University of Essex

Department of Mathematical Sciences

MA981-7-FY-CO Dissertation

Finding a causal relationship between Body Mass Index with the mental and subjective wellbeing of an individual using Mendelian Randomization

Registration number: 1906463

Supervisor: Dr. Yanchun Bao

Date of submission (August 28)

Word count: ”11038”

Contents

1 Abstract

Studies have suggested that keeping an eye on our BMI could reveal early signs of many chronic diseases[1]. Generally, our BMI is a reflection of our lifestyle, and it is itself a measure of our overall physical health and subjective well-being[2]. Studies [3][4][5] have suggested 12-item General Healthcare Questionnaire(GHQ) is a widely accepted measure of psychological stress, having a non-linear association with BMI. Individuals with normal weight are found to have a lower GHQ score and a higher score was observed in under-weighted, obese I and obese II/III categories. GWAS studies[6][7] have identified Single-nucleotide polymorphisms (SNPs) with a strong associated with BMI and could be used in further studies as a valid Instrument Variable (IV)[8][9][10] to find an estimate of a causal estimate of risk factors associated with a disease using Randomized Control Trial(RCT). We used the 'MendelianRamdomization' R package[11] for the application of Mendelian Randomization(MR) methods in this project.

Our study aimed to find an estimate of the causal effect of BMI considered as a risk factor, on mental and subjective wellbeing, using human genetic data. We have carried out a series of MR methods on three different measures of mental and subjective wellbeing: GHQ, overall life satisfaction questionnaire scores and scored for sf12mcs. In statistical analysis we have found significant association of BMI with our instrument variable (weighted and unweighted polygenic score (PGS)) to fullfill first assumption of IV. After applying various MR methods and comparing their results, our overall finding is that the intercept of Egger regression is close to zero and p-vale us is greater than the threshold in all cases, it indicates that there is no evidence that direct pleiotropic effect exists and therefore both the IVW and TSLS result are valid. Although TSLS and IVW do not provide consistent results. Using TSLS, we have found a significant association of BMI with GHQ for entire data and male subset, we observed a significant p-value for GHQ, using both weighted and unweighted PGS. One of the reasons that we failed to find an association in all other cases, is that the provided IVs are not strong enough or our sample size is small. Since we have been given with SNPs information with a very small coefficient it indicates at a weak IV. To tackle this problem, we

have added them together to make them strong IV, hoping to solve the issue, but unfortunately, it failed.

Keywords:Subjective wellbeing, mental wellbeing, Genetic variants, Mendelian Randomization (MR).

2 Introduction

Being healthy does not mean being free from diseases, but a sense of complete physical, mental, and social well being. In this busy and stressful life, we merely care about our mental and social health, but mental and social well beings are essential for feeling good. Depression is the fourth leading cause of disease burden[4]. Today with the advancement in medicine and technology, researchers are working on finding traits to establish a better understanding of the causation of a disease by studying human DNA. A healthy life is a key to a happy life, as a famous saying 'health is wealth'. A human being adopts 23 pair of chromosomes (46 total), 23 from each parent, and a genotype (the respective combination of chromosomes from each parent, know as genes) results in a phenotype (physical appearance, development, and behaviour) variation. The four bases of DNA named as Adenine (A), cytosine (C), guanine (G), and thymine (T), stores genetic information which can encode phenotype.

If the variation is present in less than 1% of the population, it is termed as a mutation, else is it known as common Single Nucleotide Polymorphism (SNP). For example, consider an SNP rs123456, with a variation on position 3, as shown below in image 1:

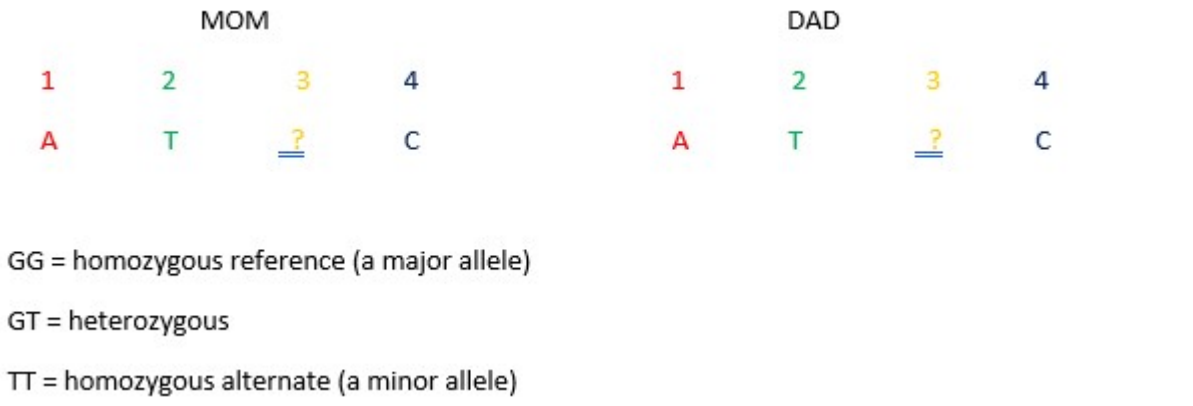


Figure 1: Image showing the three types of allele, formed during gametes

Now the first question is how do we classify an allele as a major or minor? The answer is by calculating the Minor Allele Frequency (MAF).

In epidemiology, under Genome Wide Association Studies (GWAS) and Sequencing Studies

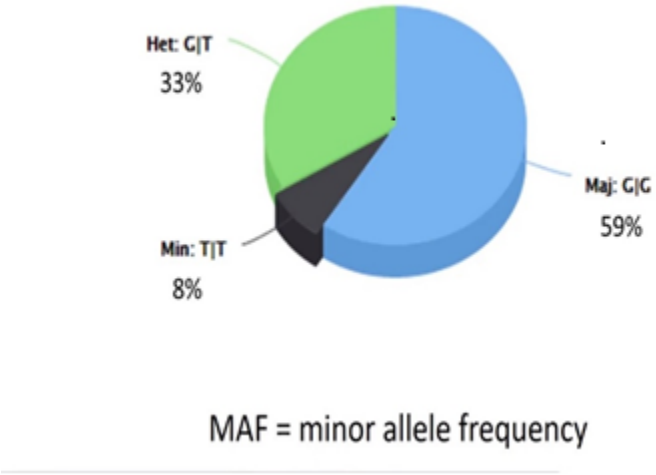


Figure 2: An exemplify showing MAF for the above example

(exome and genome), scientists are studying DNA variation to find out an association and causation(respectively) of a disease with a risk allele. GWAS study was conducted for identifying the SNP's associated with phenotypes, whereas Sequencing studies for decoding variants causing phenotypes [12]. Generally, a disease-causing/risk allele is a minor allele (as it is present in a fraction of the whole population).

In the project, we are aiming for a detailed study on the importance of mental and subjective well-being and finding a casual relationship with an individual's Body Mass Index (BMI). In the introductory section, we outline our understanding of Madelian Ranomization(MR), limitation of the observational study, and reasons why MR could produce better results in our thesis in section 2.1. In section 2.2 we discuss our understanding on Single-Nucleotide Polymorphism (SNPs) and it's association with BMI. The summary and introduction of variables of the given dataset are discussed in the final section of the introduction. In the methodology section, we outline our initial finding of preliminary data analysis and outline the assumptions for a strong IV and the mathematical implication of proposed methods to check all these assumptions. We present our findings in section 4, result section and discuss the outputs with an implication of proposed methods. In the last section of the discussion, we conclude our findings with a summary of the whole project and

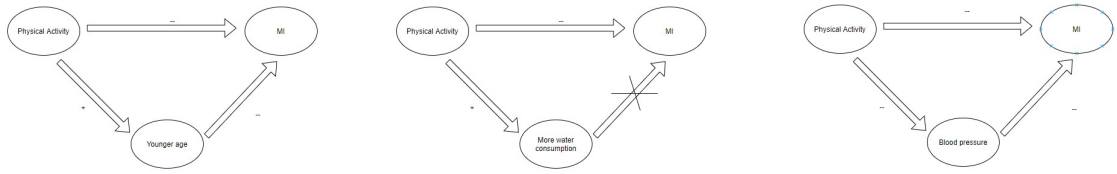
evaluate the success rate.

2.1 Mendelian Randomization

The study of the miscellaneous arrangement of alleles at the time of reproduction results in an increase of a putative risk factor causing disease is termed as Mendelian Randomization (MR) [9]. In observational epidemiology, Mendelian Randomisation (MR) is a statistical approach of using calculated variation in genes to find a causal effect of exposure on the disease [13]. For example, a decrease in the level of serum cholesterol is associated with cancer, but since the genetic variant related to cholesterol remains unaffected by any environment and behavioural change, hence further study by application of MR suggest that the allele responsible for the lower level of cholesterol is a high-risk factor for cancer[9].

Other good examples are finding a linkage between cigarette smoking with lung cancer, an association of high blood pressure with stroke, and birth weight association with the risk of type 2 diabetes [1]. The presence of confounders in the form of environmental factors and reverse causation, often results in misleading results and confusing correlations in the observational epidemiology. Mathematically, the confounder is a factor that is correlated with both outcome and exposure, causing a spurious relationship between the two, giving an alternative explanation for the causal pathway. In order to find all the risk factors associated with a disease, we need to outline and remove all confounding factors from the study. For example, studies suggested that there is a negative correlation of physical activities with Myocardial Infarction(MI), so one of the possible confounders could be age factor, as younger people are more tents to be physically active, hence less prone to develop MI. Another example could be water consumption, physically active people generally consume more water, but since there is no correlation of water consumption with MI, hence it is not considered as a confounding factor. One more example in this instinct could be blood pressure, studies have shown that blood pressure is negatively correlated with both physical

activities and MI, but it is a causal pathway or mediator, not a considerable confounder. Image 3a,3a,and 3a are showing an example of a confounder, non-confounder and mediator.



(a) An example of confounding factor. **(b)** Example showing non-confounder. **(c)** Example showing a mediator.

Figure 3: Image showing a confounder 3a, non-confounding factor3b, and a casual pathway 3c.

A phenomenon when a disease causing a risk factor instead of vice versa, is termed as a reverse causality, that is causation is having a causal effect. Mathematically, X and Y are said to be correlated if both are someway related (X to Y or Y to X), whereas X causation Y is a linkage (negative or positive) from X to Y only. For example, the statement wife generating more income than husband is a negative correlation with her husband's health, suggests the presence of some psychological stress such as ego or some other factors. Whereas the reverse causation suggests husband with a poor health condition is more likely to earn lesser than his wife due to medical reasons and more office leaves, relatively explains the case. Another explanation could be, husband's poor health leading to lesser earning and an increasing financial needs, which pushes the wife to work more efficiently to earn more. Image 4 is showing correlations diagram for the mentioned example.

Randomized Control Trials (RCTs) are an efficient method of finding an association of a risk factor with the disease, without the intervene of confounders and reverse causations. Despite giving many successive associations of exposure with increasing or decreasing a disease, RCTs have identified many failures [8]. The most suggested explanation is ignoring reverse causation and confounding factors related to lifestyle and socioeconomic factors. Two such false proven examples are the consumption of vitamin E supplements believes to prevent cancer and hormone replacements medication post menopause resulting in preventing cardiovascular diseases. In both the mentioned

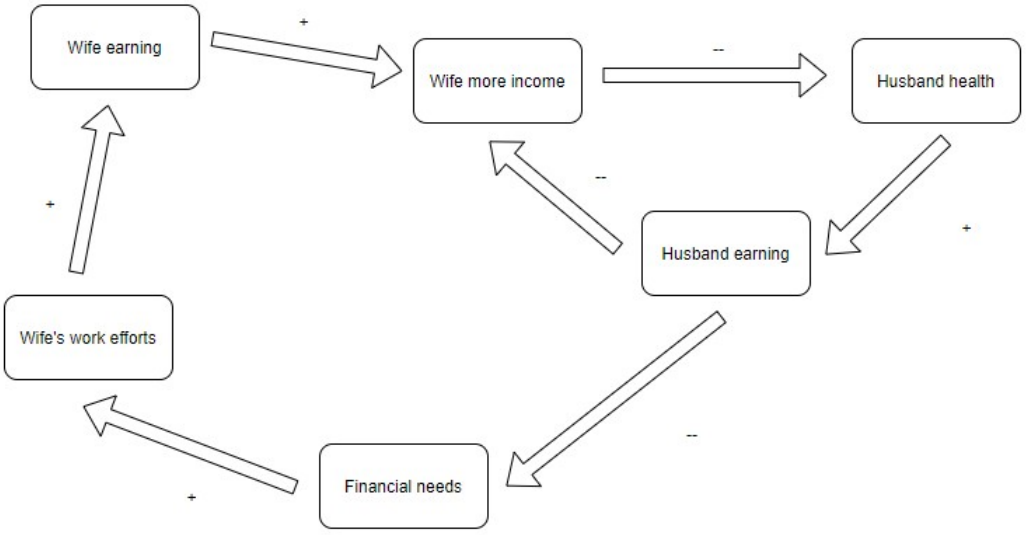


Figure 4: An exemplify showing reverse causality

examples, a strong influence of socioeconomic and lifestyle confounding factors presence was ignored.

The random distribution of alleles from parents to children, being unaffected by the presence of any confounding (socioeconomic and lifestyle factors) provides more information on genetic transmission of several diseases from parents to the offspring. The health outcomes of an offspring are randomly transmitted from parents and vary in among all biological siblings other an identical twins [8]. Also, the correlation of a risk factor with the disease is independent of reverse causality, hence RCT yields better results.

2.2 BMI associated with SNPs (IV)

Obesity can increase the risk of many diseases such as diabetes, high blood pressure, chronic heart diseases, asthma, and depression as well. Being associated with mortality and morbidity, a better understanding and study of obesity is necessary to reduce the burden on public healthcare. Studies have shown that there is approx. 40- 70% genetic variation among individuals, responsible for excess weight [12]. Studies suggested that there is a positive and significant relationship be-

tween health and subjective wellbeing[2]. BMI is the measure that uses the height and weight of an individual to define his/her healthy weight. The table 1 below is showing categories for BMI. BMI score between 18.5 to 25 is considered as a healthy weight, other categories are considered as more prone to develop a disease.

Table 1: Table showing BMI categories

BMI index(in kg/m^2)	below 18.5	18.5 - 24.99	25- 29.99	30-34.99	35 above
Category	Under weight	Normal	Over weight	Obese I	Obese II/III

A GWAS study[6] found that 941 near-independent SNPs associate with BMI, with a new threshold of $p < 1 * 10^-8$, instead of $5 * 10^-8$. Also the results of the study show the correlation of Polygenic Score on these SNPs with actual $BMI \sim 0.22$ only. For selecting the SNPs showing greater association with exposure BMI, the threshold of choosing the imputation based on the posterior probability of genotype was set 0.9 and SNPs call rate > 0.95, minor allele frequency > 0.0001, and P-value of Hardy Weinberg equilibrium threshold 10^6. So only SNPs meeting all the threshold values were selected and total 941 near-independent SNPs were identified with 480 new loci. Hence, for our project, we have considered genetic variants as a strong IV to find the causal effect of our exposure BMI on the outcome.

