costs. It synthesises and estimates evidence on cost-utility analysis. The model is used to project the differences in quality-adjusted life years (QALYs), and direct lifetime health-care costs over a specified time scale. The direct healthcare costs are presented separately in terms of hospital admissions, general practitioner costs, medication costs and social care costs. Outputs can be discounted for any specific discount rate.

This following section provides an overview of the main assumptions of the model.

Population models

Populations are implemented as instances of the TPopulation C++ class. The TPopulation class is created from a population (*.ppl) file. Usually a simulation will use only one population but it can simultaneously process multiple populations (for example, different ethnicities within a national population).

Population Editor

The Population Editor allows editing and testing of TPopulation objects. The population is created in the start-year and propagated forwards in time.

People within the model can die from specific diseases or from other causes. A disease file is created within the program to represent deaths from other causes. The following distributions are required by the population editor (**Table B**).

Table B Summary of the parameters representing the distribution component

Distribution name	symbol	note
Males by age by year	$p_m(a)$	Input in year 0 – probability of a male having age a
Females by age by year	$p_f(a)$	Input in year 0 – probability of a female having age a
Births by age of mother	$p_b(a)$	Input in year 0 – conditional probability of a birth at age a| the mother gives birth.
Number of births	$p_l(n)$	λ≡TFR, Poisson distribution, probability of giving birth to n children

Birth model

Any female in the child bearing years 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](4).

The probability that a mother (who does give birth) gives birth to a child at age a is determined from the births by age of mother 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,

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

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

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

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 $\lambda p_b(a)$,

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

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 $N_m/(N_m+N_f)$.

The Population editor' menu item Population Editor\Tools\Births\show random birthList creates an instance of the TPopulation class and uses it to generate and list a (selectable) sample of mothers and the years in which they give birth.

[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.

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.

The risk factor model

Overview of the HealthLumen assumptions on the salt risk factor module

Note that in the HealthLumen microsimulation, two risk factors are used: salt consumption and systolic blood pressure (SBP). More specifically, the microsimulation framework enables studies of salt consumption patterns and their influence on the systolic blood pressure. Changes in systolic blood pressure (ΔSBP) are dependent upon modifications in salt consumption level (ΔSAL). Based on the empirical studies discussed below, we assume that the relationship between these two variables is linear.

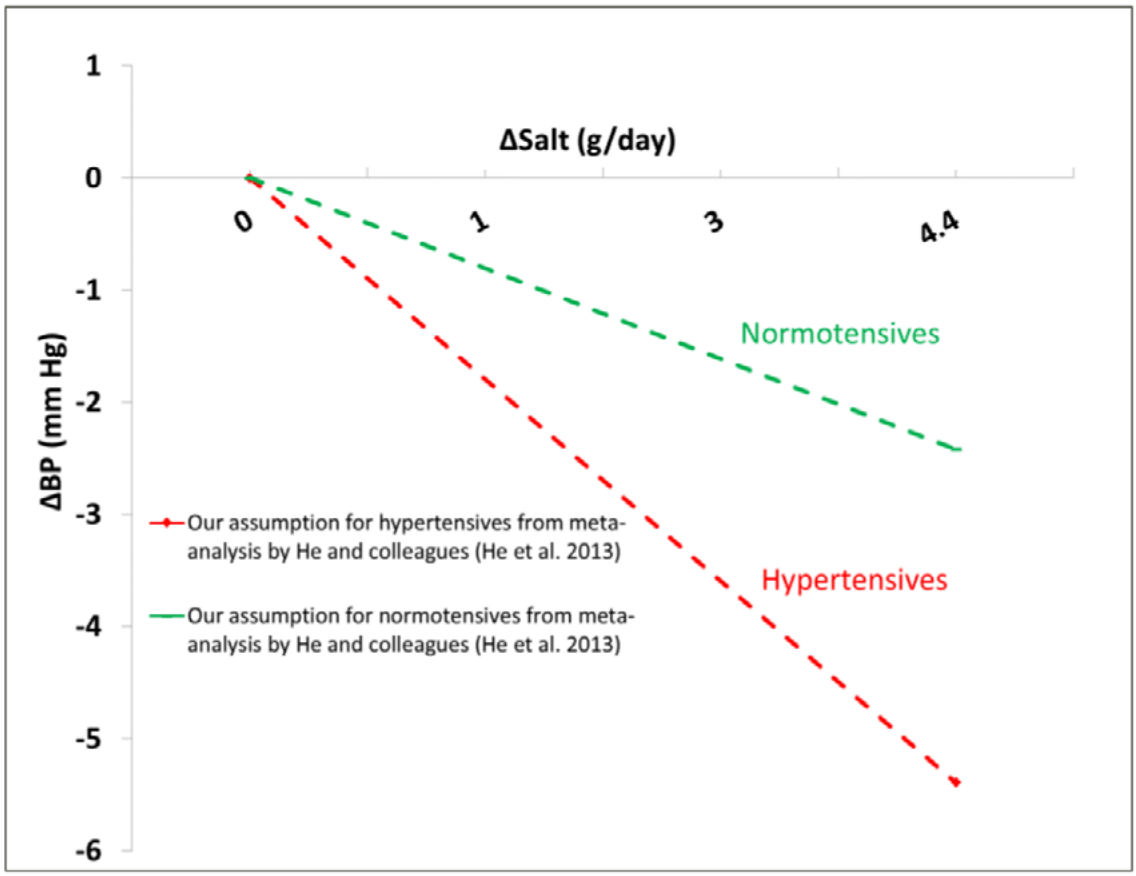
For all the population under the age of 65 we use the results from the meta-analysis by He and colleagues (1) which show that for hypertensive individuals 4.4 g/day reduction in salt intake (75 mmol sodium) resulted in 5.39 mmHg reduction in BP. For normotensive individuals the same reduction in salt would result in 2.42 mmHg reduction in BP. In addition we assume that if there are no changes in salt consumption this will incur no changes on the blood pressure. In both the hypertensive and normotensive cases straight lines have been approximated based on the blood pressure changes at 0 and 4.4 g/day.

