INTERNATIONAL JOURNAL OF MICROSIMULATION (2018) 11(3) 122-133

INTERNATIONAL MICROSIMULATION ASSOCIATION

Global Sensitivity Analysis of a Model Simulating an Individual's Health State through Their Lifetime
APPENDIX

Abbygail Jaccard, Lise Retat, Martin Brown†, Laura Webber, Zaid Chalabi

A TECHNICAL APPENDIX FOR THE INDIVIDUAL BASED MODEL

An individual *j* at age *a*, in year y has a risk factor (RF) value $rf_j(a, y)$ (e.g. BMI). The set of RF values for all possible integer ages is termed a RF trajectory $\{rf(a_0, y_0), rf(a_0+1, y_0+1), .., rf(a_{\max}, y_0 + a_{\max} - a_0)\}$. In this model, the RF trajectories are assumed to be static, so an individual's BMI value will not change over time. At any age, a person will be in one of many exclusive and exhaustive states. The state update equation (Equation A.1) is

$$p_{Si}(a+1, y+1, s) = \sum_{j=0}^{j=|S|-1} T_{ij}(a, y, s)\, p_{Sj}(a, y, s) \tag{A.1}$$

where, T is the state-transition matrix and $|S|$ are the total number of states. The set of states is complete so that, for all *a*, and *s*, for each year *y*, the probabilities of state membership are (Equation A.2)

$$\sum_{i=0}^{i=|S|-1} p_{Si}(a, y, s) = 1 \tag{A.2}$$

†This author is now deceased

where $p_{Si}(a, y, s)$ is the probability of being in state *i* at age *a* in year *y* with sex *s*. The possible health states range from alive with no disease, alive with a disease and dead. An individual can have a maximum of four different diseases at any one time. The probability that an individual acquires a disease *d* is calculated from the calibrated incidence $p_{Id}(a, y_0, s \mid rf_0)$ and relative risk (RR) $\rho^d_{RFj}(a, s)$ of the disease given the individual's BMI value. In Equation A.3 the disease incidence probabilities are given as

$$p_{Id} = p_{Id}(a, y_0, s \mid rf_j) = \rho^d_{rf_j}(a, s)\, p_{Id}(a, y_0, s \mid rf_0) \tag{A.3}$$

where the RR $\rho^d_{RFj}(a, s)$ is that appropriate to the RF group rf_j, which is the group identified by the element $rf(a, y, s)$ of the person's RF trajectory. This is assumed to hold for subsequent years.

The input disease incidence data are used to determine the probabilities of disease incidence for a zero-risk (RF group $0, rf_0$) person — the probability $p_{Id}(a, y_0 \mid rf_0)$. This is calculated as (Equation A.4)

$$p_{Id}(a_0, y_0, s \mid rf_0) = \frac{\bar{p}_{Id}(a_0, y_0, s)}{\sum_{j=0}^{j=|RF|-1} \rho_{RFj}(a_0, s)\, p_{RFj}(a_0, y_0, s)} \tag{A.4}$$

where, the probability of being in a RF group $p_{RFj}(a_0, y_0, s)$ is determined from the 5 year age and sex group trends.

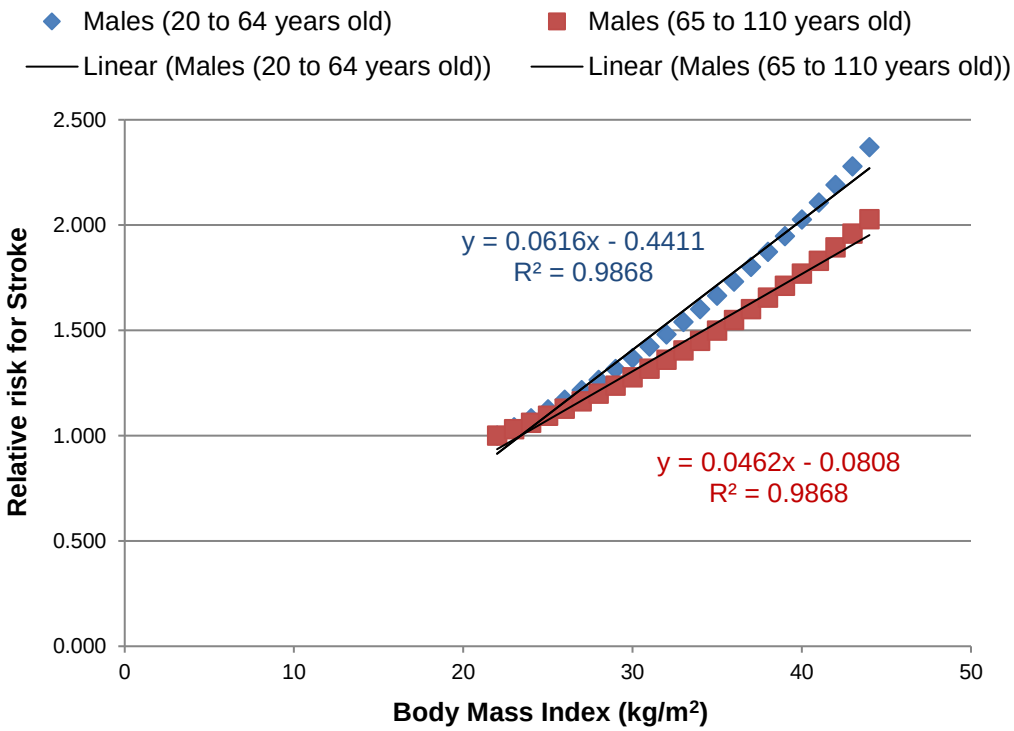
Each year an individual may die from a disease or from other causes. The probability that an individual dies from other causes is calculated from the total mortality rates by age and sex. These rates are available from the Office for National Statistics (ONS). The probability that an individual dies from a specific disease is calculated from the survival probabilities. If an individual has multiple disease their probability of dying is calculated as shown in Equation (A.5).

$$\begin{aligned} p_{\omega 1} &= p_{\Omega 0}(1 - p_{\Omega 1}) + p_{\Omega 1}(1 - p_{\Omega 0}) \\ p_{\omega 2} &= p_{\Omega 0}(1 - p_{\Omega 2}) + p_{\Omega 2}(1 - p_{\Omega 0}) \\ p_{\omega 12} &= p_{\Omega 0}(1 - p_{\Omega 1})(1 - p_{\Omega 2}) + p_{\Omega 1}(1 - p_{\Omega 2})(1 - p_{\Omega 0}) + p_{\Omega 2}(1 - p_{\Omega 0})(1 - p_{\Omega 1}) \end{aligned} \tag{A.5}$$

B ESTIMATING THE RELATIONSHIP BETWEEN RELATIVE RISKS OF COLORECTAL CANCER AND STROKE BY BODY MASS INDEX

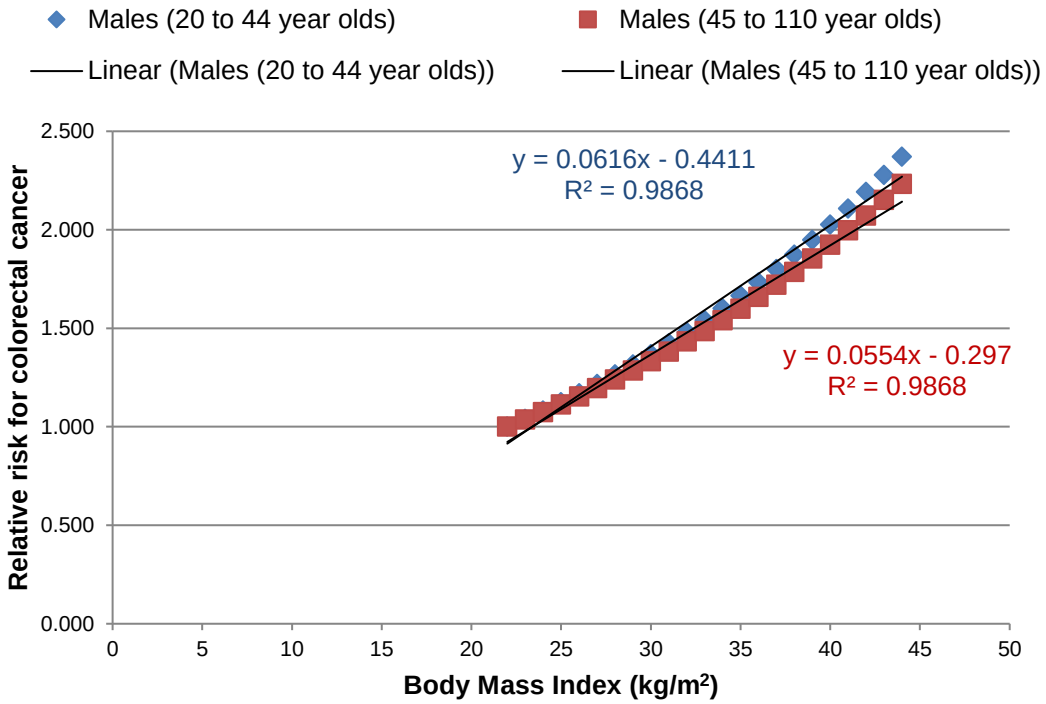
The RR for colorectal cancer and stroke were estimated from the DYNAMO-HIA project (World Obesity Federation, 2008). An equation was provided for both RRs, which described how the RR could be approximated for each BMI value for two different age groups. For each disease, the RR was plotted against the BMI value for the two different age groups (Figures B.1 and B.2). A linear equation was fitted to each age group for each disease.