2.3 BMI associated with psychological distress

General Health Questionnaire(GHQ), is a scale used for screening minor psychiatric disorders. GHQ is more suitable for short-term psychological illness, and appropriate for participants of all age, above adolescent. GHQ-12, GHQ-20, GHQ-28, GHQ-30, and GHQ-60 are five available versions of this questionnaire, enabling clinicians and researchers to customise according to their needs. It is a unidimensional instrument but suggested to consider three factors: Anxiety Depression, Social Dysfunction, and Loss of Confidence [4]. For this study 12-item, GHQ has

been used as a measure for mental wellbeing, which is a unidirectional measure of psychological morbidity[5]. By calculation 12-item GHQ, scoring range between 0-3 per question, we can calculate the psychological well being of a person range between 0 to 36, where a higher score indicates unwell.

Previous studies have suggested that there is a linear relationship between BMI and subjective wellbeing, however, some studies have claimed that there is a non-linear relationship between them[14][3]. Under the FinnTwin16 study, using GHQ-20 with a few other indicators of subjective, an inverse U-shaped association was found for men participants (2151 sample size), whereas no overall relationship was found for female participants (sample size 2422). A few highlighted limitations of the study were considering self-reported BMI of the participants for research, unable to generalise the result for all participants, and is a cross-sectional study, unable to find causality.

In another study by Mark H et al. (2017) on a sample of 114218 participants were taken from 10 general population household-based surveys, using GHQ-12[3]. The results showed compared to participants with normal BMI, a higher level of distress for the population with underweight and stage II/III obese, and show a lower level of distress for overweight and stage I obese. The limitation of this study is that the data was collected only one-time point, hence this research is weak to define the direction of the study.

In another comparative study with a sample of 120 participants, to understand the usage and properties of GHQ-12, comparing the Beck Anxiety Inventory and the short form 36 [4]. It was found that though GHQ-12 is widely used as a unidimensional instrument variable, it contains three factors named as Anxiety and Depression, Social Dysfunction, and Loss of confidence. The limitations of the study were working on a small sample and all the participants were outpatients of psychological and mental illness.

There are other measures of mental and social wellbeing, widely used. For this project, we are aiming to study three different aspects or parameters of subjective wellbeing and do a combined study on casual effect. We are giving three variables GHQ, SF12, and overall life satisfaction. GHQ consider very short term behavioural changes, last up to 2-week times time prior to the participation to fill a questionnaire. SF12 is another health-related quality of life questionnaire, containing 12 questions, covering both mental and physical functioning. It also reflects changes in the short term, only within the span of 4 weeks time. Overall life satisfaction is a scale to positively evaluate long term changes and quality of life.

2.4 Dataset summary

For this project, we are working with three datasets Understanding Society Genotype and Survey Data, provided by The UK Household Longitudinal Study (UKHLS). The Understanding Society, UKHLS conducted a longitudinal survey, where data has been collected in several waves. In the first wave, approximately 40000 participants have participated in the study via an online survey, age 10-15 for youth questionnaire, and aged over 15 filled adult surveys [15]. First wave onwards, a team of trained interviewers or a medical nurse conducts a house visit to collect updated data such as weight, height, BMI, blood pressure, and more. The main aim of the study was to develop a better understanding of how life long changes in social, economical, and environmental factor affect our health by altering the genes.

The first dataset is containing the UKHLS Survey data of 9921 participants, collected in 9 waves. In wave 1, participants were asked some of the basic health-related questions about their age, weight, long-standing medical condition, physical disability, ethnicity, gender, GHQ score, life satisfaction, higher education, income, and more. During each follow-up session with a registered nurse visit, blood sample collection and other tests were done to monitor the complete health status of the participant. Also, the BMI is collected in waves 2 and 3 during nurse visits.

We are labeling some of the important variables form the first dataset:

Short Warwick-Edinburgh Mental Well Being Scale (SWEMWBS): WEMWBS is an instrument to measure mental well being, comprises of 14 items[16]. To remove bias for gender and bias for age, a few items were deleted and now 'SWEMWBS', a shorter version with 7-items is used to measure psychological well being. The correlation between the full 14-items and short 7-items version was 0.954[16]. Scored on a 5-point positively phrased Likert-style scale, score range between 7 to 35[17]. A higher score indicates better well being.

General Healthcare Questionnaire (GHQ): There are two variables showing different aspects of GHQ, 'scghq1_dv' and 'scghq2_dv'. The variable 'scghq1_dv' is likert subjective well being (GHQ), 12-items questionnaire scales between 0 to 3 (instead of 1 to 4) and summing them up on the scale 0 to 36[18]. The variable 'scghq2_dv' is caseness subjective well being (GHQ), 12-items questionnaire scaling 1 and 2 values to 0, and scaling 3 and 4 values to 1[18]. The sum of caseness GHQ varies from 0 to 12.

Satisfaction with life overall: We have variable 'sclfsato', which is the satisfaction with life overall. In each wave, a different set of questions were presented to each household and information was collected on a volunteering basis. Verbal consent was taken from the parents or guardian of the participant aged 10-15 years. Participants were asked to answer on a scale of 1 to 7, where 1 represents completely dissatisfied and 7 represents completely satisfied.

Health-Related Quality of Life (HRQOL): HRQOL refers to functioning and well-being in the physical, social, and mental dimensions of life. SF-36 is a multi-item HRQOL instrument, using an 8 multi-item scales to measure quality of life[19]. SF-12 is a widely used subset of SF-36, composed of 2 multi-item scales, measuring mental and physical functioning scores. The variable 'sf12mcs' is SF-12 mental component summary (MCS) and variable 'sf12pcs' is SF-12 physical component summary (PCS). The value range from 0 to 100, where higher score shows better func-

tioning.

'fimngrs_dv': Shows the figure of the total personal monthly income in pounds.

'age_dv': Shows the age of the participants at the time of the interview. The age has been derived from the provided date of birth and derived from the date of the interview.

'bmival': The variable shows the value of BMI, which is measured and collected during nurse visits in waves 2 and 3 only.

Principal Component Analysis (PCA): Principal Component Analysis (PCA) is a statistical method to find correlation between explanatory variables and convert them into linearly independent orthogonal principal components to deal with complexity of dealing with too many variables and to extract maximum information. Since we have about 10,000 people each with millions of SNPs, a principal component analysis had been conducted and the first 10 PCA has been provided by UKHLS study to be used in model to account for the possible genetic re-lateness among the people. For our study, we are only going to use first 5 PCA variables, represented as PCA1, PCA2, PCA3, PCA4 and PCA5.

So the given dataset is a combination of two different groups of participants. The first group of participants were people who joined the British Household Panel Study(BHPS) among 1990-2010 and then continue to participant the new survey study UK Household Longitudinal Study (UKHLS)in 2011. This group of participants were called BHPL + UKHLS participants. The second group of participants were people who joined the UKHLS study in 2010. This group of participants were called UKHLS participants. Apart from regular yearly survey questionnaires, some participants have additional nurse visits survey in wave 2 (for UKHLS participants) and wave 3 (for BHPS + UKHLS participants). And BMI was collected in nurse visit survey.

The second dataset we are using is containing the genetic information of 9920 participants,

with 480 SNPs information. Blood samples were collected during nurse visits, to run several tests including extraction of participant's genetic data. Several studies have suggested the impact of environmental and socioeconomic factors on the physical as well as the mental wellbeing of an individual, genetic data was collected for a further study in this field only. The dataset containing various information of 480 SNPs related to our study. We are given with entries like Effect Allele Frequency (EAF), Minor Allele Frequency (MAF), Risk Allele Frequency (RAF), beta, and more. MAF is the measure to calculate the variation in the given size of population for an SNP, to check how common or rare an SNP is. In the context of a disease, the risk allele is the one with a significant association with the risk of developing a disease in the cohort. For SNPs correctly oriented, all BMI increasing we assigned EAF = MAF, else EAF = 1- MAF.

The PGS is the estimated effect of all genetic variants that affect a phenotype. According to the GWAS study[6] only SNPs with significant marginal genetic effect with a threshold p-value $<$ $5 * 10^{-8}$ are considered. Then weighted and unweighted PGS was calculated as:

$PGS\_unwt = \sum_{j=1}^{J} G_j$

where PGS_unwt is the unweighted PGS, which is the sum of all risk alleles and PGS_exwt is the weighted PGS, which is the sum of numbers of the risk allele. For N participants and of J genetic variants, $G_{i1}, G_{i2}, ..., G_{iJ}$ for any $i \in N$ such that:

$PGS\_exwt_i = \sum_{j=1}^{J} G_{ij} * \gamma_j$, where γ_j is the corresponding weight and we use the effect size of BMI with SNPs from GWAS study (cite Yengo paper here) as our weight.

3 Methodology

In this project, we are trying to find a casual association of BMI with the mental wellbeing of an individual, using genetic variants Single Nucleotide Polymorphism (SNPs) as an Instrument Variable (IV). We are working with MobaXterm to get access to encrypted datasets from CERES with maintaining it's confidentiality and to use R interactively. MendelianRandomization is an R

package, that perform Mendelian Randomization analysis [11]. We first use two stage least square (TSLS) method where the generated PGS score is used as instrument. Then to find a consistent estimate of the causal effect of the exposure on the outcome using multiple genetic variants, we are aiming to apply methods like inverse-variance weighted (IVW), MREgger, and weighted median methods. Within this package, mr_allmethods() is used to apply several methods at once. mr_plot() function using mr_allmethods() function to plot static graphs depicting the various casual estimates.

3.1 Assumptions for a valid IV

In MR, genetic variants are consider as a good IV candidate [8], as genes are deeply studied and understood, genetic variants are inherited independently which results in a specific association, and it's nature of remained unchanged under the influence of environmental factors thus avoid reverse causation [10]. For each individual, SNPs, represented as a random variable values 0, 1, or 2, by the number of variant alleles. There are several methods available for an IV estimation, such as ratio estimate methods, two-stage least square (TSLS) methods, likelihood-based methods, and semi-parametric methods. For this study, we are going to use ration estimate and TSLS methods.

Conditions of a strong IV are [13][20]:

- IV1: IV G is a factor correlated with the exposure X, mathematically

$$cov(G, X) \neq 0$$

- IV2: G is independent of confounder U of the exposure-outcome association i.e.

$$cov(G, U) = 0$$

- IV3: G is independent of outcome Y, conditionally on exposure X and confounder U i.e.

$$cov[Y, G|X] = 0$$

3.2 Mathematical understanding of MR methods

Two stage least square(TSLS)

Consider a linear regression model, outcome Y is a linear combination of exposure X and confounder U

$$Y = \beta_0 + X * \beta + \epsilon^* \tag{1}$$

where β is the casual effect of X on Y and ϵ^* is the error term, such that

$$\epsilon^* = U + \epsilon^Y$$

where U represents unknown confounding factors and ϵ^Y is associated error terms. When there exists unobserved confounder, the exposure X is correlated to ϵ^*, then

$$cov(X|\epsilon^*) \neq 0$$

Two stage least square (TSLS) is the a linear regression model to estimate the cause effect if instrument variable(s) is valid instrument, i.e. satisfy the above mentioned three condition. TSLS encapsulate two stages, for a continuous outcome. In the first-stage of regression the exposure X is regressed on the IV G to get fitted values and the second stage regression is applied on outcome Y, regressing to fitting in the values of exposure from the first-stage model[13]. The result of second stage regression is the casual estimate of change in Y, due to a unit change in X. TSLS with a single IV is same as ratio estimate and with multiple IVs it is a weighted average of ratio estimate of all IV. For N participants and J number of genetic variants, let us suppose for any $i \in N$, genetic variants $G_{i1}, G_{i2}, ..., G_{ij}$, and exposure X_i:

$$X = \gamma_0 + \sum_{j=1}^{J} G_{ij} * \gamma_j + \epsilon^X \tag{2}$$

where ϵ_i^X is the error term and γ_j is coefficient for genetic variants effect on the exposure. Equation

(2) is regression model for stage-one. The fitted value of (2) is

$$\hat{X} = \sum_{j=1}^{J} G_{ij} * \hat{\gamma}_j \tag{3}$$

fitting the value of exposure X in equation (3) into equation (1), we get:

$$Y = \beta_0 + \beta_1 * \hat{X} + \epsilon^* = \beta_0 + \beta_1 * \sum_{j=1}^{J} G_{ij} * \hat{\gamma}_j + \epsilon^* \tag{4}$$

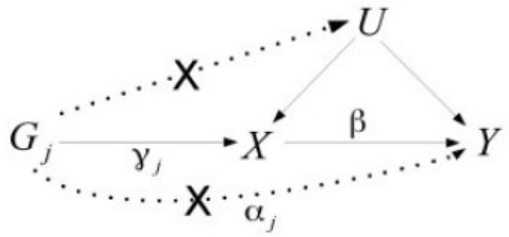


Figure 5: Showing a strong valid instrument variable

Hence the revised assumptions for TSLS, referencing to the figure 5 can be given as:

1.$\gamma_j \neq 0$
2.$\alpha_j = 0 \iff cov(G_{ij}, \epsilon^Y) = 0$, i.e. G_{ij} is not directly associated with ϵ^Y.
3.$\theta_j = 0 \iff cov(G_{ij}, U) = 0$, i.e. G_{ij} is not directly associated with U.

For $\epsilon^* = \epsilon^Y + U$, therefore by assumption 2 and 3, $cov(G_{ij}, \epsilon^*) = 0$.

MR Egger

Pleiotropic variants could bias the result of casual effect and increase type I error rate for testing casual null hypothesis. We use Egger Regression, to detect the study bias due to presence of pleiotropy[20]. The coefficient of the slope of Egger regression provides an estimate of the casual effect. This is considered as a robust method to find casuality in case of exist invalid instruemt variables. Hence the loose assumptions for an invalid IV referencing to the figure 5 can be given as:

1.$\gamma_j \neq 0$
2.$\alpha_j \neq 0$
3.$\theta_j = 0$

In the presence of pleitropy, to get an unbias estimate of the casual effect, we apply an Instrument Strength Independent of Direct Effect (InSIDE) assumption. An InSIDE assumption is apply to balance the pleitropic effects of the variants, assuming that the magnitude of the SNP-exposure associations across all variants is independent. For MR-Egger method, we assume that the association of each genetic variant with exposure is independent of pleiotropic effect of the variant not via the exposure. So the revised IV assumptions are applied, allowing $\alpha_j \neq 0$ with an InSIDE condition of $\alpha_j \perp \gamma_j$. For any $i \in N$, the regression model is:

$$
\begin{aligned}
Y_i &= \sum_{j=1}^{J} \alpha_j * G_{ij} + \beta * X_i + U_i + \epsilon_i^Y \\
&= \sum_{j=1}^{J} \alpha_j * G_{ij} + (\sum_{j=1}^{J} G_{ij} * \gamma_j) * \beta + U_i + \epsilon_i^Y \\
&= \sum_{j=1}^{J} (\alpha_j + \gamma_j * \beta) * G_{ij} + U_i + \epsilon_i^Y
\end{aligned} \tag{5}
$$

where β is the casual effect of X on Y, α_j is coefficient for genetic variants effect on the outcome directly, ϵ_i^Y is the error term. The reduced for of equation (5) is:

$$
Y_i = \Gamma_j * G_{ij} + \epsilon_{ij}^{'Y}
$$

where $\Gamma_j = (\alpha_j + \gamma_j * \beta)$ and $\epsilon_{ij}^{'Y}$ = $U_i + \epsilon_i^Y$. $\hat{\Gamma_j}$ is the coefficient of regression of outcome Y_j on genetic variant G_{ij} and $\hat{\gamma_j}$ is the coefficient of regression of exposure X_j on genetic variant G_{ij}. and

$$
\hat{\Gamma_j} = \beta_{OE} + \beta * \hat{\gamma_j}
$$

where β_{OE} is the estimate of average of α_j, which is the average of direct plieotropy effect.

Inverse-variance weighted methods (IVW)

IVW is one of the other proposed methods to calcuate an estimate of multiple genetic variants

as a single casual estimate and it gives similar result as TSLS[21]. Hence we applied both TSLS and IVW for comparison to check if we are a getting significant casual estimate, assuming all IVs are valid. Now for a valid instrument variable, $\alpha_j = 0$such that $\sum \alpha_j = 0$, implies that $\beta_{OE} = 0$, the the IVW regression model is :

$$\hat{\Gamma}_j = \beta * \hat{\gamma}_j \tag{6}$$

Where $\hat{\Gamma}_j$ is the coefficient of regression of outcome Y_j on genetic variant G_{ij} and $\hat{\gamma}_j$ is the coefficient of regression of exposure X_j on genetic variant G_{ij}. $\alpha_j = 0$ for a valid IV, then by ratio method, the casual effect of exposure on the outcome for equation (6) is:

$$\frac{\Gamma_j}{\gamma_j} = \frac{\beta * \gamma_j}{\gamma_j} = \beta$$

Median method

In the case of an invalid IV, there is a high probability of type I error and bias in the outcome, hence we use Median method, which is a robust methods to deal with these situation in the case of invalid IVs[22]. In the case of a study with all valid genetic variants, IVW is an efficient method to calculate the estimate. But unfortunately this have a 0% of break-down as at fails even if a single genetic variant is invalid. However median method is considered as a robust method to deal with the stituation with 50% break-down level. So the loose IV assumptions for median are:

1.$\gamma_j \neq 0$
2.$\alpha_j \neq 0, j < \frac{J}{2}$
3.$\theta_j \neq 0, j < \frac{J}{2}$

So we are allowing 2nd and 3rd assumptions to be true for at-least half of the IVs satisfy the condition, i.e. for $j < \frac{J}{2}$. Now the simple median is the median of the ratio estimate given by:

$$\delta_j = \frac{\Gamma_j}{\gamma_j}$$

δ_j is the ordered ratio estimate, arranges in ascending order, for J is odd number (J=2j+1), the simple median estimator would the δ_{j+1} and for even J (=2j), median estimate would be $\frac{\delta_j + \delta_{j+1}}{2}$. Since the simple median estimator is insufficient in some cases, hence a weighted median is intro-

duced. Let us suppose w_j is the weight of jth ordered ratio estimate, such that $\sum_{j=1}^{J} w_j = 1$ and :

$$\delta_j = \frac{\Gamma_j}{\gamma_j} * w_j$$

Although the simple median provides a consistent estimate of causal effect if at least 50% of IVs are valid, the weighted median will provide a consistent estimate if at least 50% of the weight comes from valid IVs.

To check if the SNPs are a strong IV and meeting all above mentioned conditions, we will follow the below mentioned steps:

- Check for association of BMI with weighted and unweighted Polygenic Score (PGS), using ordinary least square regression.
- Check the aboved associations but also control age,sex and PCA1-PCA5.
- Check for an association of outcome Y with exposure X. For our study we are consider GHQ, life satisfaction and sf12mcs as outcome and BMI as exposure.
- Check for an association of outcome Y and exposure X with other important variables such as PCA, age and sex.
- Applying regression with TSLS and other MR methods such as MR Egger, IVW, MR simple median and weighted median.
- We are aiming for a comparitive study with TSLS with IVW methods for validation of our results and checking results of MR-Egger regression and weighted medium to provide more robust estimations.

4 Results

In this section, we are going to summarise our finding upon applying various methods used above. To establish a better understand with our dataset we do some preliminary data analysis and

discuss our finding here. We are also visualising our dataset here to see correlation between different variables. In the next subsection, we are going to outline our finding after applying different MR methods to find the causality between the exposure and outcome.

4.1 Statistical data analysis

We are given with three datasets, first is Survey data with dimension 9920 * 107, SNP dataset with dimension 9921 * 481 and PGS with dimension 9921 * 4, containing weighted and unweighted polygeneic score. We merge Survey and PGS dataset as Survey_n with dimension 9920*110 for further analysis. There are 43.94% of men and 56.05% of women participants in the Survey. Since the data has been collected in nine different waves, we used waves two and three for initial analysis. We redefine new variable bmi, age, ghq, income, lifesat, sf12mcs and sf12pcs from wave b and c.

Table 2: Table showing output of preliminary analysis

Variable name	Overall mean	Overall SD	Men mean	Men SD	Women mean	Women SD	NA values Total
Age	51.77	16.93	52.2	17.2	51.4	16.7	0
GHQ	11.11	5.44	10.3	4.99	11.7	5.7	170
bmi	28.03	5.33	28.1	4.69	28.0	5.78	291
lifesat	4.92	1.58	4.94	1.56	4.90	1.59	168
sf12mcs	50.39	9.37	51.7	8.56	49.4	9.82	276
income	1623.17	1430.16	2042	1672.13	1297	1980.8	0

We have total 9920 participants, with 5561 females and 4359 males. In the exploratory analysis we found that the mean age for female and male participants are nearly the same 51.4 and 52.2 respectively, with overall average of 51.77. Similary the standard deviation for male, female and overall data are 17.2, 16.7, and 16.93 respectively. Mean GHQ score for female participants are comparatively slightly higher than the male candidates, 11.7 and 10.3 respectively. Table 2 showing that a slightly bigger percentage of female candidates are having higher GHQ score comparing to male, which indicates that the former is somehow feeling more stressed and suffering from a poor health condition. Similary the standard deviation for females and males are 5.70 and 4.99

respectively. Hence we are considering Sex as an important variable in our study, assuming it could play an important role in find causality.

Table 3: Table showing GHQ score percentage distribution among male and female participants

Gender	$score : 0 - 10$	$score : 11 - 20$	$score : 21 - 30$	$score : 31 - 36$
Female	49.79%	41.57%	7.30%	1.33%
Male	62.04%	32.63%	4.77%	0.54%

Similarly, uneven distribution of BMI is present in the data, as shown in table 4. We can see that the a greater number of men participants were categories under over-weight and obese type I with 45.74% and 21.9% respectively, comparing to 34.56% and 19.15% respectively in women candidates. However, a slightly dramatic percentage increase in female fraction as under-weight and obese type II/III categories with 1.20% and 12.34% respectively, comparing with 0.52% and 7.50% of men in the respective categories. However, the mean BMI for the both gender is nearly equal as shown in table2.

Table 4: Table showing percentage distribution of BMI among male and female participants

	under weight	normal	over weight	Obese I	Obese II/III
Gender	$BMI < 18.5$	$18.5 - 24.99$	$25 - 29.99$	$30 - 34.99$	> 35
Female	1.2%	32.65%	34.56%	19.15%	12.43%
Male	0.52%	24.20%	45.74%	21.90%	7.5%

Table 5: Table showing percentage distribution of sf12mcs on gender basis for Survey data

	score	score	score	score	score
Gender	< 20	$21 - 40.99$	$41 - 60.99$	$61 - 80.99$	> 81
Female	1.46%	16.39%	77.10%	5.05%	0%
Male	0.43%	11.15%	80.19%	8.23%	0%

Similarly, the mean life-satisfaction for men and women is slightly equal, with 4.94 and 4.90 respectively, while the overall mean for the entire dataset is 4.92. As we know the higher score means more overall satisfaction with life. In the case of sf12mcs, the average scores for men is slightly higher than women with 51.7 49.4 respectively. The same could be observed in the table 5 that the percentage of men in 61-81 score category is 8.23%, comparing to women with 5.05%, similarly in another category of score 41-61 with 80.19% and 77.10% respectively. This points out that female participants have scored lesser in Mental Health Summary than male participants. Hence with considering 'sf12mcs' as an important variable, we are using it for further analysis and to check correlation with gender and BMI.

On the basis of gender there is a dramatic difference in average income. The overall average income is 1623, with 2042 for men and 1287 average for women participants, as shown in table 2. Hence we are considering Income as another important variable in our study. Other than BMI, Sex, Income, GHQ, sf12mcs-score and overall life-satisfaction, we are considering PCA1 to PCA5, in our predicting model. For prediction we considering BMI as our risk factor or exposure aiming to find a casual effect on mental wellbeing of an individual, considering GHQ, sf12mcs-score and overall life-satisfaction as outcomes to check with various different angles.

4.2 Output of MR methods on overall data

```
Call:
lm(formula = bmi ~ PGS_unwt, data = Survey_n)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8513  -3.6829  -0.6533   2.8332  28.6469

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.812893   1.886749  -1.491    0.136
PGS_unwt     0.064457   0.003942  16.350   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.258 on 9325 degrees of freedom
  (593 observations deleted due to missingness)
Multiple R-squared:  0.02787,   Adjusted R-squared:  0.02776
F-statistic: 267.3 on 1 and 9325 DF,  p-value: < 2.2e-16
```

Figure 6: Linear regression model to check significant association of BMI with PGS_unwt

Moving forward to further analysis aiming to check whether IV assumptions(mentioned in section 3.1) are valid. First we apply linear regression model for BMI with PGS weighted and unweighted, to check for IV:1 assumption, to check for a correlation of exposure X (BMI) with Instrument (PGS). Figure 6 showing BMI is having a negative intercept of -2.81 with the unweighted polygenic score. As we know that a negative intercept shown a inverse relation between the explanatory variable X and predicted variable Y. So the model predicts that for 0 units of PGS_unwt, there will be -2.81 units of BMI. The slope of the model is 0.064, which represent a unit increase in PGS_unwt, will result in 0.064 unit increment in BMI.

```
> summary(lm(bmi~PGS_exwt, data=Survey_n))

Call:
lm(formula = bmi ~ PGS_exwt, data = Survey_n)

Residuals:
     Min       1Q   Median       3Q      Max
-14.2745  -3.6407  -0.6184   2.8268  28.4139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0910     1.8215  -3.344 0.000829 ***
PGS_exwt      4.6194     0.2465  18.737  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.235 on 9325 degrees of freedom
  (593 observations deleted due to missingness)
Multiple R-squared:  0.03628,   Adjusted R-squared:  0.03618
F-statistic: 351.1 on 1 and 9325 DF,  p-value: < 2.2e-16
```

Figure 7: Linear regression model to check significant association of BMI with PGS_exwt

Figure 7 showing BMI is having a negative intercept of -6.09 with the weighted polygenic score. So the model predicts that for 0 units of PGS_exwt, there will be -6.09 units of BMI. The slope of the model is 4.62, which represent a unit increase in PGS_exwt, will result in 4.62 unit increment in BMI.

Moving forward, we apply a linear regression model to check for a significant association of BMI with both PGS_exwt and PGS_unwt along with other crucial variables such as sex, age and PCA. Figure 8 showing BMI is having a negative intercept of -4.95 with the unweighted polygenic score. And the slope is 0.064, 0.038, 0.039 with PGS_unwt, age and sex-male respectively.

```
Call:
lm(formula = bmi ~ PGS_unwt + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5, data = Survey_n)

Residuals:
    Min      1Q  Median      3Q     Max
-14.961  -3.660  -0.698   2.782  28.169

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.954367   1.883682  -2.630  0.00855 **
PGS_unwt     0.064838   0.003916  16.555  < 2e-16 ***
age          0.037712   0.003218  11.720  < 2e-16 ***
sexmale      0.039172   0.108977   0.359  0.71926
PCA1        -3.785843   5.414691  -0.699  0.48446
PCA2        -0.781558   5.388087  -0.145  0.88467
PCA3        -4.036512   5.405177  -0.747  0.45521
PCA4         3.892909   5.407005   0.720  0.47156
PCA5        -2.713797   5.381157  -0.504  0.61405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.221 on 9318 degrees of freedom
  (593 observations deleted due to missingness)
Multiple R-squared:  0.0422,    Adjusted R-squared:  0.04138
F-statistic: 51.32 on 8 and 9318 DF,  p-value: < 2.2e-16
```

Figure 8: Linear regression model to check significant association of BMI with PGS_unwt, sex, age and PCA

```
Call:
lm(formula = bmi ~ PGS_exwt + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5, data = Survey_n)

Residuals:
     Min       1Q   Median       3Q      Max
-15.0097  -3.6267  -0.6596   2.7792  28.2491

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.255226   1.819458  -4.537 5.77e-06 ***
PGS_exwt     4.646840   0.244988  18.968  < 2e-16 ***
age          0.037717   0.003203  11.774  < 2e-16 ***
sexmale      0.043886   0.108494   0.404    0.686
PCA1        -4.232455   5.390725  -0.785    0.432
PCA2        -0.636318   5.364202  -0.119    0.906
PCA3        -5.670280   5.383138  -1.053    0.292
PCA4         3.733735   5.382915   0.694    0.488
PCA5        -2.329869   5.357350  -0.435    0.664
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.198 on 9318 degrees of freedom
  (593 observations deleted due to missingness)
Multiple R-squared:  0.05068,   Adjusted R-squared:  0.04987
F-statistic: 62.18 on 8 and 9318 DF,  p-value: < 2.2e-16
```

Figure 9: Linear regression model to check significant association of BMI with PGS_exwt, sex, age and PCA

A similar result can been observed for weighted polygenic score. Figure 9 showing BMI is having a slightly greater negative intercept of -8.26 with the weighted polygenic score, comparing to unweighted one. Also, the slope for PGS_exwt is 4.65, showing higher increase in BMI with a unit increase in PGS_exwt, than PGS_unwt. However the slope for age and sex-male are nearly the

same with 0.038, 0.044 respectively.

```
Call:
ivreg(formula = ghq ~ bmi + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5 | age + sex + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 +
    PGS_unwt, data = Survey_n)

Residuals:
    Min      1Q  Median      3Q     Max
-12.914  -3.570  -1.191   1.811  25.947

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.435414   1.623512   5.812 6.39e-09 ***
bmi           0.115532   0.061730   1.872   0.0613 .
age          -0.019484   0.004036  -4.828 1.40e-06 ***
sexmale      -1.377004   0.112406 -12.250  < 2e-16 ***
PCA1         -0.565869   5.576275  -0.101   0.9192
PCA2         -2.654876   5.556903  -0.478   0.6328
PCA3        -13.196394   5.569703  -2.369   0.0178 *
PCA4          6.986000   5.577943   1.252   0.2104
PCA5          0.948175   5.544274   0.171   0.8642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.338 on 9161 degrees of freedom
Multiple R-Squared: 0.02512,    Adjusted R-squared: 0.02427
Wald test: 22.89 on 8 and 9161 DF,  p-value: < 2.2e-16
```

Figure 10: Result TSLS to estimate the casual effect of PGS_unwt with other variable such as bmi, sex, age and PCA on explanatory variable GHQ

```
Call:
ivreg(formula = ghq ~ bmi + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5 | age + sex + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 +
    PGS_unwt, data = Survey_n)

Residuals:
    Min      1Q  Median      3Q     Max
-12.914  -3.570  -1.191   1.811  25.947

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.435414   1.623512   5.812 6.39e-09 ***
bmi           0.115532   0.061730   1.872   0.0613 .
age          -0.019484   0.004036  -4.828 1.40e-06 ***
sexmale      -1.377004   0.112406 -12.250  < 2e-16 ***
PCA1         -0.565869   5.576275  -0.101   0.9192
PCA2         -2.654876   5.556903  -0.478   0.6328
PCA3        -13.196394   5.569703  -2.369   0.0178 *
PCA4          6.986000   5.577943   1.252   0.2104
PCA5          0.948175   5.544274   0.171   0.8642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.338 on 9161 degrees of freedom
Multiple R-Squared: 0.02512,    Adjusted R-squared: 0.02427
Wald test: 22.89 on 8 and 9161 DF,  p-value: < 2.2e-16
```

Figure 11: Result TSLS to estimate the casual effect of PGS_exwt with other variable such as bmi, sex, age and PCA on explanatory variable GHQ

For further analysis we are going to apply MR methods, such as TSLS, MR Egger, IVW estimate, simple median and weighted median using genetic data. Applying TSLS to the estimate casual effect of BMI on outcome GHQ using genetic variants. The coefficient estimate for both unweighted and weighted are nearly equal, 9.4354 and 9.4757 respectively. Figures 10 and 11 shows the result with PGS_unwt and PGS_exwt respectively. The respective slope for BMI is 0.1155 and 0.1139 both with p-value¿0.05; for age the slope is equal to -0.0194 and for gender-male is equal to -1.3770. Also, the p-value for both the cases is $2.2\epsilon^{-16}$

We used 'MendelianRandomization' R package to apply mr_input(), mr_allmethods() and mr_plot() for the purpose[11]. mr_allmethods() is the one single function, which gives a table of output for MR-Egger, IVW, weighted and simple median methods. mr_plot() is the an useful function to produce an interactive plot to visualise results given by mr_allmethods(). For MR analysis using mentioned function, we use variables such as: β_E: coefficient for association with the exposure, ϵ_E: standard error term for exposure, β_O: coefficient for association with the outcome, ϵ_0: standard error term for outcome. In order to calculate β_O and ϵ_0, we apply linear regression model for outcome (GHQ in first case, and then sf12mcs and overall life-satisfaction) considering SNPs info, BMI, age, sex and PCA, as explanatory variables.

