ISARIC (International Severe Acute Respiratory and Emerging Infections Consortium)

A global federation of clinical research networks, providing a proficient, coordinated, and agile research response to outbreak-prone infectious disease

Analysis Plan for ISARIC International COVID-19 Patients

Please complete the following sections:

Title of proposed research
The relative influence of clinical and sociodemographic variables on the long-term COVID-19-associated QALYs lost
Version: (Date: Day/Month/Year)
12/05/22
Working Group Chair (name, ORCID ID, email, institution, country)
Tigist Menkir, 0000-0001-6070-8017, tmenkir@hsph.harvard.edu, Harvard University and Oxford University, US and UK
[1]Working group co-chair (name, ORCID ID, email, institution, country)
Louise Sigfrid, MD, PhD, FFPH, Clinical Research Fellow, Public Health specialist, ISARIC, Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK 0000-0003-2764-1177 Christl Donnelly, PhD, 0000-0002-0195-2463 , christl.donnelly@stats.ox.ac.uk, Oxford University and Imperial College London, UK Dr Waasila Jassat, South Africa [to complete] Luis Felipe Reyes, MD. PhD; Universidad de La Sabana, Colombia.

[1] Either chair and/or co-chair are based in an institution in an LMIC. If you would like to be connected with an eligible co-chair please let us know at ncov@isaric.org.

Arne Søraas, MD, PhD, 0000-0003-1622-591X, arne@meg.no, Oslo University Hospital, Norway Anders B. Nygaard, PhD, 0000-0003-1922-0751, anders.b.nygaard@gmail.com, Oslo University Hospital, Norway
Statistician (name, ORCID ID, email, institution, country)
Tigist Menkir

Final draft SAPs will be circulated to all ISARIC partners for their input with an invitation to participate. ISARIC can help to set up collaborator meetings; form a working group; support communications; and accessing data. Please note that the details of all approved applications will be made publicly available on the ISARIC website. Please complete all sections of this form fully and return to ncov@isaric.org

Introduction

This document details the initial analysis plan for publication on a subset of COVID-19 patients in the global cohort in the ISARIC database, as of 30 April 2021. There are currently 64 countries (as of 10 February 2021) contributing data and these have so far contributed data on 305,241 patients. This data will represent the global experience of ~12 months of this pandemic.

A subset of sites in five countries have followed up patients to characterise long term COVID-19 outcomes in a subset of patients. The ISARIC global adult follow up protocol and associated data collection form was activated in August 2020. The follow up protocol was developed for implementation in any resourced settings, and wide dissemination via self-assessment (online, paper form) or clinical led assessment (in-clinic, telephone) to facilitate wide dissemination during the pandemic circumstances. The follow up data will be linked with the acute data for the analysis.

Recognising that people who were not hospitalised during the acute phase were also impacted by prolonged recovery, an aligned form was developed to enable inclusion of both hospitalised and non-hospitalised patients. The follow up data collection forms were developed by the ISARIC global follow up working group and clinicians from different specialties to capture core key symptoms and complications relating to different organ systems. Validated tools (EQ5D5L, UN/Washington disability score, MRC Dyspnea Scale) were incorporated to facilitate comparison between different settings and studies. In addition data on demographic variables, pre-existing risk factors, and impact on occupational activities, and socioeconomic variables.

Much of the literature thus far has focused on identifying a myriad of clinical risk factors for long-term COVID-19 consequences, including co-infections and pre-existing conditions, a number of lab measures, severe acute infection, vaccination, as well as demographic factors such as age and sex [include UKHSA short summary report]. Clinical factors that have been consistently determined to be key correlates of long-term sequelae include obesity, asthma, Type 2 diabetes, HIV/AIDS, liver disease, reactivated Epstein-Barr infection, many of which have been testified to increase the risk of disease severity.[1–7]