Figure B.1: A graph illustrating the relationship between the RR for stroke and BMI for males for two different age groups.



Notes: Both plots were fitted with a linear equation.

Figure B.2: A graph illustrating the relationship between the RR for colorectal cancer and BMI for males for two different age groups.



Notes: Both plots were fitted with a linear equation.

A mean RR was calculated for the overweight (25-30 kg/m^2) and obese (30-45 kg/m^2) groups for each age group and each disease. A uniform distribution was assumed for each mean. The upper and lower bounds were calculated from the mean using the standard deviation from Guh et al. (2009).

Due to the dependency of BMI on the RR constraints were used within PSUADE. These constraints were calculated by taking the average of the two linear functions for each age group for each disease.

C COMPUTING BMI TRENDS

BMI is analysed within the model as a RF, as described in Table C.1.

Table C.1: Description of the categories used for the RF BMI.

Risk factor (RF)	Number of categories (N)	Categories
BMI	3	BMI < 25 kg/m^2 (normal weight)
		BMI from 25 to 29.99 kg/m^{-2} (overweight)
		BMI ≥ 30 kg/m^2 (obesity)

For the RF, let N be the number of categories for a given RF, e.g. $N = 3$ for BMI. 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 (Equation C.1)

$$\ln\left(\frac{p_k(t)}{p_1(t)}\right) = \beta_0^k + \beta_1^k t \tag{C.1}$$

The prevalence of the first category is obtained by using the normalisation constraint $\sum_{k=1}^{N} p_k(t) = 1$. Solving Equation C.2 for $p_k(t)$, we obtain

$$p_k(t) = \frac{\exp\left(\beta_0^k + \beta_1^k t\right)}{1 + \sum_{k'=1}^{N} \exp\left(\beta_0^{k'} + \beta_1^{k'} t\right)}, \tag{C.2}$$

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

C.1 Multinomial logistic regression for risk factor BMI

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 normal weight, overweight and obese as they are extracted from the survey data set. Each data point is treated as a normally distributed random variable; together they are a set of N groups (number of years) of K probabilities $\{\{t_i, \mu_{ki}, \sigma_{ki} \mid k \in [0,K\text{-}1]\} \mid i \in [0,N\text{-}1]\}$. 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 $\{p_k(\mathbf{a}, \mathbf{b}, t) \mid k \in [0,K\text{-}1]\}$ 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 normal weight over time, $k = 1$ the probability of being overweight, and $k = 2$ the probability of being obese.

