Analysis of Health Survey for England (HSE) 2019

Candidate Numbers Here

March 13, 2024

Abstract

This report provides an analysis of data related to health, age, socio-economic factors and lifestyle habits in adults (from the age of 16) from the population in England, derived from the Health Survey for England 2019.

Summary (Non-Technical)

Introduction

In the UK, smoking, "vaping" and alcohol consumption are widespread habits among adults, particularly younger adults. In fact, it is estimated that around 14.1% of UK adults identified as cigarette smokers (ONS 2019a), and a similar study found that there were 7,565 deaths attributed to alcohol-specific causes in 2019 (ONS 2019b). Thus, it is crucial to be aware of the predictors and consequences of these habits, which is the motivation for this analysis. We conducted this analysis on a subset containing 8,204 adults living in England (ages 16+), who were interviewed in their homes about their demographics and their smoking and drinking habits as part of the Health Survey for England (HSE) 2019. 4,947 of these participants who consented to an at-home nurse visit had vitals such as blood pressure, weight and height measured. We investigated which features of a participant are associated with smoking habits, and which habits are associated with high blood pressure. The mechanisms which drive lifestyle habits are part of a complex and ever-changing field in behavioural psychology. What is known, however, is that such habits are driven by cues and cravings (ask Cam for the citation). The HSE 2019 captures demographic and socioeconomic characteristics that have the potential to make such cues more visible and cravings harder to resist. For example, living in a region where smoking is more prevalent, like ones with higher deprivation [citation needed], or not having a husband/wife for an accountability partner to help you quit or resist these habits. In this report, we will use weighting variables to estimate prevalence of cigarette, e-cig, and alcohol users in our population, and compare it with the larger population of adults in England. We will then attempt to identify associations with a participants current smoking status, explore whether predictive modelling can be used to identify smokers with a high accuracy, and investigate which lifestyle habits are associated with increased values of systolic blood pressure. *This will be crucial in enabling healthcare providers to implement preventative care for people at risk of developing hypertension.*

Exploratory Analysis

The full HSE 2019 cross-sectional data set contains 10,299 observations across thousands of variables [cite the dataset here], but we only studied patients over the age of 16 among key variables. Our subset includes 8,204 participants and 19 variables, which are grouped into five categories and described in Table ??.

Table ??

We found 36 pairs of observations with exactly equal variables (excluding ID variables and lab measurements), but we did not remove these from our analysis because the supporting documentation doesn't state a protocol for repeated visits. We assumed these were genuine observations from different participants and thus included them in our dataset. However, there were 3 pairs of observations that had duplicate lab variables also. We believed that these were true duplicates and removed one observation from each of the three pairs resulting in a dataset of 8201 observations.

All variables are coded as numeric in our dataset, so we recoded all factor variables accordingly. Some of the variables used for analysis were dichotomised by us for easier interpretation. We coded a binary variable indicating current smoking status, which will serve as our *primary* outcome variable. Also, we dichotomised urbanity into two levels being 'Rural' and 'Urban', and finally marital status was dichotomised into 'Married' and 'Not Married'. We group respondents into 'Higher Education', 'Further Education', 'A-Level equiv.', 'GCSE equiv.', 'No qualification' and 'Foreign/other'.

The number of missing observations for each are shown in Table 1:

Table 1: Missing values in the training dataset

Variable	Missing Values	% Missing
omsysval	4036	61.5%
BMIVal	1519	23.2%
dnoft_19	1496	22.8%
cigdyal_19	57	0.869%

Variable	Missing Values	% Missing
cigsta3_19	56	0.854%
NDPNow_19	53	0.808%
d7many3_19	52	0.793%
drinkYN_19	51	0.777%
topqual2	46	0.701%
origin2	29	0.442%
marstatD	1	0.0152%

There is significant missingness in the lab values, which can be explained by the fact additional consent was needed from the participant to allow a nurse visit. One variable with a substantial amount of data missing was the frequency of drinking in the last 12 months, which due to the retrospective and sensitive nature of the question could be explained by either recall bias or a participant's refusal to answer. Education level, marital status and ethnicity are the only demographic variables with any missing entries, with 46 observations (0.701% of the data) missing at least one of the three. These are not necessary for identifiability and as they are sensitive, we do not expect every participant to answer these questions. Therefore, we did not remove these observations.