While there have been efforts to examine social factors potentially linked to the manifestation of long-term symptoms, findings on these relationships, particularly with respect to race/ethnicity, have been generally mixed.[5,8] It is important to highlight, however, that some of these studies may be affected by poor recruitment of socially disadvantaged populations, due to poor linkage to post-acute care systems, among other factors.[9–11] For instance, while a meta-analysis by Thompson et al.[6] suggests that the odds of post-acute COVID symptoms is non-significantly lower among ethnic minorities, a majority of the studies considered in this summary measure included low proportions of minority populations (median=4.1%)[6] and thus were potentially ill-suited to detect any meaningful differences. Additionally, many studies rely on self-reported measures of COVID recovery and experiences with continued symptomology[6,8,12], some of which may be highly subjective classifications, with potential between-group differences in the proclivity to report such experiences. In contrast, more well-defined measures such as ability to and time taken to return to work have shown a clear signal with ethnic minority status[12,13], suggesting that the extent of the burden of long-term sequelae may be underestimated in these groups.

Given this context, we aim to complement co-existing efforts[14] centered on uncovering disparities in long-term COVID outcomes, leveraging a large dataset collected in three diverse country settings: the United Kingdom, South Africa, and Norway. Specifically, we aim to explicitly disentangle the relative influence of social disparities on reducing the quality of life following COVID infection. Our work strives to address two of the aforementioned challenges by 1) including a representative proportion of racial/ethnic minorities and individuals in more deprived communities 2) focusing on a well-established outcome measure that, while self-reported, incorporates information on a range of physical and mental health consequences. To do so, we will first conduct a general exploration of diverse groups of risk factors – representing both underlying biological and social vulnerabilities – to identify those which may be most predictive of reductions in QALYs associated with long-term symptoms. We will subsequently focus on the relationship between the social factors socio-economic status and race/ethnicity with QALY losses and evaluate the extent to which clinical and behavioral risk factors contribute to any observed disparities.

Participatory Approach

All contributors to the ISARIC database are invited to participate in this analysis through review and input on the statistical analysis plan and resulting publication. The outputs of this work will be disseminated as widely as possible to inform patient care and public health policy, this will include submission for publication in an international, peer-reviewed journal. ISARIC aims to include the names of all those who contribute data in the cited authorship of this publication, subject to the submission of contact details and confirmation of acceptance of the final manuscript within the required timelines, per ICMJE policies and the ISARIC publication policy.

Research Plan

Summary of Research Objectives

The first aim of this study is to estimate Quality-adjusted life years (QALYs) lost associated with long-term COVID-19 symptoms as a function of a collection of clinical and social characteristics, and identify which factors contribute most significantly to QALYs lost.

The second aim of this study is to determine the extent to which age (group), sex, and comorbidities may explain observed 'effects' of socio-demographic factors* on QALYs lost attributed to long-term COVID-19 symptoms and by extension, imply other key non-clinical drivers

*Note: for the purposes of distinction, here we refer to variables like SES and race/ethnicity as socio-demographic variables and variables like age and sex as demographic variables

Proposed Target Population

Inclusion criteria

- People aged 18 years and older.
- From the United Kingdom, South Africa, Norway, India, and Colombia
- Laboratory-confirmed SARS-CoV-2 infection or physician-confirmed COVID-19
- At least 1-month post-onset of COVID-19, or post-SARS-CoV-2 positive test.
- Person (or family member/carer for patients who lack capacity) consent to participate

Clinical Questions/Descriptive Analyses

1. What are the main predictors (among *categories* of medical and socio-demographic factors) of reductions in quality of life attributed to long-term COVID-19 symptoms?

2. What proportion of the effect of socio-demographic factors on QALYs lost due to long-term COVID consequences is mediated by co-morbidities and risk behaviors like smoking?

*hereafter referred to as long COVID QALYs lost

Planned Statistical Analyses, Methodology and Representation

AIM #1

Each of the subsequent analyses will be run independently for each country, due to cohort differences in patient selection and data quality, as well as the underlying relationships between our variables of interest and long COVID QALYs lost.

