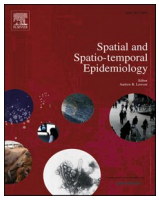
Spatial and Spatio-temporal Epidemiology 48 (2024) 100633

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



Chronic back pain prevalence at small area level in England - the design and validation of a 2-stage static spatial microsimulation model

Harrison Smalley *, Kimberley Edwards

School of Medicine, Queens Medical Centre, University of Nottingham, Nottingham, United Kingdom

ARTICLE INFO

Keywords:
Spatial microsimulation
Deterministic reweighting
Chronic back pain
Health inequality
Small area estimation
England

ABSTRACT

Spatially disaggregated estimates provide valuable insights into the nature of a disease. They highlight inequalities, aid public health planning and identify avenues for further research. Spatial microsimulation is advantageous in that it can be used to create large microdata sets with intact microlevel relationships between variables, which allows analysis of relationships between variables locally. This methodological paper outlines the design and validation of a 2-stage static spatial microsimulation model for chronic back pain prevalence across England, suitable for policy modelling. Data used was obtained from the Health Survey for England and the 2011 Census. Microsimulation was performed using SimObesity, a previously validated static deterministic program, and the synthetic chronic back pain microdataset was internally validated. The paper also highlights modelling considerations for researchers embarking on similar work, as well as future directions for research in this area of microsimulation.

1. Background

Spatial epidemiology is a field of epidemiology focusing on the spatial patterns and processes of health and disease. It is closely related to health geography, a field of human geography that utilises geographic information for the study of health, disease and health services. Single global estimates of disease, for example, national prevalence, are useful but possess different benefits to spatially disaggregated estimates, for example, the prevalence of a disease in each region within a nation. Possessing spatially disaggregated estimates allows for the study of the spatial pattern of a disease, increasing understanding of the processes underlying it, highlighting inequalities, enabling public health planning and identifying avenues for further research (Edwards and Clarke, 2009; D Ballas et al., 2006; Procter et al., 2008). When working with spatially disaggregated data it is important to consider how finely estimates are disaggregated (if at all) depending on the study question attempting to be answered. That is, the spatial scale at which the disease is studied, is an important consideration. If the scale is too large, processes acting at a smaller spatial scale will be unappreciable, for example studying a disease at county level may show no variation across a nation despite clusters of the disease existing around city centres. Smaller scales can identify these more local variations and allow finer gradients to be established. However, too small a scale may bring difficulties due to confidentiality, resource intensity and small number issues (Twigg et al., 2000; Pearce et al., 2003).

A spatially disaggregated dataset can be obtained from primary data collection or via modelling. Primary data collection at a small area level over a large study area, e.g., a country, is extremely resource-intensive. A large total sample size would be required to accurately represent each small area. Small area estimation (SAE) provides a solution to this. There are two commonly used methods of SAE – statistical e.g. regression modelling or geographical (spatial microsimulation) (Koh et al., 2018). A statistical example relevant to chronic back pain (CBP) is the Versus Arthritis Musculoskeletal Calculator (Versus Arthritis, 2021) which uses logistic regression modelling (Adomaviciute et al., 2018). This approach can prove useful but only provides an output for the outcome variable. It is therefore not possible to analyse the relationship between the outcome variable and its covariates at a small area level or to predict the impact of modification of such variables.

Spatial microsimulation (SMS) is a category of microsimulation used to create large spatially disaggregated microdata sets containing variables of interest for which previously only aggregate data was possessed (Tanton and Clarke, 2014). Unlike statistical approaches to SAE, SMS produces a microdata set/ synthetic population. This is a dataset approximately equal in size to and representative of the true population where each synthetic individual is characterised by all the variables

* Corresponding author.
E-mail address: harrison.smalley@nottingham.ac.uk (H. Smalley).

https://doi.org/10.1016/j.sste.2023.100633
Received 13 March 2023; Received in revised form 20 December 2023; Accepted 30 December 2023
Available online 31 December 2023

entered into the model. Relationships between variables are intact at the microlevel. As such, local relationships between variables can be analysed. As well as SAE, SMS can be used for small area projection and small area policy modelling (Tanton and Edwards, 2012). The multivariate nature of SMS applies well to the study of chronic disease as many chronic diseases are multifactorial. However, the use of SMS in the field of health is still growing. Its uses have included the study of obesity (Edwards and Clarke, 2009; Koh et al., 2018), COVID-19 (Spooner et al., 2021), osteoarthritis (Ifesemen et al., 2019), health behaviours (Smith et al., 2011) and health inequality (Campbell and Ballas, 2016; D Ballas et al., 2006).

Various methods of SMS exist and have been discussed in detail by Tanton (Tanton, 2013). They typically involve taking a national survey dataset which includes variables of interest, and a census dataset, which includes local demographic data and matching the survey data to the census data. Individuals from the survey are allocated to geographic areas based on how well their characteristics fit the demographics of that area. The demographic variables on which individuals are matched to areas are known as constraint variables. Additional (non-constraint) variables can be included from the survey dataset to create a synthetic population in which each individual is characterised by more than just the constraint variables. These non-constraint variables do not dictate the matching process in any way but can be used in later analysis.

SimObesity (Edwards and Clarke, 2009) the SMS program used in this study utilises a static deterministic method which is described briefly below and in more detail by Edwards et al. (Edwards and Clarke, 2012). The main advantage of a deterministic method is that the same result is produced each time the model is run for a given input dataset. This means that any change seen in the output is a result of the change in the input rather than occurring due to chance. This is a particularly relevant consideration when performing what-if analyses.

CBP carries a significant burden in England and worldwide (Bridges, 2012; Hoy et al., 2014). CBP is typically defined as constant or intermittent pain in the posterior thoracic (upper back) or posterior lumbar (lower back) region lasting 3 months or more (The World Health Organization 2020). Back pain can be broadly categorised as specific where the pain is caused by a known pathophysiological mechanism, for example, a sprain/strain, disc herniation or fracture, or non-specific in which there is no known pathophysiological mechanism. Non-specific back pain is a diagnosis of exclusion and is commonly cited as making up 90 % of low back pain cases (Maher et al., 2017). Back pain is more common in women and black individuals (Fillingim et al., 2009; Knox et al., 2012; Carey et al., 2010). Its prevalence increases up to the age of 60 and then decreases (Hoy et al., 2012; Björck-van Dijken et al., 2008; Papageorgiou et al., 1995). Prevalence is also higher amongst those with occupations that involve a high physical workload as well as in those with a lower educational status (Björck-van Dijken et al., 2008; Coenen et al., 2014; Lötters et al., 2003; Costa-Black et al., 2010; Dionne et al., 2001). Smoking and obesity have also both been associated with an increased risk of back pain (R Shiri et al., 2010; Green et al., 2016; Alkherayf et al., 2010; R Shiri et al., 2010; Leboeuf-Yde, 2000). Low physical activity (PA) is another possible risk factor and presents a target for consideration by public health planners (Shiri and Falah-Hassani, 2017). Producing a validated model for CBP prevalence across England, with the ability to simulate the effect of policies to increase PA, could lead to an enhanced understanding of the disease and aid public health planning.