We analysed the two lab measurement variables in Figure 1, finding potential outliers to be present in both. There were many outliers for BMI, which we note used self-estimated weight from the participant in the calculation when it exceeded 130kg. Additionally, weight measurements were taken in inconsistent environments such as on different flooring in the participants' homes which has been shown in previous research to impact measurements accuracy [citation needed]. We believed this could have explained the extremely high BMI values such as *LIST EXAMPLES*, and so coded them as missing in the analysis dataset.

The readings for systolic blood pressure were collected by taking an average of three readings, each five minutes apart, performed by a trained nurse who had to declare each reading to be valid. For this reason, it seemed highly unlikely that these readings were mistakes, and so we included them in our analysis. We did, however, note that values of systolic blood pressure that were this consistently high (>140mmHg) were indicators for hypertensive crisis, and so a part of our population may have serious underlying health conditions that could affect the generalisability of our findings [citation needed].

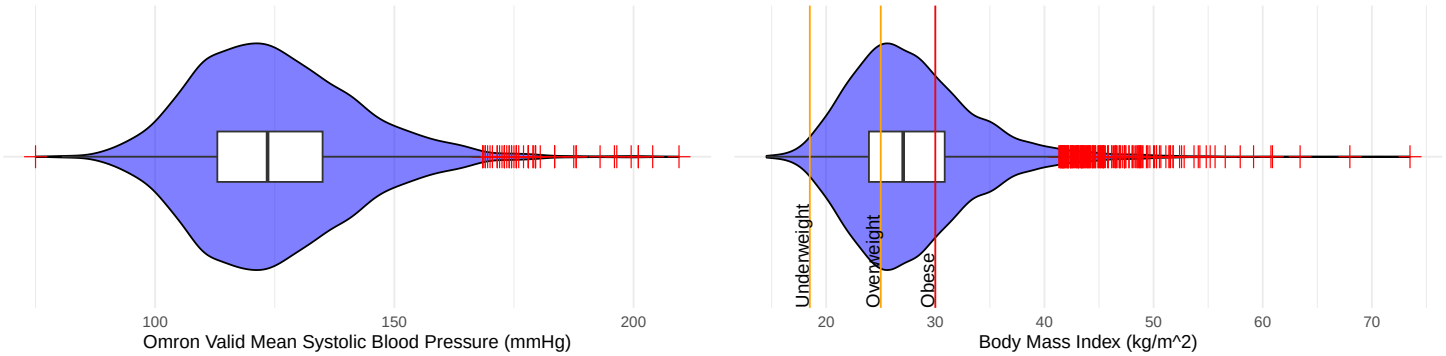

Figure 1: Distribution of BMI and Mean Systolic Blood Pressure

Methology

What is the prevelance of drinking, smoking and E-cig usage?

To calculate the prevalence of each habit we assumed each of the n observations, $x_1, \ldots, x_n$, to be independent, identically distributed (iid) random variables (RVs) where $x_i \sim Bern(p)\, \forall i = 1, \ldots, n$ and p denotes the probability of an observation having the relevant habit. We use the household weights to calculate a weighted Maximum Likelihood Estimate (MLE) of p. That is, letting w_i denote the weight of the i^{th} observation, we altered the

standard likelihood function of a Bernoulli distribution as below:

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} (p^{x_i}(1-p)^{1-x_i})^{w_i}$$

From this, we calculated our weighted MLE as $\widehat{p} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$. It can also be shown that this MLE has variance given by $\text{var}(\widehat{p}) = \frac{p(1-p)}{\sum_{i=1}^{n} w_i}$, which we estimated using $\widehat{p}$. We used large sample properties of the MLE to get a normal approximation and estimated 95% confidence intervals for each habit, which are shown in Table 2.

