Modeling the relative influence of socio-demographic variables on post-acute COVID-19 quality of life

Authors: Tigist F. Menkir[1,2], Barbara Wanjiru Citarella[2], Louise Sigfrid[2,3], Yash Doshi[4], Luis Felipe Reyes[2,5,6], Jose A. Calvache[7,8], Anders Benjamin Kildal[9,10], Anders B. Nygaard[11], Jan Cato Holter[11,12], Prasan Kumar Panda[13], Waasila Jassat[14,15], Laura Merson[2], Christl A. Donnelly[16,17], Mauricio Santillana[1,18], Caroline Buckee[1], Stéphane Verguet[19], Nima S. Hejazi[20], The ISARIC Clinical Characterisation Group*

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard TH Chan School of Public Health, Harvard University, USA
[2]ISARIC, Pandemic Sciences Institute, University of Oxford, UK
[3]Policy and Practice Research Group, Pandemic Sciences Institute, University of Oxford, Oxford UK
[4]Terna Speciality Hospital & Research Centre, Mumbai, India
[5]Universidad de La Sabana, Chia, Colombia
[6]Clinica Universidad de La Sabana, Chia, Colombia
[7]Departamento de Anestesiología, Universidad del Cauca, Colombia
[8]Department of Anesthesiology, Erasmus University Medical Center, Netherlands
[9]Department of Anesthesiology and Intensive Care, University Hospital of North Norway, Tromsø, Norway
[10]Department of Clinical Medicine, Faculty of Health Sciences, UIT The Arctic University of Norway, Tromsø, Norway
[11]Department of Microbiology, Oslo University Hospital, Oslo, Norway
[12]Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[13]All India Institute of Medical Sciences (AIIMS), Rishikesh, India
[14]National Institute for Communicable Diseases, South Africa
[15]Right to Care, South Africa
[16]Department of Statistics, University of Oxford, Oxford, UK
[17]MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics and Department of Infectious Disease Epidemiology, Imperial College London, London, UK
[18]Machine Intelligence Group for the Betterment of Health and the Environment, Network Science Institute, Northeastern University, Boston, MA, USA
[19]Department of Global Health and Population, Harvard TH Chan School of Public Health, Harvard University, USA
[20]Department of Biostatistics, Harvard TH Chan School of Public Health, Harvard University, USA
*The complete list of members and their affiliations is listed at the end of the manuscript

Key Points

Question: How do social and medical factors compare in predicting differences in quality of life (QoL) with long COVID and to what extent do clinical mediators explain social variables' relationships with long COVID QoL?

Findings: Socio-economic proxies employment status and educational attainment and female sex ranked on par with or above age and neuropsychological and rheumatological comorbidities as predictors of variation in long COVID QoL across participants. Additionally, estimated adjusted associations between each of these social factors and long COVID QoL were largely unexplained by a set of key comorbidities.

Meaning: Long COVID-based interventions may be more broadly beneficial if they account for social disparities as important risk factors for differential long COVID burden and, in addition to clinical targets, address broader structural determinants of health.

Abstract

Importance

Post-acute sequelae of SARS-CoV-2, referred to as "long COVID", are a globally pervasive threat. While their many clinical determinants are commonly considered, their plausible social correlates are often overlooked.

Objective

To compare social and clinical predictors of differences in quality of life (QoL) with long COVID. Additionally, to measure how much adjusted associations between social factors and long COVID-associated quality of life are unexplained by important clinical intermediates.

Design, Setting, and Participants

Data from the ISARIC long COVID multi-country prospective cohort study. Subjects from Norway, the United Kingdom (UK), and Russia, aged 16 and above, with confirmed acute SARS-CoV-2 infection reporting >= 1 long COVID-associated symptoms 1+ month following infection.

Exposure

The social exposures considered were educational attainment (Norway), employment status (UK and Russia), and female vs male sex (all countries).

Main outcome and measures

Quality of life-adjusted days, or QALDs, with long COVID.