This paper outlines the development and validation of a static two-stage spatial microsimulation model for chronic back pain prevalence across England.

2. Method

A two stage SMS approach using SimObesity (Edwards and Clarke, 2009) was adopted as seen in work by Ifesemen et al. (Ifesemen et al., 2019). This approach was chosen to enable later 'what-if' analysis, due to requirements for policy-useful constraint variables. Stage 1 matched individuals from a national survey dataset, containing PA, to wards of the 2011 Census based on shared demographic data. Stage 2 then matched individuals from a national survey dataset containing chronic back pain (CBP) and PA to the stage 1 output PA geofile. See Fig. 1.

2.1. Data sources

This study used data from the Health Survey for England (HSE), obtained from the UK Data Service (University College London 2017) and the UK Census, obtained from InFuse (Office for National Statistics 2017). The HSE is an annual cross-sectional survey of the health of adults and children living in England (University College London 2017). This simulation used HSE years 2013 (NatCen Social Research 2015), 2014 (NatCen Social Research 2018), 2015 (NatCen Social Research 2019) and 2017 (University College London 2017). It is common in microsimulation methodology to amalgamate multiple years of a survey to achieve a larger sample size. However, doing so depends on the relevance of year to the outcome of interest, and data used must be present in all years. The HSE has a set of core questions and then has additional questions which vary annually depending on the chosen focus for that year. The 2017 HSE is the only year containing both CBP and PA data needed for the stage 2 simulation, whilst the 2013, 2014 and 2015 HSEs provided data on PA for the stage 1 simulation. PA data is in minutes of moderate-to-vigorous physical activity per week (MVPA), whilst CBP is binary. The 2011 UK Census (Office for National Statistics 2017) provided small area level demographic data for the stage 1 simulation (the 2021 Census was not available at the time of this analysis). The Census is performed every 10 years and gathers data on the UK population required for resource planning and allocation. The Census surveys households and presents the results aggregated to various levels. The smallest level is the output area, which are formed from combining adjacent postcodes (Office for National Statistics 2021). Estimates for all larger area sizes are produced by best-fitting from output areas to the level desired (Office for National Statistics 2021). This study used data at the census merged ward level.

2.2. Data handling

The four HSE datasets were downloaded from the UK Data Service (University College London 2017). The datasets were reduced to variables of interest only. For the stage 1 dataset, the three years were merged and variables of interest not present or inconsistently defined across all three years were removed. Variables required from the 2011 Census were selected and downloaded from InFuse (Office for National Statistics 2017). Where necessary, potential constraint variables were manipulated in the HSE and/or Census datasets by collapsing categories to achieve consistency between the HSE and census definitions. For variables with >5 % missing data, missing values were imputed in R version 4.1.0 (R Core Team 2021) using the MICE (Multivariate Imputation by Chained Equations) package (van Buuren and Groothuis-Oudshoorn, 2011). For variables with <5 % missing data, missing values were dealt with via listwise deletion. This method was chosen as a less resource-intensive option, compared with imputing all missing values. Non-spatial statistical analyses were performed using IBM SPSS Statistics version 27 (Corp, 2020).

2.3. Model selection

A shortlist of potential constraint variables was determined from a literature review and was based on their ability to predict CBP and PA. The strength of these predictors in the HSE datasets was then assessed further using regression analysis. For the PA simulation, shortlisted constraint variables in the stage 1 dataset first underwent univariate multinomial logistic regression. Statistically insignificant and variables with small effect sizes were discarded. To assess for collinearity,

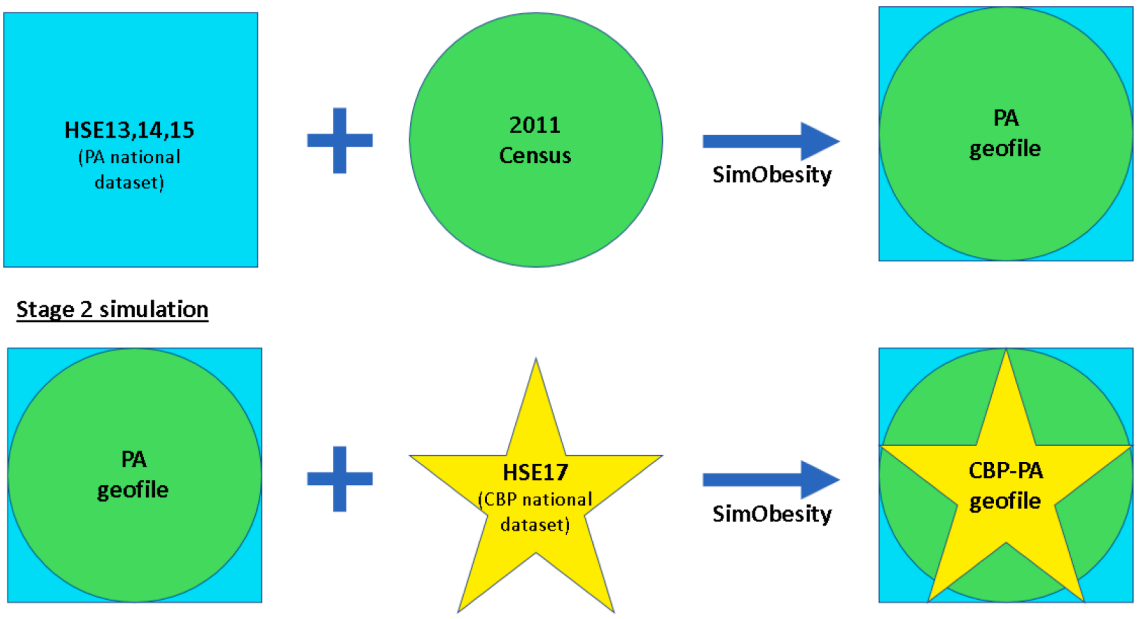


Fig. 1. Two-stage spatial microsimulation process.

pairwise independence was examined using the chi-squared test. The remaining variables then underwent multivariate multinomial logistic regression, with the most predictive variables selected. This process was repeated for the stage 2 dataset using univariate then multivariate binary logistic regression. To determine which combination of constraints simulated most accurately, several simulations were ran and internally validated (see below).

2.4. Simulation - SimObesity

The algorithms used in SimObesity have been described in detail by Edwards et al. (Edwards and Clarke, 2012). In brief, there are two main stages involved in estimation using SimObesity. Firstly, for each area a deterministic reweighting algorithm is used to alter the weights applied to individuals in the survey dataset so that the aggregate of the survey data matches the census data for that area. This is done for each of the constraint variables. Following this, the weights are converted to integers, using a cumulative process on ranked weights, so that the output simulated dataset is comprised of whole individuals only.