Table 6: MR result for GHQ as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	0.366	0.384	-0.387 1.119	0.341
Weighted median	0.264	0.425	-0.569 1.098	0.534
IVW	0.210	0.251	-0.283 0.703	0.404
MR-Egger	0.274	0.660	-1.020 1.568	0.678
(intercept)	-0.001	0.011	-0.022 0.020	0.916

Mr-Egger is used for check for bias in conventional IV analysis, that violating IV assumptions and also to find an estimate of casual effect[23]. The table 6 shows the estimate for Simple median is 0.366, which is slightly higher than estimate for weighted median and MR-Egger which is 0.264

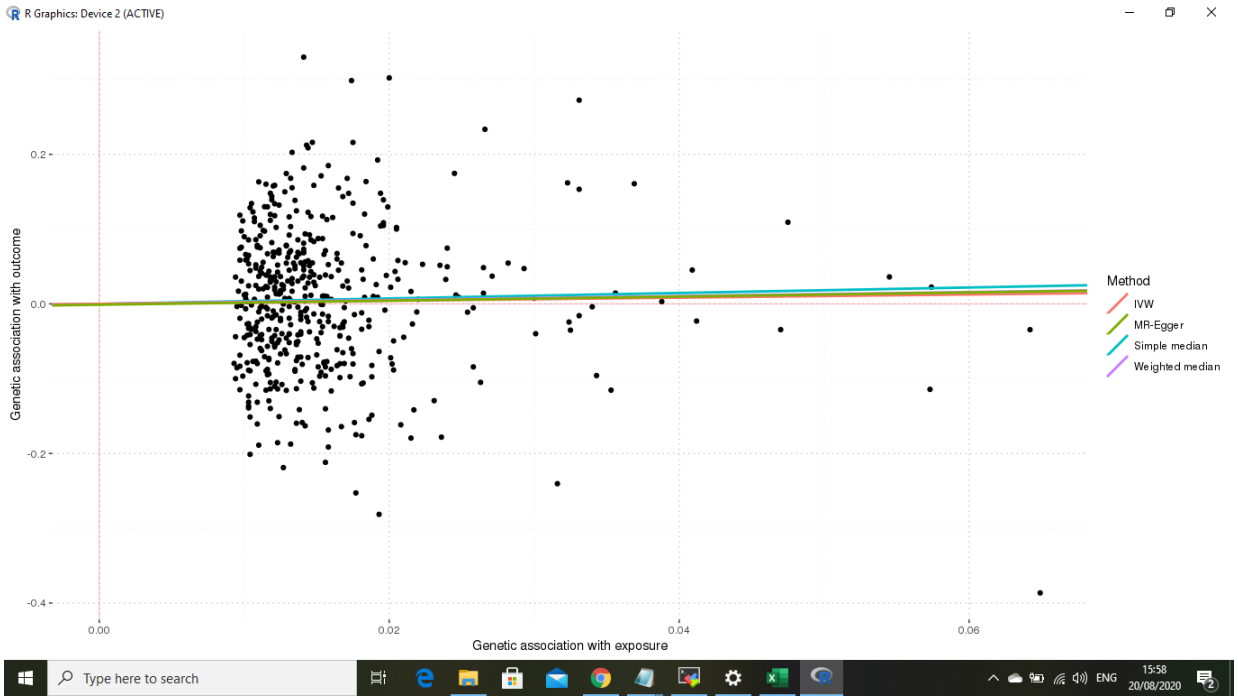


Figure 12: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on GHQ

and 0.274 respectively. The estimate for IVW is 0.210. The intercept is -0.001 and the p-value for this estimation is 0.916 which indicates that there is no evidence that direct pleotropic effect exist and therefore both TSLS and IVW can provide consistent estimates. Although the IVW give a slightly higher estimate comparing to TSLS estimates, the p-values for these estimates are larger than 0.05 and show no evidence of a causal relationship from BMI to GHQ. Figure 12 is an interactive plot showing best fitting line for the outcome of MR-Egger, weighted median, IVW and simple median methods.

Following the same procedure of applying all MR methods, considering life-satisfaction as an output. The coefficient of intercept for TSLS with unweighted and weighted PGS are 5.4426 and 5.1948 respectively. The respective slopes for BMI are 0.0152 and 0.0246, for age are -0.0112 and -0.0116, and for sex-male are 0.0782 and 0.0779. Figures 13 and 14 shows the estimate of casual estimate of BMI on overall life-satisfaction using weighted and unweighted genetic variants, with the help of TSLS method.

```
Call:
ivreg(formula = lifesat ~ bmi + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5 | age + sex + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 +
    PGS_unwt, data = Survey_n)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8180 -0.5788 -0.2476  0.8031  3.3224

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.442661   0.476964  11.411   <2e-16 ***
bmi          0.015158   0.018135   0.836   0.4032
age         -0.011240   0.001183  -9.500   <2e-16 ***
sexmale      0.078230   0.032968   2.373   0.0177 *
PCA1         0.375960   1.638126   0.230   0.8185
PCA2        -1.275081   1.629233  -0.783   0.4339
PCA3         0.227565   1.633560   0.139   0.8892
PCA4         0.428114   1.636693   0.262   0.7937
PCA5         2.626913   1.627616   1.614   0.1066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.566 on 9165 degrees of freedom
Multiple R-Squared: 0.01513,    Adjusted R-squared: 0.01427
Wald test: 15.97 on 8 and 9165 DF,  p-value: < 2.2e-16
```

Figure 13: Figure showing the outcome for TSLS considering life-satisfaction as outcome with unweighted PGS

```
Call:
ivreg(formula = lifesat ~ bmi + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5 | age + sex + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 +
    PGS_exwt, data = Survey_n)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9073 -0.5969 -0.2395  0.8210  3.4403

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.194844   0.420929  12.341   <2e-16 ***
bmi          0.024643   0.015974   1.543   0.1229
age         -0.011588   0.001141 -10.156   <2e-16 ***
sexmale      0.077951   0.032996   2.362   0.0182 *
PCA1         0.399664   1.639400   0.244   0.8074
PCA2        -1.259822   1.630581  -0.773   0.4398
PCA3         0.245751   1.634887   0.150   0.8805
PCA4         0.384854   1.637635   0.235   0.8142
PCA5         2.663964   1.628675   1.636   0.1019
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.567 on 9165 degrees of freedom
Multiple R-Squared: 0.01343,    Adjusted R-squared: 0.01257
Wald test: 16.16 on 8 and 9165 DF,  p-value: < 2.2e-16
```

Figure 14: Figure showing the outcome for TSLS considering life-satisfaction as outcome with unweighted PGS

Table 7: MR result for lifesatisfaction as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	0.007	0.137	-0.261 0.275	0.961
Weighted median	0.219	0.156	-0.087 0.525	0.161
IVW	0.092	0.090	-0.085 0.269	0.308
MR-Egger	0.445	0.236	-0.017 0.908	0.059
(intercept)	-0.006	0.004	-0.014 0.001	0.105

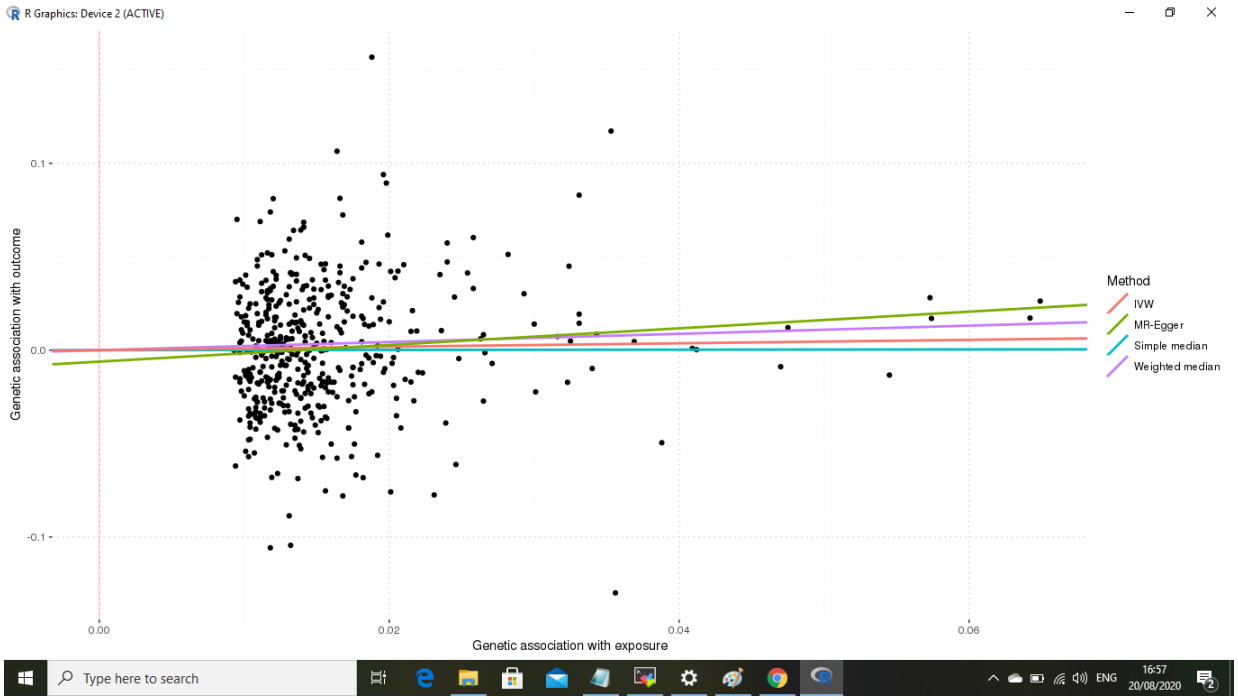


Figure 15: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on Life satisfaction

Table 7 shows the result for rest of the MR methods.The slope of simple median is 0.007 which is very close to zero and very less than weighted median, which is 0.219. MR-Egger slope is 0.445 and IVW is 0.092. The intercept is -0.006 with p-value is 0.105 and show no significant direct pleitropy effect. And all estimates from MR methods including TSLS, IVW, Egger and median show no significant causal effect from BMI to life-satisfaction. Figure 15 is an interactive plot showing best fitting line for the outcome of MR-Egger, weighted median, IVW and simple median methods.

We applied all 5 proposed MR methods to estimate casual effect of exposure BMI on sf12mcs. The coefficient of intercept for TSLS with unweighted and weighted PGS are 48.9785 and 48.5226 respectively. The respective slopes for BMI are -0.1587 and -0.1412, for age are 0.0958 and 0.0952, and for sex-male are 2.3034 and 2.3032. Figures 16 and 17 shows the estimate of casual estimate of BMI on sf12mcs using weighted and unweighted genetic variants, with the help of TSLS method.

```
Call:
ivreg(formula = sf12mcs ~ bmi + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5 | age + sex + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 +
    PGS_unwt, data = Survey_n)

Residuals:
    Min      1Q  Median      3Q     Max
-45.010  -4.573   2.172   6.351  25.097

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.978582   2.766837  17.702   <2e-16 ***
bmi          -0.158695   0.105236  -1.508   0.1316
age           0.095823   0.006941  13.805   <2e-16 ***
sexmale       2.303420   0.192552  11.963   <2e-16 ***
PCA1         -2.834957   9.556869  -0.297   0.7667
PCA2         -9.863718   9.508295  -1.037   0.2996
PCA3         21.169983   9.532933   2.221   0.0264 *
PCA4        -16.621366   9.559684  -1.739   0.0821 .
PCA5          7.451582   9.498678   0.784   0.4328
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.096 on 9067 degrees of freedom
Multiple R-Squared: 0.04557,    Adjusted R-squared: 0.04473
Wald test: 51.06 on 8 and 9067 DF,  p-value: < 2.2e-16
```

Figure 16: Figure showing the outcome for TSLS considering sf12mcs as outcome with unweighted PGS

```
Call:
ivreg(formula = sf12mcs ~ bmi + age + sex + PCA1 + PCA2 + PCA3 +
    PCA4 + PCA5 | age + sex + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 +
    PGS_exwt, data = Survey_n)

Residuals:
    Min      1Q  Median      3Q     Max
-44.949  -4.578   2.206   6.365  24.914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.522605   2.434073  19.935   <2e-16 ***
bmi          -0.141236   0.092398  -1.529   0.1264
age           0.095174   0.006683  14.242   <2e-16 ***
sexmale       2.303263   0.192503  11.965   <2e-16 ***
PCA1         -2.802646   9.553978  -0.293   0.7693
PCA2         -9.840753   9.505640  -1.035   0.3006
PCA3         21.197058   9.530182   2.224   0.0262 *
PCA4        -16.692491   9.555048  -1.747   0.0807 .
PCA5          7.503192   9.495091   0.790   0.4294
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.093 on 9067 degrees of freedom
Multiple R-Squared: 0.04606,    Adjusted R-squared: 0.04521
Wald test:  51.1 on 8 and 9067 DF,  p-value: < 2.2e-16
```

Figure 17: Figure showing the outcome for TSLS considering sf12mcs as outcome with unweighted PGS

Table 8 shows the result for rest of the MR methods.The slope of simple median is -0.346, whereas weighted median is -0.119. MR-Egger slope is 0.085 and IVW is -0.291. The intercept is -0.007. Similar finding is here that no evidence for direct pleitropic effect and no evidence for

Table 8: MR result for sf12mcs as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	-0.346	0.647	-1.613 0.921	0.593
Weighted median	-0.119	0.728	-1.546 1.307	0.870
IVW	-0.291	0.425	-1.124 0.542	0.494
MR-Egger	0.085	1.115	-2.100 2.270	0.939
(intercept)	-0.007	0.018	-0.041 0.028	0.715

