medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Collider bias undermines our understanding of COVID-19 disease risk and severity

Gareth Griffith[1,2,^], Tim T Morris[1,2,^], Matt Tudball[1,2,^], Annie Herbert[1,2,^], Giulia Mancano[1,2,^], Lindsey Pike[1,2], Gemma C Sharp[1,2], Tom M Palmer[1,2], George Davey Smith[1,2], Kate Tilling[1,2], Luisa Zuccolo[1,2], Neil M Davies[1,2,3,], Gibran Hemani[1,2,^*]

**Affiliations:**

[1] Medical Research Council Integrative Epidemiology Unit, University of Bristol, BS8 2BN, United Kingdom.

[2] Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom.

[3] K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Norway.

^ Contributed equally

* Corresponding authors (email: g.hemani@bristol.ac.uk)

**Key Words**: Selection; sample; coronavirus; epidemiology

Word count: 3967



NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Standfirst

Observational data on COVID-19 including hypothesised risk factors for infection and progression are accruing rapidly. Here, we highlight the challenge of interpreting observational evidence from non-random samples of the population, which may be affected by collider bias. We illustrate these issues using data from the UK Biobank in which individuals tested for COVID-19 are highly selected for a wide range of genetic, behavioural, cardiovascular, demographic, and anthropometric traits. We discuss the sampling mechanisms that leave aetiological studies of COVID-19 infection and progression particularly susceptible to collider bias. We also describe several tools and strategies that could help mitigate the effects of collider bias in extant studies of COVID-19 and make available a web app for performing sensitivity analyses. While bias due to non-random sampling should be explored in existing studies, the optimal way to mitigate the problem is to use appropriate sampling strategies at the study design stage.

# Key messages

- Collider bias can occur in studies that non-randomly sample people from the population of interest. This bias can distort associations between variables or induce spurious associations.
- It may be possible to estimate the underlying selection model or run sensitivity analyses to examine the credibility of the threat of collider bias, but it is difficult to prove that bias has been reduced or eliminated.
- Tested samples in the UK Biobank cohort are highly selected for a range of traits.
- Sampling strategies that are resilient to collider bias issues should be used at the design stage of data collection where possible.
- Where this is not possible, linkage or collection of data on the target population can help in sensitivity and validation analyses.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Introduction

Government health organisations, researchers and private companies, amongst others, are generating data on the COVID-19 status of millions of people, along with measures of health and behaviour, for the purpose of using these samples to understand the *risk factors* (see **Box 1** for scope) relevant to the disease in the general population. Numerous studies have reported risk factors associated with COVID-19 infection and subsequent disease severity, such as age, sex, occupation, smoking, and ACE-inhibitor use (1–7). But if we are to make reliable inference about the causes of infection and severity, we need to be aware that there are serious limitations to such observational data. Of particular importance to understanding the aetiology of COVID-19 or developing predictors for infection or severity is the problem of collider bias (sometimes referred to as selection bias, sampling bias, ascertainment bias). Emerging datasets relating to COVID-19 may be particularly susceptible to this issue, having serious implications for the reliability of causal inference and generalisability of predictors.

Collider bias can be counter-intuitive and its implications highly context-specific, but several illustrative examples have been published to aid with understanding this issue (8–10). Consider the situation where we want to test whether a hypothesised risk factor (e.g. tobacco smoke) causes an outcome (e.g. contracting COVID-19). If the hypothesised risk factor and the outcome each influence a third variable, conditioning on that third variable will induce an artifactual association between the risk factor and the outcome even if the risk factor does not cause the outcome(**Figure 1A**). Collider bias can induce associations where there is no true causal effect in the general population, attenuate or inflate true causal effects, or reverse the sign of true causal effects.

In the context of this paper, conditioning upon the third variable can mean examining the effect of the risk factor in only the subset of individuals with that particular characteristic (e.g. only analysing disease cases for disease progression), or non-randomly selecting samples from the target population. An intuitive example of collider bias is as follows. Suppose we want to test the hypothesis that being a health worker is a risk factor for severe COVID-19 symptoms. Our target population for hypothesis is all adults in the general population. However, our study sample will be restricted only to those who are tested for active COVID-19 infection. If we take the UK as an example (until late April 2020), the majority of tests were performed either on health workers, or members of the general public who had symptoms severe enough to require hospitalisation. In this testing environment, our sample of participants will be selected for both the hypothesised risk factor (being a healthcare worker) and the outcome of interest (severe symptoms). In this strata of the population, healthcare workers will generally appear to have relatively low severity (inducing a negative observational association, **Figure 1B**). In reality, there are real occupational hazards of being a healthcare worker, and the true causal effect is likely to be in the opposite direction, as healthcare workers are likely exposed to higher viral loads. In this paper, we discuss why collider bias should be of particular concern to observational studies of COVID-19, and show how sample selection can lead to dramatic biases. We then go on to describe the approaches that are available to explore and mitigate this problem.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Why observational COVID-19 research is particularly susceptible to collider bias

Though unquestionably valuable, observational datasets can be something of a black box because the associations to which they give rise can be due to many different mechanisms. Suppose we wish to draw inferences that can be generalised to a wider population such as the UK (the population). To conduct our observational study, we must first define a group of people that we wish to sample (the target population) who are representative of the target population. The members of the target population who respond to the invitation and participate in the study form the study sample. If individual characteristics cause people to be more likely to respond to an invitation to participate in the study, the study sample will not be representative of the target population. To give context on how serious a problem collider bias can be, there is a continuing debate in the literature about the extent to which it is appropriate to adjust for covariates in observational associations (11–14). If we assume that a given covariate influences both the hypothesised risk factor and the outcome, it is appropriate to condition on that covariate to remove bias induced by the confounding structure. However, if the covariate is a common consequence rather than a common cause, then we risk inducing, rather than reducing bias (15). *A priori* knowledge of what the hidden causal structure truly is can be hard to deduce, and it is appropriate to treat collider bias with a similar level of caution to confounding bias.

There are multiple ways in which data are being collected on COVID-19, and they can introduce unintentional conditioning in the selected sample in various ways. The characteristics of participants recruited are related to a range of factors including policy decisions, cost limitations, technological access, and testing methods. It is also widely acknowledged that the true prevalence of disease in the population remains unknown (16). Here we describe the forms of data collection for COVID-19 and then go on to detail the circumstances surrounding COVID-19 that make its analysis susceptible to collider bias.

## COVID-19 sampling strategies and case definitions

### Sampling conditional on voluntary participation (Case definition: probable COVID-19, Figure 1A)

Probable COVID-19 status can be determined through studies that require voluntary participation. These may include, for example, surveys conducted by existing cohort and longitudinal studies (17,18), data linkage to administrative records is also available in some cohort studies such as the UK Biobank (19), or mobile phone based app programmes (20,21). Participation in scientific studies has been shown to be strongly non-random (e.g. participants are disproportionately likely to be highly educated, health conscious, and non-smokers), so the volunteers in these samples are likely to differ substantially from the general population (22–24). See **Box 2** for a vignette on how one study (21) explored collider bias in this context.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Sampling conditional on being tested for active COVID-19 infection (Case definition: positive test for COVID-19, Figure 1B)**

Polymerase chain reaction (PCR) antigen tests are used to confirm a suspected (currently active) COVID-19 infection. Studies that aim to determine the risk factors for confirmed current COVID-19 infection therefore rely on participants having received a COVID-19 antigen test (hereafter for simplicity: COVID-19 test or test). Unless a random sample or the entire population are tested, these studies do not provide an unbiased estimate of active COVID-19 infection prevalence in the general population. As testing is a resource limited endeavour, different countries have been using different (pragmatic) strategies for prioritising testing, including on the basis of characteristics such as occupation, symptom presentation and perceived risk. See **Box 3** for an investigation into the extent to which testing is non-random with respect to a range of measurable potential risk factors, using the recently released COVID-19 test data in the UK-Biobank.