Results

This cohort study included a total of 3891 participants. In all three countries, educational attainment, employment status, and female sex were important predictors of long COVID QALDs. Furthermore, a majority of the estimated relationships between each of these social correlates and long COVID QALDs could not be attributed to key long COVID-predicting comorbidities. In Norway, 90% (95% CI: 77%, 100%) of the adjusted association between the top two quintiles of educational attainment and long COVID QALDs was not explained by clinical intermediates. The same was true for 86% (73%, 100%) and 93% (80%,100%) of the adjusted associations between full-time employment and long COVID QALDs in the United Kingdom (UK) and Russia. Additionally, 77% (46%,100%) and 73% (52%, 94%) of the adjusted associations between female sex and long COVID QALDs in Norway and the UK were unexplained by the clinical mediators.

Conclusions and Relevance

This study highlights the role of socio-economic status indicators and female sex, in line with or beyond commonly cited clinical conditions, as predictors of long COVID-associated QoL, and further reveal that other (non-clinical) mechanisms likely drive their observed relationships. Our findings point to the importance of COVID interventions which go further than an exclusive focus on comorbidity management in order to help redress inequalities in experiences with this chronic disease.

Manuscript word count: 2999 words

Introduction

Long-term COVID-19 sequelae, referred to as long COVID, have resulted in a pressing public health crisis since early 2020. As defined by the World Health Organization, long COVID encompasses unexplainable symptoms which persist at least three months after an infection, occurring over two or more months.[1] Its widespread presence and impacts have been immense: a multinational study found that nearly half of individuals who were previously infected with SARS-CoV-2 went on to experience long-term symptoms around four months post infection[2]; further an estimated 59% of previously infected subjects reported a reduced quality of life (QoL).[3] Prior work has focused on identifying a myriad of clinical risk factors for post-COVID-19 conditions, including co-infections and pre-existing conditions, vaccination status, age, and female sex.[4-7] Conditions that have been consistently identified as key correlates of long COVID sequelae include obesity, asthma and other pulmonary diseases, chronic cardiac disease, diabetes, and smoking.[6,7]

Beyond clinical factors, social vulnerabilities are often critical determinants of differential disease burden overall.[8-11] Such inequities are attributed to broader challenges in access to health services and an array of health-threatening exposures, including but not limited to food and housing insecurity, financial discrimination, and air pollution.[8-11]

While there have been efforts to examine social factors potentially linked to long-term symptoms of COVID-19, findings on these relationships have been somewhat mixed.[7,12-17] For instance, a 2021 study in the United Kingdom (UK) found that living in high-deprivation settings was associated with both a higher and lower odds of symptom persistence, depending on the measure of deprivation index used[18], and a 2021 study in Michigan (USA) found that lower income was both significantly associated and not associated with long COVID symptoms' prevalence, depending on the post-illness duration considered.[13] It is also important to highlight that many of these studies rely on self-reported binary measures of post-COVID recovery, which do not capture nuanced experiences in recovery.

Given this context, we aimed to complement efforts centered on uncovering disparities in long COVID outcomes by leveraging a large, multi-national prospective cohort study. Specifically, we formally assessed the relationship between a diverse group of exposures and long COVID QoL, reasoning that factors like socio-economic status (SES) would be as much or more critical risk factors than important comorbidities, as has been recently illustrated for related outcomes, such as "healthy aging".[19] We further evaluated the extent to which clinical intermediates contribute to any observed disparities, hypothesizing that they may only partially explain these differences. Our analysis applies a similar mediation-centered lens to that of Vahidy et al.[17] and Lu et al.[20] However, rather than independently evaluating various mediators in the relationship between a given social factor and long COVID risk[17] or exploring the role of social factors as mediators[20] we measured the degree to which social variables' associations with long COVID QoL *cannot* be explained by a collective set of comorbidities.

Methods

Study design

This study uses data from the International Severe Acute Respiratory and emerging Infection Consortium's (ISARIC) multi-cohort consortium.[21] This prospective study across 76 countries collected demographic and medical data during acute SARS-CoV-2 infection, with a subset of sites assessing participants any time from one-three months post-infection and periodically thereafter.[22] Complete details on the study design and recruitment procedures can be found in the published follow-up protocol.[22] The first iteration of the statistical analysis plan for this work is also accessible online.[23]

We focused our analysis on subjects reporting one or more long COVID-associated symptom that was not present prior to illness and countries with data available on SES, age, and sex, QoL, *and* comorbidities, with combined demographic, comorbidity, and QoL datasets yielding sample sizes of at least n=1000 subjects. The countries meeting this criteria were Norway (n=1672), the UK (n=1064), and Russia (n=1155).