Table 2: Estimates and 95% Confidence Intervals for % of Population

Habit	Estimate	C.I.
Drinking	80.4%	(79.5%, 81.2%)
Smoking	16.5%	(15.7%, 17.3%)
Smoking E-cigarettes	4.28%	(3.84%, 4.72%)

We found that the usage of e-cigarette usage among adults is relatively low, making it challenging to dissect any significant trends within the data. *Maybe a wee bit more here*

Next, we worked to uncover factors that have possible associations with smoking prevalence. We started with the demographic factors of age and gender, plotting the smoking prevalence across combinations of these groups to identify any patterns. Figure 3 suggests a negative correlation between age and the proportion of current cigarette smokers, across both genders. Interestingly, the proportion of males who quit smoking in later-life was much greater than that of females (75yrs+; M: 50%, F: 31%), which could be attributed to men being at a higher risk of smoking-related diseases [citation needed].

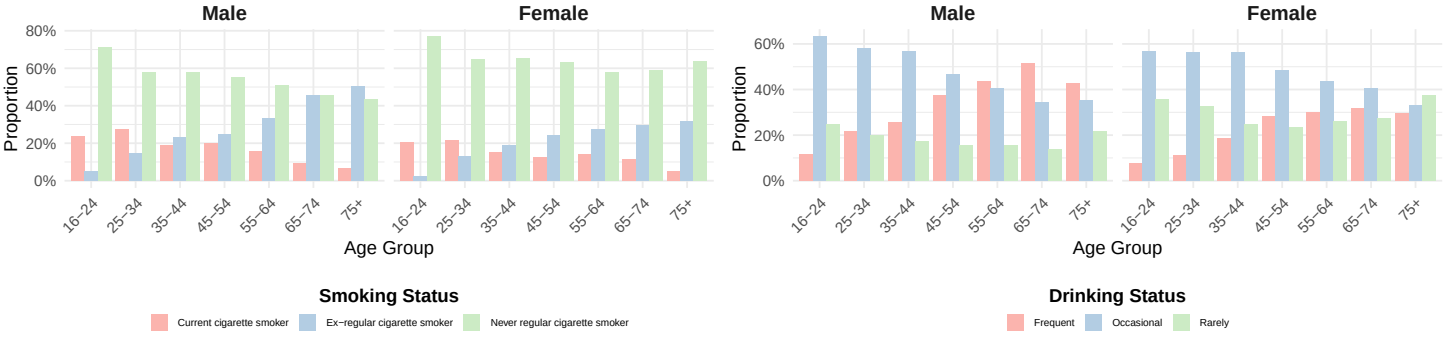

Figure 2: Smoking and drinking status proportions by age group and gender

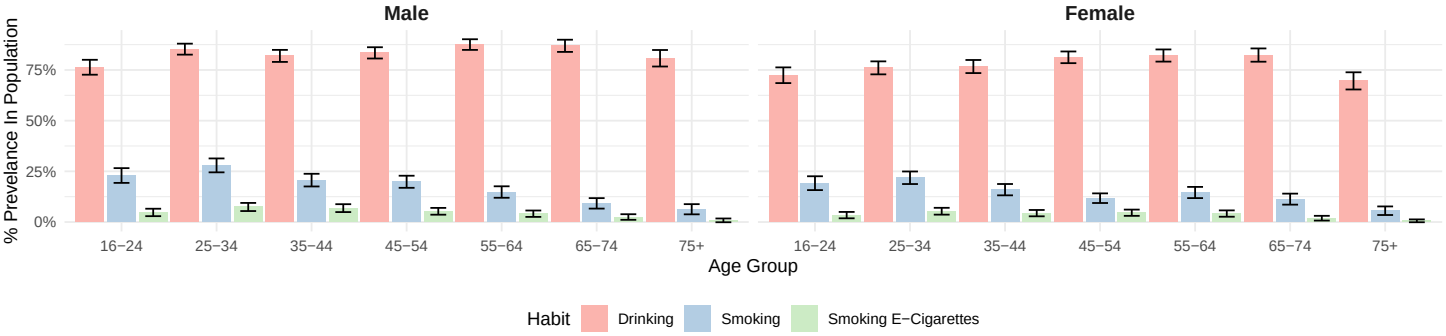

Figure 3: Estimation of the prevelance of smokers by deprivation

Additionally, males demonstrate a higher prevalence of drinking across nearly all demographic categories in comparison to females. We can clearly see that smoking prevalence increases with our deprivation variable, meaning

that smokers tend to be from lower levels of deprivation. This correlation may be caused by the hefty tobacco duties the UK impose on its' residents (GOV.UK 2014). Less deprived individuals they can take-up more unnecessary, expensive habits, with smoking being one of the main candidates.

How is smoking associated with socioeconomic factors and age?

We reserved 80% of our data to train the model, and used the remaining 20% to test the model, which was made possible due to the large size of our dataset. The training set contained 6561 observations and the test set contained 1640. *Table Z summarises the key variables between the test and train dataset to illustrate that both are representative of the whole dataset.* This reduced the risk of overfitting and allowed us to test the predictive power of each model we proposed.