**Sampling conditional on having a positive test for active COVID-19 infection (Case definition: COVID-19 severity, Figure 1B)**

Studies that aim to determine the risk factors for severity of confirmed current COVID-19 infection therefore rely on participants having received a COVID-19 antigen test (hereafter for simplicity: COVID-19 test or test), and that the result of the test was positive. As above, testing is unlikely to be random, and conditioning on the positive result will also mean bias can be induced by all factors causing infection, as well as those causing increased likelihood of testing.

**Prognosis and mortality sampling conditional on hospitalisation (Case definition: COVID-19 death, Figure 1C)**

Many studies have started analysing the influences on disease progression once individuals are infected, or infected and then admitted to hospital (i.e. the factors that influence survival). Such datasets necessarily condition upon a positive test. **Figure 1C** illustrates how this so-called 'index event bias' is a special case of collider bias (25–27). If we accept that COVID-19 increases mortality, and there are risk factors for infection of COVID-19, then in a representative sample of the target population, any cause of infection would also exert a causal influence on mortality, mediated by infection. However, once we condition on being infected, all factors for infection become correlated with each other. If some of those factors influence both infection and progression then the association between a factor for infection and death in the selected sample will be biased. This could lead to factors that increase risk of infection falsely appearing to be protective for severe progression (1,28). An example of this relevant to COVID-19 is discussed in **Box 2**. How different directions of selective sampling influence the direction of bias is discussed in **Figures 1 and 2.**

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Sample selection pressures for COVID-19 testing

While some of the factors that impact the sampling processes may be common across all modes of sampling listed above, some will be mode-specific. In general, these factors will differ across national and healthcare system contexts. Here we list a series of possible selection pressures acting upon COVID-19 testing and case identification/definition and detail how they may bias inference if left unexplored.

**Symptom severity**

With few notable exceptions (e.g. (3)), population testing for COVID-19 is not generally performed in random samples. Several countries adopted the strategy of offering tests predominantly to patients experiencing symptoms severe enough to require medical attention, e.g. hospitalisation, as is the case in the UK until the end of April 2020. Many true positive cases in the population will therefore remain undetected and be subject to negative sample selection if enrollment is dependent upon test status. High rates of asymptomatic virus carriers or cases with atypical presentation will further compound this issue.

**Symptom recognition**

Related to but distinct from symptom severity, inclusion in COVID-19 datasets will vary based upon symptom recognition (29). If an individual fails to recognise the correct symptoms or deems their symptoms to be nonsevere, they are less likely to seek medical attention and therefore be tested for COVID-19. People will also assess their symptom severity differently; those with health related anxiety may be more likely to over-report symptoms, while those with less awareness or access to health advice may be under-represented. This problem may be compounded by changing symptom guidelines which could induce systematic relationships between symptom presentation and testing (29,30).

**Occupation**

In many countries, frontline healthcare workers are far more likely to be tested for COVID-19 than the general population (5,31) due to their proximity to the virus and the potential consequences of infection related transmission (32). As such, they will be heavily over-represented in samples conditional on test status. Other key workers may be at high risk of infection due to large numbers of contacts relative to non-key workers, and may therefore be over-represented in samples conditional on test status or cause of death. Any factors related to these occupations (e.g. ethnicity, socio-economic position, age and baseline health) will therefore also be associated with sample selection. **Figure 1B** illustrates an example where the hypothesised risk factor does not need to influence sample selection causally, it could simply be associated due to a confounding between the risk factor and sample selection.

**Place of residence and social connectedness**

A number of more distal or indirect influences on sample selection likely exist. People with better access to healthcare services may be more likely to be tested than those with poorer access. Those in areas with a greater number of medical services or better public transport may find it easier to access services for testing, while those in areas with lower local medical service

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

utilisation may be more likely to be tested as a function of service capacity (33). People living in areas with stronger spatial or social ties to existing outbreaks may also be more likely to be tested due to increased medical vigilance in those areas. Family and community support networks are also likely to influence access to medical care, for instance, those with caring responsibilities and weak support networks may be less able to seek medical attention (34).

**Frailty**

Some groups of the population, such as elderly in care homes, are treated differently in terms of reporting on COVID-19 in different countries (35). For example in the UK early reports of deaths "due to COVID-19" may have been conflated with deaths "while infected with COVID-19" (36). Individuals at high risk are more likely to be tested in general, but specific demographics at high risk such as those in care homes have been liable to under-representation. A challenge that arises with trying to evaluate the problem of collider bias is that it may be difficult to ascertain if particular groups with COVID-19 are being over or under represented in the selected sample, making sensitivity analysis difficult.

## Sample selection pressures for voluntary self-reporting

Sample selection pressures for voluntary self-reporting COVID-19 efforts are likely distinct from those for COVID-19 testing.

**Internet access and Technological Engagement**

Sample recruitment via internet applications has been shown to under-represent certain groups (23,37). Furthermore, voluntary "pull-in" data collection methods have been shown to produce more engaged but less representative samples than "push out" advertisement methods (24). These groups likely have greater access to electronic methods of data collection, and greater engagement in social media campaigns that are designed to recruit participants. As such, younger people are more likely to be over-represented in app based voluntary participation studies (20).

**Medical and scientific interest**

Voluntary participation studies are likely to contain a disproportionate amount of people who have a strong medical or scientific interest. It is likely that these people will themselves have greater health awareness, healthier behaviour, be more educated, and have higher incomes (22,38).

**Ethnicity**

Some groups may experience barriers into voluntary participation of scientific studies due to many factors, for example language, cultural norms or access to information.

Many of the factors for being tested or being included in datasets described here are borne out in the analysis of the UK Biobank test data (Box 2).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Methods for overcoming collider bias

In this section we describe methods to either overcome bias or evaluate how sensitive any associations could be to collider bias. The primary task in any analysis is to evaluate the extent to which sample selection is likely to have actually occurred. This can be done by comparing means and prevalences in the selected sample against those obtained from external data that represents the target population. Ideally, this would be done for the hypothesised risk factor and outcome, as well as any related variables. If there are even subtle departures in the characteristics of the study sample from the general population then this provides evidence of selective sampling. With respect to analysis of COVID-19 disease risk, one major obstacle to this endeavour is that in most cases the actual prevalence of infection in the general population is unknown, making it impossible to prove an absence of selection through validation.

If a study is at risk of selective sampling, the unfortunate truth is that it is very difficult to prove that any method has resolved issues with collider bias. Sensitivity analyses are therefore crucial in exploring factors that could be related to selection, and examining robustness of conclusions to plausible selection mechanisms.

Several methods exist that do attempt to adjust for collider bias or examine how sensitive the study is to collider bias. The likelihood and extent of collider bias induced by sample selection can be evaluated by comparing distributions of variables in the sample with those in the target population (or a representative sample of the target population). This provides information about the profile of individuals selected into the sample from the target population of interest, such as whether they tend to be older or more likely to have comorbidities. It is particularly valuable to report these comparisons for key variables in the analysis, such as the hypothesised risk factor and outcome, and other variables related to these.

The applicability of different methods depends on the data that are available on non-participants. These methods can broadly be split into two categories: a) where the selected sample is nested within a larger dataset that comprises samples believed to be representative of the target population, or b) where the entire dataset comprises only the selected samples used for hypothesis testing (stand-alone).