Long COVID QALYs lost will be calculated using estimated EQ-5D values from the Tier 1 follow-up surveys (using responses of experiences at follow-up and recalled experiences prior to infection). We will then multiply our period of interest (one year) by estimates of the change in QALYs to arrive at long COVID QALYs one year following infection

(1) Clinical variables, demographic variables (age and sex), and socio-demographic variables socio-economic status (SES), as proxied by zip code, educational attainment, and/or employment status, and ethnic group, will be grouped depending on subject matter knowledge. Clinical variables include asthma, chronic cardiac disease, hypertension, chronic kidney disease, chronic neurological disorder, malignant neoplasm, obesity, chronic liver disease, chronic pulmonary disease, HIV/AIDS, diabetes, tuberculosis, smoking, and vaccination status, which have been found to be associated with long COVID.
(2) Alternatively, we may decide to have variables grouped algorithmically, as described below. Variables will be assigned to groups in order to facilitate the identification of ‘sets’ of risks factors (representing common potential underlying mechanisms, and are thus likely highly correlated) for long COVID QALYs lost.

If we select grouping procedure (1), we will then implement a sparse-group lasso regression as described in Simon et al.[16] in which groups of socio-demographic, demographic, and the aforementioned clinical variables, as well as each of the variables included within each group, are either kept or discarded in the model based on their relative influence on overall and within-group model predictions of long COVID QALYs lost, fit directly to our estimates of long COVID QALYs lost.[16] Consequently, we may be able to evaluate the relative role of categories of demographic, socio-demographic, and medical factors, and variables within these categories, in predicting the emergence of long-term COVID-19 symptoms. All groups selected in the model may be considered meaningful drivers of the manifestation of long COVID QALYs lost.

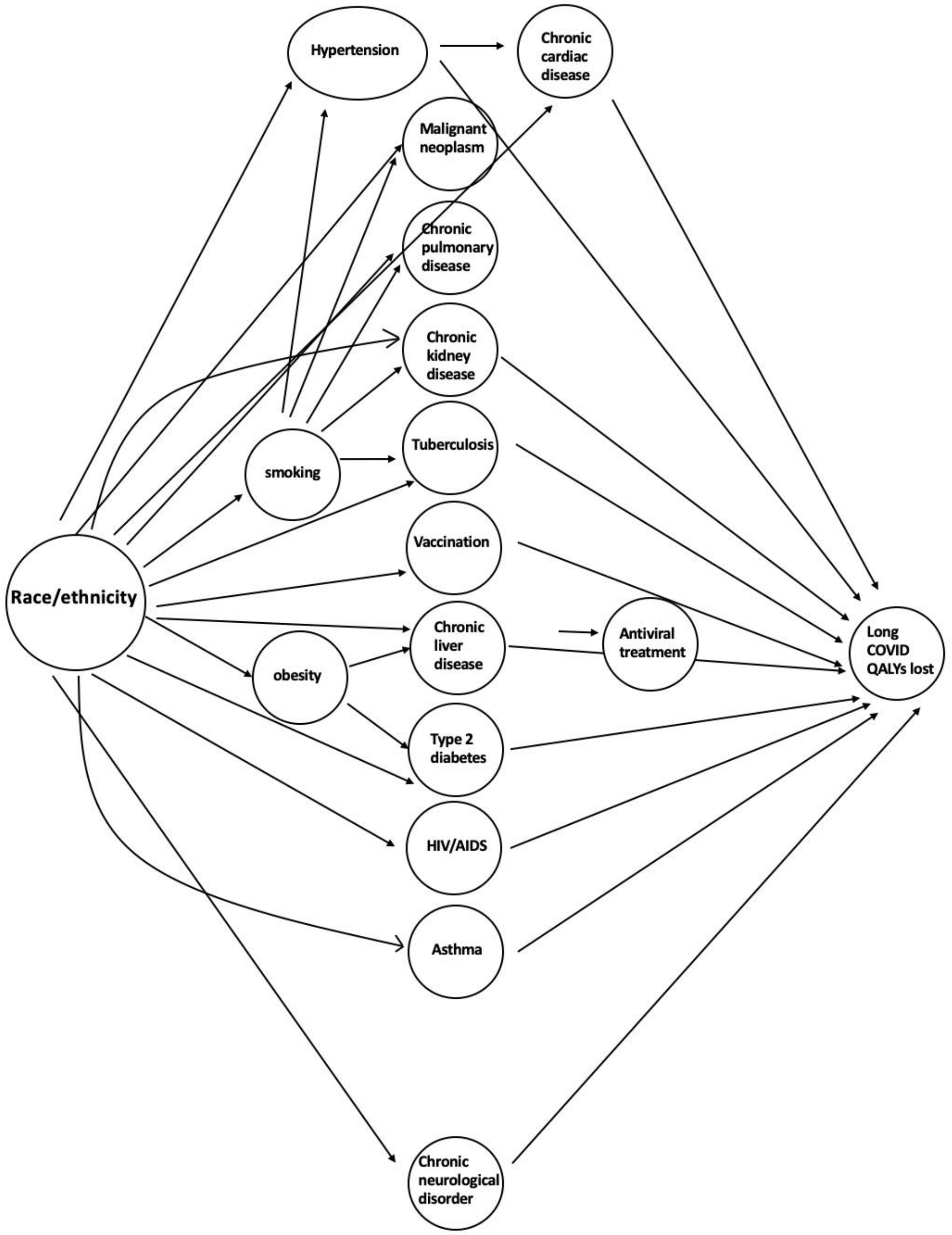
Alternatively, if we select grouping procedure (1), we will assign a “synthetic variable”[17,18] to each group summarizing the information contained within them, which are subsequently used as predictors for the random forest regression model.