2.5. Validation

Validation was undertaken to determine whether the simulated dataset was representative of the real population (at ward level). This can be achieved by comparing values from the simulated dataset to corresponding known values in the real population. It is essential that all SMS models are internally validated (Edwards et al., 2011). In this study, internal validation was achieved by plotting scatter plots for each category of each constraint variable, with the simulated prevalence on the y-axis and the census (real) prevalence on the x-axis. A simulated dataset identical to the census dataset should result in a regression line $y = x$. Then, for each constraint variable total absolute error (TAE) of all wards was calculated and divided by the number of wards to give mean absolute error (MAE) (Timmins and Edwards, 2016; Lovelace et al., 2015). Finally, the total number of simulated areas for each of the constraint categories with >5 % error (E5) and >10 % error (E10) was also calculated (Timmins and Edwards, 2016; Lovelace et al., 2015).

3. Results

3.1. Input datasets (HSE and census)

3.1.1. Missing values

The amalgamated HSE 2013–2015 (PA) dataset contained responses from 24,906 adults. Most variables had a small amount of missing data (<1 %). National Statistics Socio-economic Classification (NS-SEC) had 1.9 % and alcohol intake had 2.7 % missing data. These cases were dealt with via listwise deletion. 1240 individuals were deleted. Variables with large amounts of missing data (14.1–16.7 %) included variables related to PA, sedentary time and BMI. These missing values were imputed. The 2017 HSE (CBP) dataset contained responses from 7997 adults. Most variables had <1 % missing data. NS-SEC had 1.9 % and alcohol intake had 2.5 %. These cases were dealt with via listwise deletion. 419 individuals were deleted. Variables with large amounts of missing data (9.2–18.3 %) included variables related to PA, sedentary time, BMI, anxiety/depression and life satisfaction. These missing values were imputed.

3.1.2. Potential constraint variables

Most potential constraints were defined in the same way in the Census as the HSEs. These variables were categorised in the same way or were able to be matched by collapsing categories. The only exception was the variable disability, which had to be excluded. See Fig. 2.

3.1.3. Non-constraint variables

Most of the extra variables intended for inclusion were present and defined consistently over all three years for the PA dataset. Fruit and vegetable consumption, cardiovascular disease, anxiety/depression and life satisfaction were exceptions to this (Fig. 3) and so could not be included in the stage 1 simulation. All extra variables intended for inclusion, present in the 2017 survey (CBP dataset), were still included in the stage 2 simulation as alignment of these variables with stage 1 was not necessary.

3.2. Stage 1 (PA) simulation

3.2.1. Analysis of potential constraint variables

Univariate multinomial logistic regression showed that all the potential constraints (seen in Fig. 2) were statistically significant predictors

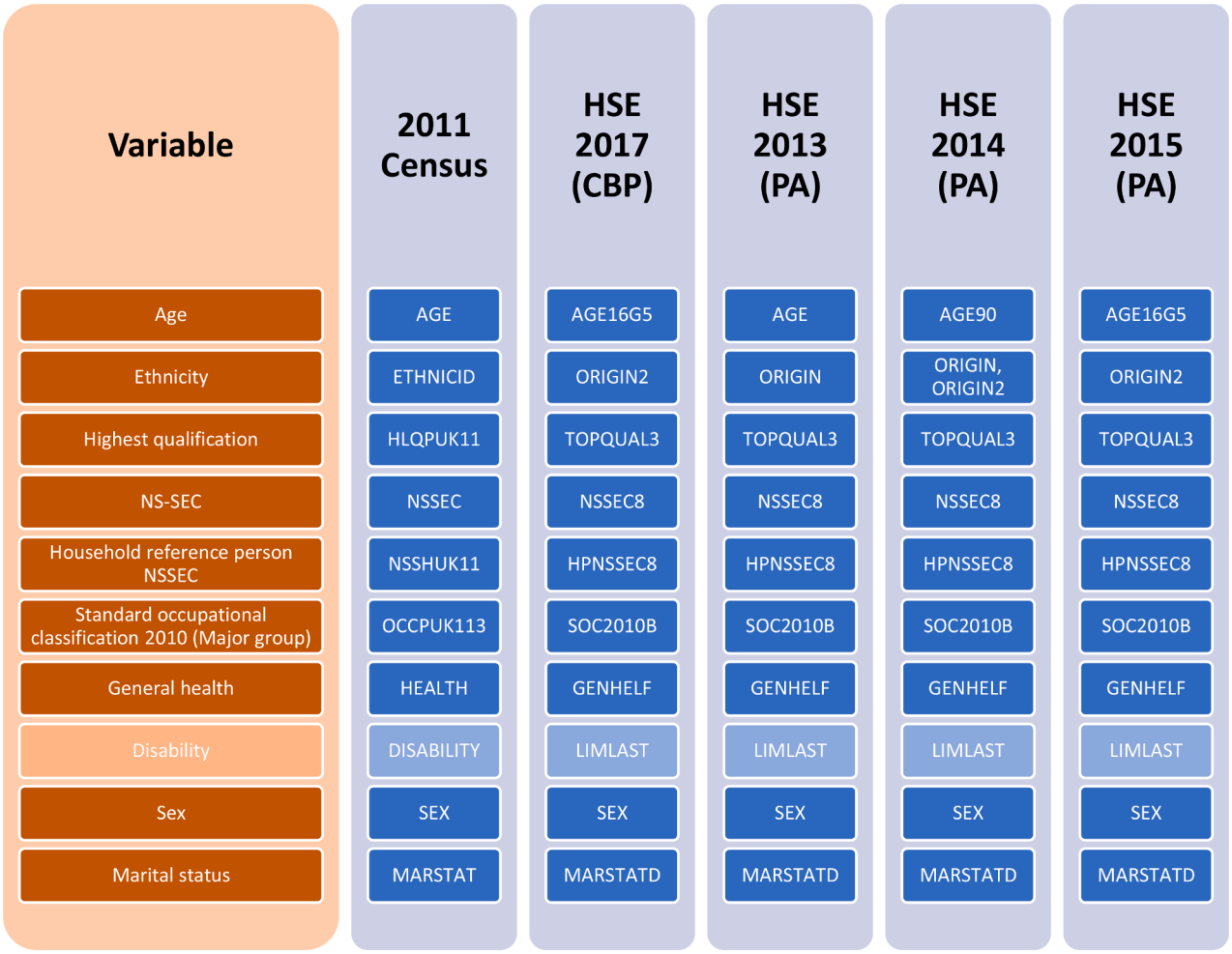


Fig. 2. Potential constraint variables.

of PA (Table 1). General health and age were the strongest individual predictors of PA. Ethnicity and marital status were relatively weak predictors and so were not included in the multivariate analysis. Assessing the data for collinearity revealed high collinearity between NS-SEC and Standard Occupational Classification, which is to be expected as NS-SEC is derived from Standard Occupational Classification. Accordingly, these two variables were not included together in any of the multivariate models. 10 multivariate multinomial logistic regression models were produced to assess potential constraint variable combinations. Sex and age were included in all multivariate models. Limited improvement was seen between four-variable and five-variable models. Excluding the five-variable models, Models 6 and 8 performed the best (Table 2).

