Framework to aid analysis and interpretation of ongoing COVID-19 research [version 1; peer review: 1 approved with reservations]

Thomas Yates, Francesco Zaccardi, Cameron Razieh, Clare L. Gillies, Alex Rowlands, David E. Kloecker, Yogini V. Chudasama, Melanie J. Davies, Kamlesh Khunti

Abstract
The global coronavirus pandemic has precipitated a rapid unprecedented research response, including investigations into risk factors for COVID-19 infection, severity, or death. However, results from this research have produced heterogeneous findings, including articles published in Wellcome Open Research. Here, we use ethnicity, obesity, and smoking as illustrative examples to demonstrate how a research question can produce very different answers depending on how it is framed. For example, these factors can be both strongly associated or have a null association with death due to COVID-19, even when using the same dataset and statistical modelling. Highlighting the reasons underpinning this apparent paradox provides an important framework for reporting and interpreting ongoing COVID-19 research.

Keywords
coronavirus disease-2019, ethnicity, epidemiology, obesity, severe acute respiratory syndrome coronavirus 2, smoking
Corresponding author: Thomas Yates (ty20@le.ac.uk)

Author roles: Yates T: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Visualization, Writing – Original Draft Preparation; Zaccardi F: Conceptualization, Formal Analysis, Methodology, Visualization, Writing – Review & Editing; Razieh C: Project Administration, Writing – Review & Editing; Gillies CL: Methodology, Visualization, Writing – Review & Editing; Rowlands A: Writing – Review & Editing; Kloecker DE: Writing – Review & Editing; Chudasama YV: Writing – Review & Editing; Davies MJ: Writing – Review & Editing; Khunti K: Conceptualization, Writing – Review & Editing

Competing interests: Prof. Kamlesh Khunti is a member of the independent SAGE group

Grant information: Prof Kamlesh Khunti is an applicant on the current Wellcome Trust Institutional Strategic Support Fund (WTISSF) awarded to the University of Leicester and has received direct funding from WTISSF for the Centre for BME Health, which in turn supported this research. This research was funded by a grant from the UKRI COVID-19 Rapid Response Call (MR/V020536/1) and supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre and the NIHR Applied Research Collaboration – East Midlands. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Yates T et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Yates T, Zaccardi F, Razieh C et al. Framework to aid analysis and interpretation of ongoing COVID-19 research [version 1; peer review: 1 approved with reservations] Wellcome Open Research 2020, 5:208 https://doi.org/10.12688/wellcomeopenres.16227.1

First published: 04 Sep 2020, 5:208 https://doi.org/10.12688/wellcomeopenres.16227.1
Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes coronavirus disease-2019 (COVID-19), is devastating global economies and placing an unprecedented strain on health care services and communities. The mortality and economic burden caused by the virus have expedited unprecedented research activity, including investigating factors that identify those at greatest risk.

Over recent months, obesity and ethnicity have emerged as widely characterised risk factors, generating extensive media interest and policy debate. However, it is also becoming increasingly clear that the strength of associations between the investigated risk factors and COVID-19 outcomes vary substantially from one study to the next, particularly for mortality outcomes, with some studies showing no or weak associations. Other markers of global health status, such as smoking, even appear to be protective in some studies.

This heterogeneity is demonstrated by two recent articles published in Wellcome Open Research. An analysis of in-hospital data suggested that there was no ethnic difference in the risk of COVID-19 mortality in those with South Asian ethnicity compared to white Europeans, whereas another study found South Asian and other minority ethnic groups had a higher risk of COVID-19 mortality on a population level.

We argue this heterogeneity is strongly influenced by how data are selected and analysed.

To date, research into risk factors for COVID-19 has broadly used and addressed one of two approaches or questions:

The Tested Population Approach (TPA): ‘In those that have been tested for SARS-CoV-2, what are the risk factors for testing positive, and in those that have tested positive, what are the risk factors for developing subsequent complications, including death?’

The Whole Population Approach (WPA): ‘What are the risk factors for SARS-CoV-2 positivity and resulting complications in the overall general population?’

The aim of this article is to discuss how the inherent differences in these questions can produce apparently paradoxical findings for the same outcome. We hope that highlighting this concept will inform how future research is presented and interpreted. Whilst others have provided more technical insights into some of the key concepts discussed here, we believe it is important to illustrate empirically the impact on risk factor associations for a wider audience.

Methods

Demonstrator dataset

In order to demonstrate our argument, we incorporate an illustrative example from UK Biobank, a cohort of over 500,000 adults. UK Biobank has been linked to national COVID-19 testing (via Public Health England’s Second Generation Surveillance System) and ONS mortality data in a well-phenotyped cohort allowing for novel risk factor research. As of July 14th 2020, 918 articles searchable on Google Scholar included the terms (UK BIOBANK) AND (COVID-19 OR CORONAVIRUS OR SARS-COV-2). Furthermore, few other available datasets allow for an analysis of both the TPA and WPA within the same cohort. Therefore, UK Biobank provides an instructive example of the key issues related to COVID-19 risk factor research.

