Title: The relationship between BMI and COVID-19: an exploration of misclassification and selection bias in a two-sample Mendelian randomisation study

Authors: Gemma L Clayton[1,2*], Ana Gonçalves Soares[1,2*], Neil Goulding[1,2], Maria Carolina Borges[1,2], Michael V Holmes[4-6,], George Davey Smith[1,2,3], Kate Tilling[1,2,3], Deborah A Lawlor[1,2,3†], Alice R Carter[1,2†]

Affiliations

[1] MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, UK

[2] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[3] National Institute for Health Research Bristol Biomedical Research Centre (NIHR Bristol BRC) at University Hospitals Bristol NHS Foundation Trust and University of Bristol, Bristol, UK

[4] Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK;

[5] Medical Research Council Population Health Research Unit at the University of Oxford, Oxford, UK;

[6] National Institute for Health Research, Oxford Biomedical Research Centre, Oxford University Hospital, Oxford, UK;

* Authors contributed equally

† Authors contributed equally

Motivation

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus that causes coronavirus disease 2019 (Covid-19). There have been reports of those with a higher cardiovascular disease (CVD) risk being associated with worse Covid-19 outcomes (1, 2). Some observational studies suggest that hypertension, diabetes, and other obesity-related and cardiovascular disease traits are associated with COVID-19 (3). Mendelian randomization (MR) studies suggest that genetically predicted higher BMI is associated with an increased odds of both COVID-19 infection and severe COVID-19 (4-7). However, whether these associations are causal or explained by bias or residual confounding is unclear. MR uses genetic variants related to potential exposures to explore their causal effects (8, 9). It can be implemented as instrumental variable (IV) analysis that uses genetic variants associated with modifiable risk factors to mitigate confounding (e.g. by socioeconomic and behavioural factors).

In this study we use body mass index (BMI) as an example to highlight potential sources of bias in current MR studies. These biases may arise from the definition of cases (and controls) used and from which control population is chosen based on what the specific causal question are trying to be answered.

Aim

To conduct a two sample MR analysis of BMI and COVID-19 susceptibility and severity and demonstrate ways in which selection and misclassification bias could be explored in MR studies of risk factors for COVID-19, as well as general sources of bias in MR, such as from horizontal pleiotropy and population stratification.

This will include:

- (i) assessing the effect of BMI on Covid-19 susceptibility and severity: using a range of case/control definitions to assess both selection and misclassification bias
- (ii) a no- relevance study testing the association between Covid-19 and BMI to test for selection bias in the Covid-19 data and
- (iii) sensitivity analyses such as MR-egger and weighted median to assess general assumptions of MR.

Data

Sample 1 exposure GWAS:

Complete summary GWAS results for BMI will be obtained from publicly available online GWAS summary data repositories, with majority of them retrieved via MR Base (10):

- GIANT – the most recent one including UKBB. Sensitivity analysis will be done excluding UKBB from the analyses.

Sample 2 outcome GWAS:

Genetic variants that are robustly (genome wide significant (p <5x10-8 and replicated)) associated with of Covid-19 will be extracted from the following data sources and used as instrumental variables for Covid-19:

- Source: COVID-19 Host Genetics Initiative (largest one to date) (11) – includes both 23 and me and UKBB and a range of case and control definitions.
  Link: https://www.covid19hg.org/results/r5/

Case and control definitions and causal questions

Case and control definitions vary by GWAS:

Table Case and control definition (from Host Genetics) and causal questions

Phenotype		**Notes**	**Casual question answered**
Case definition*	**Control**		
Very severe respiratory confirmed Covid	All Population control (anyone)	Release 5 (Jan 21)	Severity and susceptibility
Hospitalised Covid	Non hospitalised Covid	Release 5	Severity

Hospitalised Covid	All Population control	Release 5	Severity and susceptibility
Covid (positive)	All Population control	Release 5	Susceptibility
Very severe respiratory confirmed Covid	Non hospitalised Covid	Release 4 (Oct 20)	Severity
Covid (positive)	lab/self-reported negative (questionnaire?)	Release 4	Susceptibility
Predicted Covid from self-reported symptoms	Predicted or self-reported non-covid	Release 4	Susceptibility

*Footnote for more details on how cases were defined. Restricted to European ancestry

In this study we use publicly-available GWAS results from relevant publications and database (https://gwas.mrcieu.ac.uk/). No individual participant data were collected or used. Details of ethical approval and participant consent for each of the studies that contributed to the GWAS can be found in the original publications (*Ethical approval*).

Methods

MR is a statistical approach that uses genetic instruments to provide information about the relationship between an exposure and an outcome. The relationship between a genetic instrument and an exposure is known as the genetic instrumental variable (IV)-exposure association. 2SMR is a MR approach in which the genetic IV-exposure associations and the genetic IV-outcome association comes from two non-overlapping samples that are from the same underlying population.

Study population

Individuals of European (or mixed) ancestry.

Statistical analyses

We will use two-sample summary data MR to assess the effect of BMI on Covid-19 outcomes. 2SMR assumes that the two samples are independent of each other. For each of the outcome GWAS we will determine whether any of the sample 1 (Covid-19 GWAS) cohorts were included in those outcome GWAS and use that to estimate the percentage overlap.