3.2.2. *Simulating and validating outputs*

The best performing multivariate regression models, shown in Table 2, were run as simulations. The results of the internal validation of these simulations are summarised in Table 3. S1.2, S1.4 and S1.8 performed the best over the three internal validation measures. These three models were taken forward and used to produce inputs for the stage 2 (CBP) simulations for further validation.

In each simulation containing general health, for the general health variable there were a high number of areas with >10 % error. In an attempt to improve these simulations, the general health variable was collapsed from 5 categories to 3 (GENHELF2). This resulted in a modest improvement to the stage 1 simulations (Table 4).

3.3. *Stage 2 (CBP) simulation*

The process to determine the optimum constraint variable combinations was repeated for stage 2, starting with the analyses of potential constraint variables with univariate and then multivariate regression models. The stage 2 (CBP) simulations were performed using the same constraints as their corresponding stage 1 (PA) simulation with the addition of the PA variable MVPA as a constraint. For example, S1.4 AGE+SEX+SOC2010 became S2.4 AGE+SEX+SOC2010+MVPA.

3.3.1. *Analysis of potential constraint variables*

Univariate binary logistic regression showed all potential constraints except ethnicity and marital status (which were then excluded) to be statistically significant predictors of CBP (Table 5). Multivariate binary logistic regression showed that the models containing general health were superior predictors of CBP (Table 6). As in Stage 1, a collapsed 3-category general health variable (GENHELF2) was also trialled.

3.3.2. *Simulating and validating outputs*

The internal validation results showed an overall reduction in accuracy when comparing the stage 2 simulations with the stage 1 simulations (see Tables 3, 4 and 7). Improvements seen through modification of the general health variable in stage 1 were not conferred to stage 2, still leaving a sizeable number of >10 % outliers (Table 7).

3.3.3. *The final model*

Of the stage 2 simulations S2.4 performed the best on internal

Fig. 3. Non-constraint variables included in the model (useful for subsequent analyses).

validation and was selected as the final model. That is, age, sex and standard occupational classification 2010 (major group) were chosen as the final model constraint variables. Scatterplots of the validation of stages 1 and 2 of this final model, by constraint variable category, are attached as e-appendices. They show that the model validates adequately. Additionally, the fit of the constraint variables in the final simulation models at the national level can be seen in Fig. 4. It shows that overall, the demographics of the synthetic population are close to that of the true population (2011 census). The simulations were robust and reliable.

4. Discussion

This paper has outlined the production and validation of a 2-stage static SMS model of CBP at ward level in England. The use of a SMS methodology allowed simulation of all variables of interest at the small area level, with robust validation. These data will allow analysis of the relationship between CBP and potentially confounding variables locally.

Table 1
Results of univariate multinomial logistic regression for potential stage 1 constraint variables.

Independent variable	Chi-Squared	Degrees of freedom	P value	Nagelkerke's R^2
Age	1022	24	<0.001	0.046
Ethnicity	67	12	<0.001	0.003
Highest qualification	776	15	<0.001	0.035
Eight-class NS-SEC	546	24	<0.001	0.025
Household reference person eight-class NS-SEC	405	24	<0.001	0.019
Standard occupational classification 2010 (Major group)	586	27	<0.001	0.027
General health	1714	12	<0.001	0.076
Sex	392	3	<0.001	0.018
Marital status	48	3	<0.001	0.002

Table 5
Results of univariate binary logistic regression of potential stage 2 constraint variables.

Independent variable	Chi-Squared	Degrees of freedom	P value	Nagelkerke's R^2
Age	116	8	<0.001	0.026
Ethnicity	4	4	0.377	0.001
Highest qualification	59	5	<0.001	0.013
Eight-class NS-SEC	48	8	<0.001	0.011
Household reference person eight-class NS-SEC	34	8	<0.001	0.008
Standard occupational classification 2010 (Major group)	41	9	<0.001	0.009
General health	589	4	<0.001	0.128
Sex	36	1	<0.001	0.008
Marital status	2	1	0.162	0.000

Table 2
Results of multivariate multinomial logistic regression of potential stage 1 constraint variable combinations.
NSSEC8 - Eight-class NS-SEC, TOPQUAL - Highest qualification, SOC2010 - Standard occupational classification 2010, GENHELF - General health.

Model	Variables	Chi-squared	Degrees of freedom	P value	Nagelkerke's R^2
M1.1	SEX+AGE	1439	27	<0.001	0.065
M1.2	SEX+AGE+NSSEC8	1861	57	<0.001	0.083
M1.3	SEX+AGE+TOPQUAL	1756	42	<0.001	0.078
M1.4	SEX+AGE+SOC2010	1856	54	<0.001	0.083
M1.5	SEX+AGE+GENHELF	2706	39	<0.001	0.118
M1.6	SEX+AGE+GENHELF+NSSEC8	3108	63	<0.001	0.135
M1.7	SEX+AGE+GENHELF+TOPQUAL	2900	54	<0.001	0.126
M1.8	SEX+AGE+GENHELF+SOC2010	3097	66	<0.001	0.134
M1.9	SEX+AGE+GENHELF+NSSEC8+TOPQUAL	3202	78	<0.001	0.138
M1.10	SEX+AGE+GENHELF+SOC2010+TOPQUAL	3178	81	<0.001	0.138

Table 3
Results of the internal validation of stage 1 simulations.

Sim	Variables	Statistic	AGE	SEX	GENHELF	NSSEC8	SOC2010	TOPQUAL	Whole model mean
S1.2	AGE+SEX+NSSEC8	MAE	2.032	3.475		2.021			2.509
		E5	3012	2920		3997			3310
		E10	244	2		293			180
S1.4	AGE+SEX+SOC2010	MAE	1.997	3.397			1.217		2.204
		E5	3377	3122			822		2440
		E10	43	0			12		18
S1.5	AGE+SEX+GENHELF	MAE	1.943	3.811	3.168				2.974
		E5	3440	0	8404				3948
		E10	5	0	892				299
S1.6	AGE+SEX+GENHELF+NSSEC8	MAE	2.133	3.467	3.159	2.056			2.703
		E5	4112	2869	8363	4572			4979
		E10	204	2	970	273			362
S1.8	AGE+SEX+GENHELF+SOC2010	MAE	2.083	3.386	3.146		1.274		2.472
		E5	4149	3060	8299		1238		4187
		E10	27	0	750		14		198
S1.9	AGE+SEX+GENHELF+NSSEC8+TOPQUAL	MAE	2.490	3.472	3.378	2.552		6.440	4.583
		E5	7389	3744	9289	9503		21,278	12,801
		E10	275	2	2415	748		11,726	3792
S1.10	AGE+SEX+GENHELF+SOC210+TOPQUAL	MAE	2.431	3.339	3.341		1.548	6.422	3.416
		E5	7201	2948	9148		4650	21,206	9031
		E10	150	0	2297		1103	11,658	3042

Table 4
Results of the internal validation of stage 1 simulations containing the modified general health variable (GENHELF2).