If we select grouping procedure (2), we will then implement a Clustering of Variables-Variable Selection Procedure (CoV-VSURF).[17,18] in which variables are first assigned to clusters using a hierarchical clustering algorithm, and then summarized via synthetic variables. As before, we fit our random forest regression to our estimates of long COVID QALYs lost as a function of the derived synthetic variables. Importantly, we can identify the selection of “informative”[17] clusters which contribute most significantly to predicting long COVID QALYs lost, with which we may then use to make our final predictions on the test dataset.

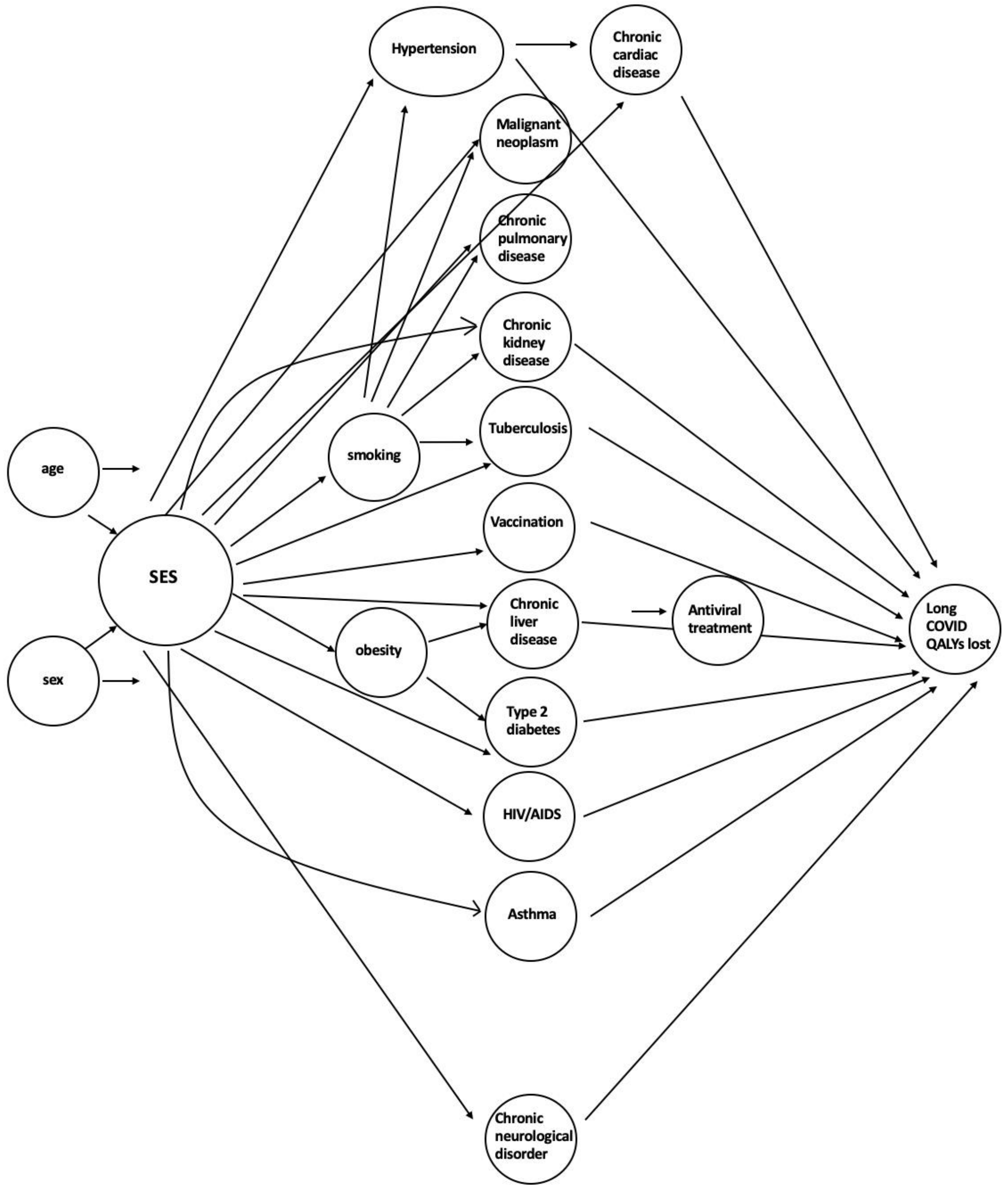
AIM #2

For each socio-demographic variable of interest, SES and ethnic group, we will estimate both the natural indirect effect (NIE) and controlled direct effect (CDE) of that variable on long COVID QALYs lost, through a pre-defined collection of comorbidities and behavioral risk factors. The natural indirect effect here describes the proportion of the total effect of the socio-demographic variable on the presence of long COVID QALYs lost that can be linked to the aforementioned comorbidities and risk behaviors considered.[19] The controlled direct effect here describes the proportion of the effect of the socio-demographic variable on long COVID QALYs lost that would persist after fixing values of each of these mediators to a certain level.[19]