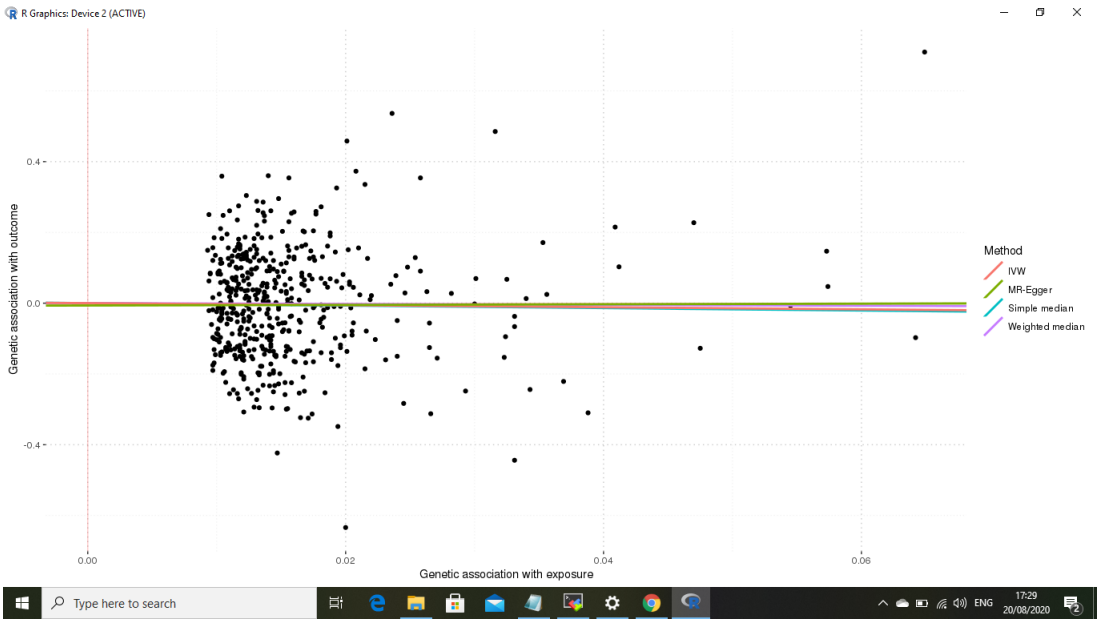


Figure 18: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on sf12mcs

causal effect from BMI to sf12mcs. Figure18 is an interactive plot showing best fitting line for the outcome of MR-Egger, weighted median, IVW and simple median methods.

4.3 Gender based stratification

As we have observed in preliminary data analysis that the overall distribution of data is not homogeneous among both the genders. Table 2 shows the non-homogeneous distribution of data among male and female participants. We have observed in the provided data the average GHQ score for women is 11.7, whereas for men is 10.3. As we know higher GHQ score refers to greater psychological stress level. Similarly there is a slight variation in score for overall life satisfaction, male to female respective average scores are 4.94 and 4.90. Where, a higher score indicates more

satisfaction, or considering a happy life. The same was noticed for sf12mcs score as well, the observed average scores for male vs females are 51.7 and 49.4 respectively, whereas higher score refers to better functioning, and lower mental stress/illness. Since, the overall sum up of the initial data analysis indicates presence of a greater level of psychological stress and higher chance of mental illness in our female participants. Hence, considering sex a crucial variable, we have decided to check for estimate of causal effect of BMI on subjective and mental wellbeing, with a gender based stratification.

```
Call:
ivreg(formula = ghq ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_unwt,
    data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-12.246  -3.872  -1.144   1.963  24.915

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.113655   2.103229   5.760 8.92e-09 ***
bmi           0.016654   0.079489   0.210  0.83405
age          -0.017694   0.005477  -3.231  0.00124 **
PCA1         -3.288265   7.921492  -0.415  0.67808
PCA2         -5.202504   7.852433  -0.663  0.50766
PCA3        -12.655096   7.870459  -1.608  0.10791
PCA4          7.846802   7.874307   0.997  0.31905
PCA5         10.703242   7.870245   1.360  0.17390
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.643 on 5151 degrees of freedom
Multiple R-Squared: 0.006493,   Adjusted R-squared: 0.005143
Wald test: 2.758 on 7 and 5151 DF,  p-value: 0.007356
```

Figure 19: Figure showing the outcome for TSLS considering GHQ as outcome with unweighted PGS

4.3.1 Stratified subset- Female

Figures 19 and 20 are showing TSLS result for considering GHQ as outcome, with unweighted and weighted polygenic scores respectively. For the PGS_unwt, the value of intercept is 12.1137, the slope for BMI is 0.0167 with p-value 0.834. Whereas in the case of PGS_exwt, the value of intercept for TSLS is 11.7112, the slope for BMI is 0.0319 with p-value 0.652. Since the $p = value > 0.05$ in both the cases, hence there is no significant association.

Figures 21 and 22 are showing TSLS result for considering overall life satisfaction as outcome, with unweighted and weighted polygenic scores respectively. For the PGS_exwt, the value of intercept is 4.9896, the slope for BMI is 0.0299 with p-value 0.134. Whereas in the case of PGS_unwt,

```
Call:
ivreg(formula = ghq ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_exwt,
    data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-12.173  -3.889  -1.158   1.944  25.042

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.711266   1.879832   6.230 5.04e-10 ***
bmi           0.031976   0.070913   0.451  0.65207
age          -0.018219   0.005333  -3.416  0.00064 ***
PCA1         -3.245395   7.912690  -0.410  0.68171
PCA2         -5.092657   7.840150  -0.650  0.51600
PCA3        -12.718092   7.860969  -1.618  0.10575
PCA4          7.829593   7.866083   0.995  0.31961
PCA5         10.751697   7.861315   1.368  0.17147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.637 on 5151 degrees of freedom
Multiple R-Squared: 0.008542,   Adjusted R-squared: 0.007194
Wald test: 2.786 on 7 and 5151 DF,  p-value: 0.006815
```

Figure 20: Figure showing the outcome for TSLS considering GHQ as outcome with weighted PGS

```
Call:
ivreg(formula = lifesat ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_unwt,
    data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8664 -0.5835 -0.2217  0.8383  3.4212

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.175945   0.592163   8.741  < 2e-16 ***
bmi          0.022893   0.022364   1.024   0.3060
age         -0.010263   0.001534  -6.689 2.49e-11 ***
PCA1        -0.399530   2.225010  -0.180   0.8575
PCA2        -1.521068   2.206166  -0.689   0.4906
PCA3         0.206143   2.209430   0.093   0.9257
PCA4         0.962195   2.210639   0.435   0.6634
PCA5         4.569728   2.208213   2.069   0.0386 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.585 on 5154 degrees of freedom
Multiple R-Squared: 0.009991,   Adjusted R-squared: 0.008646
Wald test: 8.071 on 7 and 5154 DF,  p-value: 8.566e-10
```

Figure 21: Figure showing the outcome for TSLS considering life satisfaction as outcome with unweighted PGS

the value of intercept for TSLS is 5.1759, the slope for BMI is 0.0229 with p-value 0.306. In both the cases the $p - value > 0.05$, not meeting the threshold, hence there is no significant association found.

Figures 23 and 24 are showing TSLS result for considering score of sf12mcs as outcome, with unweighted and weighted polygenic scores respectively. For the PGS_exwt, the value of intercept is 46.1326, the slope for BMI is -0.0657 with p-value 0.67. In the case of PGS_unwt, the value of intercept for TSLS is 45.9107, the slope for BMI is -0.0572 with p-vale 0.586. Like previous cases,

```
Call:
ivreg(formula = lifesat ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_exwt,
    data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9733 -0.6216 -0.2111  0.8491  3.5095

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.989595   0.531018   9.396  < 2e-16 ***
bmi          0.029984   0.020018   1.498   0.1342
age         -0.010502   0.001499  -7.007 2.75e-12 ***
PCA1        -0.387188   2.228287  -0.174   0.8621
PCA2        -1.471169   2.208351  -0.666   0.5053
PCA3         0.176153   2.212344   0.080   0.9365
PCA4         0.954317   2.213935   0.431   0.6664
PCA5         4.594206   2.211261   2.078   0.0378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.588 on 5154 degrees of freedom
Multiple R-Squared: 0.007012,   Adjusted R-squared: 0.005663
Wald test: 8.218 on 7 and 5154 DF,  p-value: 5.375e-10
```

Figure 22: Figure showing the outcome for TSLS considering life satisfaction as outcome with weighted PGS

```
Call:
ivreg(formula = sf12mcs ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_unwt,
    data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-44.808  -4.951   2.363   6.925  22.416

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.910744   3.572826  12.850   <2e-16 ***
bmi          -0.057247   0.135331  -0.423    0.672
age           0.100240   0.009508  10.543   <2e-16 ***
PCA1         -5.925316  13.591655  -0.436    0.663
PCA2        -13.640724  13.446619  -1.014    0.310
PCA3         20.178513  13.484086   1.496    0.135
PCA4         -3.291890  13.516985  -0.244    0.808
PCA5         -1.381887  13.509886  -0.102    0.919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.634 on 5103 degrees of freedom
Multiple R-Squared: 0.03194,    Adjusted R-squared: 0.03061
Wald test: 21.19 on 7 and 5103 DF,  p-value: < 2.2e-16
```

Figure 23: Figure showing the outcome for TSLS considering sf12mcs as outcome with unweighted PGS

no significant association is found with BMI.

Table 9 shows a table of result for MR Egger, IVW, simple median and weighted median, considering GHQ as outcome. The coefficient estimate for simple and weighted median are -0.276 and -0.382 respectively. The estimate for IVW is -0.156 with a p-value of 0.663. The intercept for MR Egger is -0.008 and the p-value for this estimation is 0.750 which indicates that there is no ev-

```
Call:
ivreg(formula = sf12mcs ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_exwt,
    data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-44.790  -4.930   2.369   6.931  22.612

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.132593   3.193783  14.444   <2e-16 ***
bmi          -0.065715   0.120740  -0.544    0.586
age           0.100541   0.009255  10.863   <2e-16 ***
PCA1         -5.944255  13.588325  -0.437    0.662
PCA2        -13.694487  13.438412  -1.019    0.308
PCA3         20.217231  13.478572   1.500    0.134
PCA4         -3.279612  13.514066  -0.243    0.808
PCA5         -1.412779  13.505421  -0.105    0.917
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.632 on 5103 degrees of freedom
Multiple R-Squared: 0.03232,    Adjusted R-squared: 0.03099
Wald test: 21.21 on 7 and 5103 DF,  p-value: < 2.2e-16
```

Figure 24: Figure showing the outcome for TSLS considering sf12mcs as outcome with weighted PGS

idence that direct pleotropic effect exist and therefore both TSLS and IVW can provide consistent estimates. Although the IVW give a slightly higher estimate comparsing to TSLS estimates, the p-values for these estimates are larger than 0.05 and show no evidence of a causal relationship from BMI to GHQ. Figure 25 is an interactive plot, showing the line of best fit for the result generated by simple median, weighted median, IVW and Egger regression methods.

Table 9: Coefficient estimate for simple median, weighted median, IVW and Egger regression, with GHQ as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	-0.276	0.545	-1.344 0.792	0.612
Weighted median	-0.382	0.592	-1.543 0.779	0.519
IVW	-0.156	0.358	-0.858 0.545	0.663
MR-Egger	0.299	0.941	-1.545 2.143	0.750
(intercept)	-0.008	0.015	-0.037 0.022	0.601

Table 10 shows a table of result for MR Egger, IVW, simple median and weighted median, considering overall life satisfaction score as an outcome. The coefficient estimate for simple and weighted median are -0.037 and 0.184 respectively. The estimate for IVW is 0.101 with a p-value

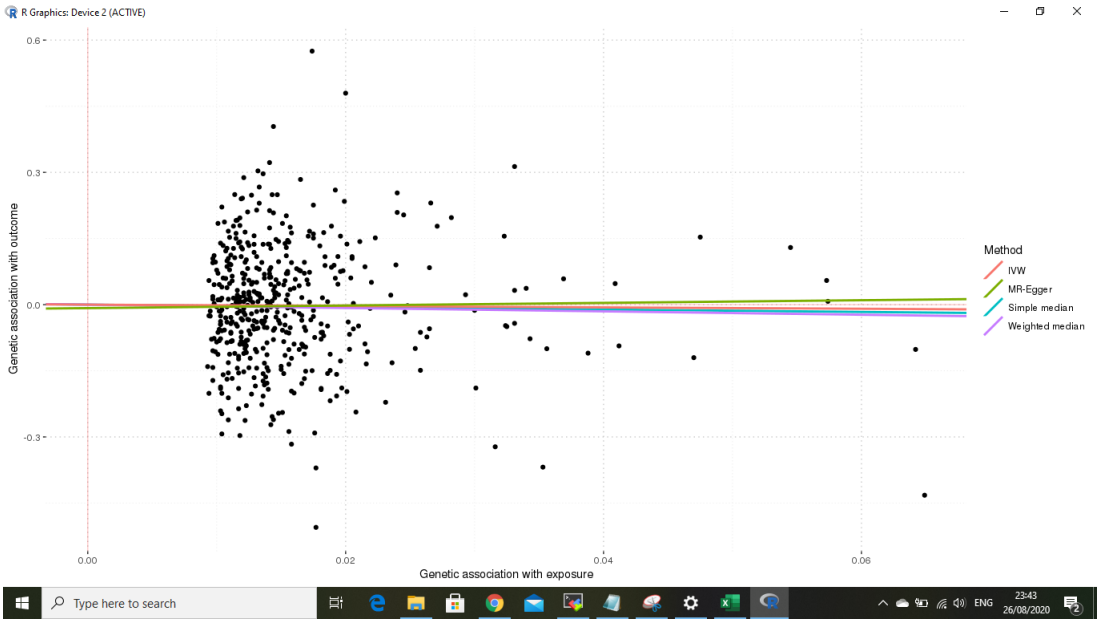


Figure 25: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on GHQ

of 0.304. The MR Egger intercept is -0.005 and the pvalue for this estimation is 0.142 which indicates that there is no evidence that direct pleotropic effect exist and therefore both TSLS and IVW can provide consistent estimates. Although the IVW give a slightly higher estimate comparsing to TSLS estimates, the p-values for these estimates are larger than 0.05 and show no evidence of a causal relationship from BMI to overall life satisfaction. Figure 26 is an interactive plot, showing the line of best fit for the result generated by simple median, weighted median, IVW and Egger regression methods.

Table 10: Coefficient estimate for simple median, weighted median, IVW and Egger regression, with life satisfaction as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	-0.037	0.150	-0.330 0.256	0.805
Weighted median	0.184	0.171	-0.151 0.519	0.282
IVW	0.101	0.098	-0.092 0.294	0.304
MR-Egger	0.379	0.258	-0.127 0.885	0.142
(intercept)	-0.005	0.004	-0.013 0.003	0.245

Table 11 shows a table of result for MR Egger, IVW, simple median and weighted median,

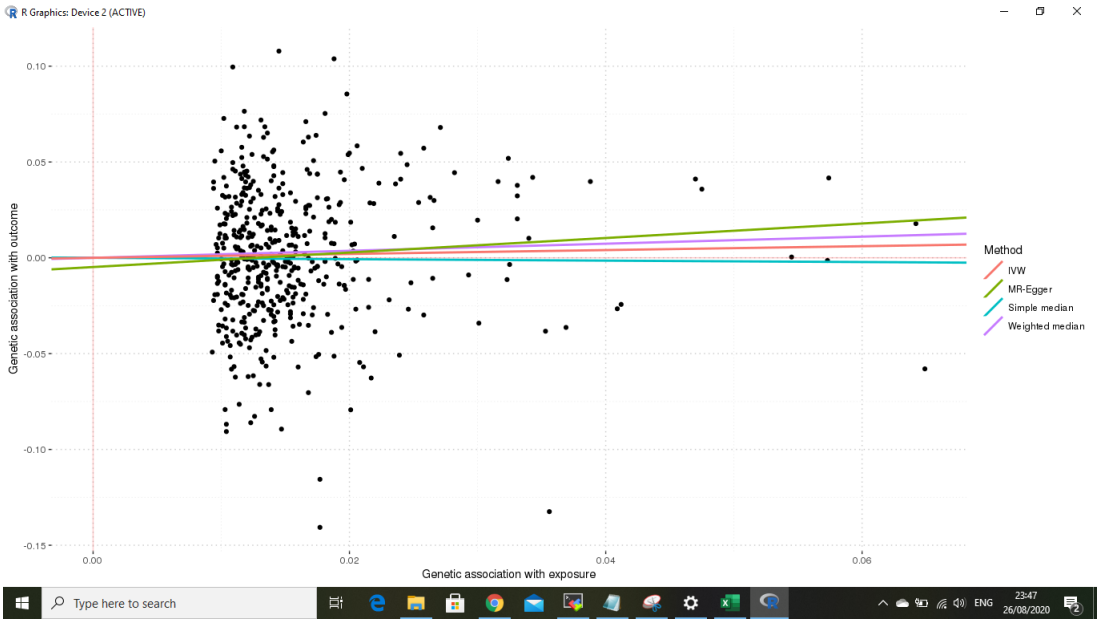


Figure 26: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on life satisfaction

considering overall life satisfaction score as an outcome. The coefficient estimate for simple and weighted median are 0.983 and -0.099 respectively. The estimate for IVW is 0.089 with a p-value of 0.882. The intercept for MR-Egger is 0.004 and the pvalue for this estimation is 0.922 which indicates that there is no evidence that direct pleotropic effect exist and therefore both TSLS and IVW can provide consistent estimates. Although the IVW give a slightly higher estimate compars-ing to TSLS estimates, the p-values for these estimates are larger than 0.05 and show no evidence of a causal relationship from BMI to sf12mcs. Figure 27 is an interactive plot, showing the line of best fit for the result generated by simple median, weighted median, IVW and Egger regression methods.

4.3.2 Stratified subset- Male

Figure 28 and 29 are showing TSLS result for considering GHQ as outcome, with unweighted and weighted polygenic scores respectively. For the PGS_unwt, the value of intercept is 3.7254,

Table 11: Coefficient estimate for simple median, weighted median, IVW and Egger regression, with sf12mcs score as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	0.983	0.927	-0.833 2.800	0.289
Weighted median	-0.099	1.048	-2.154 1.956	0.925
IVW	0.089	0.600	-1.087 1.264	0.882
MR-Egger	-0.154	1.575	-3.240 2.933	0.922
(intercept)	0.004	0.025	-0.045 0.054	0.868

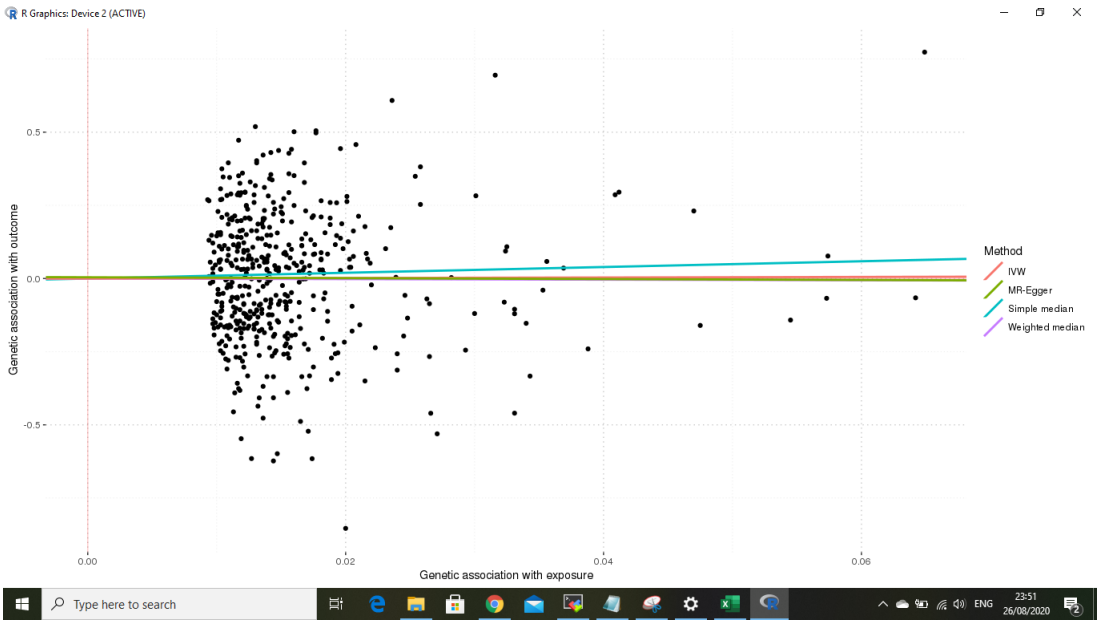


Figure 27: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on sf12mcs

```
Call:
ivreg(formula = ghq ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_unwt,
    data = datam)

Residuals:
    Min      1Q  Median      3Q     Max
-11.537  -3.251  -1.105   1.894  26.491

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.725429   2.612043   1.426  0.15387
bmi          0.278163   0.100022   2.781  0.00544 **
age         -0.023939   0.006146  -3.895 9.98e-05 ***
PCA1         3.344478   7.861034   0.425  0.67053
PCA2        -1.375505   7.923568  -0.174  0.86219
PCA3       -11.798343   7.965965  -1.481  0.13866
PCA4         4.143172   7.948278   0.521  0.60221
PCA5       -11.221741   7.822303  -1.435  0.15148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.023 on 4003 degrees of freedom
Multiple R-Squared: -0.0266,    Adjusted R-squared: -0.0284
Wald test: 3.075 on 7 and 4003 DF,  p-value: 0.00312
```

Figure 28: Figure showing the outcome for TSLS considering GHQ as outcome with unweighted PGS

```
Call:
ivreg(formula = ghq ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_exwt,
    data = datam)

Residuals:
    Min      1Q  Median      3Q     Max
-11.441  -3.245  -1.129   1.869  26.230

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.620070   2.240066   2.062  0.03923 *
bmi           0.243741   0.085636   2.846  0.00445 **
age          -0.022565   0.005769  -3.911 9.34e-05 ***
PCA1          3.203620   7.815996   0.410  0.68192
PCA2         -1.199362   7.876679  -0.152  0.87898
PCA3        -12.130841   7.907999  -1.534  0.12511
PCA4          4.438788   7.893549   0.562  0.57392
PCA5        -11.329095   7.778616  -1.456  0.14535
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.996 on 4003 degrees of freedom
Multiple R-Squared: -0.01559,   Adjusted R-squared: -0.01736
Wald test: 3.149 on 7 and 4003 DF,  p-value: 0.002546
```

Figure 29: Figure showing the outcome for TSLS considering GHQ as outcome with weighted PGS

the slope for BMI is 0.2782 with p-value 0.005. For PGS_exwt, the value of intercept for TSLS is 4.6201, the slope for BMI is 0.2437 with p-value 0.004. Since the $p - value < 0.05$ for weighted PGS, hence it shown significant association of BMI to GHQ

```
Call:
ivreg(formula = lifesat ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_unwt,
    data = datam)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7845 -0.6083 -0.2569  0.7901  3.0720

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.8899175  0.8035426   7.330 2.77e-13 ***
bmi          0.0039228  0.0307852   0.127    0.899
age         -0.0122828  0.0018958  -6.479 1.03e-10 ***
PCA1         1.4562863  2.4245849   0.601    0.548
PCA2        -0.8605359  2.4319060  -0.354    0.723
PCA3        -0.0001194  2.4490924   0.000    1.000
PCA4        -0.1824032  2.4468304  -0.075    0.941
PCA5         0.1746496  2.4126844   0.072    0.942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.544 on 4004 degrees of freedom
Multiple R-Squared: 0.01843,    Adjusted R-squared: 0.01671
Wald test: 10.21 on 7 and 4004 DF,  p-value: 9.796e-13
```

Figure 30: Figure showing the outcome for TSLS considering life satisfaction as outcome with unweighted PGS

Figures 30 and 31 are showing TSLS result for considering life satisfaction as outcome, with unweighted and weighted polygenic scores respectively. For the PGS_unwt, the value of intercept

```
Call:
ivreg(formula = lifesat ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_exwt,
    data = datam)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8972 -0.6171 -0.2548  0.7891  3.1671

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.544132   0.692206   8.009 1.50e-15 ***
bmi          0.017234   0.026476   0.651    0.515
age         -0.012818   0.001787  -7.173 8.69e-13 ***
PCA1         1.504152   2.422606   0.621    0.535
PCA2        -0.926997   2.429319  -0.382    0.703
PCA3         0.124916   2.443323   0.051    0.959
PCA4        -0.297700   2.441723  -0.122    0.903
PCA5         0.232821   2.410395   0.097    0.923
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.543 on 4004 degrees of freedom
Multiple R-Squared: 0.0195,     Adjusted R-squared: 0.01778
Wald test: 10.28 on 7 and 4004 DF,  p-value: 7.843e-13
```

Figure 31: Figure showing the outcome for TSLS considering life satisfaction as outcome with weighted PGS

is 5.8898, the slope for BMI is 0.0039 with p-value 0.899. For PGS_exwt, the value of intercept for TSLS is 5.5441, the slope for BMI is 0.0172 with p-value 0.515. Since the $p - valve > 0.05$ in both the cases, hence there is no significant association observed for life satisfaction with BMI

```
Call:
ivreg(formula = sf12mcs ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_unwt,
    data = datam)

Residuals:
    Min      1Q  Median      3Q     Max
-44.394  -4.208   1.905   5.730  27.020

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.89499    4.41222  12.668   <2e-16 ***
bmi          -0.31678    0.16867  -1.878   0.0604 .
age           0.09233    0.01028   8.982   <2e-16 ***
PCA1          1.20331   13.27822   0.091   0.9278
PCA2         -2.59264   13.36278  -0.194   0.8462
PCA3         20.12407   13.44138   1.497   0.1344
PCA4        -31.96395   13.40851  -2.384   0.0172 *
PCA5         18.43807   13.19948   1.397   0.1625
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.429 on 3957 degrees of freedom
Multiple R-Squared: 0.0113,     Adjusted R-squared: 0.009556
Wald test: 16.72 on 7 and 3957 DF,  p-value: < 2.2e-16
```

Figure 32: Figure showing the outcome for TSLS considering sf12mcs as outcome with unweighted PGS

Figures 32 and 33 are showing TSLS result for considering sf12mcs score as outcome, with unweighted and weighted polygenic scores respectively. For the PGS_unwt, the value of intercept

```
Call:
ivreg(formula = sf12mcs ~ bmi + age + PCA1 + PCA2 + PCA3 + PCA4 +
    PCA5 | age + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + PGS_exwt,
    data = datam)

Residuals:
    Min      1Q  Median      3Q     Max
-44.371  -4.147   1.947   5.719  26.345

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.21831    3.77519  14.362   <2e-16 ***
bmi          -0.25238    0.14407  -1.752   0.0799 .
age           0.08982    0.00967   9.289   <2e-16 ***
PCA1          1.29594   13.22171   0.098   0.9219
PCA2         -2.90775   13.29979  -0.219   0.8269
PCA3         20.71015   13.36168   1.550   0.1212
PCA4        -32.43065   13.33737  -2.432   0.0151 *
PCA5         18.56830   13.14273   1.413   0.1578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.394 on 3957 degrees of freedom
Multiple R-Squared: 0.01962,    Adjusted R-squared: 0.01788
Wald test: 16.79 on 7 and 3957 DF,  p-value: < 2.2e-16
```

Figure 33: Figure showing the outcome for TSLS considering sf12mcs as outcome with weighted PGS

is 55.8950, the slope for BMI is -0.3168 with p-value 0.060. For PGS_exwt, the value of intercept for TSLS is 54.2183, the slope for BMI is -0.2524 with p-value 0.079. No significant association is found for sf12mcs with BMU, as $p - value > 0.05$ in both cases.

Table 12 shows a table of result for MR Egger, IVW, simple median and weighted median, considering overall life satisfaction score as an outcome. The coefficient estimate for simple and weighted median are 0.935 and 0.820 respectively. The estimate for IVW is 0.698 with a p-value of 0.044. The Egger intercept is 0.008 and the pvalue for this estimation is 0.792 which indicates that there is no evidence that direct pleotropic effect exist and therefore both TSLS and IVW can provide consistent estimates. Although the IVW give a slightly higher estimate comparsing to TSLS estimates, the p-values for these estimates are larger than 0.05 and show no evidence of a causal relationship from BMI to GHQ. Figure 34 is an interactive plot, showing the line of best fit for the result generated by simple median, weighted median, IVW and Egger regression methods.

Table 12: Coefficient estimate for simple median, weighted median, IVW and Egger regression, with GHQ score as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	0.935	0.540	-0.122 1.993	0.083
Weighted median	0.820	0.612	-0.380 2.020	0.181
IVW	0.698	0.347	0.017 1.379	0.044
MR-Egger	0.239	0.910	-1.544 2.023	0.792
(intercept)	0.008	0.015	-0.021 0.037	0.585

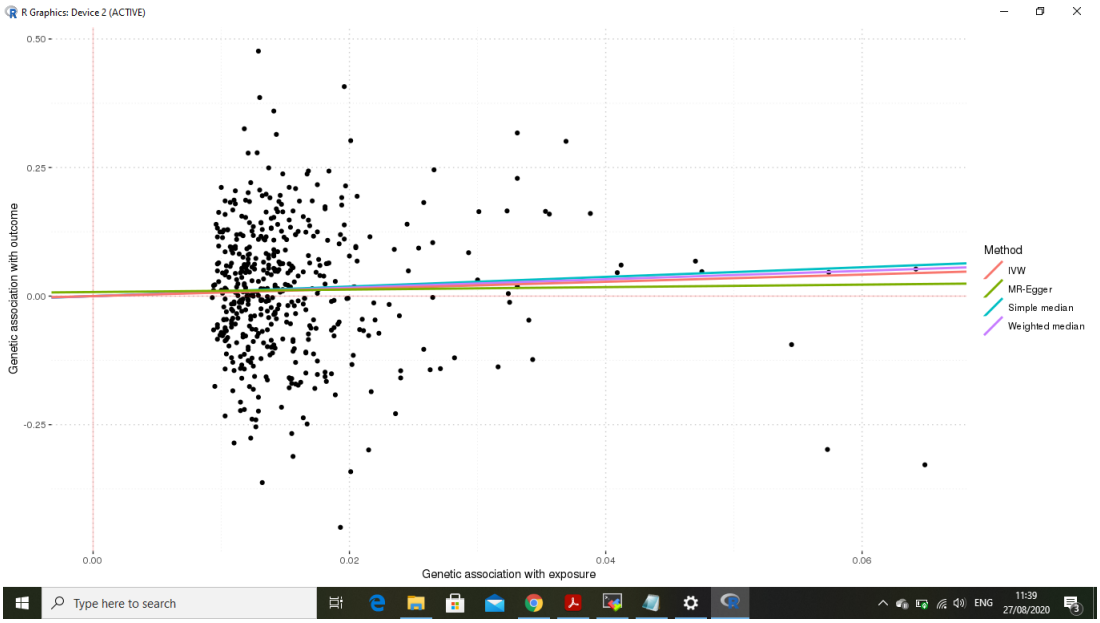


Figure 34: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on GHQ

Table 13 shows a table of result for MR Egger, IVW, simple median and weighted median, considering overall life satisfaction score as an outcome. The coefficient estimate for simple and weighted median are 0.007 and 0.173 respectively. The estimate for IVW is 0.020 with a p-value of 0.857. The intercept for MR-Egger is -0.006 and the pvalue for this estimation is 0.210 which indicates that there is no evidence that direct pleotropic effect exist and therefore both TSLS and IVW can provide consistent estimates. Although the IVW give a slightly higher estimate compars-ing to TSLS estimates, the p-values for IVW estimate is larger than 0.05 and show no evidence of a causal relationship from BMI to GHQ. Figure 35 is an interactive plot, showing the line of best fit for the result generated by simple median, weighted median, IVW and Egger regression methods.

Table 13: Coefficient estimate for simple median, weighted median, IVW and Egger regression, with life satisfaction as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	0.007	0.168	-0.322 0.336	0.968
Weighted median	0.173	0.187	-0.193 0.539	0.353
IVW	0.020	0.112	-0.200 0.240	0.857
MR-Egger	0.368	0.294	-0.207 0.944	0.210
(intercept)	-0.006	0.005	-0.015 0.003	0.200

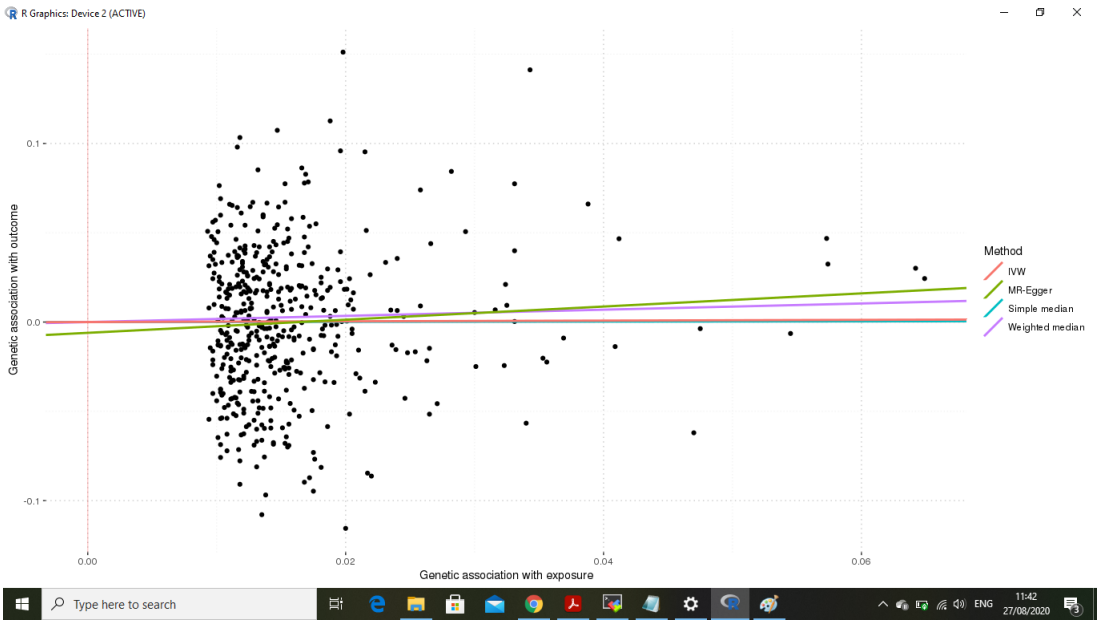


Figure 35: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on life satisfaction

Table 14 shows a table of result for MR Egger, IVW, simple median and weighted median, considering overall life satisfaction score as an outcome. The coefficient estimate for simple and weighted median are -0.360 and -0.213 respectively. The estimate for IVW is -0.801 with a p-value of 0.195. The Egger regression intercept is -0.020 and the pvalue for this estimation is 0.834 which indicates that there is no evidence that direct pleotropic effect exist and therefore both TSLS and

IVW can provide consistent estimates. Although the IVW give a slightly lower estimate comparsing to TSLS estimates, the p-values for these estimates are larger than 0.05 and show no evidence of a causal relationship from BMI to sf12mcs. Figure 36 is an interactive plot, showing the line of best fit for the result generated by simple median, weighted median, IVW and Egger regression methods.

Table 14: Coefficient estimate for simple median, weighted median, IVW and Egger regression, with sf12mcs as an outcome

Method	Estimate	Std Error	95% CI	P-value
Simple median	-0.360	0.928	-2.179 1.458	0.698
Weighted median	-0.213	1.028	-2.228 1.802	0.836
IVW	-0.801	0.618	-2.013 0.411	0.195
MR-Egger	0.340	1.620	-2.835 3.514	0.834
(intercept)	-0.020	0.026	-0.071 0.031	0.446

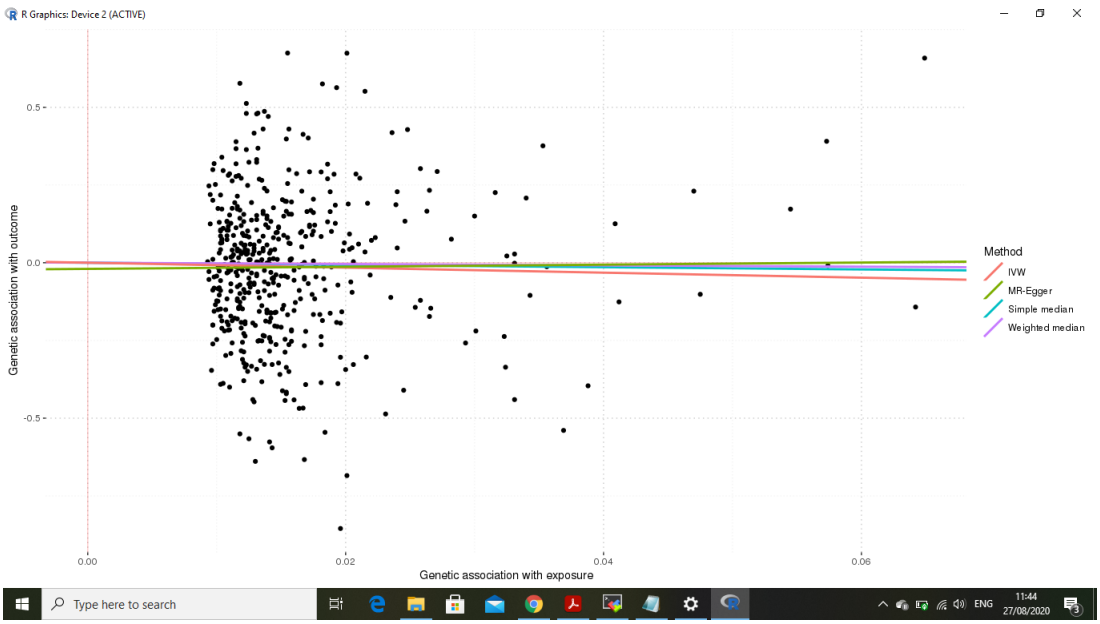


Figure 36: Graph showing the best fitted line for MR Egger, IVW, simple and weighted median, to estimate casual effect of BMI on sf12mcs

5 Discussion

The aim of this study was to analyse the data of 9920 participants, containing containing results of various physical and mental wellbeing scores. The aim was to find an estimate of casual effect of risk factors causing a disease with their genetic variants. We considered BMI as a predictor variable with other factors such as gender, age and first five PCA of all these variables. As a parameter to measure status of subjective well-being we consider overall life-satisfaction for and mental well-being, we used GHQ score and sf12mcs score, as our three prime response variables. We aimed to apply Mendelian Randomization methods to check for IV assumptions and then finding required results.

Firstly we have checked for the correlation of BMI with weighted and unweighted PGS. The concept of considering unweighted PGS indicates that all genetic variants contributes equally in the study and the weighted one is picked from GWAS study, which suggest that the some of SNPs weights more in the overall estimate, than others. By applying a linear regression model we found that the p-value for both PGS_unwt and PGS_exwt is $2.2e^{-16}$, which seems very promising, hence we considered the both for our project, to check which one is giving more significant results with the data in this case. During the preliminary statistical analysis of data, we observed that there is uneven distribution of data and a marginal difference in average GHQ, sf12mcs and life satisfaction scores. So, in the next step, we have applied all 5 proposed MR methods on the whole dataset and then on the gender based stratified data subsets to explore all sort of possible association.

GHQ

With TSLS on the overall data, the p-value for PGS_unwt and PGS_exwt are 0.40 and 0.035 respectively. Hence for the entire dataset there is a significant association of BMI with GHQ, considering weighted PGS, as the p-value is less than 0.05. And the association is positive, i.e increase in BMI indicates increase in GHQ score. Similarly, in the case of the male subset, for TSLS the p-values for both PGS_unwt and PGS_exwt are less than 0.05, indicates significant and positive

association. However, for the female subset, by TSLS the p-values are greater than 0.05, indicates no significant association. Also, the result of all other MR methods are insignificant in each of these cases.

Overall life satisfaction

By applying TSLS on overall dataset, the p-value for PGS_unwt and PGS_exwt are 0.410 and 0.124 respectively. Hence, there is no significant association of life satisfaction with BMI. By applying the same method on male subset the p-value are greater than the threshold value, hence no significant association. Similarly for female subset TSLS findings indicates greater p-value, hence no significant. Also, the results of other MR methods shows insignificant results to prove an association between the two.

sf12mcs

The summary of TSLS on the overall dataset, the p-values are 0.131 and 0.126 respectively with unweighted and weighted PGS as IV, indicates no significant association of the outcome with BMI. MR Egger intercept is -0.007 with p-value 0.939 suggests the significance of IVW and TSLS. However, the p-value of IVW 0.494 points out to insignificant result. For the female subset, TSLS summary p-values are 0.672 and 0.586 with PGS_unwt and PGS_exwt respectively. The intercept for Egger and p-values are 0.004 and 0.922 respectively and for IVW is 0.882. Since for the female subset all methods are showing insignificant results, hence no association. Similarly for the male subset the p-values for TSLS are greater than 0.05, no association found. For the Egger regression, intercept is 0.008 with p-value 0.792, which indicates that there is no evidence that direct pleotropic effect exist and therefore both TSLS and IVW can provide consistent estimates. Although the IVW give a slightly higher estimate comparsing to TSLS estimates, the p-values for these estimates are larger than 0.05 and show no evidence of a causal relationship from BMI to sf12mcs.

Hence our overall finding is that the intercept of Egger regression is close to zero and p-vale us

is greater than the threshold, it indicates that there is no evidence that direct pleiotropic effect exists and therefore both the IVW and TSLS result are valid. Although TSLS and IVW do not provide the same estimates consistent with each other. Using TSLS, we have found a significant association of BMI using on entire data and male subset, we observed a significant p-value for GHQ, using both weighted and unweighted PGS. One of the reasons that we failed to find significant association when GHQ as an outcome for female and life satisfaction and sc12mcs as an outcome for all three data, is that the provided IVs are not strong enough or our sample size is small. Since we have been given with SNPs information with a very small coefficient it indicates at a weak IV. To tackle this problem, we have added them together to make them strong IV, hoping to solve the issue, but unfortunately, it failed. Hence further study in this area is required, considering the case of weak IV and use some of the techniques to get more precise results.

Our body is considered one of the most complex machineries that work on its own. Birth of an offspring from the fertilization of two single cells is more than just a biological process. Our body is consist of eleven major organ systems, each responsible for a specific function, inter-depending to each other. A disorder or malfunctioning that interrupts its vital function is termed as a disease. The age of childhood till the late '20s is considered as the growth years, early '30s till the mid '50s is known as maintenance age and after that, we enter into the deteriorated phase. At a younger age, a person is more capable to follow a healthier diet to maintaining good health and increasing mortality. Hence the whole concept of MR study is to find the traits of disease causation by observing the risk factors.

One of the limitations of MR is that the study could take several years to decades to monitor the causal effect of a risk factor and turn into a disease. However, the development of a coronary disease takes more decades, before the symptoms started appearing. For example, in the case of cardiac heart disease, thickening, and stuffing of artery walls interrupt the blood flow to our tissues and organs, which causes a life-threatening situation. The formation and development of plaque

in artery walls happen slowly and gradually over the span of many years and generally started at a young age only. Hence there might be a possibility in the future that biomedical could introduce a test, which reads genetic data of an individual at a young age and predicts the probability of potentially developing a life-threatening illness. This would be a stepping stone that in human history, which will allow us to start working on our overall wellbeing and to minimise those risk factors by adopting positive changes in our lifestyle. My grandfather used to say to me, the richest person in the world is the one who is free from any form of illness, follows healthy rhyme, and able to enjoy a plethora of foods, without any medical restriction.

References

[1] Tao Huang, Tiange Wang, Yan Zheng, Christina Ellervik, Xiang Li, Meng Gao, Zhe Fang, Jin-Fang Chai, Yujie Wang, Trudy Voortman, et al. Association of birth weight with type 2 diabetes and glycemic traits: A mendelian randomization study. *JAMA network open*, 2(9): e1910915–e1910915, 2019.

[2] Marilyn JH Morris AO, William AS and Robert AW. Health and subjective well-being: A meta-analyis. *The International journal of aging and human development*, 19(2):111–132, 1984.

[3] Emmanuel S Mark H. U-shaped association between body mass index and psychological distress in a population sample of 114,218 british adults. In *Mayo Clinic Proceedings*, volume 92, pages 1865–1866. Elsevier, 2017.

[4] Julian T Calvin F Shu-Chuen L Fei G, Nan L and Yin-Bun C. Does the 12-item general health questionnaire contain multiple factors and do we need them? *Health and quality of life outcomes*, 2(1):63, 2004.

[5] Josep ML Jaume V José MT Maria FR, Gabriel MJ. Method effects associated with negatively and positively worded items on the 12-item general health questionnaire (ghq-12): results from a cross-sectional survey with a representative sample of catalonian workers. *BMJ open*, 9(11), 2019.

[6] Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Human molecular genetics*, 27(20):3641–3649, 2018.

[7] Jianhua Zhao, Jonathan P Bradfield, Mingyao Li, Kai Wang, Haitao Zhang, Cecilia E Kim, Kiran Annaiah, Joseph T Glessner, Kelly Thomas, Maria Garris, et al. The role of obesity-

associated loci identified in genome-wide association studies in the determination of pediatric bmi. *Obesity*, 17(12):2254–2257, 2009.

[8] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

[9] George Davey Smith and Shah Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International journal of epidemiology*, 33(1):30–42, 2004.

[10] George Davey Smith and Shah Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22, 2003.

[11] Olena O Yavorska and Stephen Burgess. Mendelianrandomization: an r package for performing mendelian randomization analyses using summarized data. *International journal of epidemiology*, 46(6):1734–1739, 2017.

[12] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[13] Simon GT Stephen B, Dylan SS. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

[14] Anu R Elina S Anna KR Aila R Milla SL, Jaakko K. Body mass index and subjective well-being in young adults: a twin population study. *BMC public health*, 13(1):231, 2013.

[15] Gundi Knies. Understanding society: waves 1–7, 2009–2016 and harmonised bhps: waves 1–18, 1991–2009, user guide. *Colchester: University of Essex*, 2017.

[16] Sarah Stewart-Brown, Alan Tennant, Ruth Tennant, Stephen Platt, Jane Parkinson, and Scott Weich. Internal construct validity of the warwick-edinburgh mental well-being scale

(wemwbs): a rasch analysis using data from the scottish health education population survey. *Health and quality of life outcomes*, 7(1):15, 2009.

[17] Ruth Tennant, Louise Hiller, Ruth Fishwick, Stephen Platt, Stephen Joseph, Scott Weich, Jane Parkinson, Jenny Secker, and Sarah Stewart-Brown. The warwick-edinburgh mental well-being scale (wemwbs): development and uk validation. *Health and Quality of life Outcomes*, 5(1):63, 2007.

[18] Simon Chapman. Commentary: Tobacco industry health research blood money ': The british health promotion research trust. *Community health studies*, 11(2):139–142, 1987.

[19] Sepideh S Farivar, William E Cunningham, and Ron D Hays. Correlated physical and mental health summary scores for the sf-36 and sf-12 health survey, v. 1. *Health and quality of life outcomes*, 5(1):54, 2007.

[20] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

[21] Stephen Burgess and Jack Bowden. Integrating summarized data from multiple genetic variants in mendelian randomization: bias and coverage properties of inverse-variance weighted methods. *arXiv preprint arXiv:1512.04486*, 2015.

[22] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

[23] Stephen Burgess and Simon G Thompson. Interpreting findings from mendelian randomization using the mr-egger method. *European journal of epidemiology*, 32(5):377–389, 2017.

Appendix

```
install.packages("AER")
library(AER)
install.packages("systemfit")
library(systemfit)
install.packages("MendelianRandomization"
library(MendelianRandomization)
library(dplyr)
PGS <- read.delim("PGSdata_Nirupma.txt", head=T)
SNP <- read.delim("SNPdata_Nirupma.txt", head=T)
Survey <- read.delim("Surveydata_Nirupma.txt", head=T)
Survey_n <- merge(PGS, Survey, by="id")
Survey_n$wave=ifelse(is.na(Survey_n$b_bmival), 3, 2)
Survey_n$bmi=ifelse(Survey_n$wave==2, Survey_n$b_bmival, Survey_n$c_bmival)
Survey_n$bmi[Survey_n$bmi > 60] <- NA
sum(is.na(Survey_n$bmi))
Survey_n$wave=ifelse(is.na(Survey_n$b_age_dv),3, 2)
Survey_n$age=ifelse(Survey_n$wave==2, Survey_n$b_age_dv, Survey_n$c_age_dv)
sum(is.na(Survey_n$age))
Survey_n$wave=ifelse(is.na(Survey_n$b_scghq1_dv),3, 2)
Survey_n$ghq=ifelse(Survey_n$wave==2, Survey_n$b_scghq1_dv, Survey_n$c_scghq1_dv)
sum(is.na(Survey_n$ghq))
Survey_n$wave=ifelse(is.na(Survey_n$a_fimngrs_dv), 3, 2)
Survey_n$income=ifelse(Survey_n$wave==2, Survey_n$b_fimngrs_dv , Survey_n$c_fimngrs_dv )
sum(is.na(Survey_n$income))
Survey_n$b_sclfsato[Survey_n$b_sclfsato == ""]<- NA
sum(is.na(Survey_n$b_sclfsato))
```

```
Survey_n$c_sclfsato[Survey_n$c_sclfsato == ""]<- NA
sum(is.na(Survey_n$c_sclfsato))
Survey_n$wave=ifelse(is.na(Survey_n$b_sclfsato), 3, 2)
Survey_n$lifesat=ifelse(Survey_n$wave==2, Survey_n$b_sclfsato , Survey_n$c_sclfsato )
sum(is.na(Survey_n$lifesat))
Survey_n$wave=ifelse(is.na(Survey_n$b_sf12mcs_dv), 3, 2)
Survey_n$sf12mcs=ifelse(Survey_n$wave==2, Survey_n$b_sf12mcs_dv , Survey_n$c_sf12mcs_dv)
sum(is.na(Survey_n$sf12mcs))
mydata<-merge(SNP, Survey_n, by="id")
dim(mydata)
##### Using GHQ as output for MR ########
Beta<- vector("double", 481)
ex.outcome<- vector("double", 481)
for (i in 2:ncol(SNP))
Beta[[i]] <-summary(lm(ghq~ mydata[,i]+ age +bmi +sex+PCA1 +PCA2+PCA3+PCA4+PCA5,
data= mydata))$ coefficient[2,1]
ex.outcome[[i]] <- summary(lm(ghq~ mydata[,i]+ age +bmi +sex+PCA1+PCA2+PCA3+PCA4
+PCA5, data= mydata))$ coefficient[2,2]
SNPname <-colnames(mydata[2:481])
mat<- cbind(SNPname, Beta, ex.outcome)
setwd("/home/np19443/")
SNPinfo <- read.csv("SNP_information_UKHLS.csv", head=T)
dim(SNPinfo)
names(SNPinfo)[names(SNPinfo)== "SNP"]<- "SNPname"
Matrix <- merge(mat, SNPinfo, by="SNPname")
dim(Matrix)
input<- mr_input(bx=C$ beta, bxse=C$ se.exposure, by=C$ Beta, byse=C$ se.outcome)
```

```
MRInput<- mr_input (bx= Matrix$ beta, bxse= Matrix$ se.exposure, by=Matrix$ betaY, byse=Matrix$
erY)
All<- mr_allmethods(MRInput, method =”main”)

    ##### Using GHQ as output for TSLS ########
olsreg<- lm(ghq  bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5, data = Survey_n)
ivreg<- ivreg(ghq  bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5—age +sex+PCA1+ PCA2+PCA3
+PCA4+PCA5+PGS_unwt, data = Survey_n)
summary(ivreg)
ivreg2<- ivreg(ghq∼ bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5—age+sex+PCA1+PCA2+
PCA3+PCA4+PCA5+PGS_exwt, data = Survey_n)
summary(ivreg2)

    ##### Using lifesat as output for MR ########
X1<- vector(”double”, 481)
X2<- vector(”double”, 481)
for (i in 2:ncol(SNP)) X1[[i]]¡-summary(lm(lifesat∼ mydata[,i]+age+bmi+ sex+PCA1+ PCA2+PCA3+PCA4+PC
data=mydata))$ coefficient[2,1]
X2[[i]] <-summary(lm(lifesat∼ mydata[,i]+ age+bmi +sex+PCA1+PCA2+PCA3+PCA4+ PCA5,
data= mydata))$ coefficient[2,2]
matI<- cbind(SNPname, X1, X2)
MatrixI<- merge(matI, SNPinfo, by=”SNPname”)
MRInputI<- mr_input (bx= MatrixI$ beta, bxse= MatrixI$ se.exposure, by=MatrixI$ X1, byse=MatrixI$
X2)
All¡- mr_allmethods(MRInput, method =”main”)

    ##### Using lifesat as output for TSLS ########
```

olsreg3<- lm(lifesat~ bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5, data = Survey_n)

summary(olsreg3)

ivreg3<- ivreg(lifesat~ bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5—age +sex+PCA1+PCA2+ PCA3+PCA4+PCA5+PGS_unwt, data = Survey_n)

summary(ivreg3)

ivreg4¡- ivreg(lifesat~ bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5—age +sex+PCA1+PCA2+ PCA3+PCA4+PCA5+PGS_exwt, data = Survey_n)

summary(ivreg4)

Using sf12mcs as output for MR

s1<- vector("double", 481)

s2<- vector("double", 481)

for (i in 2:ncol(SNP)) s1[[i]]<-summary(lm(sf12mcs~ mydata[,i]+ age+bmi +sex+PCA1+PCA2+ PCA3+PCA4+PCA5, data= mydata))$ coefficient[2,1]

s2[[i]] <- summary(lm(sf12mcs~ mydata[,i]+ age+bmi +sex+PCA1+PCA2+ PCA3+PCA4+PCA5, data= mydata))$ coefficient[2,2]

mat2<- cbind(SNPname, s1, s2)

Matrix2 <- merge(mat2, SNPinfo, by="SNPname")

MRInput<- mr_input (bx= Matrix2$ beta, bxse= Matrix2$ se.exposure, by=Matrix2$ s1, byse=Matrix2$ s2)

All<- mr_allmethods(MRInput, method ="main")

Using sf12mcs as output for TSLS

olsreg5<- lm(sf12mcs~ bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5, data = Survey_n)

summary(olsreg5)

ivreg5<- ivreg(sf12mcs~ bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5—age +sex+PCA1+

```
PCA2+PCA3+PCA4+PCA5+PGS_unwt, data = Survey_n)
summary(ivreg5)
ivreg6<- ivreg(sf12mcs~ bmi+age +sex+PCA1+PCA2+PCA3+PCA4+PCA5—age +sex+PCA1+
PCA2+PCA3+PCA4+PCA5+PGS_exwt, data = Survey_n)
summary(ivreg6)

    ##### Stratified subset-female ########

    dataf <- mydata[ which(mydata$sex=='female'), ]
olsreg<- lm(ghq~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5, data = dataf)
ivreg<- ivreg(ghq~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age +PCA1+PCA2+PCA3+PCA4+PCA5+P
data = dataf)
summary(ivreg)
ivreg1<- ivreg(ghq~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age+PCA1+PCA2+PCA3+PCA4+PCA5+
data = dataf)
summary(ivreg1)
olsreg<- lm(lifesat~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5, data = dataf)
ivreg<- ivreg(lifesat~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age +PCA1+PCA2+PCA3+PCA4+PCA5
data = dataf)
summary(ivreg)
ivreg1<- ivreg(lifesat~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age+PCA1+PCA2+PCA3+PCA4+PCA5
data = dataf)
summary(ivreg1)
olsreg<- lm(sf12mcs~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5, data = dataf)
ivreg<- ivreg(sf12mcs~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age +PCA1+PCA2+PCA3+PCA4+PCA
data = dataf)
summary(ivreg)
```

```
ivreg1<- ivreg(sf12mcs~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age+PCA1+PCA2+PCA3+PCA4+PC
data = dataf)
summary(ivreg1)
g1<- vector("double", 481)
g2<- vector("double", 481)
l1<- vector("double", 481)
l2<- vector("double", 481)
s1<- vector("double", 481)
s2<- vector("double", 481)
for (i in 1:ncol(SNP))
g1[[i]]<-summary(lm(ghq~ dataf[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,1]
g2[[i]] <- summary(lm(ghq~ dataf[,i]+ age+bmi+PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,2]
l1[[i]]<-summary(lm(lifesat~ dataf[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,1]
l2[[i]] <- summary(lm(lifesat~ dataf[,i]+ age+bmi+PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,2]
s1[[i]]<-summary(lm(sf12mcs~ dataf[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data=
data))$ coefficient[2,1]
s2[[i]]<-summary(lm(sf12mcs~ dataf[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data=
data))$ coefficient[2,2]
mat<- cbind(SNPname, g1, g2,l1, l2, s1, s2)
Matrix<- merge(mat, SNPinfo, by="SNPname")
MRInput<- mr_input (bx= Matrix$ beta, bxse= Matrix$ se.exposure, by=Matrix$ g1, byse=Matrix$
g2)
All<- mr_allmethods(MRInput, method ="main")
```

```
mr_plot(mr_allmethods(MRInput, method =”main”))
MRInput<- mr_input (bx= Matrix$ beta, bxse= Matrix$ se.exposure, by=Matrix$ l1, byse=Matrix$
l2)
All<- mr_allmethods(MRInput, method =”main”)
mr_plot(mr_allmethods(MRInput, method =”main”))
MRInput<- mr_input (bx= Matrix$ beta, bxse= Matrix$ se.exposure, by=Matrix$ s1, byse=Matrix$
s2)
All<- mr_allmethods(MRInput, method =”main”)
mr_plot(mr_allmethods(MRInput, method =”main”))

    ##### Stratified subset-male ########
data <- mydata[ which(mydata$sex==’male’), ]
olsreg<- lm(ghq∼ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5, data = data)
ivreg<- ivreg(ghq∼ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age +PCA1+PCA2+PCA3+PCA4+PCA5+P
data = data)
summary(ivreg)
ivreg1<- ivreg(ghq∼ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age+PCA1+PCA2+PCA3+PCA4+PCA5+
data = data)
summary(ivreg1)
olsreg<- lm(lifesat∼ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5, data = data)
ivreg<- ivreg(lifesat∼ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age +PCA1+PCA2+PCA3+PCA4+PCA5
data = data)
summary(ivreg)
ivreg1<- ivreg(lifesat∼ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age+PCA1+PCA2+PCA3+PCA4+PCA5
data = data)
summary(ivreg1)
olsreg<- lm(sf12mcs∼ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5, data = data)
```

```
ivreg<- ivreg(sf12mcs~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age +PCA1+PCA2+PCA3+PCA4+PCA
data = data)
summary(ivreg)
ivreg1<- ivreg(sf12mcs~ bmi+age+PCA1+PCA2+PCA3+PCA4+PCA5—age+PCA1+PCA2+PCA3+PCA4+PC
data = data)
summary(ivreg1)
G1<- vector("double", 481)
G2<- vector("double", 481)
L1<- vector("double", 481)
L2<- vector("double", 481)
S1<- vector("double", 481)
S2<- vector("double", 481)
for (i in 1:ncol(SNP))
G1[[i]]<-summary(lm(ghq~ data[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,1]
G2[[i]]<- summary(lm(ghq~ data[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,2]
L1[[i]]<-summary(lm(lifesat~ data[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,1]
L2[[i]]<- summary(lm(lifesat~ data[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data= data))$
coefficient[2,2]
S1[[i]]<-summary(lm(sf12mcs~ data[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data=
data))$ coefficient[2,1]
S2[[i]]<-summary(lm(sf12mcs~ data[,i]+ age+bmi +PCA1+PCA2+PCA3+PCA4+PCA5, data=
data))$ coefficient[2,2]
mat<- cbind(SNPname, G1, G2,L1, L2, S1, S2)
Matrix <- merge(mat, SNPinfo, by="SNPname")
```

```
MRInput<- mr_input (bx= Matrix$ beta, bxse= Matrix$ se.exposure, by=Matrix$ G1, byse=Matrix$
G2)
All<- mr_allmethods(MRInput, method =”main”)
mr_plot(mr_allmethods(MRInput, method =”main”))
MRInput<- mr_input (bx= Matrix$ beta, bxse= Matrix$ se.exposure, by=Matrix$ L1, byse=Matrix$
L2)
All<- mr_allmethods(MRInput, method =”main”)
mr_plot(mr_allmethods(MRInput, method =”main”))
MRInput<- mr_input (bx= Matrix$ beta, bxse= Matrix$ se.exposure, by=Matrix$ S1, byse=Matrix$
S2)
All<- mr_allmethods(MRInput, method =”main”)
mr_plot(mr_allmethods(MRInput, method =”main”))
```