Sim	Variables	Statistic	AGE	SEX	GENHELF2	NSSEC8	SOC2010	TOPQUAL	Whole model mean
S1.5b	AGE+SEX+GENHELF2	MAE	1.932	3.818	1.625				2.459
		E5	3395	4	3138				2179
		E10	7	0	104				37
S1.8b	AGE+SEX+GENHELF2+SOC2010	MAE	2.064	3.389	1.732		1.267		2.113
		E5	3958	3052	3968		1106		3021
		E10	30	0	218		13		65

Table 6
Results of multivariate binary logistic regression of potential stage 2 constraint variable combinations.

Sim	Variables	Chi-squared	Degrees of freedom	P value	Nagelkerke's R^2
M2.2	AGE+SEX+NSSEC8+MVPA	263	20	<0.001	0.058
M2.4	AGE+SEX+SOC2010+MVPA	262	21	<0.001	0.058
M2.5b	AGE+SEX+GENHELF2+MVPA	630	14	<0.001	0.137
M2.8	AGE+SEX+GENHELF+SOC2010+MVPA	679	25	<0.001	0.147
M2.8b	AGE+SEX+GENHELF2+ SOC2010+MVPA	636	23	<0.001	0.138

Table 7
Results of the internal validation of stage 2 simulations:.

Sim	Variables	Statistic	AGE	SEX	GENHELF(2)	NSSEC	SOC2010	TOPQUAL	Whole model mean
S2.4	AGE+SEX+SOC2010+MVPA	MAE	2.835	4.244			1.775		2.213
		E5	11,937	5938			3791		7222
		E10	293	282			106		227
S2.5b	AGE+SEX+GENHELF2+MVPA	MAE	2.668	4.146	2.495				3.103
		E5	8916	1676	7080				5891
		E10	147	0	1499				549
S2.8	AGE+SEX+GENHELF+SOC2010+MVPA	MAE	3.057	4.267	3.837		1.918		3.269
		E5	12,121	6029	9712		5253		8279
		E10	659	246	3514		392		1203
S2.8b	AGE+SEX+GENHELF2+SOC2010+MVPA	MAE	3.006	4.251	2.803		1.888		2.987
		E5	11,803	6002	8238		5128		7793
		E10	585	246	2520		253		901

4.1. Choice of method

SimObesity was used in this study, a previously validated SMS program with a demonstrated ability to simulate health data (Edwards and Clarke, 2009; Ifesemen et al., 2019; Edwards and Clarke, 2012). The use of a pre-existing SMS program such as this can substantially reduce project workload. In this study, a 2-stage SMS approach previously seen in work by Ifesemen et al. (Ifesemen et al., 2019) was adopted with success. The use of a 2-stage approach will allow for later 'what-if' analysis. In Ifesemen et al.'s work (Ifesemen et al., 2019), in stage 2 they used a survey data file from The English Longitudinal Study of Aging (ELSA) which included their outcome of interest knee osteoarthritis. As BMI was not included in the ELSA, they incorporated BMI as a constraint variable by using a HSE dataset in stage 1. In contrast, in this study, the HSE contains both data on our outcome of interest CBP and our main predictor of interest, PA. Nevertheless, a 2-stage approach was still chosen to enable later what-if analysis, highlighting the value of a 2-stage approach beyond just variable imputation.

4.2. Input data

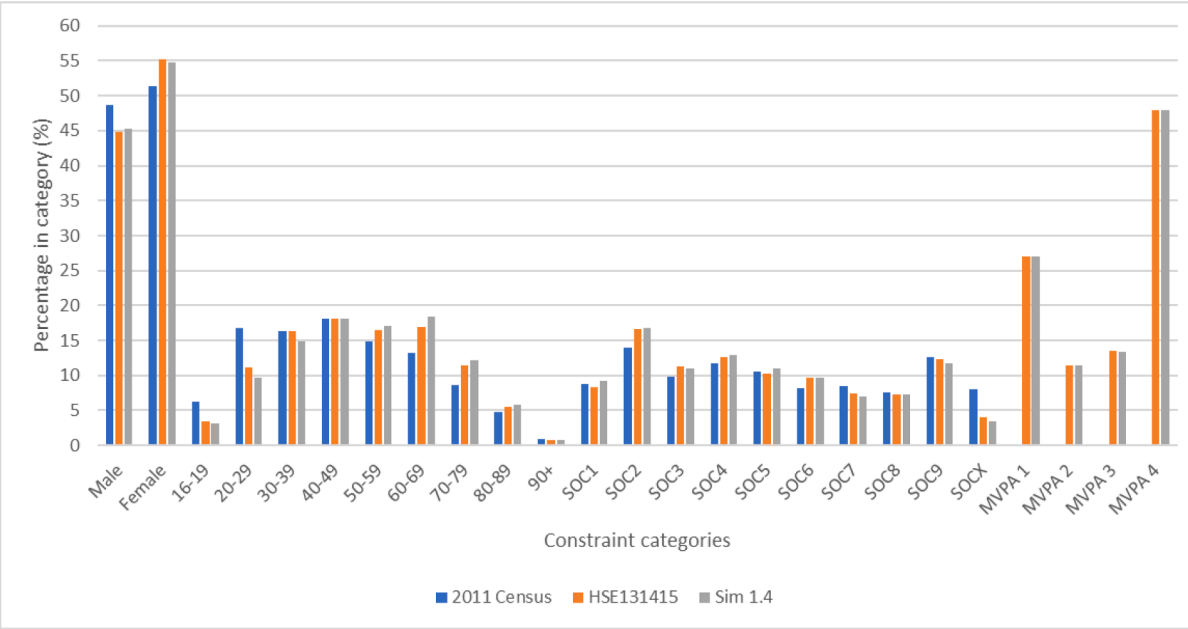
Our final model constraint variables were age, sex, standard occupational classification 2010 and MVPA (as an elective constraint). Age, sex and markers of socioeconomic status are commonly used constraints in SMS models (Edwards and Clarke, 2009; Koh et al., 2018; Spooner et al., 2021; Ifesemen et al., 2019; Smith et al., 2011; Campbell and Ballas, 2016; D Ballas et al., 2006). Specific to CBP estimation, Adomaviciute et al.'s (Adomaviciute et al., 2018) final CBP model, like our model, included age, sex and socioeconomic status. For socioeconomic status they used eight-class NS-SEC. Standard occupational classification appears not to have been trialled in their models. Whilst NS-SEC is derived from standard occupational classification, we found standard occupational classification to perform better than NS-SEC on internal validation. In addition to these variables, they also included BMI, smoking and education. Whilst we found these variables to be associated with CBP in our literature review, variables for BMI and smoking are not included in the 2011 census and so could not be considered for use as constraint variables. The education variable 'Highest qualification' was trialled but was excluded in favour of standard occupational classification. Another example relevant to CBP estimation, this time at an individual level, is Mukasa et al.'s (Mukasa and Sung, 2020) Cox proportional hazard model of first onset LBP, developed on large-scale Korean prospective cohort data. Similarly, to our model, they included age, sex, income and PA. However, their model comprised a total of 11 variables, with BMI, alcohol consumption, total cholesterol, blood pressure, bone mineral density disorders, disc degeneration and spinal stenosis also included.

Our model used relatively few predictor (constraint) variables compared to these statistical models. Increasing the number of constraints tended to worsen the internal validation measures. This issue has been noted previously by Chin et al. (Chin and Harding, 2006). This results from the model having to 'compromise' between more constraints and so each individual constraint is less accurately represented in the synthetic population. However, this does not necessarily mean that the synthetic population is overall less representative of the actual population.

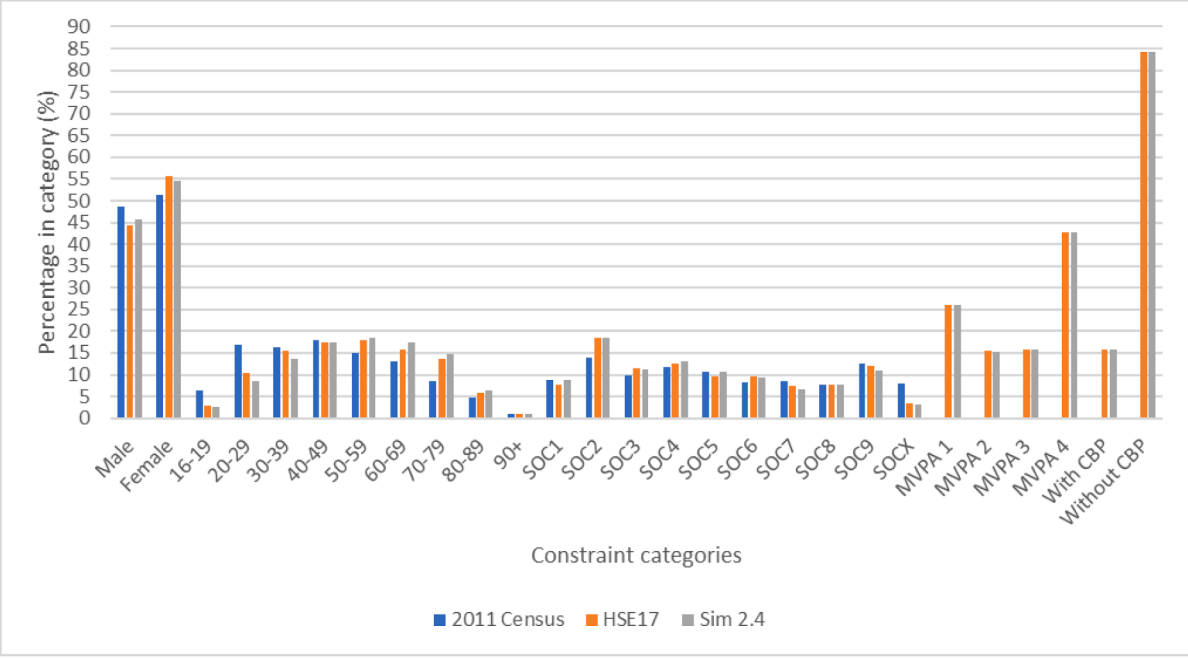
The HSE and Census datasets used in the construction of this model are two well-recognised high-quality datasets. MVPA was estimated at ward level using large HSE samples ($n = 24{,}906$ Stage 1; $n = 7997$ Stage 2). Whilst the use of a 2-stage approach may seem to increase sample size (Ifesemen et al., 2019), individuals in the stage 1 survey dataset are not available for selection in stage 2 and vice versa. It is therefore the case that the chain is as strong as its weakest link. A low sample size in one stage cannot be made up for by a large sample size in another stage.

In the preparation of the study datasets, a multiple imputation approach was taken using MICE in R. The imputations were averaged to produce a single dataset to allow input into SimObesity. This effectively reduces the approach to a single imputation method (van Buuren, 2018). In future it may be possible to explore a solution to this, for example, running simulations for each of the m imputations. Albeit, this would increase workload.

Another consideration is that of the socio-economic variables' classification. In SOC2010 used in the final simulation, both students and people who have never worked fall outside the classification system and are grouped in a 'not applicable' category. However, these two populations are not necessarily similar and the proportions of the two will vary largely by area, e.g., mainly made up of students in areas close to Higher Education Institutions for example. Having these two populations grouped relies on other constraints, such as age, to differentiate the two populations and select the correct 'type' of individuals to populate a synthetic area. This may explain why SOCX performed relatively poorly on internal validation.



a. Stage 1 results



b. Stage 2 results

Fig. 4. Summary of constraint categories at national level for both stages of the SMS.

4.3. Output considerations

This simulation was carried out at ward level. Whilst smaller and larger scales would have been possible, this was deemed a good compromise of factors previously mentioned (appreciating more local variation versus avoiding issues with confidentiality, resource intensity and small numbers). Nevertheless, by not using a finer scale, patterns present within wards will inevitably be missed.

The model created in this study is a static SMS model. The time period of such a model is dictated by the datasets used to construct it. The census file used was the 2011 Census, whilst the 2017 HSE contained the outcome, CBP. Individuals were therefore chosen to match the distribution of constraint variables in 2011 wards (confirmed on internal validation). However, as the synthetic individuals are from the 2017 HSE, the prevalence of CBP in each area will be determined by the relationship between CBP and its predictors for that 2017 population. The demographics of areas may have changed by varying amounts since the 2011 Census, but the relationship between CBP and its predictors is likely more stable. Therefore, this simulation likely represents CBP prevalence across England in 2011.

4.4. Validation

In this study a combination of published internal validation methods

were used; TAE and MAE were used to assess overall accuracy and E5 and E10 were used to assess outliers. Mean validation statistics were calculated to allow comparison between models of varying sizes. In the evaluation of SMS models, TAE has been suggested and utilised by various authors (Voas and Williamson, 2001; Williamson et al., 1998; Tanton and Vidyattama, 2009; Lovelace and Dumont, 2018). The E5 statistic has also been previously suggested by Lovelace et al. (Lovelace et al., 2015). In addition to E5, in this study E10 was also used. This was done to evaluate more severe outliers. The use of MAE, E5 and E10, combined with visualisation through scatter plots, creates an easily accessible comprehensive assessment of internal validity with intra and intermodel comparability. Overall, the final model performed well on internal validation. However, there was some reduction in performance between stage 1 and stage 2. This pattern has also been noted by Ifesemen et al. (Ifesemen et al., 2019) in their 2-stage simulation and may be a factor associated with the reduction in population sample size. As highlighted by Huang et al. (Huang and Williamson, 2001), when employing a combinatorial optimisation method, having a larger survey sample size allows for more possible combinations of individuals in the synthetic population and thus a higher likelihood of an accurate model fit. However, this degradation may also be a feature of the 2-stage approach as the second stage is simulating based on an already synthetic population, therefore compounding the error.