## Nested sample

In the case that we have a selected sample with COVID-19 measures, which is a subset of a sample that is representative of the target population, one approach is to use inverse probability weighting (39,40). Here the causal effect of risk factor on outcome is examined using a weighted regression, where the participants who are overrepresented are down-weighted and the participants who are underrepresented are up-weighted. In practice, we estimate the probability of different individuals selecting into the sample from the population-representative sample based on their measured covariates, using a statistical model (the “sample selection model”), and use this to create a weight for each participant (41). An example is where the study sample is those with a positive covid test, nested within the UK Biobank study. If we

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

assume that UK Biobank is representative of the target population (the general population of the UK), then we can use data from UK Biobank to estimate the probability of having a covid test for each individual in UK Biobank. We can then appropriately re-weight the study sample to represent the population of UK Biobank.

Seaman and White (2013) provide a detailed overview of the practical considerations and assumptions for inverse probability weighting, such as correct specification of the sample selection model, variable selection and approaches for handling unstable weights (i.e. weights which are zero or near-zero). An additional assumption for inverse probability weighting is that each individual in the target population must have a non-zero probability of being selected into the sample. Neither this assumption, nor the assumption that the selection model has been correctly specified, are testable using the observed data. A conceptually related approach, using propensity score matching, is sometimes used to avoid index event bias (42,43).

## Stand-alone samples

When we only have data on the study sample (e.g. only data on participants who were tested for COVID-19) it is not possible to estimate the selection model directly since non-selected (untested) individuals are unobserved. Instead, it is important to apply sensitivity analyses to assess the plausibility that sample selection induces collider bias.

**Bounds and parameter searches**

It is possible to infer the extent of collider bias given knowledge of the likely size and direction of influences of risk factor and outcome on sample selection (whether these are direct, or via other factors) (12,44,45). However, this approach depends on the size and direction being correct, and there being no other factors influencing selection. It is therefore important to explore different possible sample selection mechanisms and examine their impact on study conclusions. We created a simple web application guided by these assumptions to allow researchers to explore simple patterns of selection that would be required to induce an observational association: http://apps.mrcieu.ac.uk/ascrtain/. In **Figure 3** we use a recent report of a protective association of smoking on COVID-19 infection (46) to explore the magnitude of collider bias that can be induced due to selected sampling, under the null hypothesis of no causal effect.

Several other approaches have also been implemented into convenient online web apps (Appendix). For example Smith and VanderWeele (2019) proposed a sensitivity analysis which allows researchers to bound their estimates by specifying sensitivity parameters representing the strength of sample selection (in terms of relative risk ratios). They also provide an ‘E-value’, which is the smallest magnitude of these parameters that would explain away an observed association (47). Aronow and Lee (2013) proposed a sensitivity analysis for sample averages based on inverse probability weighting when the weights cannot be estimated but are assumed to be bounded between two researcher-specified values (48). This work has been generalised to allow regression models to incorporate relevant external information (e.g. summary statistics from the census) (49). Zhao et al (2017) developed a sensitivity analysis for the degree to which

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

estimated probability weights differ from the true probability weights due to misspecification (50). This approach is particularly useful when we can estimate probability weights including some, but not necessarily all, of the relevant predictors of sample selection. For example, where we have the study sample being those with a covid test, nested within the UK Biobank, we may have data within UK Biobank on some predictors of testing, but there could be other factors that are not recorded within UK Biobank (e.g. general predisposition for seeking healthcare).

These sensitivity analysis approaches allow researchers to explore whether there are credible collider structures that could explain away observational associations. However, they do not represent an exhaustive set of models that could give rise to bias, nor do they necessarily prove that collider bias influences the results. If the risk factor for selection is itself the result of further upstream causes then it is important that the impact of these upstream selection effects are considered (i.e. not only how the risk factor influences selection but also how the causes of the risk factor and/or the causes of the outcome influence selection e.g. **Figure 1B**). While these upstream causes may individually have a small effect on selection, it is possible that lots of factors with individually small effects could jointly have a large selection effect and introduce collider bias (Groenwold et al. 2016).

**Negative control analyses**

If there are factors measured in the selected sample that are known to have no influence on the outcome, then testing these factors for association with the outcome within the selected sample can serve as a negative control (51,52). By virtue, negative control associations should be null and they are therefore useful as a tool to provide evidence in support of selection. If we observe associations with larger magnitudes than expected then this indicates that the sample is selected on both the negative control and the outcome of interest (53,54).

**Correlation analyses**

Conceptually similar to the negative controls approach above, when a sample is selected, all the features that influenced selection become correlated within the sample (except for the highly unlikely case that causes are perfectly multiplicative). Testing for correlations amongst hypothesised risk factors where it is expected that there should be no relationship can indicate the presence and magnitude of sampling bias, and therefore the likelihood of collider bias distorting the primary analysis (55).

# Implications

The majority of scientific evidence informing policy and clinical decision making during the COVID-19 pandemic has come from observational studies (56). We have illustrated how these observational studies are particularly susceptible to non-random sampling. Randomised clinical trials will provide experimental evidence for treatment, but experimental studies of infection will not be possible for ethical reasons. The impact of collider bias on inferences from observational studies could be considerable, not only for disease transmission modelling (57,58), but also for causal inference (7) and prediction modelling (2).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

While many approaches exist that attempt to ameliorate the problem of collider bias, they rely on untestable assumptions. It is difficult to know the extent of sample selection, and even if that were known it cannot be proven that it has been fully accounted for by any method. Representative population surveys or sampling strategies that avoid the problems of collider bias (59) are urgently required to provide reliable evidence. Results from samples that are likely not representative of the target population should be treated with caution by scientists and policy makers.

## Appendix

Exploring bounds and spaces that could explain an observational association can easily be achieved using a range of packages and apps:

- AscRtain app: http://apps.mrcieu.ac.uk/ascrtain/
- CollideR app (10): https://watzilei.com/shiny/collider/
- Selection bias app (47): https://selection-bias.herokuapp.com/
- Bias app (45): https://remlapmot.shinyapps.io/bias-app/
- Lavaan R package (60): http://lavaan.ugent.be/
- Dagitty R package (61): http://www.dagitty.net/
- simMixedDAG: https://github.com/IyarLin/simMixedDAG

## Acknowledgements

We are grateful to Josephine Walker for helpful comments on this manuscript. This research has been conducted using the UK Biobank Resource under Application Number 16729. The Medical Research Council (MRC) and the University of Bristol support the MRC Integrative Epidemiology Unit [MC_UU_12013/1, MC_UU_12013/9, MC_UU_00011/1]. NMD is supported by a Norwegian Research Council Grant number 295989. GH is supported by the Wellcome Trust and Royal Society [208806/Z/17/Z].

## Author contributions

G.H., N.M.D., L.Z. conceived the idea
G.H. performed the analysis
G.G., G.H., and T.P. wrote the software
All authors discussed the results and contributed to the final manuscript

## Competing interests

None

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Box 1: Collider bias in the context of prediction and aetiological studies**

An aetiological study seeks to identify causes of the outcome of interest ("causal factors"), whereas a predictive study aims to predict the outcome from a range of variables ("predictors") which need not be causal. The term "risk factor" has been used synonymously for both causal factors and predictors in the literature (62,63).

Risk factors measured in observational studies, may associate with outcomes of interest (e.g. hospitalised with COVID-19), for many reasons. For example, the factor may affect the outcome (true causal interpretation), statistical evidence of association may be purely due to chance, the outcome may affect the factor (reverse causation), there may be a third factor that causes both the exposure and the outcome (confounding), or the exposure and outcome (or causes of the exposure and/or outcome) may influence likelihood of being selected into the study (collider bias).

Aetiological studies are in principle only concerned with the causal effect, and aim to avoid all forms of bias. By contrast, some forms of bias such as confounding or reverse causation can actually improve the performance of a prediction study. As long as the causal structure by which the study sample is drawn from the target population is the same as in the population in which predictions will be made, it can be of benefit to leverage these distinct association mechanisms to improve prediction accuracy (64,65).