It is empirically proven that for the same reduction in salt intake, greater fall in BP is seen in older individuals (2). To account for this fact we have agreed to assume that the population aged over 65 years have the same fall in BP with salt reduction as the hypertensive population. The relationship between changes in salt consumption level and blood pressure is summarised in Figure 1 and the equation below.

$\Delta BP = 0.55\Delta SAL$, BP ≤ 140 and age < 65

$\Delta BP = 1.225\Delta SAL$, BP > 140 or age ≥ 65

Figure 1. Assumed linear relationship between changes in salt consumption and blood pressure for both normotensives and hypertensives



We assume that by using this method, the effect of salt reduction on blood pressure will be conservative. Any influence from antihypertensive drug treatments on the changes in blood pressure with changing salt intake is not considered.

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 = {0-4, 5-9, ..., 70-74, 75+}. 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. Salt is modelled as a continuous risk factor.

Continuous risk factors

Salt consumption is modelled as a continuous variable.

In the case of a continuous RF such as salt consumption, 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

$$p_k(t|A,S) = \int_{\beta \in k} f(\beta|A,S,t)\, d\beta. \quad (0.14)$$

Equation (0.2) and (0.14) both refer to the same quantity. Equation (0.14) 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 for all $k = 0, \ldots, N$. The cumulative distribution function of β is

$$F(\beta|A,S,t) = \int_0^{\beta} f(\beta|A,S,t)\, d\beta. \quad (0.15)$$

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 our earlier work (1) is based on the assumption that a 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. 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 SAL distribution. The SAL 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 years older will at some time move into the next age group A' and will have a SAL 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 SAL exposure level of that individual is on the p-th percentile of the SAL distribution, their SAL exposure level will change from β to β' so that

$$\beta = F^{-1}\left(\frac{p}{100}|A,S,t\right) \quad (0.16)$$

$$\beta' = F^{-1}\left(\frac{p}{100}|A',S,t'\right) \Rightarrow \beta' = F^{-1}\left(F(\beta \mid A,S,t)|A',S,t'\right) \quad (0.17)$$

Where F^{-1} is the inverse of the cumulative distribution function of β, which we model with a continuous uniform distribution within the RF categories. Equation (0.17) 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 (0.17) as a suitably fast algorithm.

Relative risks

Suppose that α is a risk factor state of some risk factor A and denoted 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 (0.18).

$$p_A(d|\alpha,a,s) = \rho_{A|d}(\alpha|a,s)\, p_A(d|\alpha_0,a,s)$$
$$\rho_{A|d}(\alpha_0|a,s) \equiv 1 \qquad (0.18)$$

Where α_0 is the zero risk state.