When selecting genetic instruments from the exposure GWAS we will identify genetic instruments for Covid-19 at genome wide significant p value ($<5\text{x}10^{-8}$), and exclude instruments which have high linkage disequilibrium (LD) with other instruments ($r^2<0.001$). We will then search for the genetic instruments in the outcome datasets. For genetic instruments not available for an outcome, a proxy instrument in high LD with the original instrument ($r^2>0.8$) will be identified via MR-Base based on 1000 Genomes catalogue (CEU reference population). No proxy instruments will be identified for outcomes not available in MR-Base. We will align each genetic association for exposure and outcome on the same effect allele.

For our main Mendelian randomization analyses, we use inverse variance weighting (IVW) with multiplicative random effects to obtain the causal effect of BMI on Covid-19 outcomes and their risk factors. This method generates a causal estimate of BMI on Covid-19 outcomes by regressing the SNP-BMI association on the SNP- Covid-19 outcomes association, weighted by the inverse of the SNP-Covid-19 outcomes association, and constraining the intercept of this regression to zero. Standard errors are corrected to take into account any between SNP heterogeneity and assumes that there is no directional horizontal pleiotropy.

Analyses to explore and account for possible violation of MR assumptions

To check the relevance assumption and weak instrument bias we estimate the mean F statistics and total R2 overall and by each case control comparison.. We will check for between-SNP heterogeneity using the Cochran's Q test and undertake sensitivity analyses using MR-Egger (22), and weighted median.

To assess bias caused by population stratification we use skin tanning as a negative control outcome to compare the association observed between BMI and Covid-19 with the association observed between BMI and skin tanning. Evidence of an association when using the negative control outcome could indicate bias from population stratification in the BMI GWAS. We similarly explore bias in the Covid-19 GWAS by using Covid-19 as the exposure.

Methods to explore potential selection and misclassification bias

Whilst not exclusively tests for selection bias, and indeed not an exhaustive range of tests for selection bias, we will carry out a range of analyses to help understand whether selection bias may be present in 2SMR analyses of BMI and Covid-19 susceptibility and severity. We will explore the following:

- Use different case-control definitions (of Covid-19) to explore different sources of misclassification or selection bias. Similar results across susceptibility/severity questions would give us more confidence in determining causality.
- Use Covid-19 as the exposure and BMI as the outcome in a no relevance study to determine whether the genetic instruments for Covid-19 were related to BMI. Given Covid-19 could not influence BMI assessed prior to 2019, plausibly we would expect null findings. If any effects of Covid-19 on BMI are observed, this suggests selection bias and it would be likely that effects of BMI on Covid-19 (main analysis) are potentially similarly biased.
- We test the genetic correlation using LD score regression between Covid-19 SNPs and SNPs associated with predictors of getting a test.
- We will use multivariable MR (MVMR) to adjust for potential predictors of selection and therefore estimate a direct effect of BMI on Covid-19 independent of selection into the study.

Evidence of a direct effect which is different to the total effect in the main IVW would support the presence of selection bias.

- To assess bias caused by population stratification we use skin tanning as a negative control outcome to compare the association observed between BMI and Covid-19 with the association observed between BMI and skin tanning. Evidence of an association when using the negative control outcome could indicate bias from population stratification in the BMI GWAS. We similarly explored bias in the Covid-19 GWAS by using Covid-19 as the exposure.

Multiple testing

No formal adjustment will be made for multiple testing. Consideration will be taken in interpretation of results to reflect the number of statistical tests performed and the consistency, magnitude and direction of effect estimates for different outcomes.

References

1. Clerkin KJ, Fried JA, Raikhelkar J, Sayer G, Griffin JM, Masoumi A, et al. COVID-19 and Cardiovascular Disease. Circulation. 2020;141(20):1648-55.

2. Nishiga M, Wang DW, Han Y, Lewis DB, Wu JC. COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives. Nature Reviews Cardiology. 2020;17(9):543-58.

3. Bansal M. Cardiovascular disease and COVID-19. Diabetes & metabolic syndrome. 2020;14(3):247-50.

4. Leong A, Cole JB, Brenner LN, Meigs JB, Florez JC, Mercader JM. Cardiometabolic risk factors for COVID-19 susceptibility and severity: A Mendelian randomization analysis. PLOS Medicine. 2021;18(3):e1003553.

5. Aung N, Khanji MY, Munroe PB, Petersen SE. Causal Inference for Genetic Obesity, Cardiometabolic Profile and COVID-19 Susceptibility: A Mendelian Randomization Study. Frontiers in genetics. 2020;11:586308-.

6. Ponsford Mark J, Gkatzionis A, Walker Venexia M, Grant Andrew J, Wootton Robyn E, Moore Luke SP, et al. Cardiometabolic Traits, Sepsis, and Severe COVID-19. Circulation. 2020;142(18):1791-3.

7. Freuer D, Linseisen J, Meisinger C. Impact of body composition on COVID-19 susceptibility and severity: A two-sample multivariable Mendelian randomization study. Metabolism. 2021;118:154732.

8. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1-22.

9. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23(R1):R89-R98.

10. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7:e34408.

11. The C-HGI. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. European Journal of Human Genetics. 2020;28(6):715-8.

12. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. Genet Epidemiol. 2016;40(7):597-608.