To estimate the NIEs and CDEs, we will implement a weighting-based method, as described in VanderWeele and Vansteelandt. The central advantage of this approach is that it simply requires models for the exposure and outcome, thereby imposing no assumptions about exposure-mediator and mediator-mediator interactions, which may be complex and unwieldy in the presence of multiple mediators, as is the case here.[20] Please find below our assumed DAGs, subject to change.



DAG 1: Potential mediation of the effect of race/ethnicity on long COVID QALYs lost by a selection of clinical and behavioral risk factors. Note: for the purposes of visual simplification, we include one short arrow going into antiviral treatment to represent associations between all the preceding clinical variables (co-morbidities considered for antiviral prescribing), as well as race/ethnicity.



DAG 2: Potential mediation of the effect of socio-economic status on long COVID QALYs lost by a selection of clinical and behavioral risk factors. Note: for the purposes of visual simplification, as above, we also include short arrows emanating from age and sex to represent associations between age and sex with the clinical factors considered, and long COVID QALYs lost.

Specifically, when considering the effect of SES, we will fit a logistic regression to estimate the conditional odds of level = 1 for SES (dichotomized as high/low) as a function of subjects' age and sex, a linear model for long COVID QALYs lost, given SES, age, sex, and mediators considered, and a linear model for long COVID QALYs lost solely as a function of SES. That is, we will fit the following models:

- Logit(Pr[SES=1| age, sex]) to weight the observed long COVID QALYs lost among subjects with SES=1, which yields $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES*,M_{SES*}}}]$ Specifically weights equal $\frac{\boldsymbol{Pr}[\boldsymbol{SES=1}]}{\boldsymbol{Pr}[\boldsymbol{SES=1|age,sex}]}$, where M denotes the list of mediators and Y denotes long COVID QALYs lost
- Logit(Pr[SES=0| age, sex]) to weight the observed long COVID QALYs lost among subjects with SES=0, which yields $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES}}}]$. Specifically, weights equal $\frac{\boldsymbol{Pr}[\boldsymbol{SES=0}]}{\boldsymbol{Pr}[\boldsymbol{SES=0|age,sex}]}$
- E[long COVID QALYs lost| SES=0, mediators, age, sex] among subjects with SES=1, which is weighted by $\frac{\boldsymbol{Pr}[\boldsymbol{SES=1}]}{\boldsymbol{Pr}[\boldsymbol{SES=1|age,sex}]}$ among subjects with SES=1 to yield $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES*}}}]$
- E[long COVID QALYs lost|SES] for all subjects to yield the total effect of long COVID QALYs lost ascribed to SES (include Bellavia ref), i.e. E[Y|SES]

The natural direct effect can then be estimated by taking the difference between $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES*}}}]$ *and* $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES*,M_{SES*}}}]$*. The corresponding proportion can be estimated as* $\frac{\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES*}}}]-\ \boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES*,M_{SES*}}}]}{\boldsymbol{E|Y|SES}]}$*.*

The natural indirect effect can then be estimated by taking the difference between $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES}}}]$ *and* $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES*}}}]$*. The corresponding proportion can be estimated as* $\frac{\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES}}}]-\ \boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{SES,M_{SES*}}}]}{\boldsymbol{E|Y|SES}]}$*.*

When considering the effects of race/ethnicity, we will fit a linear model for long COVID QALYs lost, given race/ethnicity and the mediators considered and a linear model for long COVID QALYs lost solely as a function of race/ethnicity, which we denote as eth. That is, we will fit the following models:

- E[long COVID QALYs lost| eth=0, mediators, age, sex] among subjects with eth=1, which is weighted by $\frac{\boldsymbol{Pr}[\boldsymbol{eth=1}]}{\boldsymbol{Pr}[\boldsymbol{eth=1|age,sex}]}$ among subjects with eth=1 to yield $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{eth,M_{eth*}}}]$
- E[long COVID QALYs lost|eth] for all subjects to yield the total effect of long COVID QALYs lost ascribed to race/ethnicity, i.e. E[Y|eth]

To obtain $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{eth*,M_{eth*}}}]$, we simply take the mean of the observed long COVID QALYs lost among subjects with eth = 1, as the weights = 1 (race/ethnicity is not predicted by any other variable)
To obtain $\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{eth,M_{eth}}}]$, we simply take the mean of the observed long COVID QALYs lost among subjects with eth = 0, again because the weights = 1

As before,

The natural direct effect can then be estimated by taking the difference between $E[Y^{eth,M_{eth*}}]$ *and* $E[Y^{eth*,M_{eth*}}]$. *The corresponding proportion can be estimated as* $\frac{E[Y^{eth,M_{eth*}}]- E[Y^{eth*,M_{eth*}}]}{E|Y|eth]}$.

The natural indirect effect can then be estimated by taking the difference between $E[Y^{eth,M_{eth}}]$ *and* $E[Y^{eth,M_{eth*}}]$. *The corresponding proportion can be estimated as* $\frac{E[Y^{eth,M_{eth}}]- E[Y^{eth,M_{eth*}}]}{E|Y|eth]}$.

**alternatively, we may choose to implement separate models for each race/ethnicity group vs non-Hispanic white.

We anticipate that selection bias will arise because ethnic minority groups – who tend to be over-represented in essential work and other high-demand positions – may face competing priorities and limited interest in being involved in Tier 1 follow-up surveys. To address this, we will fit inverse probability of censoring weights, for which we will consider a proxy measure of essential work (socio-economic deprivation) as our 'measured' covariate reflecting this driver of non-participation.