Our aim is not to undertake primary research, as previous studies have reported associations of smoking, ethnicity, and obesity with selected COVID-19 outcomes, rather, to illustrate how the same dataset can produce apparently paradoxical findings for the same outcome.

We used data that were linked to COVID-19 test results for the period 16th March 2020 to 20th June 2020; mortality data were available up to 26th April 2020.

There were 7645 test results available within the UK Biobank cohort, of which 1545 (20.2%) were SARS-CoV-2 positive. There were 250 deaths with COVID-19 listed as the primary cause of death, of which 187 had a positive test available within the dataset.

The tested population approach (TPA)

Using the UK Biobank participants who have been tested for SARS-COV-2, logistic regression adjusted for age and sex were used to estimate the risk of the test being positive and the risk of mortality due to COVID-19 (defined as ICD-10 code U07.1 or U07.2 as the primary cause of death on the death certificate) in those with a positive test. Positive test results were only considered up to 26th April 2020 for the mortality analysis.

Data for ethnicity, smoking and obesity status were included on a complete case basis.

The whole population approach (WPA)

Data were included from all participants within UK Biobank that were eligible for data linkage. This encompassed 420,619 individuals that were recruited from English centres and were alive on 16th March 2020. Logistic regression, adjusted for age and sex, were used to quantify the risk of having confirmed SARS-COV-2 infection and the risk of death due to COVID-19 within this wider cohort. Data for ethnicity, smoking and obesity status were included on a complete case basis.

Results and discussion

Figure 1 shows the selection and cohort characteristics for each outcome addressed.

Figure 2 shows the age and sex adjusted odds of having a test across demonstrator risk factors of obesity, ethnicity and smoking. The odds of being tested were between 1.5 to 1.6 times greater for those that were obese (compared to normal weight), smokers (compared to non-smokers), or those of non-white ethnicity (compared to white ethnicity).
For the TPA approach, Figure 2 shows that in those that were tested, the odds of a positive test were moderately higher for obesity and non-white ethnicity but lower for smokers. No risk factor was associated with an increased risk of mortality from COVID-19 in those that tested positive.

Using the WPA approach, all risk factors were associated with the odds of testing positive and COVID-19 mortality (Figure 2), with odds between 2 to 3 times greater in those with obesity or those with non-white ethnicity and between 1.3 to 1.8 times greater in smokers compared to non-smokers.

Resolving the apparent paradox: the impact of non-random testing

When looking at one outcome in isolation, for example mortality, these findings appear to provide a paradoxical interpretation while in fact there need be no paradox when looking at the overall steps required to gain these results. For example, in the TPA, by definition groups that are more likely to be tested, and more likely to test positive, are also more likely to have an overall risk of mortality on a whole population level, even when an increased risk is not present in those that test positive. Conversely for the WPA, it is possible that smokers, non-white ethnicities or those with obesity have a greater exposure to SARS-CoV-2 at a population level, making the absolute risk of subsequent complications higher, even if the risk is similar once exposed.

This apparent paradox is essentially related to the testing bias within UK Biobank, with those from non-white ethnicities, obesity, or smokers being more likely to be tested. As testing in the UK during the linkage period for UK Biobank has predominately been collected through Pillar 1 (testing for those with clinical need and health care workers), this data should be interpreted as reflecting this testing reality. Whilst population biases in testing procedures are inevitable in anything other than universal or random testing policies, they produce some important challenges for research. One widely discussed challenge is collider bias, of which smoking is a good example. For example, we showed using the TPA (Figure 2) that, whilst smokers were more likely to be tested, they were less likely to test positive in those that were tested. This could potentially be explained by collider bias, as smoking may directly cause symptoms that overlap with those of COVID-19 (e.g. sore throat, cough) which increases the risk of being tested but reduces the risk of the test being positive using the TPA.

Testing bias also ensures findings using the WPA are a function of the TPA. Figure 3 illustrates this concept by modelling how the odds of a positive test for COVID-19 in non-white ethnicities compared to white Europeans on a population level is directly related to the odds of being tested, even when assuming no difference in the test being positive once tested between the two ethnic groups. That is, where there is ethnic bias in the odds of being tested, a null ethnic association
Figure 2. Empirical example of risk factor associations from UK Biobank.
for a positive test addressed through the TPA will *ipso facto* result in an association in the WPA, which becomes more pronounced the greater the testing bias. In the instance where non-white ethnicities are more likely to be tested, they will appear to have a higher odds of testing positive when analysed using the WPA. This phenomenon is unaffected by the proportion of ethnic minorities within a population.