Under certain circumstances collider bias can improve prediction performance if the training sample and the sample to be predicted have the same patterns of sample selection. For example, if the factors causing having a test for COVID-19 are the same/similar across the UK, a predictive model for the result being positive that was developed in London will perform well in the North East if those samples are both selected in the same way. However, collider bias is a problem for the generalisability of both causal inference and prediction in the target population when the training sample is selected, because it induces artifactual associations that are idiosyncratic to that dataset. If the intention is to predict COVID-19 status, rather than COVID-19 status conditional on being tested, the prediction will underperform.

While the term 'risk factor' can be ambiguous and refer to either a hypothesised causal determinant or a predictor of the disease, we intentionally use it throughout this paper for the sake of brevity as causal inference and prediction analyses both share a vulnerability to the detrimental impacts of collider bias in the COVID-19 context - where typically the selected samples are being used to develop models relevant to the general population.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Box 2. The potential association between ACE inhibitors and COVID-19: why sampling bias matters**

One research question that has gained attention is whether blood pressure lowering drugs such as ACE inhibitors (ACE-i) and angiotensin-receptor blockers (ARBs), which act on the Renin–Angiotensin–Aldosterone System (RAAS) system, make patients more susceptible to COVID-19 infection (66–70).

Relationships between ACE-i/ARBs and COVID-19 are to be investigated in clinical trials (71,72), but in the meantime have been rapidly investigated through observational studies (73–75). One such recent analysis used data from a UK COVID-19 symptom tracker app (76), which was released in March just before the UK Lockdown policy was implemented to increase social distancing. The app allows members of the public to contribute to research through self-reporting data including demographics, conditions, medications, symptoms and COVID-19 test results. The researchers observed that people reporting ACE-i use were twice as likely to report COVID-19 symptoms, even after adjusting for differences in age, BMI, sex, diabetes, and heart disease (21).

The researchers investigated whether sampling bias may play a role. If taking ACE-i and having COVID-19 symptoms would lead to being either less or more likely to sign up to the app or contribute data, this could induce an association between these factors (Figure 1A). Since ACE-is are prescribed to those with diabetes, heart disease, or hypertension, ACE-i users are likely to be considered high-risk for COVID-19 (77). They are therefore potentially more sensitised to their current health status and may be more likely to use the app (78,79). People who are COVID-19 symptomatic may also be more likely to remember to contribute data than asymptomatic people. Taken together, this could result in a false or inflated association between taking ACE-i and COVID-19. However, in reality, deciding in which direction ACE-i and COVID-19 symptoms would influence participation is complicated. For example, people with severe COVID-19 symptoms who are hospitalised could be too ill to contribute data.

Careful consideration is required for each set of exposures and outcomes that are studied. Amongst those participants who were actually tested in the COVID-19 symptom tracker app study, there was no evidence for an association between ACE-i use and COVID-19 positive status (21). In this analysis there are joint selection pressures of a) factors underlying being tested and b) factors underlying app participation.

Should ACE-i use truly increase risk of COVID-19 infection, it could imply that observational results for disease progression studies are influenced by collider bias. For example, it has been reported that ACE-i/ARB use may be protective against severe symptoms, conditional on already being infected (80,81), which is consistent with index event bias as illustrated in **Figure 1C**.

It is important to consider the plausibility of the different selection pathways, both statistically (for example, through methods such as bounds and parameter searches) and biologically. Such considerations will ensure that data interpretation is at least robust to known biases of unknown

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

magnitude, and policy decisions are based on the best interpretation of the scientific evidence. Indeed, in consideration of the benefits that ACE-i/ARBs have on the cardio-respiratory system, current guidelines should continue to recommend use of these drugs until there is sufficiently reliable scientific evidence against this (82,83).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Box 3. Factors influencing being tested in UK Biobank**

In April 2020, General Practices across the UK released primary care data on COVID-19 testing for linkage to the participants in the UK Biobank project (84) and analyses are already appearing (85). Of the 486,967 participants, 1,410 currently have data on COVID-19 testing. While it may be tempting to look for factors that influence whether an individual tests positive, it is crucial to evaluate the potential that those tested are not a random sample of the UK-Biobank participants (who are themselves not a random sample of the UK population).

We examined 2,556 different characteristics for association with whether or not a UK Biobank participant had been tested for COVID-19. There was very large enrichment for associations (Figure 2), with 811 of the phenotypes (32%) giving rise to a false discovery rate < 0.05. These associations involved a wide range of traits, including measures of frailty, medications used, genetic principal components, air pollution, socio-economic status, hypertension and other cardiovascular traits, anthropometric measures, psychological measures, behavioural traits, and nutritional measures. A full list of all traits assessed and their associations with whether a participant had COVID-19 test data are available in Supplementary Table 1. The first genetic principal component, which relates to major ethnic groups, was one of the strongest associations with being tested, which may have implications for understanding the association of race on testing positive for COVID-19 (85).

We can not know the actual COVID-19 prevalence amongst all participants, but if it is different from the prevalence amongst those tested, then every one of the traits listed above could be associated with COVID-19 in the dataset solely due to collider bias, or at least the magnitude of those associations could be biased as a result. The fact that the UK Biobank data are already a non-random sample of the UK population further complicates the matter (26).

Ideally, inverse-probability weighted regressions would be performed to minimise any such bias. However, because we can not know the COVID-19 status of participants outside the tested group (sampling fractions), such weights will be impossible to calculate without strong assumptions that are currently untestable (50). Inverse-probability-weighting also depends on the selection model being correctly specified, including that all characteristics predicting selection (that are related to variables in the analysis model) have been included, and in the right functional form. As with unmeasured confounding, there is always the possibility of having unmeasured selection factors.

**Methods:** UK-Biobank phenotypes were processed using the PHESANT pipeline (86) and filtered to include only quantitative traits or case-control traits that had at least 10,000 cases. In addition, sex, genotype chip and the first 40 genetic principal components were included for analysis (2,556 traits in total). A 'tested' variable was generated that indicated whether an individual had been tested for COVID-19 or not within UK Biobank, and logistic regression was performed for each of the 2,556 traits against the 'tested' variable. Code: https://github.com/explodecomputer/covid_ascertainment

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Figures

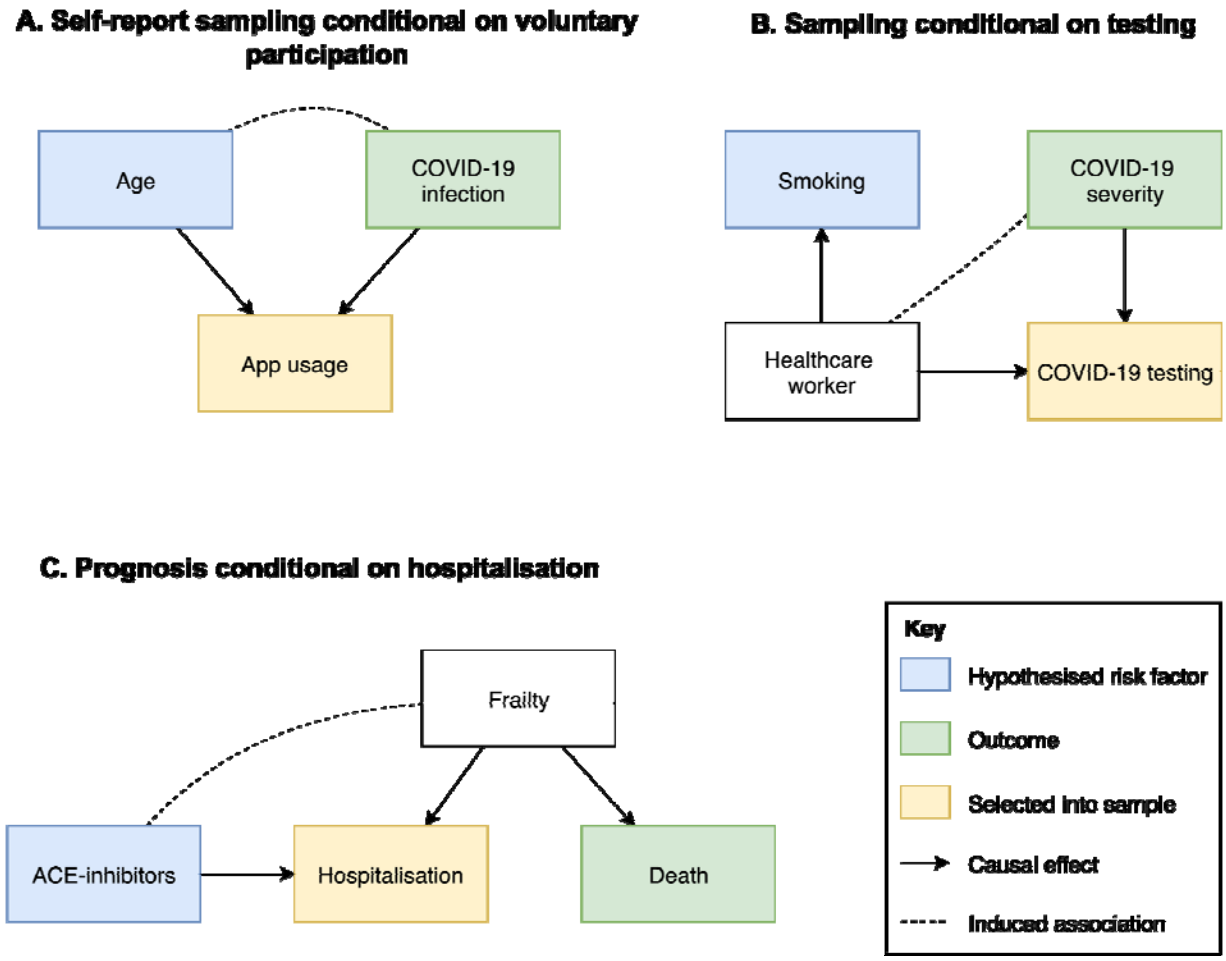


**Figure 1: Collider bias induced by conditioning on a collider in three scenarios relating to COVID-19 analysis.** These are simplified Directed Acyclic Diagrams where only the main variables of interest have been represented for sake of illustrating collider bias scenarios. All assume no unspecified confounding or other biases. Rectangles represent observed variables and solid directed arrows represent causal effects. The dashed line represents an induced association when conditioning on the collider, which in these scenarios are variables that indicate whether an individual is selected into the sample. (A) When some hypothesised risk factor (e.g. age) and outcome (e.g. COVID-19 infection) each associate with sample selection (e.g. voluntary data collection via mobile-phone apps), the hypothesised risk factor and outcome will be associated within the sample. The presence and direction of these biases are model dependent; where causes are supra-multiplicative they will be positively associated in the sample; where they are sub-multiplicative they will be negatively correlated; and where they are exactly multiplicative they will remain unassociated. We extend this scenario in (B) where the association between the hypothesised risk factor and the collider does not need to be causal. (C) When inferring the influence of some hypothesised risk factor on mortality, in an unselected sample the risk factor for infection is a causal factor for death (mediated by COVID-19 infection).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

However, if analysed only amongst individuals who are known to have COVID-19 (i.e. we condition on the COVID-19 infection variable) then the risk factor for infection will appear to be associated with any other variable that influences both infection and progression. In many circumstances this can lead to a risk factor for disease onset that appears to be protective for disease progression. Each of these scenarios represent those described in the main text.

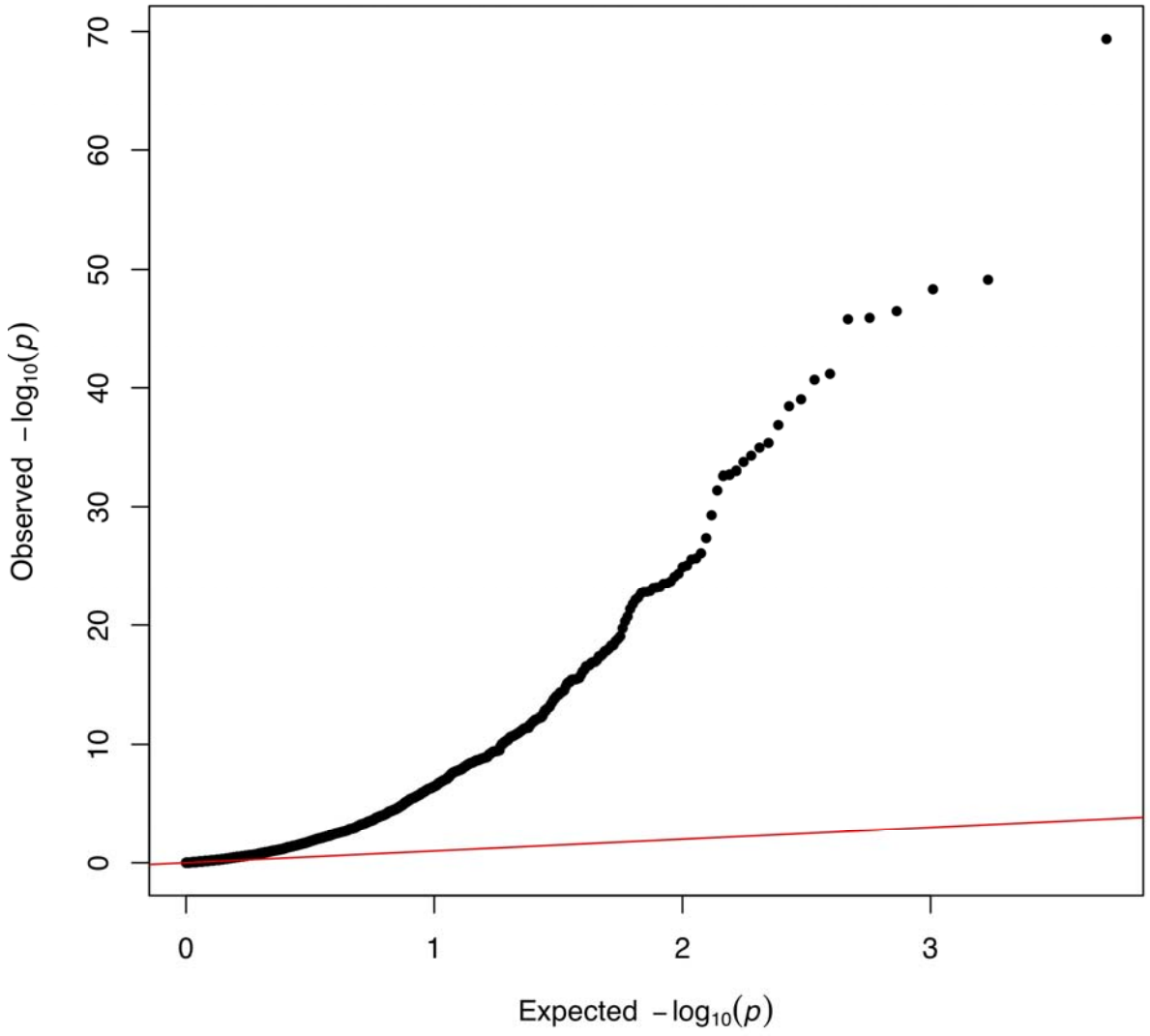


**Figure 2: Quantile-Quantile plot of -log10 p-values for factors influencing being tested for COVID-19 in UK Biobank.** The x-axis represents the expected p-value for 2,556 hypothesis tests and y-axis represents the observed p-values. The red line represents the expected relationship under the null hypothesis of no associations.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