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

$$p(d|a,s) = \sum_{\alpha} p_A(d|\alpha,a,s)\pi_A(\alpha|a,s)$$
$$= p_A(d|\alpha_0,a,s)\sum_{\alpha}\rho_{A|d}(\alpha|a,s)\pi_A(\alpha|a,s) \qquad (0.19)$$

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

$$p(d|\alpha,a,s) = \rho_{A|d}(\alpha|a,s)\frac{p(d|a,s)}{\sum_{\beta}\rho_{A|d}(\beta|a,s)\pi_A(\alpha|a,s)} \qquad (0.20)$$

Previous to any series of Monte Carlo trials the microsimulation program pre-processes the set of diseases and stores the *calibrated* incidence statistics $p_A(d|\alpha_0, a, s)$. These incidence statistics are calibrated to national level data. In this project the risk factor distributions and incidence risks for England are used to calculate the calibrated risks.

Modelling diseases

Disease modelling relies heavily on the sets of incidence, mortality, survival, relative risk and prevalence statistics. In some cases where a data set is unavailable or not available is the specified form for the model, data has been approximated from the known sets of the data.

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.

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.

Mortality statistics

In any year, in some population, in a sample of N people who have the disease a subset N_ω will die from the disease.

Mortality statistics record the cross sectional probabilities of death as a result of the disease – possibly stratifying by age

$$p_\omega = \frac{N_\omega}{N} \qquad (0.21)$$

Within some such subset N_ω of people that die in that year from the disease, the distribution by year-of-disease is not usually recorded. This distribution would be most useful. Consider two important idealised, special cases

Suppose the true probabilities of dying in the years after some age a_0 are $\{p_{\omega 0}, p_{\omega 1}, p_{\omega 2}, p_{\omega 3}, p_{\omega 4}\}$

The probability of being alive after N years is simply that you don't die in each year

$$p_{survive}(a_0 + N) = (1 - p_{\omega 0})(1 - p_{\omega 1})(1 - p_{\omega 2})..(1 - p_{\omega N-1}) \qquad (0.22)$$

Survival rates

It is common practice to describe survival in terms of a survival rate R, supposing an exponential death-distribution. In this formulation the probability of surviving t years from some time t_0 is given as

$$p_{\text{survival}}(t) = 1 - R\int_0^t du e^{-Ru} = e^{-Rt} \quad (0.23)$$

For a time period of 1 year

$$p_{survival}(1) = e^{-R}$$
$$\Rightarrow$$
$$R = -\ln\left(p_{survival}(1)\right) = -\ln(1 - p_\omega) \quad (0.24)$$

For a time period of, for example, 4 years,

$$p_{\text{survival}}(t=4) = 1 - R^{-1}\int_0^4 du e^{-Ru} = e^{-4R} = (1 - p_\omega)^4 \quad (0.25)$$

In short, the Rate is minus the natural log of the 1-year survival probability.

The survival models

For any potentially terminal disease the model can use any of the three survival models, numbered ((0, 1, 2)). The parameters describing these models are given below.

Survival model 0

A single probability of dying $\{p_{\omega 0}\}$, where $p_{\omega 0}$ is valid for all years. Given the 1-year survival probability $p_{survival}(1)$

The model uses 1 parameter ((R))

$$R = -\ln\left(p_{survival}(1)\right) \quad (0.26)$$

Survival model 1

Two different probabilities of dying $\{p_{\omega 0}, p_{\omega 1}\}$, where $p_{\omega 0}$ is valid for the first year; $p_{\omega 1}$ thereafter. The model uses two parameters ((p1, R)). Given the 1-year survival probability $p_{survival}(1)$ and the 5-year survival probability $p_{survival}(5)$

$$p_1 = 1 - p_{survival}(1)$$
$$R = -\frac{1}{4}\ln\left(\frac{p_{survival}(5)}{p_{survival}(1)}\right) \quad (0.27)$$

Survival model 2

Three different probabilities of dying $\{p_{\omega 0}, p_{\omega 1}, p_{\omega 5}\}$, where $p_{\omega 0}$ is valid for the first year; $p_{\omega 1}$ for the second to the fifth year; $p_{\omega 5}$ thereafter. The model uses three parameters ((p1, R, R>5))

Given the 1-year survival probability $p_{survival}(1)$ and the 5-year survival probability $p_{survival}(5)$

$$p_1 = 1 - p_{survival}(1)$$
$$R = -\frac{1}{4}\ln\left(\frac{p_{survival}(5)}{p_{survival}(1)}\right)$$
$$R_{>5} = -\frac{1}{5}\ln\left(\frac{p_{survival}(10)}{p_{survival}(5)}\right) \quad (0.28)$$

Different probabilities will apply to different age and gender groups. Typically the data might be divided into 10 year age groups.