Measures

Our study incorporates information on both self-reported continued symptomatology and QoL. Health utility values were obtained using standard QoL-adjustment estimation procedures, based on subjects' responses to the EQ-5D-5L survey in follow-up forms, eliciting self-reported rankings of the intensity of problems experienced with each of five dimensions of health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression)[24], prior to and their COVID-19 illness and in the present.

Utility scores were computed following standard practice and then time-transformed (Supplementary Appendix: Methods). Similar to Sandmann et al.[25], we used a measure of quality-adjusted life days or QALDs (See Supplementary Appendix: Methods) and additionally focused on QoL at least three months following infection. Issues with recalling experiences several months in the past are likely to bias measures of pre-COVID QALDs.

We incorporated covariate data on age, socio-demographic variables, female sex at birth and SES proxies, a set of clinical comorbidities, treatments, and COVID-19 severity. Indicators of SES were selected depending on the data available in each cohort.

Statistical Analysis

To identify social predictors of long COVID QALDs, we applied a series of random forest ensemble learners[26] for each country, fit to all available clinical and demographic data, where variables were either treated individually (RF #1), pre-grouped based on subject matter knowledge (RF #2), or grouped algorithmically via hierarchical clustering (RF #3). We implemented a pre-grouped procedure, which incorporates subject matter knowledge, as an

alternative to approaches agnostic to such context. For RF #1 and RF #2, the percent increase in mean squared error (MSE) associated with each variable was reported as a measure of importance, while, for RF #3, the frequency of variable selection was reported.

To estimate the natural direct effects (NDE) and natural indirect effects (NIE) of binary SES proxies or female sex on long COVID QALDs in each cohort, we applied a flexible semi-parametric statistical approach (Supplementary Appendix: Methods).[27,28] The NDE and NIE arise from a decomposition of the average treatment effect or total effect (TE), as first described by Robins and Greenland.[29] In this context, the NDE describes the relationship between a given social variable and long COVID QALDs, operating through all pathways excluding the mediators of interest, while the NIE describes this exposure/outcome relationship *through* the mediators. We define the proportion non-mediated as NDE/the total effect (TE), i.e. the share of the TE of the social variable that cannot be explained by the clinical intermediates. In order for these measures to be interpreted causally, several assumptions are necessary, including exposure and mediator positivity, well-defined exposures and potential outcomes, no interference between study units, and no unmeasured confounding.[29,30] Rather than make the strict, and possibly untenable assumption, that these criteria are all met, we eschew any claims about causality, using this analysis as a framework to arrive at *interpretable, model-free* statistical (i.e., non-causal) target parameters. In other words, we report adjusted associations, where the NDE, NIE, and proportion non-mediated are used to communicate, in a clearly interpretable manner, how mediators play a role in any observed disparities.

Our code is publicly available at: https://github.com/goshgondar2018/social_long_covid.

Results

Norway

Long COVID QALDs in this cohort reported a median of 345 (Interquartile range/IQR: 313-360). There was no broadly consistent trend in long COVID QALDs across quintiles of educational attainment, although the lowest mean QALDs occurred in the bottom two quintiles. The greatest differences in long COVID QALDs, in order of increasing magnitude, occurred between quintiles 3 and 1, 5 and 1, and 4 and 1. Estimated long COVID QALDs among males slightly exceeded that of females ($p<0.001$).

This cohort was the youngest, with a mean age of 51.8 years (SD: 13.6 years). The most commonly reported comorbidity was asthma (22%).

Among the leading individual predictors of long COVID QALDs from RF #1, anxiety/depression ranked first, followed by educational attainment, rheumatological disorder, and age (Figure 1a). For RF #2, the first and second PCs of the cluster containing all socio-demographic variables, i.e., educational attainment indicators and sex, ranked below the first and second PCs of the cluster containing psychological disorder and chronic neurologic disorder (Figure 1b). RF #3 largely corroborated these orderings, where psychological disorder, rheumatological disorder, chronic neurological disorder, and asthma were the most consistently selected variables within

identified important clusters, followed by educational attainment (in years) and a dummy educational attainment indicator for quintile 5 (vs 1) (Figure 1c).

We estimated that falling in the top two quintiles of educational attainment was significantly associated with 12.3 (6.49,18.2) additional long COVID QALDs, on average, via the NDE, and 0.67 (-0.982, 2.32) additional long COVID QALDs on average, via the NIE (non-significant), corresponding to a proportion non-mediated of 0.897 (0.773, 1). That is, 89.7% of the adjusted association between high educational attainment and long COVID QALDs could not be explained by the included mediators and must thus be attributed to other mechanisms. The exact relationship between the NDE/NIE and proportion non-mediated may not hold for computational reasons, because CIs are estimated separately for each of the measures using cross-validation. We obtained consistent directionality in findings for pairwise comparisons of quintiles 3 and 1, 4 and 1, and 5 and 1, with the greatest proportion non-mediated for the quintile 5 versus 1 comparison, although we cannot make any conclusions about significance as multiple testing corrections are warranted.

A clear and statistically significant negative association was also observed between female sex and long COVID QALDs, with an estimated NDE of -6.79 (-12.8, -0.723), NIE of -3.05 (-5.89, -0.215) and proportion non-mediated of 0.773 (0.455,1).

UK

The median (IQR) of long COVID QALDs was 295 (233, 342). Employment status was markedly skewed towards full-employment (51%), retirement (30%), part-time employment (10%) and unemployment (7%). Estimated long COVID QALDs were greatest among participants who reported being furloughed, students, or full-time employees and lowest among those in the unemployed and retired categories. Estimated long COVID QALDs were also slightly higher among males ($p<0.001$).

This cohort was skewed towards older adults (mean (SD): 59.0 (12.6) years) and the most commonly reported comorbidity was hypertension (36%).

Employment status was the leading predictor for long COVID QALDs in the UK, followed by psychological disorder, age, employment status category, chronic neurological disorder, and rheumatological disorder, based on RF #1 (Figure 2a). Sex followed in the rankings, which, along with the acute COVID-19 severity indicator, fell among the top ten predictors (Figure 2a). RF #2 further supported the predictive role of employment status and sex as a group, with the PCs of the socio-demographic variables leading, closely followed by the PCs of the group of mental health and neurological disorders (Figure 2b). Age alone ranked highly, even in comparison to grouped factors (Figure 2b). Findings from RF #3 aligned well with these results, with age, chronic neurological disorder, employment status indicators, and psychological disorder the most commonly selected variables, followed by rheumatological disorder, across key clusters (Figure 2c).

Increased income/job stability, as proxied by employment status, was consistently associated with increased long COVID QALDs, irrespective of the binary designation. We found that self-reported full-time employment versus any other employment status category (excluding retirement) was associated with, on average, 31.7 (14.2, 49.3) higher long COVID QALDs, via the NDE, and 4.90 (-0.0652, 9.86) higher long COVID QALDs, via the NIE (narrowly non-significant), with a proportion non-mediated of 0.862 (0.729, 0.996). We obtained an even stronger and significant relationship between self-reported full-time employment versus unemployment, with, on average, 79.5 (50.0, 109) increased long COVID QALDs among the full-time employed relative to the unemployed (NDE) and 9.50 (1.04, 18.0) increased long COVID QALDs among the full-time employed versus unemployed (NIE). The proportion non-mediated was 0.905 (0.829, 0.981), suggesting that 90.5% of the adjusted association between full-time employment versus unemployment on long COVID QALDs does not operate through the considered mediators.

Female sex was associated with lower expected long COVID QALDs, with an NDE and NIE of -24.2 (-37.8, -10.7) and -9.61 (-16.3, -2.95), respectively. The corresponding proportion non-mediated was the lowest observed among all contrasts, with 72.9% (51.9%, 93.5%) of the adjusted association between female sex and long COVID QALDs being unexplained by the clinical intermediates.

Russia

The median (IQR) of long COVID QALDs was 353 (334-365). Employment status was markedly skewed towards full-employment (55%) and retirement (39%). Long COVID QALDs were highest among students, part-time employees, full-time employees, and carers, and lowest among those in the retired and unemployed categories. Males reported higher long COVID QALDs than females ($p<0.001$).

The mean age of participants was 59.6 years (SD: 14.4 years) and hypertension was the most frequently reported comorbidity (59%).

According to RF#1, age, followed by employment status indicators, hypertension, and chronic neurological disorder, outranked all other variables in predicting long COVID QALDs in this cohort (Figure 3a). RF #2 generally supported these findings. The cluster containing solely age led the rankings. The principal components (PCs) of the cluster containing the socio-demographic variables, the first PC of the cluster containing hypertension and other cardiac disease, and the first PC of the cluster containing dementia and chronic neurological disorder (Figure 3b) then followed. Similarly, for RF #3, age, other chronic cardiac disease, chronic neurological disorder, as well as dementia, employment status indicators, hypertension, rheumatological disorder, and sex led the set of most frequently selected variables (Figure 3c).

Full-time employment was associated with higher long COVID QALDs compared to all other employment status categories. 12.9 (95% CI: 5.28, 20.5) more long COVID QALDs were expected among subjects self-reporting full-time employment compared to all other employment

categories, via the NDE, which was significant. An additional 4.03 (-1.37, 1.56) long COVID QALDs were expected among subjects self-reporting full-time employment compared to all other employment categories, via the NIE. The proportion non-mediated was estimated to be 0.93 (0.80, 1).

Female sex was associated with lower COVID QALDs, with an estimated NDE of -7.49 (-13.3, -1.69) and NIE of -0.0547 (-2.02, 1.91). The estimated proportion mediated was negative, which corresponds to a proportion non-mediated exceeding 1. This result is intuitive given our estimates of the NIE, where the upper bound of the CI fell markedly above 0, indicating insufficient evidence in favor of a *positive* NIE. In other words, the NDE and NIE for female sex may act in opposite directions in this cohort.

Discussion

In this study, we provided a quantitative assessment of the extent to which social factors, compared to commonly highlighted clinical conditions, may relate to varying experiences with long COVID.

Of note, we observed that educational attainment or employment status and sex at birth were generally as or more predictive of long COVID QALDs as highly ranked neuropsychological and rheumatological comorbidities and age. Our mediation analyses further suggest that not only are indicators of social disadvantage notably predictive of lower long COVID QALDs, but also that the connection between these variables and long COVID QALDs could only be partially explained by key long COVID-predicting comorbidities. This general finding, i.e., that disparities are not solely attributable to underlying differences in comorbidity rates across various demographic groups, has been validated in other studies conducted over the course of the pandemic.[12,17,31]

Our study benefited from the use of a sizable and multi-national cohort with long-term QoL data, providing more nuanced information on post-acute COVID-19 experiences beyond simply whether patients experienced long COVID symptoms. Given the reasonably large sample sizes of our study cohorts, we were able to apply data-adaptable machine learning tools, including recent developments in causal machine learning.[27,28] The variable selection methods we used avoid strict modeling assumptions and further accommodate inherent variable groupings. The causal mediation approaches we applied integrate flexible, but simultaneously relatively precise, algorithms[27,28], providing a promising alternative to strictly parametric approaches.[32]

There are several important limitations of our analysis. First, we note that we did not have information on subject-specific duration of long COVID, and instead assumed a uniform duration, consistent with examples in the literature.[33,34] Additionally, we were unable to examine the varying roles of the different social and clinical factors on *changes* in QoL, due to the aforementioned issues with recall. However, to informally assess whether differences in post-COVID QALDs between socioeconomic groups and self-reported sex were simply artifacts of baseline QoL differences, we compared pre-COVID QoL scores across groups. We found

that for Norway and Russia, cross-group variability in estimated QALDs was much higher post- vs pre-COVID, suggesting that pre-COVID heterogeneity in QoL did not fully drive the differences we observed across groups. The UK cohort had pre-COVID QoL measures for only 26% of subjects, so we could not make a standardized comparison. For future work, it is imperative to collect information on QoL measures at all stages of illness, not simply ex post facto, by positioning readily implementable study protocols at the outset of an outbreak.

It can also be concluded that survivorship bias[35] might affect our results, as only subjects who completed a follow-up survey at any follow-up interval can have their QoL measures recorded, and those lost to follow up due to debilitating medical events are also likely to have a far reduced QoL prior to this event. However, we note that no participants in the three cohorts died at any point during follow-up.

Finally, we were limited to specific socio-economic variables that may not fully reflect participants' levels of socio-economic deprivation, which further varied by country. Thus, we can only draw conclusions about the role of socio-economic status in relation to the specific measures defined for each country. For future work, it would be useful to emphasize the collection of more proximate *shared* indicators of socio-economic status. We were also limited to data on sex at birth, which does not capture important gender-based disparities that exist beyond this binary.[36]

Conclusion

Here, we provide a robust statistical framework for highlighting the contribution of social disparities to chronic ill-health. Our approach can be used to compare a collective of diverse variables in predicting post-acute COVID-19 QoL and to distill the unique role of a given social variable. Our data highlights the multifactorial relationship between pre-existing risk factors and socio-economic factors and long COVID QoL. As such, we demonstrate that accounting for social vulnerabilities when evaluating determinants of post-acute COVID-19 trajectories is essential and that studies and interventions focusing solely on clinical targets may not be sufficient. Conversely, transformational societal interventions, which address access to care, education, employment, etc., have the opportunity to lead to potentially more comprehensive benefits and improve overall well-being in marginalized communities.

References

1. World Health Organization. Post COVID-19 condition (Long COVID). (2022).
2. O'Mahoney, L. L. *et al.* The prevalence and long-term health effects of Long Covid among hospitalised and non-hospitalised populations: a systematic review and meta-analysis. *eClinicalMedicine* **55**, 101762 (2023).
3. Malik, P. *et al.* Post-acute COVID-19 syndrome (PCS) and health-related quality of life (HRQoL)—A systematic review and meta-analysis. *J Med Virol* **94**, 253–262 (2022).
4. Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nat Med* **27**, 626–631 (2021).
5. Thompson, E. J. *et al.* Long COVID burden and risk factors in 10 UK longitudinal studies and electronic health records. *Nat Commun* **13**, 3528 (2022).
6. Tsampasian, V. *et al.* Risk Factors Associated With Post−COVID-19 Condition: A Systematic Review and Meta-analysis. *JAMA Intern Med* **183**, 566 (2023).
7. Subramanian, A. *et al.* Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat Med* **28**, 1706–1714 (2022).
8. Office of Disease Prevention and Health Promotion, Office of the Assistant Secretary for Health, Office of the Secretary, U.S. Department of Health and Human Services. Social Determinants of Health.
9. Jones, C. P., Jones, C. Y., Perry, G. S., Barclay, G. & Jones, C. A. Addressing the Social determinants of children's Health: A cliff Analogy. *JHCPU* **20**, 1–12 (2009).
10. Berger, Z., Altiery De Jesus, V., Assoumou, S. A. & Greenhalgh, T. Long COVID and Health Inequities: The Role of Primary Care. *Milbank Q* **99**, 519–541 (2021).
11. Bibbins-Domingo, K. Integrating Social Care Into the Delivery of Health Care. *JAMA* **322** (2019).
12. Shabnam, S. *et al.* Socioeconomic inequalities of Long COVID: a retrospective population-based cohort study in the United Kingdom. *J R Soc Med* **116**, 263–273 (2023).
13. Hirschtick, J. L. *et al.* Population-based estimates of post-acute sequelae of SARS-CoV-2 infection (PASC) prevalence and characteristics. *Clin Infect Dis* **73,** 2055-2064 (2021).
14. Müller, S. A. *et al.* Prevalence and risk factors for long COVID and post-COVID-19 condition in Africa: a systematic review. *Lancet Glob Health* **11**, e1713–e1724 (2023).
15. Robinson-Lane, S. G. *et al.* Race, Ethnicity, and 60-Day Outcomes After Hospitalization With COVID-19. *J Am Med Dir Assoc* **22**, 2245–2250 (2021).
16. 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. *Thorax.* **76**, A141 (2021).
17. