


RESEARCH ARTICLE

Open Access



Trends in the ability of socioeconomic position to predict individual body mass index: an analysis of repeated cross-sectional data, 1991–2019

Liam Wright^{1*} , Charis Bridger Staatz¹, Richard J. Silverwood¹ and David Bann¹

Abstract

Background The widening of group-level socioeconomic differences in body mass index (BMI) has received considerable research attention. However, the predictive power of socioeconomic position (SEP) indicators at the individual level remains uncertain, as does the potential temporal variation in their predictive value. Examining this is important given the increasing incorporation of SEP indicators into predictive algorithms and calls to reduce social inequality to tackle the obesity epidemic. We thus investigated SEP differences in BMI over three decades of the obesity epidemic in England, comparing population-wide (SEP group differences in mean BMI) and individual-level (out-of-sample prediction of individuals' BMI) approaches to understanding social inequalities.

Methods We used repeated cross-sectional data from the Health Survey for England, 1991–2019. BMI (kg/m²) was measured objectively, and SEP was measured via educational attainment, occupational class, and neighbourhood index of deprivation. We ran random forest models for each survey year and measure of SEP adjusting for age and sex.

Results The mean and variance of BMI increased within each SEP group over the study period. Mean differences in BMI by SEP group also increased: differences between lowest and highest education groups were 1.0 kg/m² (0.4, 1.6) in 1991 and 1.3 kg/m² (0.7, 1.8) in 2019. At the individual level, the predictive capacity of SEP was low, though increased in later years: including education in models improved predictive accuracy (mean absolute error) by 0.14% (−0.9, 1.08) in 1991 and 1.05% (0.18, 1.82) in 2019. Similar patterns were obtained for occupational class and neighbourhood deprivation and when analysing obesity as an outcome.

Conclusions SEP has become increasingly important at the population (group difference) and individual (prediction) levels. However, predictive ability remains low, suggesting limited utility of including SEP in prediction algorithms. Assuming links are causal, abolishing SEP differences in BMI could have a large effect on population health but would neither reverse the obesity epidemic nor reduce much of the variation in BMI.

Keywords Variation explained, Predictive accuracy, Body mass index, Obesity, Social inequalities, Social epidemiology

Background

Obesity rates have more than tripled among adults in England since 1980 [1]. Average body mass index (BMI) has also increased, but the population distribution of BMI has become more spread and more skewed [2], implying that individuals have not been equally affected

*Correspondence:

Liam Wright
liam.wright@ucl.ac.uk

¹ Centre for Longitudinal Studies, Social Research Institute, University College London, 55-59 Gordon Square, London WC1H 0NT, UK



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

by the obesity epidemic. Given the substantial health and economic costs associated with obesity [3], identifying solutions to the obesity epidemic continues to be an area of significant policy and research interest.

A large amount of research has focused on social inequalities in obesity and BMI (see, e.g. [4, 5] for reviews). Recent evidence finds that adults in the most deprived areas of England are twice as likely to have obesity as those in the least deprived areas [6]; a similar difference is observed comparing highest and lowest education groups [6]. Evidence further suggests that, in England, inequalities in obesity and BMI according to education level have widened — in absolute terms — alongside the development of the obesity epidemic [7], a pattern observed in multiple other countries [8], though not all [9].

Research on social inequalities in BMI has typically taken a population-level approach and focused on estimating associations — for instance, examining the mean difference in BMI according to educational attainment. Less attention has been paid to the explanatory power of socioeconomic factors at the individual level — for example, the proportion of between-person variability in BMI that can be predicted by socioeconomic position (SEP) [10]. Though measures of SEP have been included in predictive algorithms for BMI [11] and reducing social inequality has been proposed as a way to tackle high obesity rates [12], SEP appears to explain only a small amount (<6%) of between-person variability in BMI [9, 13–17]. This is the case even when multiple indicators of SEP across life are used [13, 14].

The comparatively low explanatory power of SEP accords with more general observations. The variance in adult BMI explained by environmental factors shared between twins (such as parental SEP) is very low, in contrast to the proportion explained by genetics and non-shared environmental factors [18]. This low explanatory power is observed across almost all traits and is known as the ‘gloomy prospect’ in behavioural genetics [19, 20]. Attempts to directly predict individual life outcomes using SEP and other survey data have produced humbling results. For example, a recent scientific mass collaboration showed that several socioeconomic outcomes were largely unpredictable using a range of sophisticated predictive models and unusually rich survey data (including socioeconomic histories) [21].

While the explanatory power of SEP on BMI may be lower than perhaps expected [12], it could have systematically changed across time. The increasing variation of population BMI partly reflects increasing inequalities *between* SEP groups, but it reflects increasing variation *within* these groups, too [2, 15, 22–25]. If the increasing variation *within* groups exceeds the increasing variation *between* groups, the explanatory power of SEP — already

low — may have fallen further still. Determining whether this is the case is important for understanding the role of SEP as a contributor to the obesity epidemic [22] and for understanding the (continuing) potential for using SEP in predictive algorithms. However, research on this question is limited. Studies from the USA [9] and Indonesia [15] find the explanatory power of SEP on BMI has decreased over time, but social inequalities declined in these countries over the periods assessed. Thus, results may not generalize to England or other countries that have experienced widening inequalities across time.

Existing research is further limited by a focus on individual-level (education) and not area-level (e.g. neighbourhood deprivation) measures of SEP which may capture area-based factors, such as neighbourhood walkability and fast food outlet density [26]. Existing research is also limited by the use of methods not tailored for prediction. In particular, studies have used linear regression models of limited flexibility, which may not have captured interactions and other non-linearities. They have also assessed explanatory power within the same sample as used to estimate models (thus biasing towards more optimistic results) and have not assessed predictive ability, specifically — a metric of particular importance for creating accurate prediction algorithms for BMI.

We examined trends in the explanatory and predictive power of individual- and area-level SEP on BMI more formally by adopting principles and methods from machine learning. We used random forest models and repeated cross-sectional data from the Health Survey for England (HSE) to examine changes in the predictive ability of educational attainment and neighbourhood deprivation for BMI and obesity between the years 1991 and 2019, a period in which obesity rates doubled in England [1].

Methods

Participants

The HSE is an ongoing series of annual nationally representative cross-sectional health surveys that began in 1991 [27]. A detailed description of the survey is available elsewhere [28]. The HSE uses a multi-stage sampling design with households drawn from a list of postcode sectors. Non-response weights are provided with the data from 2002 onwards, due to increasing refusal rates (household response rates fell from 77% in 1994 to 60% in 2019; see Additional file 1: Fig. S1) [28, 29]. We used these weights where available, assuming weights of 1 in other survey years. We limited our analysis to individuals aged 25–64 — the lower bound chosen to focus on ages with few members (1–8%) in full-time education (whose eventual education level is not known) and the upper bound chosen to reduce selection biases that could

arise due to higher mortality rates among high BMI individuals [30]. We further limited our sample to those of White ethnicity to create comparable populations less liable to changes in composition due to inflow and outflow of migration. For similar reasons, we also excluded a small number of individuals whose highest qualification was obtained abroad as well as individuals currently in full-time education (4.2% of observations). There was only a small amount of missingness on our covariate data (<0.1%), so we analysed complete cases only. Our final sample size was 143,094. This excluded 10.7% of the eligible sample who had missing BMI data. The sample size each year ranged from 1813 in 1991 to 9556 in 1993.

Measures

Body mass index

BMI was calculated by dividing weight in kilogrammes by height in metres squared. Height and weight were measured directly by interviewers. From 1995, individuals weighing more than 130 kg were asked to give an estimate of their weight due to limitations with the scales, so measurements for these individuals are based on self-report.

Socioeconomic position

The HSE contains few measures of SEP that are measured consistently in each wave. We focus on educational attainment, occupational social class, and neighbourhood deprivation; each captures different dimensions of SEP [31], has been widely used in the social inequalities literature [32], and is related to obesity in the UK [6]. (HSE also contains data on income quintile, but we did not use this here as it is missing in a sizeable number [~15%] of cases, with missingness increasing over the survey period.) Education was recorded using the national vocational qualification schema to categorize qualifications according to skill level (high to low: NVQ 4/5, higher education below degree level, NVQ 3, NVQ 2, NVQ1, none). [NVQ 4/5 is equivalent to degree or above; see [33] for further example qualifications.] Occupational social class was captured using the Registrar General Social Class Schema (high to low: I Professional, II Managerial and Technical, III Skilled Non-Manual, III Skilled Manual, IV Partly Skilled, V Unskilled). Data on social class are available from 1994 onwards, except 2010 and 2011. Social class is missing in a small number of cases (<3%) where occupation was not categorizable within the schema (e.g. employees of the Armed Forces) or where the participant was long-term unemployed.

Neighbourhood deprivation was measured using the index of multiple deprivation (IMD) and was categorized into quintiles (1st least deprived–5th most deprived). The IMD combines deprivation across seven domains

(income, employment, education, health, crime, barriers to housing and services, and living environment). In the HSE, IMD data are available from 2001 only; at the electoral ward level from 2001 to 2002 and lower super output area (LSOA) level thereafter (LSOAs comprise 400–1200 households). New versions of the IMD are released intermittently. The IMD2000 is available from 2001 to 2002, the IMD2004 from 2003 to 2007, the IMD2007 from 2008 to 2010, the IMD2010 from 2011 to 2014, the IMD2015 from 2015 to 2018, and the IMD2019 in 2019. We use the IMD quintile data as supplied, as it precluded further harmonization.

Covariates

We included age and sex as covariates in our prediction models as the relationship between age, sex, and SEP (particularly education) has changed strongly over time (with, e.g. the population becoming increasingly highly educated) and as age and sex may confound the association between education and IMD and BMI [7, 34]. Age was available in single years prior to 2015, but only in 5-year categories from 2015 onwards. For consistency with earlier years, for years 2015–2019, for each individual, we randomly drew a single-year age (with equal probability) from their respective 5-year age category. Mean age increased in our sample between 1991 and 2014 (average age ~43 in 1991 and ~45 in 2014).

Statistical analysis

To maximize predictive ability, we used random forest models, known to provide similar or superior predictions to traditional regression approaches in multiple settings [35, 36]. Our analysis consisted of fitting random forest models and assessing their predictive accuracy and explanatory power. Random forests are a decision tree-based method in which data are recursively split according to decision rules invoking individual predictor variables (e.g. male or female, age < 45). Decision rules are chosen such that splits minimize heterogeneity in the target variable (here, BMI). To avoid overfitting, random forests use an ensemble approach where the results of multiple decision trees are averaged, with each tree being fit on a subset of predictor variables and a random sample of observations. As predictions are generated via successive binary splits, random forests can account for non-linearities or interactions between independent variables (e.g. between age and education) without requiring their explicit parameterization, an advantage here given previously observed differences in social inequalities in BMI between males and females, across cohorts, and over the life course [7].

We fit a random forest (500 trees) for BMI for each year of data collection and measure of SEP, using SEP,

age, and sex as predictor variables. We then extracted model predictions and used these to calculate three metrics of explanatory power and predictive accuracy: variation explained (R^2), mean absolute error (the difference between observed and predicted BMI), and probability of superiority. (In this setting, the probability of superiority is the probability that among two randomly chosen participants, the participant with the higher predicted BMI score has the higher observed BMI.) Importantly, to avoid overfitting, we generated model predictions using a portion of our data that was not used to estimate the random forest model (procedure explained further below). R^2 provides a (relative) measure of how well SEP can predict between-person differences, while mean absolute error and probability of superiority provide summaries of how well SEP can predict individuals.

We compared the three metrics to (a) baseline predictions where mean BMI was used and (b) the results of random forest models including only sex and age as predictor variables. We also calculated the magnitude of the association between educational attainment and BMI by using the results of the random forest models to predict mean BMI assuming everyone in the population had the same SEP. We defined the size of the association between SEP and BMI as the difference in predicted population mean BMI for the most advantaged and disadvantaged SEP categories (NVQ 4/5 vs no qualifications for education; I Professional vs V Non-Skilled for social class; and highest vs lowest quintile for IMD). To calculate confidence intervals, we used bootstrapping accounting for the complex survey design (Rao and Wu method [37], 500 bootstrap samples). For the predictive accuracy and explanatory power metrics, we generated predictions using the observations not selected within a given bootstrap in order to avoid overfitting.

As the random forest models were estimated for each year separately, to more easily ascertain trends in (a) the proportion of prediction error explained by each SEP variable and (b) the size of the association between BMI and each SEP variable, we smoothed the bootstrap estimates by regressing estimates upon year splines using generalized additive models (GAMs) — GAMs allow for flexible, smooth non-linear associations between independent and dependent variables. The change in the age variable to 5-year categories from 2015 onwards may have artificially increased the relative incremental predictive power of including SEP in models. Consequently, we also ran the GAM models using data only up to 2014 to assess whether trends were observable prior to the change in the data.

We performed a series of further analyses. First, as social inequalities in BMI are typically found to be stronger among females than males [4], we repeated the

analysis stratifying by sex. Second, as age was imputed in later years, we re-ran models with age inputted as 5-year categories (25–29, 30–34, ..., 60–64 years old). Third, as obesity ($\text{BMI} \geq 30 \text{ kg/m}^2$) is of particular research and policy interest, we repeated the analysis using obesity as the outcome measure (see Additional file 1: Results S1 for further detail on methods used). Fourth, as random forests could potentially overfit the data, we repeated the BMI analysis using simple linear regression. In these models, predictors were included as linear (age) or categorical (sex, education, occupational class, IMD) terms with no interactions included.

The organization used to conduct the HSE changed in 1994. Some previous studies using HSE have accordingly focused on data from 1994 onwards [38]. We present results from 1991 to 2019, but in the text report results from 1994 where results from 1991 to 1993 depart considerably from those in later years.

Results

Descriptive statistics

There was an increase in the overall mean and variance of BMI and the prevalence of obesity between 1991 and 2019 (Fig. 1a–c; see also Additional file 1: Fig. S2). Education levels generally increased across time; the proportion of individuals with the highest education level increased from 11.7% in 1991 to 37.5% in 2019 (Additional file 1: Fig. S3). Increasing education levels led to non-linear changes in the variance of the education measure; variance decreased overall between 1991 and 2019 but peaked in 2002 (Fig. 1d). Descriptive statistics for the SEP measures and covariates are also shown in Additional file 2: Table S1.

Predicting BMI

Mean BMI increased among all education groups, social classes, and IMD quintiles across the survey period, including among those with the highest SEP (Fig. 2a, b) — for instance, predicted mean BMI increased for the most highly educated group (NVQ 4/5) from 26.2 kg/m^2 (95% CI = 25.6, 26.7) in 1991 to 28.2 kg/m^2 (27.7, 28.5) in 2019. More disadvantaged SEP was generally related to higher BMI and there was some evidence that social inequalities widened over time. The difference between the lowest and highest educated groups was 1.0 kg/m^2 (0.4, 1.6) in 1991 and 1.3 kg/m^2 (0.7, 1.8) in 2019, while the difference between individuals in the most and least deprived neighbourhoods was 0.6 kg/m^2 (0.3, 0.8) in 2001 and 1.3 kg/m^2 (0.7, 1.8) in 2019 (see Additional file 1: Fig. S3 for smoothed results). The trend cannot be explained by changes in age composition over time — generating effect sizes using the age structure of the 2019 HSE sample similar results (results available on request).

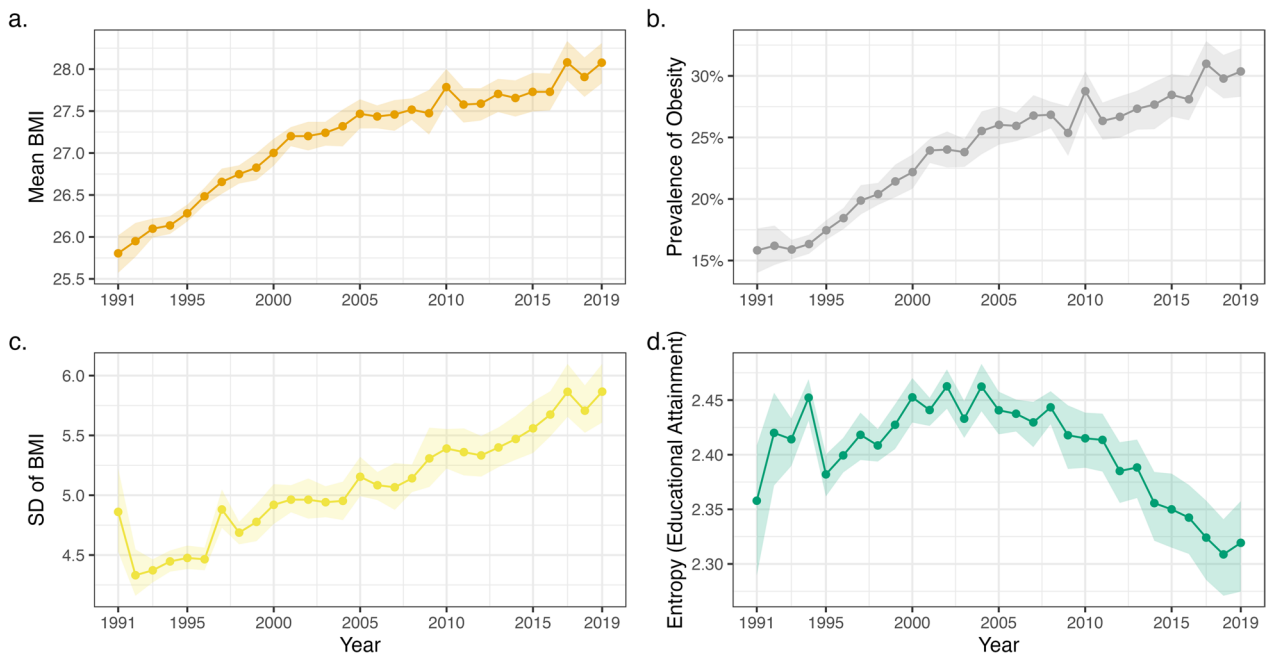


Fig. 1 Descriptive statistics (+ 95% confidence intervals) by survey year. **a** Mean body mass index. **b** Proportion of individuals who are obese (BMI ≥ 30 kg/m²). **c** Standard deviation of BMI. **d** Shannon's entropy (a measure of variability) for categorical educational attainment variable. All figures are weighted. Confidence intervals derived using the Rao and Wu bootstrap method to account for complex survey design

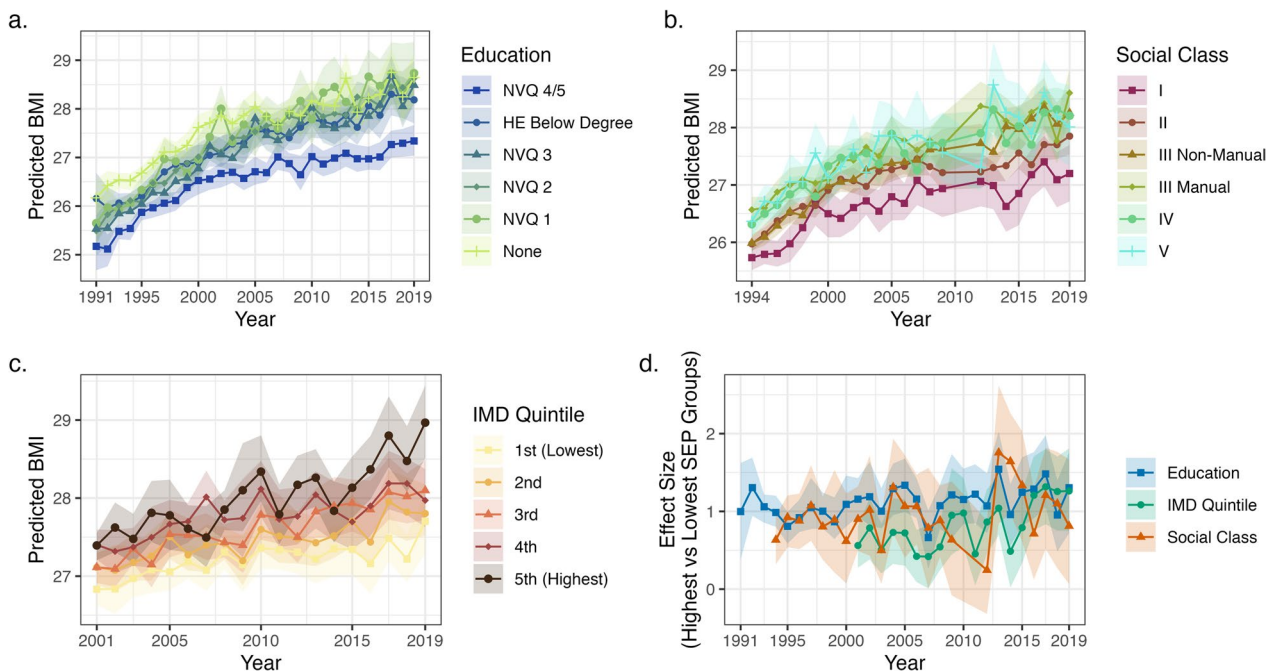


Fig. 2 Results of random forest models predicting BMI by survey year. **a** Predicted mean population BMI assuming all individuals have given educational attainment. **b** Predicted mean population BMI assuming all individuals belong to given social class. **c** Predicted mean population BMI assuming all individuals from areas in given IMD quintile. **d** Difference in mean BMI at the population level between highest (NVQ 4/5, I professional, or 1st quintile IMD) and lowest (no qualifications, V unskilled, or 5th quintile IMD) SEP groups. Confidence intervals calculated using bootstrap samples accounting for complex survey design (500 bootstraps, centile method)

While average BMI increased within SEP groups, so did its variability (Additional file 1: Fig. S5). Given this increasing variability, the total prediction error increased over time, regardless of the model used (Fig. 3a). In 1991, using age, sex, and education level to predict BMI generated an average prediction error (the difference between predicted and observed BMI) of 3.4 kg/m² (3.2, 3.6). In 2019, prediction error increased to 4.4 kg/m² (4.2, 4.6). For social class, equivalent figures were 3.3 kg/m² (3.2, 3.4) in 1994 and 4.4 kg/m² (4.3, 4.6) in 2019. For IMD, equivalent figures were 3.8 kg/m² (3.7, 3.8) in 2001 and 4.4 kg/m² (4.3, 4.6) in 2019.

While prediction errors increased in absolute size, there was some evidence that each measure of SEP explained a greater proportion of variation in BMI over time, as measured as the proportion of prediction error reduced by including education, social class, or IMD in the random forest model or, alternatively, by incremental R² (Fig. 3b, c; see Additional file 1: Fig. S4 for smoothed results). The improvement in prediction attributable to education was 0.14% (-0.90, 1.08) in 1991 and 1.05% (0.18, 1.82) in 2019 (Fig. 3b). (A trend of increasing predictive accuracy improvement from including education in models was also observed using data from 1991 to

2014 only.) Across the studied period, the total reduction in prediction error when including education, social class, or IMD in models was very small — less than 1.1% each year (see Additional file 1: Fig. S6 for model residuals). Equivalently, incremental R² was low: for education, 0.76% (0.29, 1.17) in 1994 and 1.57% (0.2, 2.62) in 2019 (Fig. 3c). Highlighting this, the ability of education, social class, or IMD to distinguish pairs of individuals at higher BMI levels was also generally poor. The probability of superiority derived from models including SEP was 0.59 or lower in each year — little different from the probability of superiority derived from models just including age and sex (Fig. 3d). Education was typically more predictive of BMI than social class or IMD (Fig. 3b), though the temporal increase in mean level differences between highest and lowest SEP groups was greatest for IMD (Fig. 2d; see also Additional file 1: Fig. S4).

Further analyses

Qualitatively similar results were obtained when linear regression was used instead of the random forest algorithm or when using the 5-year age group as a covariate, rather than the single-year age (results available on request). Qualitatively similar results were also obtained

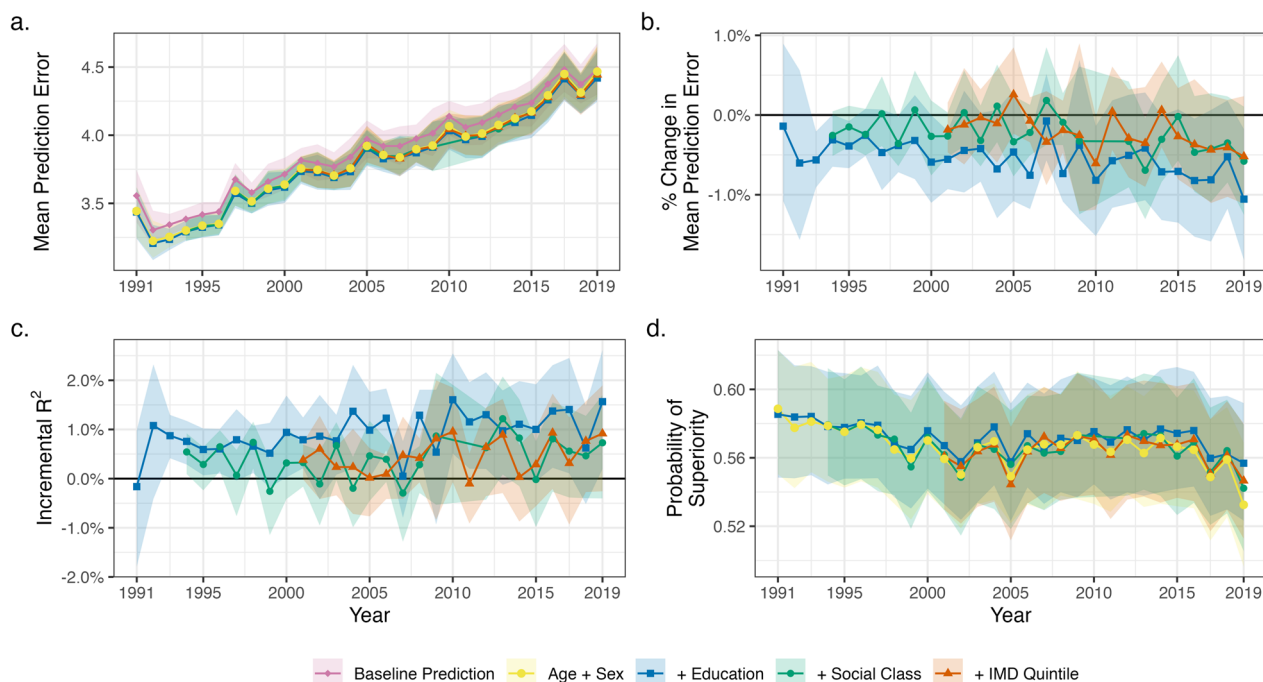


Fig. 3 Predictive accuracy of random forest models predicting individuals' BMI by survey year. **a** Mean absolute error of model predictions by model (i.e. average difference between predicted and observed BMI; baseline prediction uses sample mean, and other estimates are random forest models including stated covariates). Higher values are indicative of less accurate prediction. **b** Percentage reduction in prediction error when further including educational attainment, social class, or IMD in the random forest model (compared to the model including age and sex). **c** Incremental R² when further including educational attainment, social class, or IMD in the random forest model (compared to the model including age and sex). **d** Probability of superiority by model

when predicting obesity instead of BMI: social inequalities increased over time as did the proportional improvement in prediction when including SEP in models, but the overall predictive power of SEP was low (see Additional file 1: Results S1 and Additional file 1: Figs. S7–S10 for full detail). Larger social inequalities were found among women when stratifying the BMI analysis by sex (Additional file 1: Fig. S11). Population-level differences in mean BMI according to SEP were approximately twice as large among females compared with males. Accordingly, SEP improved individual-level predictions to a greater extent among females, though improvements in predictive accuracy remained low. The relative improvement in predictive accuracy across the study period was more clearly observed among females.

Discussion

Summary of results

The results demonstrate an increase in mean BMI and an increase in the variability of BMI between 1991 and 2019 in England, as well as an increase in the prevalence of obesity. Mean BMI and prevalence of obesity increased across all education groups and IMD quintiles, and there was an increase in social inequalities over time. However, variability in BMI *within* SEP groups also increased. While the ability of education, social class, and IMD to explain the between-person variability of BMI increased over the study period, explained variance remained low and absolute prediction errors increased in size. A broadly similar pattern of results was found when attempting to predict obesity. SEP further had limited utility in identifying, among pairs of individuals, the person with obesity or a higher BMI. Effect sizes were larger in females than males, and education was typically more predictive than social class or IMD.

Explanation of findings

These results are consistent with previous studies showing limited explanatory power of SEP for BMI [9, 13–15] and accord with studies showing increased variance within SEP groups over the obesity epidemic [2, 15, 22–25]. More generally, they are also consistent with findings that shared environmental factors explain limited variance across a wide range of behavioural and health-related traits (the ‘gloomy prospect’ of behaviour genetics [19, 20]), as well as with the results of a mass scientific collaboration study showing that socioeconomic outcomes are largely unpredictable even using rich longitudinal survey data [21]. Researchers in one study were able to predict 60% of the variance in BMI among older adults using deep learning methods and detailed socioeconomic, demographic, and other study data (> 450 variables) [39]. However, their analysis also

included several variables directly related to health, such as healthcare utilization.

Intriguingly, the observed small change in the proportion of variance explained by SEP as group-level BMI differences have increased is consistent with a model in which the effects of risk factors for high BMI have uniformly increased in strength over the obesity epidemic [40] — one study in Sweden found that genetic effects have similarly increased, while heritability has remained almost stable [41]. However, there are reasons to expect changes in the variation explained by education, including the changing distribution of education itself as the population has become more highly educated (see Fig. 1) and variation in the returns to education (i.e. through period and cohort effects in the effect of education on earnings) which could lead to differences in effect size, e.g. from changes in relative access to healthy foodstuffs.

Our results raise the question of why such low explanatory power of SEP is observed. One reason is that low SEP is neither a necessary nor a sufficient cause of high body weight. Instead, SEP is expected to operate distally at the end of long causal chains, the steps of which may be blocked, amplified, or attenuated in the presence or absence of other exposures. For instance, at a population level, neighbourhood deprivation may lead to higher BMI by influencing physical activity via affecting walkability [42], but some individuals may compensate by travelling to surrounding areas or may get sufficient exercise if they do physically demanding jobs. The effects of SEP on BMI may thus be heterogeneous, a process that would entail greater BMI variance within lower SEP groups, which is observed in practice [2]. Furthermore, extremely strong effect sizes — stronger than those found in typical epidemiological studies — are required to obtain good predictive power at the individual level [43]. As such, while SEP had an increasingly large effect size on BMI across time, it was not sufficiently large to yield accurate predictions at the individual level.

Our results may have implications for efforts to tackle obesity rates. Assuming the link between SEP and BMI is causal (an assumption supported in some, but not all, quasi-experimental studies; [44]), our results suggest that reducing the social gradient in BMI could reduce but not reverse the obesity epidemic: consistent with other work [2], our results show that obesity rates have increased among all social groups while inequalities within these groups have also increased over time. As has been previously argued, the increasing variability of BMI could mean a one-size-fits-all approach may not be effective as increased variability may reflect distinct determinants [45]. We should, however, note that predicting the effects of intervening on SEP or its mediating pathways

is challenging, partly as it is possible that inequality itself could increase obesity rates [46].

Despite an increasing association between SEP and BMI at the population level, the results suggest limited utility of the use of SEP indicators in predictive algorithms for obesity or BMI. Algorithms to predict obesity based on high-level SEP data are likely to have an unacceptably low sensitivity and specificity — focusing only on those with low SEP would miss the majority of cases. Including SEP in models may be justified for health equity reasons, however [47]; without its inclusion, risk will be systematically underestimated for low SEP individuals.

While SEP does not explain much of the between-person variation in BMI, determining its predictive ability is important as it can motivate the development of more complex and specific theories and highlight the need for other non-standard but highly predictive data. Genetic data are increasingly available — polygenic scores for BMI now achieve R^2 of 15% [48] — but text or other ‘big’ data could also be useful. A recent study mining the content and style of essays written at age 11 explained approximately 60% of the variability in childhood cognitive ability [49], though the ability to predict BMI and other physical health measures is unlikely to be this high.

Strengths and limitations

Strengths included objective measurement of BMI and use of data spanning almost three decades of the obesity epidemic in England, though for a small number of individuals with particularly high weight, self-reports were used instead. We examined measures of individual- and area-level SEP, measures that are easy to collect (and thus may appear in predictive algorithms) and have been widely studied in the social inequality literature previously. Nevertheless, due to data limitations, some dimensions of SEP (such as income) were not examined, and the variables that were used were relatively high level and restricted to a small number of categories, limiting potential predictive accuracy. The measures were also based on current SEP; life course measures of SEP — or of body weight (e.g. ever having obesity) — may have yielded more accurate predictions (though the gloomy prospect makes us circumspect as to the degree of improvement). Improvements in predictive accuracy may also have been greater if covariates other than age and sex were included in models as this would allow for the determination of more granular interaction effects. Future work should examine a larger and more detailed suite of socioeconomic data.

Though HSE is designed to be representative, non-response increased over the study period, consistent with other cross-sectional health surveys and several

longitudinal studies [50, 51]. Non-response may have been related to BMI or SEP. Previous work has shown that, among eventual HSE participants, individuals from more deprived areas or with highest or lowest incomes required more contact attempts, on average [52], and, in a major UK birth cohort, obesity was related to lower participation in a midlife biomedical sweep [53]. While we used survey weights, differential non-response could have reduced the predictive accuracy of SEP and biased time-related changes.

We also focused on White non-student or foreign-educated participants for comparability across time — results may not generalize to other sections of the population. The HSE data are cross-sectional. Assuming that our estimates at least partly confounded (see, e.g. [54]), we are likely to have obtained optimistic estimates of predictive accuracy, relative to intervening directly on SEP. Finally, the random forest models may have been too flexible and overfit the data, producing poor out-of-sample predictions. Nevertheless, using ordinary least squares (OLS) regression yielded similar results.

Conclusions

While absolute inequalities in BMI and obesity according to education and neighbourhood deprivation increased in England between 1991 and 2019, within-group inequalities also increased and were large relative to between-group inequalities, contributing to the weak explanatory power of SEP. Though explanatory power increased over the study period, it remained low which suggests that reducing inequality is unlikely to reverse the large impact on the obesity rates which increased across all SEP groups since the beginning of the obesity epidemic. Nevertheless, the possibility of heterogeneous effects of SEP means that targeted attention within SEP groups could be fruitful.

Abbreviations

AUC	Area under the receiver operating curve
BMI	Body mass index
GAM	Generalized additive model
HSE	Health Survey for England
IMD	Index of multiple deprivation
NVQ	National Vocational Qualification Scheme
OLS	Ordinary least squares
SEP	Socioeconomic position

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12916-023-03103-2>.

Additional file 1: Fig. S1. Household response rate by survey year (Sources: [28, 29]). **Fig. S2.** Distribution of BMI by year (two years grouped together), Health Survey for England. **Fig. S3.** Proportion in each educational category by year, Health Survey for England. **Fig. S4.** Smoothed results of random forest models predicting BMI. **Fig. S5.** Standard

deviation (SD) of BMI within each SEP group by survey year. **Fig. S6**. Distribution of prediction errors by year and model. **Fig. S7**. Results of random forest models predicting (probability of) obesity by survey year. **Fig. S8**. Smoothed results of random forest models predicting obesity. **Fig. S9**. Results of random forest models predicting (probability of) obesity by survey year. **Fig. S10**. Difference in the average predicted probability of obesity between obese and non-obese participants by survey year and model. **Fig. S11**. Results of random forests predicting BMI by survey year and sex.

Additional file 2: Table S1. Distribution of Socioeconomic Position Measures and Covariates (Age and [Grouped] Sex) by Year.

Acknowledgements

The authors would like to thank participants at a Centre for Longitudinal Studies departmental seminar for comments on the work.

Authors' Twitter handles

David Bann: @davidabann; Richard Silverwood: @RJ_Silverwood; Charis Bridger Staatz: @CharisStaatz.

Authors' contributions

All authors contributed to and agreed upon the analysis plan. LW carried out the analysis and wrote the first draft of the manuscript. DB, RJ, and CBS provided critical revisions. All authors read and approved the final manuscript.

Funding

The funders had no final role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. All researchers listed as authors are independent from the funders and all final decisions about the research were taken by the investigators and were unrestricted. DB, LW and RS are supported by the Economic and Social Research Council (ES/M001660/1) and DB and LW by the Medical Research Council (MR/V002147/1). CBS is funded by the ERSC (ES/V012789/1) and the National Institute of Health Research (COV-LT-0009–28654).

Availability of data and materials

Health Survey for England data are available via the UK Data Service [27]. The code used to run the analysis is available at <https://osf.io/smd7z/>.

Declarations

Ethics approval and consent to participate

This is a secondary data analysis. Ethical approval for the Health Survey for England was obtained from East Midlands Nottingham 2 Research Ethics Committee (15/EM/0254).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 26 June 2023 Accepted: 4 October 2023

Published online: 13 November 2023

References

- Hancock C. Patterns and trends in excess weight among adults in England. UK Health Security Agency. 2021. <https://ukhsa.blog.gov.uk/2021/03/04/patterns-and-trends-in-excess-weight-among-adults-in-england/>.
- Green MA, Subramanian SV, Razak F. Population-level trends in the distribution of body mass index in England, 1992–2013. *J Epidemiol Community Health*. 2016;70:832–5.
- Tremmel M, Gerdtham U-G, Nilsson PM, Saha S. Economic burden of obesity: a systematic literature review. *Int J Environ Res Public Health*. 2017;14:435.
- Newton S, Braithwaite D, Akinyemiju TF. Socio-economic status over the life course and obesity: systematic review and meta-analysis. *PLoS ONE*. 2017;12: e0177151.
- Sommer I, Griebler U, Mahlknecht P, Thaler K, Bouskill K, Gartlehner G, et al. Socioeconomic inequalities in non-communicable diseases and their risk factors: an overview of systematic reviews. *BMC Public Health*. 2015;15:914.
- Office for Health Improvement and Disparities. Obesity Profile. Fingertips Public Health Data. 2021. https://fingertips.phe.org.uk/profile/national-child-measurement-programme/data#page/7/gid/1938133368/pat/159/par/K02000001/ati/15/are/E92000001/iid/93881/age/168/sex/4/cat/-1/ctp/-1/yr/1/cid/4/tbm/1/page-options/ine-yo-1:2020-1:-1_ine-pt-0_ine-ct-115. Accessed 2 Mar 2023.
- Bann D, Johnson W, Li L, Kuh D, Hardy R. Socioeconomic inequalities in body mass index across adulthood: coordinated analyses of individual participant data from three British birth cohort studies initiated in 1946, 1958 and 1970. *PLoS Med*. 2017;14:e1002214.
- Hoffmann K, De Gelder R, Hu Y, Bopp M, Vitrai J, Lahelma E, et al. Trends in educational inequalities in obesity in 15 European countries between 1990 and 2010. *Int J Behav Nutr Phys Act*. 2017;14:63.
- Zhang Q, Wang Y. Trends in the association between obesity and socioeconomic status in U.S. adults, to 2000. *Obes Res*. 1971;2004(12):1622–32.
- Kino S, Hsu YT, Shiba K, Chien YS, Mita C, Kawachi I, et al. A scoping review on the use of machine learning in research on social determinants of health: trends and research prospects. *SSM Popul Health*. 2021;15:100836.
- Katsoulis M, Lai AG, Diaz-Ordaz K, Gomes M, Pasea L, Banerjee A, et al. Identifying adults at high-risk for change in weight and BMI in England: a longitudinal, large-scale, population-based cohort study using electronic health records. *Lancet Diabetes Endocrinol*. 2021;9:681–94.
- Health TLP. Tackling obesity seriously: the time has come. *The Lancet Public Health*. 2018;3:e153.
- Hamad R, Glymour MM, Calmasini C, Nguyen TT, Walter S, Rehkopf DH. Explaining the variance in cardiovascular disease risk factors: a comparison of demographic, socioeconomic, and genetic predictors. *Epidemiology*. 2022;33:25.
- Bann D, Wright L, Hardy R, Williams DM, Davies NM. Polygenic and socioeconomic risk for high body mass index: 69 years of follow-up across life. *PLoS Genet*. 2022;18:e1010233.
- Vaezghasemi M, Razak F, Ng N, Subramanian SV. Inter-individual inequality in BMI: an analysis of Indonesian Family Life Surveys (1993–2007). *SSM Popul Health*. 2016;2:876–88.
- Kim R, Lippert AM, Wedow R, Jimenez MP, Subramanian SV. The relative contributions of socioeconomic and genetic factors to variations in body mass index among young adults. *Am J Epidemiol*. 2020;189:1333–41.
- Kim R, Kawachi I, Coull BA, Subramanian SV. Contribution of socioeconomic factors to the variation in body-mass index in 58 low-income and middle-income countries: an econometric analysis of multilevel data. *Lancet Glob Health*. 2018;6:e777–86.
- Elks CE, den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJF, et al. Variability in the heritability of body mass index: a systematic review and meta-regression. *Front Endocrinol (Lausanne)*. 2012;3:29.
- Plomin R, Daniels D. Why are children in the same family so different from one another? *Behav Brain Sci*. 1987;10:1–16.
- Smith GD. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *Int J Epidemiol*. 2011;40:537–62.
- Salganik MJ, Lundberg I, Kindel AT, Ahearn CE, Al-Ghoneim K, Almaatouq A, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc Natl Acad Sci USA*. 2020;117:8398–403.
- Krishna A, Razak F, Lebel A, Smith GD, Subramanian SV. Trends in group inequalities and interindividual inequalities in BMI in the United States, 1993–2012. *Am J Clin Nutr*. 2015;101:598–605.
- Kim S, Subramanian SV, Oh J, Razak F. Trends in the distribution of body mass index and waist circumference among South Korean adults, 1998–2014. *Eur J Clin Nutr*. 2018;72:198–206.
- Lebel A, Subramanian SV, Hamel D, Gagnon P, Razak F. Population-level trends in the distribution of body mass index in Canada, 2000–2014. *Can J Public Health*. 2018;109:539–48.

25. Wagner KJP, Boing AF, Cembranel F, Boing AC da S, Subramanian SV. Change in the distribution of body mass index in Brazil: analysing the interindividual inequality between 1974 and 2013. *J Epidemiol Community Health*. 2019;73:544–8.
26. McAllister EJ, Dhurandhar NV, Keith SW, Aronne LJ, Barger J, Baskin M, et al. Ten putative contributors to the obesity epidemic. *Crit Rev Food Sci Nutr*. 2009;49:868–913.
27. NatCen Social Research, University College London. Health Survey for England. 2023. <https://doi.org/10.5255/UKDA-SERIES-2000021>.
28. Mindell J, Biddulph JP, Hirani V, Stamatakis E, Craig R, Nunn S, et al. Cohort profile: the health survey for England. *Int J Epidemiol*. 2012;41:1585–93.
29. NHS Digital. Health survey for England. 2023. <https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england>. Accessed 4 Sept 2023.
30. Bhaskaran K, dos-Santos-Silva I, Leon DA, Douglas IJ, Smeeth L. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3.6 million adults in the UK. *Lancet Diabetes Endocrinol*. 2018;6:944–53.
31. Galobardes B. Indicators of socioeconomic position (part 1). *J Epidemiol Community Health*. 2006;60:7–12.
32. McLaren L. Socioeconomic status and obesity. *Epidemiol Rev*. 2007;29:29–48.
33. Office for National Statistics. Education, England and Wales: Census 2021. 2023. <https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/bulletins/educationenglandandwales/census2021#measuring-the-data>. Accessed 4 Sept 2023.
34. Johnson W, Li L, Kuh D, Hardy R. How has the age-related process of overweight or obesity development changed over time? Co-ordinated analyses of individual participant data from five United Kingdom birth cohorts. *PLoS Med*. 2015;12:e1001828.
35. Kanerva N, Kontto J, Erkkola M, Nevalainen J, Männistö S. Suitability of random forest analysis for epidemiological research: exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design. *Scand J Public Health*. 2018;46:557–64.
36. Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS ONE*. 2019;14:e0214365.
37. Kolenikov S. Resampling variance estimation for complex survey data. *Stand Genomic Sci*. 2010;10:165–99.
38. Bann D, Fluharty M, Hardy R, Scholes S. Socioeconomic inequalities in blood pressure: co-ordinated analysis of 147,775 participants from repeated birth cohort and cross-sectional datasets, 1989 to 2016. *BMC Med*. 2020;18:338.
39. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Population Health*. 2018;4:95–9.
40. Domingue BW, Kanopka K, Mallard TT, Trejo S, Tucker-Drob EM. Modeling interaction and dispersion effects in the analysis of gene-by-environment interaction. *Behav Genet*. 2022;52:56–64.
41. Rokholm B, Silventoinen K, Tynelius P, Gamborg M, Sørensen TIA, Rasmussen F. Increasing genetic variance of body mass index during the Swedish obesity epidemic. *PLoS ONE*. 2011;6:e27135.
42. Zandieh R, Flacke J, Martinez J, Jones P, Van Maarseveen M. Do inequalities in neighborhood walkability drive disparities in older adults' outdoor walking? *Int J Environ Res Public Health*. 2017;14:740.
43. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159:882–90.
44. Galama T, Lleras-Muney A, van Kippersluis H. The effect of education on health and mortality: a review of experimental and quasi-experimental evidence. Cambridge, MA: National Bureau of Economic Research; 2018.
45. Razak F, Davey Smith G, Subramanian S. The idea of uniform change: is it time to revisit a central tenet of Rose's "Strategy of Preventive Medicine"? *Am J Clin Nutr*. 2016;104:1497–507.
46. Pickett KE, Kelly S, Brunner E, Lobstein T, Wilkinson RG. Wider income gaps, wider waistbands? An ecological study of obesity and income inequality. *J Epidemiol Community Health*. 2005;59:670–4.
47. Fiscella K, Tancredi D. Socioeconomic status and coronary heart disease risk prediction. *JAMA*. 2008;300:2666–8.
48. Becker J, Burik CAP, Goldman G, Wang N, Jayashankar H, Bennett M, et al. Resource profile and user guide of the Polygenic Index Repository. *Nat Hum Behav*. 2021;5:1744–58.
49. Wolfram T, Tropf FC. Man - machine - gene: predicting cognitive ability, non-cognitive traits and educational attainment from teacher assessments, short essays and the genome. preprint. SocArXiv; 2022.
50. Bann D, Johnson W, Li L, Kuh D, Hardy R. Socioeconomic inequalities in childhood and adolescent body-mass index, weight, and height from 1953 to 2015: an analysis of four longitudinal, observational, British birth cohort studies. *Lancet Public Health*. 2018;3:e194–203.
51. Mindell J, Giampaoli S, Goesswald A, Kamtsiuris P, Mann C, Mannisto S, et al. Sample selection, recruitment and participation rates in health examination surveys in Europe - experience from seven national surveys. *BMC Med Res Methodol*. 2015;15:78.
52. Boniface S, Scholes S, Shelton N, Connor J. Assessment of non-response bias in estimates of alcohol consumption: applying the continuum of resistance model in a general population survey in England. *PLoS ONE*. 2017;12:e0170892.
53. Atherton K, Fuller E, Shepherd P, Strachan DP, Power C. Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort. *J Epidemiol Community Health*. 2008;62:216–23.
54. Davies NM, Dickson M, Davey Smith G, van den Berg GJ, Windmeijer F. The causal effects of education on health outcomes in the UK Biobank. *Nat Hum Behav*. 2018;2:117–25.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