To develop a predictive model for identifying smokers, we used the binary variable for current smoking status as a response, with various socioeconomic and demographic factors as predictors. These were selected based on the associations suggested by our prior analysis.

In the dataset, there were two variables that categorised age. They differed only in the size of the bands used, with one using 10-year bands and one using 5-year bands. After comparing identical models that used one or the other, we found that modelling with smaller bands improved the effectiveness of the model.

Under the assumption that ages were approximately uniformly distributed within their respective age bands, we also defined the estimated age of the i^{th} observation as the midpoint of their respective 5-year age bracket (taking the estimation of the 90+ category as 92.5), denoted a_i. We found that age may represent a somewhat quadratic effect on the probability of smoking, leading us to include a a_i^2 term in our model. A comparison of these models are summarised in Table 3.

Table 3: Comparison of selected model evaluations

Linear Predictor	Train AIC	Test AUC	Train RMSE	Test RMSE	Test Accuracy
$\text{logit}(\mu_i) \sim a_i + a_i^2 + m_i + q_i + u_i + o_i + t_i + s_i + q_i : u_i$	4987.5	0.759	0.342	0.324	0.861
$\text{logit}(\mu_i) \sim a_i + a_i^2 + m_i + q_i + u_i + o_i + t_i + s_i$	4987.3	0.759	0.342	0.324	0.861
$\text{logit}(\mu_i) \sim a_i + m_i + q_i + u_i + o_i + t_i + s_i$	5036.4	0.740	0.343	0.327	0.861
$\text{logit}(\mu_i) \sim a_i^{(5)} + m_i + q_i + u_i + o_i + t_i + s_i$	4994.3	0.759	0.341	0.325	0.861
$\text{logit}(\mu_i) \sim a_i^{(10)} + m_i + q_i + u_i + o_i + t_i + s_i$	5005.9	0.753	0.342	0.324	0.860

We selected the model based on AIC, which enabled us to balance model complexity and model fit. This model is *change factors*

$$\text{logit}(\mu_i) \sim \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3 q + \alpha_j^m + \alpha_k^u + \alpha_l^o + \alpha_m^t + \alpha_n^s + \beta_k^* q$$

where $j \in \{Married, Not\,Married\}$, $k \in \{Urban, Not\,Urban\}$, $l \in \{White, ..., Other\}$, $n \in \{Male, Female\}$ and $m \in \{Further\,Education, ..., No\,qualification\}$.