**Implications for research**

The studies that have previously reported protective associations for smoking with COVID-19 outcomes and no association for obesity or ethnicity 6–9,11 have largely asked a form of the TPA, supporting our illustrative findings from UK Biobank. Even where positive associations have been reported, they have tended to be modest. For example, in a large study of 20,133 in-hospital patients with COVID-19, those with obesity were 1.3 times more likely to die19, which in the context of the wider literature is a relatively modest association.

Conversely, widely reported studies and surveys showing strong associations of obesity or ethnicity with COVID-19 outcomes, including national surveillance data that report the number of deaths per 100,000 of the population or the OpenSAFELY study10,20, have largely addressed a form of the WPA. This includes a study reporting a risk of death from COVID-19 up to 10 times higher in younger obese adults compared to normal weight individuals18.

Given this obvious difference in interpretation, we recommend that data generated by asking forms of the TPA and the WPA are not viewed as equivalent or comparable. This also has implications for systematic reviews and meta-analyses which should aim to stratify their results by these two types of studies. We further recommend that all original studies make it clear which type of research question they are addressing (TPA or WPA) and should make every attempt to identify and account for testing bias and model or otherwise describe the likely impact of this on their results.

**Conclusions**

As very few countries have universal SARS-CoV-2 screening or antibody testing policies, all research will be biased by the underlying testing and exposure dynamics within the population investigated. This has the potential for confusing rather than enlightening the general public and health care professionals alike.

---

**Figure 3.** Modelled example of the relationship between the probability of being tested and associations using the whole population approach. For the purposes of this example, the proportion of testing positive for SARS-CoV-2 in those who have had a test is assumed to be the same in both ethnic groups, ranging from 5% (top-left panel) to 90% (bottom-right). In this case, the odds ratio of being positive for SARS-CoV-2 in the entire (tested + non-tested) population, indicated by the z-axis, is determined by the probability of being tested in each ethnic group (x and y axis) and the probability of testing positive in those tested (each panel). The x-axis indicates the proportion of White people who have been tested for SARS-CoV-2 (from 0.1% to 0.2%) while the y-axis the proportion of non-White people who have been tested (from 0.1% to 0.5%).
Here we provide a framework for interpreting COVID-19 risk factors research which can help explain why research looking at the same risk factors can produce different results, including articles published in Wellcome Open Research, which helped prompt this Research Note.\(^1\)\(^{11,12}\) Greater insight is needed into whether existing analytical frameworks for addressing selection bias can be adapted or modified for COVID-19 research, such as through methods promoted by the AscRtain tool. In the meantime, health care professionals, researchers and policy makers alike should all be aware that the type of approach to the researched population (TPA vs WPA) is likely to affect the answer provided, even when investigating the same outcome.

**Contributors and sources**

TY, FZ & KK conceived the main argument put forth, which was expanded upon by MJD, CLG, CR, DK, and YC. TY is the guarantor.

**Data availability**

**Underlying data**

We use data from UK Biobank application number 36371 to support the case being made. All bona fide researchers can apply to use UK Biobank for health-related research that is in the public interest. Further information on the application process is available from UK Biobank [https://www.ukbiobank.ac.uk/register-apply/](https://www.ukbiobank.ac.uk/register-apply/).

**References**


Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 24 September 2020

https://doi.org/10.21956/wellcomeopenres.17823.r40298

© 2020 Bailey D. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daniel Bailey
Sedentary Behaviour, Health and Disease Research Group, Division of Sport, Health and Exercise Sciences, Brunel University London, London, UK

This is a well presented paper that is relevant for the analysis of factors related to COVID-19 infection and mortality rates. It makes a worthwhile contribution to the field. I only have relatively minor points below that would add greater clarity to the methodology and interpretation.

- Specify what ONS stands for.
- Please provide information on how ethnicity, smoking and obesity data is collected and presented in U.K. Biobank in the methods section.
- Please provide some description around the national targeted testing policies that is referred to in Figure 1.
- I find this sentence quite difficult to follow “For example, in the TPA, by definition groups that are more likely to be tested, and more likely to test positive, are also more likely to have an overall risk of mortality on a whole population level, even when an increased risk is not present in those that test positive”. Please rephrase, possibly using an additional sentence, and provide a commentary that links to Figure 2 so it is clear how this interpretation has been deduced.
- I also argue that power in the analyses would differ between the TPA and WPA approaches. This should be considered.
- Please clarify if Figure 3 has been generated from analyses of U.K. biobank data or whether it has been theoretically modelled.
- In the conclusion, state that the AscRtain tool considers collider a bias and COVID-19.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My expertise relevant to this paper are cross section analysis of risk factors and health-related outcomes

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.