The regression equations are most easily derived from a familiar least square minimization. In the following equation set (Equations C.1.1 and C.1.2) the weighted difference between the measured and predicted probabilities is written as S and 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.

$$S(\mathbf{a},\mathbf{b}) = \tfrac{1}{2}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1}\frac{\left(p_k(\mathbf{a},\mathbf{b};t_i)-\mu_{ki}\right)^2}{\sigma_{ki}^2} \tag{C.1.1}$$

$$\begin{gathered} p_k(\mathbf{a},\mathbf{b},t) \equiv \frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}} \\ \mathbf{a} \equiv (a_0,a_1,..,a_{K-1}), \quad \mathbf{b} \equiv (b_0,b_1,..,b_{K\text{-}1}) \\ A_0 \equiv 0, \quad A_k \equiv a_k + b_k t \end{gathered} \tag{C.1.2}$$

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 is a real number lying between 0 and 1. The minimum of the function S is determined from the Equations C.1.3. and C.1.4:

$$\frac{\partial S}{\partial a_j} = \frac{\partial S}{\partial b_j} = 0 \qquad \text{for j=1,2,....,k-1} \tag{C.1.3}$$

noting the relations

$$\begin{gathered} \frac{\partial p_k}{\partial A_j} = \frac{\partial}{\partial A_j}\left(\frac{e^{A_k}}{1+e^{A_1}+..+e^{A_{K-1}}}\right) = p_k\delta_{kj} - p_k p_j \\ \frac{\partial}{\partial a_j} = \frac{\partial}{\partial A_j} \\ \frac{\partial}{\partial b_j} = t\frac{\partial}{\partial A_j} \end{gathered} \tag{C.1.4}$$

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.

C.2 Bayesian interpretation

The $2K$-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 (2K-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 $S(\mathbf{a},\mathbf{b})$ 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 $\hat{\mathbf{S}} \equiv S(\hat{\mathbf{a}},\hat{\mathbf{b}})$. Hence the Equation C.2.1 is written as follows:

$$\begin{aligned} S(\mathbf{a},\mathbf{b}) &= \tfrac{1}{2}\sum_{k=0}^{k=K-1}\sum_{i=0}^{i=N-1}\frac{\left(p_k(\mathbf{a},\mathbf{b};t_i)-\mu_{ki}\right)^2}{\sigma_{ki}^2} \\ &\equiv S(\hat{a},\hat{b})+\tfrac{1}{2}(a-\hat{a},b-\hat{b})P^{-1}(a-\hat{a},b-\hat{b})+\ldots \\ &\approx S(\hat{a},\hat{b})+\tfrac{1}{2}\sum_{i,j}(a_i-\hat{a}_i)\frac{\partial^2\hat{S}}{\partial\hat{a}_i\partial\hat{a}_j}(a_j-\hat{a}_j)+\tfrac{1}{2}\sum_{i,j}(a_i-\hat{a}_i)\frac{\partial^2\hat{S}}{\partial\hat{a}_i\partial\hat{b}_j}(b_j-\hat{b}_j)+ \\ &+\tfrac{1}{2}\sum_{i,j}(b_i-\hat{b}_i)\frac{\partial^2\hat{S}}{\partial\hat{b}_i\partial\hat{a}_j}(a_j-\hat{a}_j)+\tfrac{1}{2}\sum_{i,j}(b_i-\hat{b}_i)\frac{\partial^2\hat{S}}{\partial\hat{b}_i\partial\hat{b}_j}(b_j-\hat{b}_j) \end{aligned} \quad \text{(C.2.1)}$$

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.

C.3 Estimation of the confidence intervals

The logistic regression functions $p_k(t)$ can be approximated as a normally distributed time-varying random variable $N(\hat{p}_k(t),\sigma_k^2(t))$ by expanding p_k about its maximum likelihood estimate (the trend line) $\hat{p}_k(t)=p(\hat{\mathbf{a}},\hat{\mathbf{b}},t)$ (Equation C.3.1).

$$\begin{aligned} p_k(\mathbf{a},\mathbf{b},t) &= p_k(\hat{\mathbf{a}}+\mathbf{a}-\hat{\mathbf{a}},\hat{\mathbf{b}}+\mathbf{b}-\hat{\mathbf{b}},t) \\ &= \hat{p}_k(t)+(\nabla_{\hat{a}},\nabla_{\hat{b}})\hat{p}_k(t)\begin{pmatrix}\mathbf{a}-\hat{\mathbf{a}}\\ \mathbf{b}-\hat{\mathbf{b}}\end{pmatrix}+\ldots \end{aligned} \quad \text{(C.3.1)}$$