To evaluate the predictive performance of our final model using our test data, Figure 4 demonstrates the predicted probability of each 'probability bin' against the mean actual outcomes. As we can see the calibration curve closely follows the line $y = x$, which is indicative of a well-calibrated model. *This model doesn't predict high values...*

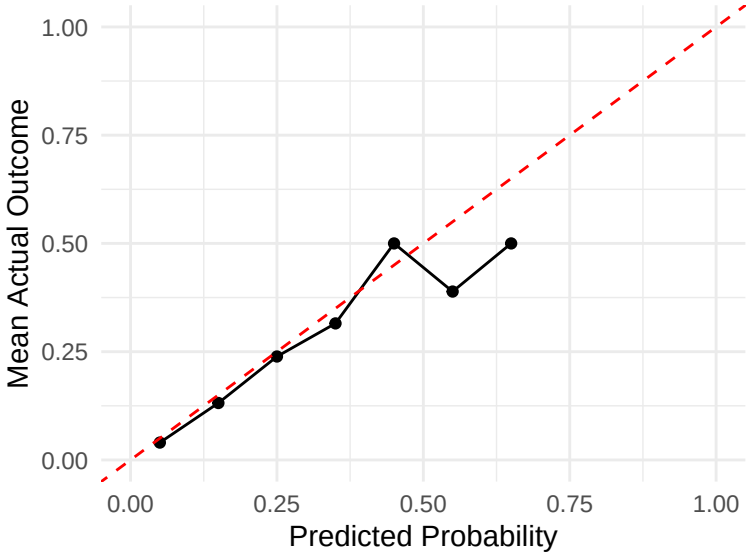

Figure 4: Calibration chart for Binomial model

Which lifestyle habits are associated with systolic blood pressure?