Approximating missing disease statistics

A number of tools have been developed in the model in order to compute missing disease statistics data such as incidence or prevalence.

Approximating survival data from mortality and prevalence

An example is provided here with a standard life-table analysis for a disease *d*.

Consider the 4 following states:

state	Description
0	alive without disease *d*
1	alive with disease *d*
2	dead from disease *d*
3	dead from another disease

p_{ik} is the probability of disease *d* incidence, aged *k*

$p_{\omega k}$ is the probability of dying from the disease *d*, aged *k*

$p_{\bar{\omega} k}$ is the probability of dying other than from disease d, aged *k*

The state transition matrix is constructed as follows

$$\begin{bmatrix} p_0(k+1) \\ p_1(k+1) \\ p_2(k+1) \\ p_3(k+1) \end{bmatrix} = \begin{bmatrix} (1-p_{\bar{\omega}k})(1-p_{ik}) & (1-p_{\bar{\omega}k}-p_{\omega k})p_{\alpha k} & 0 & 0 \\ (1-p_{\bar{\omega}k})p_{ik} & (1-p_{\bar{\omega}k}-p_{\omega k})(1-p_{\alpha k}) & 0 & 0 \\ 0 & p_{\omega k} & 1 & 0 \\ p_{\bar{\omega}k} & p_{\bar{\omega}k} & 0 & 1 \end{bmatrix} \begin{bmatrix} p_0(k) \\ p_1(k) \\ p_2(k) \\ p_3(k) \end{bmatrix} \quad (0.29)$$

It is worth noting that the separate columns correctly sum to unity.

The disease mortality equation is that for state-2,

$$p_2(k+1) = p_{\omega k} p_1(k) + p_2(k) \quad (0.30)$$

The probability of dying from the disease in the age interval [*k*, *k*+1] is $p_{\omega k} p_1(k)$ - this is otherwise the (cross-sectional) disease mortality, $p_{mor}(k)$. $p_1(k)$ is otherwise known as the disease prevalence, $p_{pre}(k)$. Hence the relation

$$p_{\omega k} = \frac{p_{mor}(k)}{p_{pre}(k)} \quad (0.31)$$

For exponential survival probabilities the probability of dying from the disease in the age-interval [k, k+1] is denoted $p_{\Omega k}$ and is given by the formula

$$p_{\omega k} = 1 - e^{-R_k} \quad \Rightarrow \quad R_k = -\ln\left(1 - p_{\omega k}\right) \qquad (0.32)$$

When, as is the case for most cancers, these survival probabilities are known the microsimulation will use them, when they are not known or are too old to be any longer of any use, the microsimulation uses survival statistics inferred from the prevalence and mortality statistics (equation (0.31)). An alternative derivation equation (0.31) is as follows. Let N_k be the number of people in the population aged k and let n_k be the number of people in the population aged k with the disease. Then, the number of deaths from the disease of people aged k can be given in two ways: as $p_{\omega k} n_k$ and, equivalently, as $p_{mor}(k) N_k$. Observing that the disease prevalence is n_k / N_k leads to the equation

$$p_{\Omega k} n_k = p_{mor}\left(k\right) N_k$$

$$p_{pre}\left(k\right) = \frac{n_k}{N_k}$$

$$\Rightarrow$$

$$p_{\Omega k} = \frac{p_{mor}\left(k\right)}{p_{pre}\left(k\right)} \qquad (0.33)$$

References

1. He FJ, Li J, MacGregor GA. Effect of longer term modest salt reduction on blood pressure: Cochrane systematic review and meta-analysis of randomised trials. BMJ : British Medical Journal. 2013;346:f1325.
2. Cappuccio FP, Markandu ND, Carney C, Sagnella GA, MacGregor GA. Double-blind randomised trial of modest salt restriction in older people. Lancet. 1997;350(9081):850-4.
3. Office for National Statistics. Birth Summary Tables - England and Wales. 2016.
4. Office for National Statistics. Birth characteristics 2015 2016 [Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datasets/birthcharacteristicsinenglandandwales.

bhf.org.uk

British Heart Foundation