Denoting mean values by angled brackets, the variance of p_k is thereby approximated as (Equation C.3.2):

$$\begin{aligned} &\sigma_k^2(t)\equiv\left\langle\left(p_k(\mathbf{a},\mathbf{b},t)-\hat{p}_k(t)\right)^2\right\rangle=\left(\nabla_{\hat{a}}\hat{p}_k(t),\nabla_{\hat{b}}\hat{p}_k(t)\right)\left\langle\begin{pmatrix}\mathbf{a}-\hat{\mathbf{a}}\\ \mathbf{b}-\hat{\mathbf{b}}\end{pmatrix}\begin{pmatrix}\mathbf{a}-\hat{\mathbf{a}}\\ \mathbf{b}-\hat{\mathbf{b}}\end{pmatrix}^T\right\rangle\times \\ &\left(\nabla_{\hat{a}}\hat{p}_k(t),\nabla_{\hat{b}}\hat{p}_k(t)\right)^T=\left(\nabla_{\hat{a}}\hat{p}_k(t),\nabla_{\hat{b}}\hat{p}_k(t)\right)P\left(\nabla_{\hat{a}}\hat{p}_k(t),\nabla_{\hat{b}}\hat{p}_k(t)\right)^T \end{aligned} \quad \text{(C.3.2)}$$

When K=3 this equation can be written as the 4-dimensional inner product (Equation C.3.3):

$$\sigma_k^2(t) = \begin{pmatrix} \frac{\partial \hat{p}_k(t)}{\partial \hat{a}_1} & \frac{\partial \hat{p}_k(t)}{\partial \hat{a}_2} & \frac{\partial \hat{p}_k(t)}{\partial \hat{b}_1} & \frac{\partial \hat{p}_k(t)}{\partial \hat{b}_2} \end{pmatrix} \begin{bmatrix} P_{aa11} & P_{aa12} & P_{ab11} & P_{ab12} \\ P_{aa21} & P_{aa22} & P_{ab21} & P_{ab22} \\ P_{ba11} & P_{ba12} & P_{bb11} & P_{bb12} \\ P_{ba21} & P_{ba22} & P_{bb21} & P_{bb22} \end{bmatrix} \begin{pmatrix} \frac{\partial \hat{p}_k(t)}{\partial \hat{a}_1} \\ \frac{\partial \hat{p}_k(t)}{\partial \hat{a}_2} \\ \frac{\partial \hat{p}_k(t)}{\partial \hat{b}_1} \\ \frac{\partial \hat{p}_k(t)}{\partial \hat{b}_2} \end{pmatrix} \quad \text{(C.3.3)}$$

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 $\left[\hat{p}_k(t) - 1.96\sigma_k(t), \mathrm{p_k}(t) + 1.96\sigma_k(t)\right]$.

DISEASE EPIDEMIOLOGICAL DATA SOURCES

Disease	Incidence	Prevalence	Mortality	Survival	Relative Risk	Utility Weight
CHD	Smolina et al 2012. Corrected data on incidence and mortality in 2013 (Smolina, Wright, Rayner, & Goldacre, 2012)	BHF, Cardiovascular Disease Statistics 2014 (British Heart Foundation, 2015)	ONS, Deaths Registrations Summary Statistics, England and Wales, 2014 (Office for National Statistics, 2014)	Computed from prevalence and mortality	World Obesity Federation (DYNAMO project) (World Obesity Federation)	Laires et al. 2015 (Laires, Ejzykowicz, Hsu, Ambegaonkar, & Davies, 2015)
Stroke	BHF, stroke statistics 2009 (British Heart Foundation, 2009)	BHF, Cardiovascular Disease Statistics 2014 (British Heart Foundation, 2015)	ONS, Deaths Registrations Summary Statistics, England and Wales, 2014 (Office for National Statistics, 2014)	Computed from prevalence and mortality	World Obesity Federation (DYNAMO project) (World Obesity Federation)	Rivero-Arias et al. 2010 (Rivero-Arias et al., 2010)
Hypertension	Derived from prevalence	Health Survey for England 2012 (Health and Social Care Information Centre, 2012)	non terminal	non terminal	World Obesity Federation (DYNAMO project) (World Obesity Federation)	Sullivan et al. 2011 (Sullivan, Slejko, Sculpher, & Ghushchyan, 2011)
Diabetes	Personal communication Dr. Craig Currie at Cardiff University	National Diabetes Audit 2015-2016(NHS Digital, 2017)	non terminal	non terminal	World Obesity Federation (DYNAMO project) (World Obesity Federation)	Sullivan et al. 2011 (Sullivan et al., 2011)
Colorectal cancer	CRUK, 2013 Statistics by cancer type (Cancer Research UK,	NA	CRUK Mortality by cancer type(Cancer Research UK, 2016a)	ONS Cancer Survival in England: adults diagnosed between 2009 and	World Obesity Federation (DYNAMO project) (World	Sullivan et al. 2011 (Sullivan et al., 2011)

2016b)	2013 and followed up to 2014 (Office for National Statistics, 2015) & ONS Cancer Survival in England: 10 year survival rates adults diagnosed between 2010-2011 and followed up to 2012 (Office for National Statistics, 2013)	Obesity Federation)

REFERENCES

Cancer Research UK. (2016a). Statistics by cancer type - Average Number of Deaths per Year and Age-Specific Mortality Rates, UK, 2010-2012. Retrieved from http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type

Cancer Research UK. (2016b). Statistics by cancer type - Average Number of New Cases Per Year and Age-Specific Incidence Rates per 100,000 Population, UK 2011-2013 from http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type

Guh, D. P., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C. L., & Anis, A. H. (2009). The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC Public Health, 9*(1), 88.

Health and Social Care Information Centre. (2012). *Health Survey for England 2012*. Retrieved from https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/health-survey-for-england-2012

Laires, P. A., Ejzykowicz, F., Hsu, T.-Y., Ambegaonkar, B., & Davies, G. (2015). Cost–effectiveness of adding ezetimibe to atorvastatin vs switching to rosuvastatin therapy in Portugal. *Journal of Medical Economics, 18*(8), 565-572.

NHS Digital. (2017). *National Diabetes Audit 2015/2016*. Retrieved from http://www.content.digital.nhs.uk/catalogue/PUB23241

Office for National Statistics. (2013). Cancer Survival in England: 10 year survival rates adults diagnosed between 2010-2011 and followed up to 2012. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancersurvivalinenglandadultsdiagnosed/2010and2014andfollowedupto2015

Office for National Statistics. (2014). Deaths Registrations Summary Statistics, England and Wales. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathregistrationssummarytablesenglandandwalesreferencetables

Office for National Statistics. (2015). Cancer Survival in England- Adults Diagnosed: 2009 to 2013, followed up to 2014. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancersurvivalinenglandadultsdiagnosed/2009to2013followedupto2014

Rivero-Arias, O., Ouellet, M., Gray, A., Wolstenholme, J., Rothwell, P. M., & Luengo-Fernandez, R. (2010). Mapping the modified Rankin scale (mRS) measurement into the generic EuroQol (EQ-5D) health outcome. *Medical Decision Making, 30*(3), 341–354.

Scarborough, P., Peto V., Bhatnagar, P., Kaur A., Leal J., Luengo-Fernandez, R. … Allender, S. Stroke Statistics 2009 edition. British Heart Foundation Statistics Database, www.heartstats.org

Smolina, K., Wright, F. L., Rayner, M., & Goldacre, M. J. (2012). Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study. Corrected data on incidence and mortality in 2013 at http://www.bmj.com/content/347/bmj.f7379.abstract. *BMJ, 344*, d8059. doi: 10.1136/bmj.d8059

Sullivan, P. W., Slejko, J. F., Sculpher, M. J., & Ghushchyan, V. (2011). Catalogue of EQ-5D scores for the United Kingdom. *Medical Decision Making, 31*(6), 800-804.

Townsend, N., Williams, J., Bhatnagar, P., Wickramasinghe, K., & Rayner, M. Cardiovascular Disease statistics 2014. British Heart Foundation. https://www.bhf.org.uk/informationsupport/publications/statistics/cardiovascular-disease-statistics-2014

World Obesity Federation. *Relative risk Assessments IASO; Prepared for DYNAMO-HIA project.* Retrieved from http://www.worldobesity.org/site_media/uploads/Appendix_Relative_Risk_Assessments_IASO.pdf

World Obesity Federation. (2008). *Relative Risk assessments: Prepared for the Dynamo-HIA project.* Retrieved from https://s3.eu-central-1.amazonaws.com/ps-wof-web-dev/site_media/uploads/Appendix_Relative_Risk_Assessments_IASO.pdf