External validity is an important factor in any predictive model. Assessing external validity in SMS is challenging as data often doesn't exist for the variable being simulated at the small area level (this is why the simulation is being performed). Hence, this step is often missed by researchers (Edwards et al., 2011). Methods have been suggested to work around this issue but there are still considerable difficulties involved (Edwards et al., 2011). As such, external validation has not yet been performed for this model.

5. Conclusion

SMS is a valuable method of small area estimation. This study successfully utilised a 2-stage SMS approach to produce a synthetic CBP microdata set which was internally validated. This will prove useful in subsequent research to understand the spatial pattern of CBP and the processes underlying it, as well as allowing 'what-if' analyses. In future, further options need to be explored for multiple imputation and external validation in this and similar SMS models.

Declarations

Ethical approval was granted by the University of Nottingham Faculty of Medicine and Health Research Ethics Committee (FMHS 199–0221; February 2021).

CRediT authorship contribution statement

Harrison Smalley: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Kimberley Edwards:** Conceptualization, Data curation, Methodology, Project administration, Software, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Funding

University of Nottingham.

Datasets are available from the corresponding author on reasonable request.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.sste.2023.100633.

References

Edwards, K.L., Clarke, G.P., 2009. The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity. Soc Sci Med 69 (7), 1127–1134.

Ballas, D., Clarke, G., Dorling, D., Rigby, J., Wheeler, B., 2006. Using geographical information systems and spatial microsimulation for the analysis of health inequalities. Health Informatics J 12 (1), 65–79.

Procter, K.L., Clarke, G.P., Ransley, J.K., Cade, J, 2008. Micro-Level Analysis of Childhood Obesity, Diet, Physical Activity, Residential Socioeconomic and Social Capital Variables: Where Are the Obesogenic Environments in Leeds? Area 40 (3), 323–340.

Twigg, L., Moon, G., Jones, K., 2000. Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators. Soc Sci Med 50 (7–8), 1109–1120.

Pearce, J., Boyle, P., Flowerdew, R., 2003. Predicting smoking behaviour in census output areas across Scotland. Health Place 9 (2), 139–149.

Koh, K., Grady, S.C., Darden, J.T., Vojnovic, I., 2018. Adult obesity prevalence at the county level in the United States, 2000–2010: Downscaling public health survey data using a spatial microsimulation approach. Spat Spatiotemporal Epidemiol 26, 153–164.

Versus Arthritis. Musculoskeletal Calculator [Internet]. [cited 2021 Mar 5]. Available from: https://www.versusarthritis.org/policy/resources-for-policy-makers/musculoskeletal-calculator/.

Adomaviciute S., Soljak M., Gardiner J., Foley K., Watt H., Newson R. Back pain prevalence models for small populations Technical document produced for Arthritis Research UK. 2018 Oct.

Tanton, R., Clarke, G., 2014. Spatial Models. Handbook of Microsimulation Modelling. Emerald Publishing Limited, Bingley, pp. 367–383.

Tanton, R., Edwards, K.L., 2012. Introduction to Spatial Microsimulation: History, Methods and Applications. Spatial Microsimulation: A Reference Guide for Users. Springer, Dordrecht, pp. 3–8.

Spooner, F., Abrams, J.F., Morrissey, K., Shaddick, G., Batty, M., Milton, R., et al., 2021. A dynamic microsimulation model for epidemics. Soc Sci Med 291, 114461.

Ifesemen, O.S., Bestwick-Stevenson, T., Edwards, K.L., 2019. Spatial microsimulation of osteoarthritis prevalence at the small area level in England-Constraint selection for a 2-stage microsimulation process. Int J Microsimul 12 (2), 36–50.

Smith, D.M., Pearce, J.R., Harland, K., 2011. Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. Health Place 17 (2), 618–624.

Campbell, M., Ballas, D., 2016. SimAlba: A Spatial Microsimulation Approach to the Analysis of Health Inequalities. Front Public Health 4 (OCT).

Ballas D., Clarke G., Dorling D., Rigby J., Wheeler B. Using geographical information systems and spatial microsimulation for the analysis of health inequalities. 2006.

Tanton, R., 2013. A Review of Spatial Microsimulation Methods. Int J Microsimul 7 (1), 4–25.

Edwards, K.L., Clarke, G., 2012. SimObesity: Combinatorial Optimisation (Deterministic) Model. editors. In: Tanton, R, Edwards, KL (Eds.), Spatial Microsimulation: A Reference Guide for Users. Springer, Netherlands, pp. 69–85.

Bridges S. Chronic pain [Internet]. Vol. 1, Health Survey for England - 2011. 2012 [cited 2021 Jul 5]. Available from: https://files.digital.nhs.uk/publicationimport/pub09xxx/pub09300/hse2011-ch9-chronic-pain.pdf.

Hoy, D., March, L., Brooks, P., Blyth, F., Woolf, A., Bain, C., et al., 2014. The global burden of low back pain: Estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis [Internet] 73 (6), 968–974 [cited 2021 Mar 5]. Available from: https://pubmed.ncbi.nlm.nih.gov/24665116/.

The World Health Organization. ICD-11 for Mortality and Morbidity Statistics [Internet]. 2020 [cited 2021 Mar 5]. Available from: https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1581976053.

Maher, C., Underwood, M., Buchbinder, R., 2017. Non-specific low back pain. The Lancet [Internet] 389 (10070), 736–747 [cited 2021 Mar 7]. Available from: https://www.sciencedirect.com/science/article/pii/S0140673616309709.

Fillingim, R.B., King, C.D., Ribeiro-Dasilva, M.C., Rahim-Williams, B., Riley, J.L., 2009. Sex, Gender, and Pain: A Review of Recent Clinical and Experimental Findings. Journal of Pain [Internet] 10 (5), 447–485 [cited 2021 Jun 11]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677686/.

Knox, J.B., Orchowski, J.R., Owens, B., 2012. Racial differences in the incidence of acute low back pain in United States military service members. Spine (Phila Pa 1976) [Internet] 37 (19), 1688–1692 [cited 2021 Jun 14]. Available from: https://pubmed.ncbi.nlm.nih.gov/22460922/.