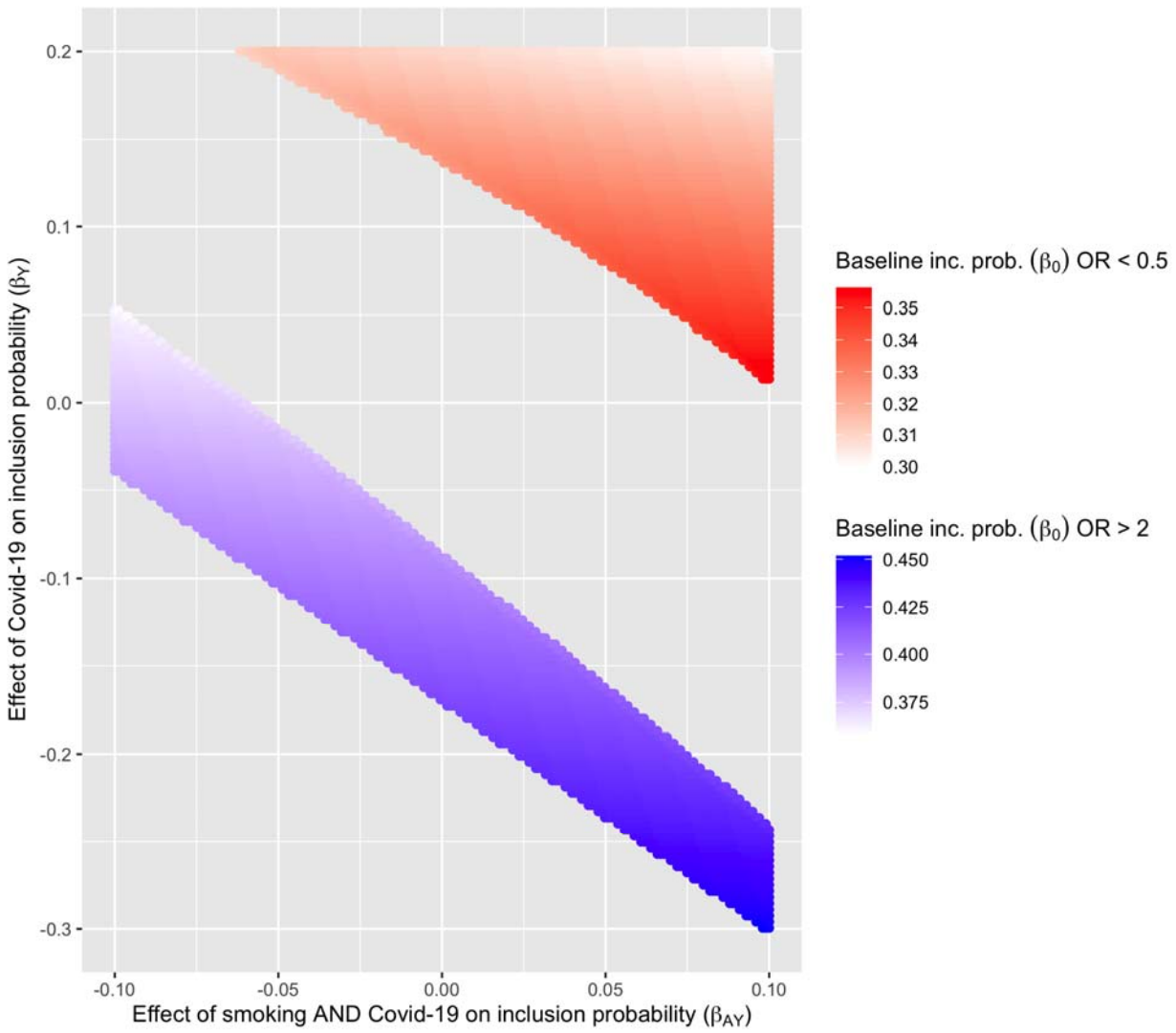


**Figure 3: Large associations can be induced by collider bias under the null hypothesis of no causal relationship, using scenarios similar to those reported for the observed protective association of smoking on COVID-19 infection.** Assume a scenario in which the hypothesised exposure (A) and outcome (Y) are both binary and each influence probability of being selected into the sample (S) e.g. where is the baseline probability of being selected, is the effect of A, is the effect of Y and is the effect of the interaction between A and Y. This plot shows which combinations of these parameters would be required to induce an apparent risk effect with magnitude OR > 2 (blue region) or an apparent protective effect with magnitude OR < 0.5 (red region) under the null hypothesis of no causal effect (45). To create a simplified scenario similar to that in Miyara et al 2020 we use a general population prevalence of smoking of 0.27 and a sample prevalence of 0.05, thus fixing at 0.22. Because the prevalence of COVID-19 is not known in the general population, we allow the sample to be over or under representative (y-axis). We also allow modest interaction effects. Calculating over this parameter space, 40% of all possible combinations lead to an artifactual 2-fold protective or risk association operating through this

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

simple model of bias alone. It is important to disclose this level of uncertainty when publishing observational estimates.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Supplementary tables

**Supplementary Table 1**: Association results for each of 2,556 variables in the UK Biobank cohort, testing for their influence on being tested for COVID-19

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# References


1. Zhang P, Zhu L, Cai J, Lei F, Qin J-J, Xie J, et al. Association of Inpatient Use of Angiotensin Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers with Mortality Among Patients With Hypertension Hospitalized With COVID-19. Circ Res [Internet]. 2020 Apr 17; Available from: http://dx.doi.org/10.1161/CIRCRESAHA.120.317134

2. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020 Apr 7;369:m1328.

3. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med [Internet]. 2020 Apr 14; Available from: http://dx.doi.org/10.1056/NEJMoa2006100

4. Chen T, Wu D, Chen H, Yan W, Yang D, Chen G, et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. BMJ. 2020 Mar 26;368:m1091.

5. Tostmann A, Bradley J, Bousema T, Yiek W-K, Holwerda M, Bleeker-Rovers C, et al. Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. Eurosurveillance. 2020 Apr 23;25(16):2000508.

6. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Med [Internet]. 2020 Mar 3; Available from: http://dx.doi.org/10.1007/s00134-020-05991-x

7. Gilmore A. Review of: “Low incidence of daily active tobacco smoking in patients with symptomatic COVID-19.” Qeios [Internet]. 2020 Apr 27; Available from: https://www.qeios.com/read/37F3UD

8. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010 Apr;39(2):417–20.

9. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu Rev Sociol. 2014 Jul;40:31–53.

10. Luque-Fernandez MA, Schomaker M, Redondo-Sanchez D, Jose Sanchez Perez M, Vaidya A, Schnitzer ME. Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application. Int J Epidemiol. 2019 Apr 1;48(2):640–53.

11. Ding P, Miratrix LW. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. Journal of Causal Inference. 2015 Mar 1;3(1):41–57.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

12. Nguyen TQ, Dafoe A, Ogburn EL. The magnitude and direction of collider bias for binary variables [Internet]. arXiv [stat.ME]. 2016. Available from: http://arxiv.org/abs/1609.00606

13. Pearl J. Myth, Confusion, and Science in Causal Analysis. 2009 May 1 [cited 2020 Apr 23]; Available from: https://escholarship.org/uc/item/6cs342k2

14. Shrier I. Letter to the Editor [Internet]. Vol. 27, Statistics in Medicine. 2008. p. 2740–1. Available from: http://dx.doi.org/10.1002/sim.3172

15. Rohrer JM. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. Advances in Methods and Practices in Psychological Science. 2018 Mar 1;1(1):27–42.