For all analyses, we will use bootstrap resampling to obtain confidence intervals for our contrasts of interest.[20]

Handling of Missing Data

Write a brief summary of how missing data will be handled. You may follow the example used in the stock text below.
[stock text, feel free to keep or omit]

Preliminary analysis would be performed to ascertain a detailed assessment of the extent of missingness in the data. This should enable the identification of variables which lack sufficient data to allow for any useful analysis to be performed on them. The type of missingness shall be considered including whether data are not missing at random and follow-up with sites will be conducted if appropriate. Variables with greater than 30% missingness will be excluded from analysis. For those variables in the analysis with less than 30% missingness, if we determine the missingness pattern to be Missing Completely at Random or Missing at Random, imputation will be performed using Multiple Imputation by Chained Equations (MICE). If we determine the missingness pattern to be Missing Not at Random, imputation will be performed using maximum likelihood estimation.

Other Information

Provide details of the timelines for dissemination of research findings.

We will aim to complete the preliminary analysis within four months of receipt of the data and submit the initial preprint up to three months afterwards with active engagement of the Working Group.

References

1. Chen, T., Song, J., Liu, H., Zheng, H. & Chen, C. Positive Epstein–Barr virus detection in coronavirus disease 2019 (COVID-19) patients. *Sci Rep* 11, 10902 (2021).
2. Cervia, C. *et al.* Immunoglobulin signature predicts risk of post-acute COVID-19 syndrome. *Nat Commun* 13, 446 (2022).
3. Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nat Med* 27, 626–631 (2021).
4. Su, Y. *et al.* Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell* 185, 881-895.e20 (2022).
5. Park, C. *et al.* Short Report on Long COVID. (2021).
6. Thompson, E. J. *et al. Risk factors for long COVID: analyses of 10 longitudinal studies and electronic health records in the UK*.

http://medrxiv.org/lookup/doi/10.1101/2021.06.24.21259277 (2021) doi:10.1101/2021.06.24.21259277.

7. Peluso, M. J. *et al.* Post-acute sequelae and adaptive immune responses in people living with HIV recovering from SARS-CoV-2 infection. http://medrxiv.org/lookup/doi/10.1101/2022.02.10.22270471 (2022) doi:10.1101/2022.02.10.22270471.
8. Hirschtick, J. L. *et al.* Population-based estimates of post-acute sequelae of SARS-CoV-2 infection (PASC) prevalence and characteristics. *Clinical Infectious Diseases* (2021).
9. Sheridan, E. 'Missing cohort': Health inequality playing out through long Covid cases, data suggests. *Hackney Citizen* (2021).
10. Cooney, E. Researchers fear people of color may be disproportionately affected by long Covid. (2021).
11. Flash, M. J. E. *et al.* Disparities in Post-Intensive Care Syndrome During the COVID-19 Pandemic: Challenges and Solutions. 15 (2020).
12. Naidu, S. *et al.* The impact of ethnicity on the long-term sequelae of COVID-19: Follow-up from the first and second waves in North London. 76, A141 (2021).
13. Robinson-Lane, S. G. *et al.* Race, Ethnicity, and 60-Day Outcomes After Hospitalization With COVID-19. *Journal of the American Medical Directors Association* 22, 2245–2250 (2021).
14. Dowling, R. Impact of long Covid on ethnic minority healthcare workers investigated. (2021).
15. Briggs, A. H. *et al.* Estimating (quality-adjusted) life-year losses associated with deaths: With application to COVID-19. *Health Economics* 30, 699–707 (2021).

16. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* 22, 231–245 (2013).

17. Chavent, M., Genuer, R. & Saracco, J. Combining clustering of variables and feature selection using random forests. *Communications in Statistics - Simulation and Computation* 50, 426–445 (2021).

18. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal* 7, 19 (2015).

19. Bellavia, A., Zota, A. R., Valeri, L. & James-Todd, T. Multiple mediators approach to study environmental chemicals as determinants of health disparities. *Environmental Epidemiology* 2, e015 (2018).

20. VanderWeele, T. & Vansteelandt, S. Mediation Analysis with Multiple Mediators. *Epidemiologic Methods* 2, (2014).