Carey, T.S., Freburger, J.K., Holmes, G.M., Jackman, A., Knauer, S., Wallace, A., et al., 2010. Race, Care Seeking, and Utilization for Chronic Back and Neck Pain:

Population Perspectives. Journal of Pain [Internet] 11 (4), 343–350 [cited 2021 Jun 12]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847652/.
Hoy, D., Bain, C., Williams, G., March, L., Brooks, P., Blyth, F., et al., 2012. A systematic review of the global prevalence of low back pain [Internet]. Arthritis and Rheumatism. Arthritis Rheum 64, 2028–2037 [cited 2021 Mar 5]. Available from: https://pubmed.ncbi.nlm.nih.gov/22231424/.
Björck-van Dijken, C., Fjellman-Wiklund, A., Hildingsson, C., 2008. Low back pain, lifestyle factors and physical activity: A population-based study. J Rehabil Med [Internet] 40 (10), 864–869 [cited 2021 Aug 8]. Available from: https://www.medicaljournals.se/jrm/content/html/10.2340/16501977-0273.
Papageorgiou, A., Croft, P., Ferry, S., Jayson, M., Silman, A, 1995. Estimating the prevalence of low back pain in the general population. Evidence from the South Manchester Back Pain Survey. Spine (Phila Pa 1976) [Internet] 20 (17), 1889–1894 [cited 2021 Aug 8]. Available from: https://pubmed.ncbi.nlm.nih.gov/8560337/.
Coenen, P., Gouttebarge, V., Van Der Burght, A.S.A.M., Van Dieën, J.H., Frings-Dresen, M.H.W., Van Der Beek, A.J., et al., 2014. The effect of lifting during work on low back pain: A health impact assessment based on a meta-analysis. Occup Environ Med [Internet] 71 (12), 871–877 [cited 2021 Jun 12]. Available from: https://oem.bmj.com/content/71/12/871.
Lötters, F., Burdorf, A., Kuiper, J., Miedema, H., 2003. Model for the work-relatedness of low-back pain. Scand J Work Environ Health [Internet] 29 (6), 431–440 [cited 2021 Jun 12]. Available from: https://www.sjweh.fi/article/749.
Costa-Black, K.M., Loisel, P., Anema, J.R., Pransky, G., 2010. Back pain and work. Best Pract Res Clin Rheumatol [Internet] 24 (2), 227–240 [cited 2021 Jun 12]. Available from: https://www.sciencedirect.com/science/article/pii/S1521694209001430.
Dionne, C.E., Von, Korff M, Koepsell, T.D., Deyo, R.A., Barlow, W.E., Checkoway, H, 2001. Formal education and back pain: a review. J Epidemiol Community Health (1978) [Internet] 55 (7), 455–468 [cited 2021 Aug 8]. Available from: https://jech.bmj.com/content/55/7/455.long.
Shiri, R., Karppinen, J., Leino-Arjas, P., Solovieva, S., Viikari-Juntura, E., 2010. The Association between Smoking and Low Back Pain: A Meta-analysis. American Journal of Medicine [Internet] 123 (1) [cited 2021 Jun 15] 87.e7-87.e35. Available from: https://www.sciencedirect.com/science/article/pii/S000293430900713X.
Green, B.N., Johnson, C.D., Snodgrass, J., Smith, M., Dunn, A.S., 2016. Association Between Smoking and Back Pain in a Cross-Section of Adult Americans. Cureus [Internet] 8 (9) [cited 2021 Aug 11]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5081254/.
Alkherayf, F., Wai, E.K., Tsai, E.C., Agbi, C., 2010. Daily smoking and lower back pain in adult Canadians: the Canadian Community Health Survey. J Pain Res [Internet] 3, 155 [cited 2021 Aug 11]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3004651/.
Shiri, R., Karppinen, J., Leino-Arjas, P., Solovieva, S., Viikari-Juntura, E., 2010. The association between obesity and low back pain: A meta-analysis. Am J Epidemiol [Internet] 171 (2), 135–154 [cited 2021 Mar 5]. Available from: https://pubmed.ncbi.nlm.nih.gov/20007994/.
Leboeuf-Yde, C., 2000. Body weight and low back pain: A systematic literature review of 56 journal articles reporting on 65 epidemiologic studies. Spine (Phila Pa 1976) [Internet] 25 (2), 226–237 [cited 2021 Mar 5]. Available from: https://pubmed.ncbi.nlm.nih.gov/10685488/.
Shiri, R., Falah-Hassani, K., 2017. Does leisure time physical activity protect against low back pain? Systematic review and meta-analysis of 36 prospective cohort studies. Br J Sports Med [Internet] 51 (19), 1410–1418 [cited 2021 Aug 10]. Available from: https://bjsm.bmj.com/content/51/19/1410.
University College London, 2017. Department of Epidemiology and Public Health NSResearch. Health Survey for England, UK Data Service, p. 2021.
Office for National Statistics, 2017. 2011 Census Aggregate Data. Service, UK Data.
NatCen Social Research, 2015. University College London D of E and PHealth. Health Survey For England, 2013. UK Data Service [data collection].
NatCen Social Research, 2018. University College London D of E and PH. Health Survey For England, 2014, 3rd Edition. UK Data Service [data collection].
NatCen Social Research, 2019. University College London D of E and PHealth. Health Survey For England, 2015, 2nd Edition. UK Data Service [data collection].
Office for National Statistics. Census geography [Internet]. 2021 [cited 2021 Aug 14]. Available from: https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography.
R Core Team, 2021. R: A language and Environment For Statistical Computing. Foundation for Statistical Computing, Vienna, Austria.
van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations. J Stat Softw 45 (3), 1–67.
Corp, I., 2020. IBM SPSS Statistics for Windows, Version 27.0. IBM Corp, Armonk, NY.
Edwards, K.L., Clarke, G.P., Thomas, J., Forman, D., 2011. Internal and External Validation of Spatial Microsimulation Models: Small Area Estimates of Adult Obesity. Appl Spat Anal Policy 4 (4), 281–300.
Timmins, K.A., Edwards, K.L., 2016. Validation of Spatial Microsimulation Models: a Proposal to Adopt the Bland-Altman Method. Int J Microsimul 9 (2), 106–122.
Lovelace, R., Birkin, M., Ballas, D., Van Leeuwen, E., 2015. Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. JASSS 18 (2), 1–15.
Mukasa, D., Sung, J., 2020. A prediction model of low back pain risk: a population based cohort study in Korea. Korean J Pain 33 (2), 165.
Chin, S., Harding, A., 2006. Regional dimensions: Creating Synthetic Small-Area Microdata and Spatial Microsimulation Models.
van Buuren, S., 2018. Analysis of imputed data. Flexible Imputation of Missing Data, 2nd ed. CRC Press.
Voas, D., Williamson, P., 2001. Evaluating Goodness-of-Fit Measures for Synthetic Microdata. Geographical and Environmental Modelling 5 (2), 177–200.
Williamson, P., Birkin, M., Rees, P.H., 1998. The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environ Plan A 30 (5), 785–816.
Tanton, R., Vidyattama, Y., 2009. Pushing it to the edge: Extending generalised regression as a spatial microsimulation method. Int J Microsimul 3 (2), 23–33.
Lovelace, R., Dumont, M., 2018. Model checking and evaluation. Spatial Microsimulation with R. CRC Press.
Huang Z., Williamson P. A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. 2001.