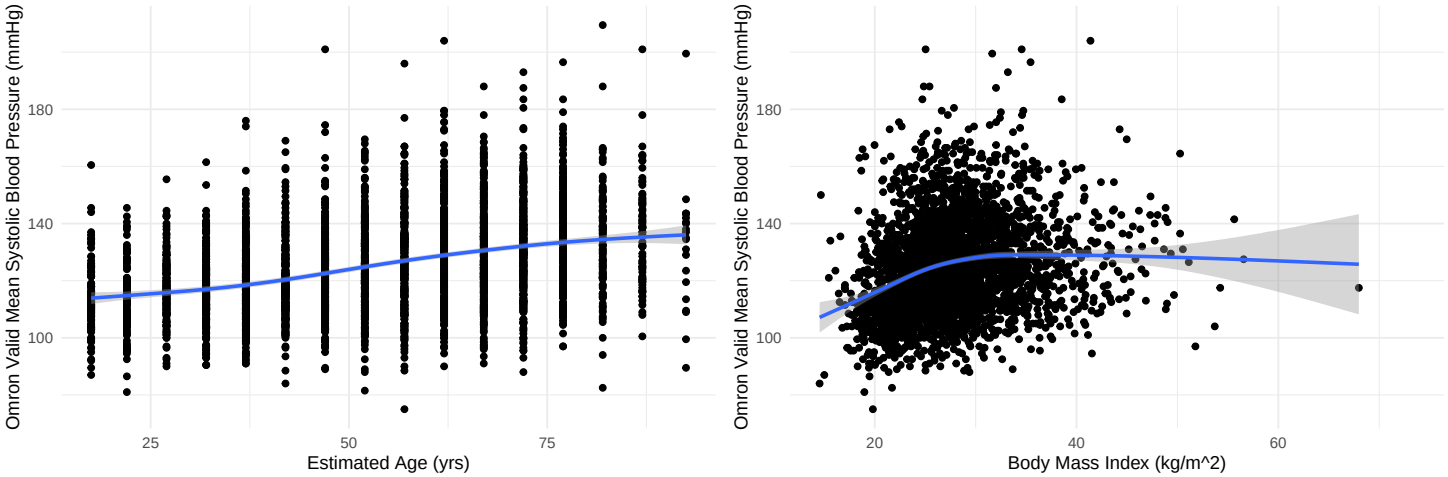

Figure 5: Relationship of BMI and Age with Mean Systolic Blood Pressure

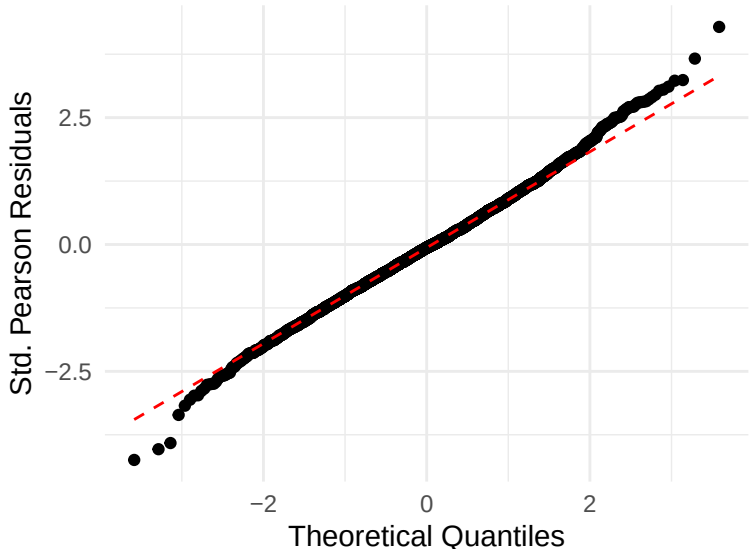

Figure 6: Q-Q plot of Inverse-Normal model residuals

Results/Conclusion

Our analysis showed that alcohol consumption is extremely common among UK adults, with approximately 80.4% being consumers, while smoking and vaping rates are lower at 16.5% and 4.3%, respectively. Older males show the highest tendencies for alcohol consumption, with 'frequent' drinking the most prevalent among males aged 65-74. Unlike 'frequent' alcohol consumption, the prevalence of smoking decreases with age. Individuals aged 16-24 appear to be the worst offenders when it comes to smoking cigarettes, indicating a significant issue within youth culture. Moreover, we found an interesting relationship between deprivation levels and smoking and drinking behaviours. Smoking prevalence seemed to increase as deprivation levels decrease, while alcohol consumption appeared to do the opposite.

To help us uncover the underlying socio-economic factors that may drive the prevalence of smoking, we first looked at the most 'comparable' respondent type in our dataset. This 'respondent' is a white, single male who has no qualifications and lives in an urban area. We found that the probability of our hypothetical respondent being a smoker is approximately 25% *Where is this from?* (notably higher than the population average of 16.5). *Maybe worth talking about confidence intervals here - is value outside 95%* The only socio-economic variables to certainly increase this probability is the deprivation level our respondent falls under. The rest of the socio-economic variables we have access to, appear to decrease the likelihood of smoking. For example, if our respondent is any of the following: Asian, black, married, or went on to further education, the chances of them being a smoker is at-least halved.

NatCen Social Research and Health (2019) *?*

References

GOV.UK. 2014. "Tax on shopping and services." https://www.gov.uk/tax-on-shopping/alcohol-tobacco.

NatCen Social Research, Department of Epidemiology, University College London, and Public Health. 2019. "Health Survey for England." http://doi.org/10.5255/UKDA-SN-8860-1.

ONS. 2019a. "Adult smoking habits in the UK: 2019." https://shorturl.at/qQW27.

———. 2019b. "Alcohol-specific deaths in the UK: registered in 2019." https://shorturl.at/gqxY1.