16. Lourenco J, Paton R, Ghafari M, Kraemer M, Thompson C, Simmonds P, et al. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.03.24.20042291v1

17. University of Bristol. 2020: COVID 19 Questionnaire PR | Avon Longitudinal Study of Parents and Children | University of Bristol [Internet]. University of Bristol. 2020 [cited 2020 Apr 23]. Available from: http://www.bris.ac.uk/alspac/news/2020/coronavirus.html

18. New Covid-19 survey from Understanding Society | Understanding Society [Internet]. [cited 2020 Apr 23]. Available from: https://www.understandingsociety.ac.uk/2020/04/23/new-covid-19-survey-from-understanding-society

19. UK BIOBANK MAKES INFECTION AND HEALTH DATA AVAILABLE TO TACKLE COVID-19 | UK Biobank [Internet]. [cited 2020 Apr 23]. Available from: https://www.ukbiobank.ac.uk/2020/04/covid/

20. Menni C, Valdes A, Freydin MB, Ganesh S, El-Sayed Moustafa J, Visconti A, et al. Loss of smell and taste in combination with other symptoms is a strong predictor of COVID-19 infection [Internet]. Epidemiology. medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.04.05.20048421v1

21. Dooley H, Lee K, Freidin M, Hemani G, Roberts A, Ni Lochlainn M, et al. ACE inhibitors, ARBs and other anti-hypertensive drugs and novel COVID-19: An association study from the COVID Symptom tracker app in 2,215,386 individuals [Internet]. 2020 [cited 2020 Apr 24]. Available from: https://papers.ssrn.com/abstract=3583469

22. Taylor AE, Jones HJ, Sallis H, Euesden J, Stergiakouli E, Davies NM, et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2018 Aug 1;47(4):1207–16.

23. Blom AG, Herzing JME, Cornesse C, Sakshaug JW, Krieger U, Bossert D. Does the

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. Soc Sci Comput Rev. 2017 Aug 1;35(4):498–520.

24. Antoun C, Zhang C, Conrad FG, Schober MF. Comparisons of Online Recruitment Strategies for Convenience Samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. Field methods. 2016 Aug 1;28(3):231–46.

25. Paternoster L, Tilling K, Davey Smith G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. PLoS Genet. 2017 Oct;13(10):e1006944.

26. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol. 2018 Feb 1;47(1):226–35.

27. Yaghootkar H, Bancks MP, Jones SE, McDaid A, Beaumont R, Donnelly L, et al. Quantifying the extent to which index event biases influence large genetic association studies. Hum Mol Genet. 2017 Mar 1;26(5):1018–30.

28. Changeux J-P, Amoura Z, Rey F, Miyara M. A nicotinic hypothesis for Covid-19 with preventive and therapeutic implications. Qeios [Internet]. 2020 Apr 22; Available from: https://www.qeios.com/read/article/581

29. Boëlle P-Y, Souty C, Launay T, Guerrisi C, Turbelin C, Behillil S, et al. Excess cases of influenza-like illnesses synchronous with coronavirus disease (COVID-19) epidemic, France, March 2020. Euro Surveill [Internet]. 2020 Apr;25(14). Available from: http://dx.doi.org/10.2807/1560-7917.ES.2020.25.14.2000326

30. Tsang TK, Wu P, Lin Y, Lau EHY, Leung GM, Cowling BJ. Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. The Lancet Public Health [Internet]. 2020 Apr; Available from: https://linkinghub.elsevier.com/retrieve/pii/S246826672030089X

31. BBC News. Health workers on frontline to be tested. BBC [Internet]. 2020 Mar 27 [cited 2020 Apr 23]; Available from: https://www.bbc.com/news/health-52070199

32. Department of Health, Care S. Coronavirus (COVID-19): scaling up our testing programmes [Internet]. GOV.UK. GOV.UK; 2020 [cited 2020 May 1]. Available from: https://www.gov.uk/government/publications/coronavirus-covid-19-scaling-up-testing-programmes/coronavirus-covid-19-scaling-up-our-testing-programmes

33. Department of Health and Social Care. Coronavirus (COVID-19): getting tested [Internet]. GOV.UK. GOV.UK; 2020 [cited 2020 Apr 29]. Available from: https://www.gov.uk/guidance/coronavirus-covid-19-getting-tested

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

34. Kuchler T, Russel D, Stroebel J. The Geographic Spread of COVID-19 Correlates with Structure of Social Networks as Measured by Facebook [Internet]. National Bureau of Economic Research; 2020. (Working Paper Series). Available from: http://www.nber.org/papers/w26990

35. Care home deaths: the untold and largely unrecorded tragedy of COVID-19 [Internet]. British Politics and Policy at LSE. 2020 [cited 2020 Apr 23]. Available from: https://blogs.lse.ac.uk/politicsandpolicy/care-home-deaths-covid19/

36. Campbell DA, Caul S. Deaths involving COVID-19, England and Wales - Office for National Statistics [Internet]. Office for National Statistics. 2020 [cited 2020 May 2]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinmarch2020

37. Revilla M, Cornilleau A, Cousteaux A-S, Legleye S, de Pedraza P. What Is the Gain in a Probability-Based Online Panel of Providing Internet Access to Sampling Units Who Previously Had No Access? Soc Sci Comput Rev. 2016 Aug 1;34(4):479–96.

38. Tyrrell J, Zheng J, Beaumont R, Hinton K, Richardson TG, Wood AR, et al. Genetic predictors of participation in optional components of UK Biobank [Internet]. bioRxiv. 2020 [cited 2020 Apr 29]. p. 2020.02.10.941328. Available from: https://www.biorxiv.org/content/10.1101/2020.02.10.941328v1

39. Mansournia MA, Altman DG. Inverse probability weighting. BMJ. 2016 Jan 15;352:i189.

40. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. BMJ. 2019 Oct 23;367:l5657.

41. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013 Jun;22(3):278–95.

42. Adamopoulos C, Meyer P, Desai RV, Karatzidou K, Ovalle F, White M, et al. Absence of obesity paradox in patients with chronic heart failure and diabetes mellitus: a propensity-matched study. Eur J Heart Fail. 2011;13(2):200–6.

43. Stensrud MJ, Valberg M, Røysland K, Aalen OO. Exploring Selection Bias by Causal Frailty Models: The Magnitude Matters. Epidemiology. 2017 May;28(3):379–86.

44. Pearl J. Linear Models: A Useful "Microscope" for Causal Analysis. Journal of Causal Inference. 2013;1(1):155–70.

45. Groenwold RHH, Palmer TM, Tilling K. Conditioning on a mediator. 2019 Dec 23 [cited 2020 Apr 24]; Available from: https://osf.io/vrcuf/

46. Miyara M, Tubach F, Pourcher V, Morelot-Panzini C, Pernet J, Haroche J, et al. Low

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

incidence of daily active tobacco smoking in patients with symptomatic COVID-19. Qeios [Internet]. 2020 Apr 21; Available from: https://www.qeios.com/read/article/574

47. Smith LH, VanderWeele TJ. Bounding Bias Due to Selection. Epidemiology. 2019 Jul;30(4):509–16.

48. Aronow PM, Lee DKK. Interval estimation of population means under unknown but bounded probabilities of sample selection. Biometrika. 2013 Mar 1;100(1):235–40.

49. Tudball M, Zhao Q, Hughes R, Tilling K, Bowden J. An Interval Estimation Approach to Sample Selection Bias [Internet]. arXiv [stat.ME]. 2019. Available from: http://arxiv.org/abs/1906.10159

50. Zhao Q, Small DS, Bhattacharya BB. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap [Internet]. arXiv [stat.ME]. 2017. Available from: http://arxiv.org/abs/1711.11286

51. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010 May;21(3):383–8.

52. Davey Smith G. Negative control exposures in epidemiologic studies. Epidemiology. 2012 Mar;23(2):350–1; author reply 351–2.

53. Arnold BF, Ercumen A, Benjamin-Chung J, Colford JM Jr. Brief Report: Negative Controls to Detect Selection Bias and Measurement Bias in Epidemiologic Studies. Epidemiology. 2016 Sep;27(5):637–41.

54. Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. Int J Epidemiol. 2006 Apr;35(2):337–44.

55. Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, et al. Genetic analyses identify widespread sex-differential participation bias [Internet]. bioRxiv. 2020 [cited 2020 May 2]. p. 2020.03.22.001453. Available from: https://www.biorxiv.org/content/biorxiv/early/2020/03/23/2020.03.22.001453

56. Moghadas SM, Shoukat A, Fitzpatrick MC, Wells CR, Sah P, Pandey A, et al. Projecting hospital utilization during the COVID-19 outbreaks in the United States. Proc Natl Acad Sci U S A. 2020 Apr 21;117(16):9122–6.

57. Zhao Q, Ju N, Bacallado S. BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic [Internet]. arXiv [stat.AP]. 2020. Available from: http://arxiv.org/abs/2004.07743

58. Pearce N, Vandenbroucke JP, VanderWeele TJ, Greenland S. Accurate Statistics on COVID-19 Are Essential for Policy Guidance and Decisions. Am J Public Health. 2020 Apr 23;e1–3.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

59. Vandenbroucke JP, Brickley EB, Christina M J, Pearce N. Analysis proposals for test-negative design and matched case-control studies during widespread testing of symptomatic persons for SARS-Cov-2 [Internet]. arXiv [q-bio.PE]. 2020. Available from: http://arxiv.org/abs/2004.06033

60. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, Articles. 2012;48(2):1–36.

61. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package "dagitty." Int J Epidemiol. 2016 Dec 1;45(6):1887–94.

62. Shader RI. Risk Factors Versus Causes. J Clin Psychopharmacol. 2019;39(4):293–4.

63. Shmueli G. To Explain or to Predict? Stat Sci. 2010 Aug;25(3):289–310.

64. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011 Dec 1;174(11):1213–22.

65. Pearl J. Invited commentary: understanding bias amplification. Am J Epidemiol. 2011 Dec 1;174(11):1223–7; discussion pg 1228–9.

66. Brown JD. Antihypertensive drugs and risk of COVID-19? Lancet Respir Med [Internet]. 2020 Mar 26; Available from: http://dx.doi.org/10.1016/S2213-2600(20)30158-2

67. Aronson JK, Ferner RE. Drugs and the renin-angiotensin system in covid-19. BMJ. 2020 Apr 2;369:m1313.

68. Kuster GM, Pfister O, Burkard T, Zhou Q, Twerenbold R, Haaf P, et al. SARS-CoV2: should inhibitors of the renin-angiotensin system be withdrawn in patients with COVID-19? Eur Heart J [Internet]. 2020 Mar 20; Available from: http://dx.doi.org/10.1093/eurheartj/ehaa235

69. Nelson DJ. Blood-pressure drugs are in the crosshairs of COVID-19 research. Reuters [Internet]. 2020 Apr 23 [cited 2020 Apr 24]; Available from: https://www.reuters.com/article/us-health-conoravirus-blood-pressure-ins-idUSKCN2251GQ

70. By Sam Blanchard Senior Health Reporter For Mailonline. High blood pressure medicines "could worsen coronavirus symptoms" [Internet]. Mail Online. Daily Mail; 2020 [cited 2020 Apr 24]. Available from: https://www.dailymail.co.uk/news/article-8108735/Medicines-high-blood-pressure-diabetes-worsen-coronavirus-symptoms.html

71. Coronavirus (COVID-19) ACEi/ARB Investigation - Full Text View - ClinicalTrials.gov [Internet]. [cited 2020 Apr 24]. Available from:

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

https://clinicaltrials.gov/ct2/show/NCT04330300?term=ace+inhibitors&cond=COVID&draw=1&rank=6

72. Prognosis of Coronavirus Disease 2019 (COVID-19) Patients Receiving Receiving Antihypertensives [Internet]. [cited 2020 Apr 24]. Available from: https://clinicaltrials.gov/ct2/show/NCT04357535?term=ace+inhibitors&cond=COVID&draw=2&rank=4

73. OHDSI. COVID-19 Updates Page [Internet]. [cited 2020 Apr 24]. Available from: https://ohdsi.org/covid-19-updates/

74. Assistance Publique-Hôpitaux de Paris. Long-term Use of Drugs That Could Prevent the Risk of Serious COVID-19 Infections or Make it Worse [Internet]. [cited 2020 Apr 24]. Available from: https://clinicaltrials.gov/ct2/show/NCT04356417?term=ace+inhibitors&cond=COVID&draw=2&rank=10

75. Payne R. Using linked primary care and viral surveilance data to develop risk stratification models to inform management of severe COVID19 [Internet]. NIHR; 2020 [cited 2020 Apr 24]. Report No.: 494. Available from: https://www.spcr.nihr.ac.uk/projects/Linked-primary-care-viral-surveillance-data-risk-stratification

76. COVID Symptom Tracker [Internet]. [cited 2020 Apr 24]. Available from: https://covid.joinzoe.com

77. Website NHS. Who's at higher risk from coronavirus - Coronavirus (COVID-19) [Internet]. nhs.uk. [cited 2020 Apr 24]. Available from: https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk-from-coronavirus/whos-at-higher-risk-from-coronavirus/

78. Kripalani S, Heerman WJ, Patel NJ, Jackson N, Goggins K, Rothman RL, et al. Association of Health Literacy and Numeracy with Interest in Research Participation. J Gen Intern Med. 2019 Apr;34(4):544–51.

79. Firmino RT, Fraiz FC, Montes GR, Paiva SM, Granville-Garcia AF, Ferreira FM. Impact of oral health literacy on self-reported missing data in epidemiological research. Community Dent Oral Epidemiol. 2018 Dec;46(6):624–30.

80. Meng J, Xiao G, Zhang J, He X, Ou M, Bi J, et al. Renin-angiotensin system inhibitors improve the clinical outcomes of COVID-19 patients with hypertension. Emerg Microbes Infect. 2020 Dec;9(1):757–60.

81. Bean D, Kraljevic Z, Searle T, Bendayan R, Pickles A, Folarin A, et al. Treatment with ACE-inhibitors is associated with less severe disease with SARS-Covid-19 infection in a multi-site UK acute Hospital Trust [Internet]. Infectious Diseases (except HIV/AIDS). medRxiv; 2020. Available from: https://www.medrxiv.org/content/10.1101/2020.04.07.20056788v1

medRxiv preprint doi: https://doi.org/10.1101/2020.05.04.20090506; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

82. Medicines and Healthcare products Regulatory Agency. Coronavirus (COVID-19) and high blood pressure medication [Internet]. GOV.UK. GOV.UK; 2020 [cited 2020 Apr 24]. Available from: https://www.gov.uk/government/news/coronavirus-covid-19-and-high-blood-pressure-medication?fbclid=IwAR1PlWny7gpN0YSF-Z9yDfrsa-HF-CG7b_bad8Mf09SkLudhe8Vrh7jL4Ws

83. International Society of Hypertension. A statement from the International Society of Hypertension on COVID-19 | The International Society of Hypertension [Internet]. [cited 2020 Apr 24]. Available from: https://ish-world.com/news/a/A-statement-from-the-International-Society-of-Hypertension-on-COVID-19/

84. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

85. Patel AP, Paranjpe MD, Kathiresan NP, Rivas MA, Khera AV. Race, Socioeconomic Deprivation, and Hospitalization for COVID-19 in English participants of a National Biobank. Epidemiology. medRxiv; 2020.

86. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol [Internet]. 2017 Oct 5; Available from: http://dx.doi.org/10.1093/ije/dyx204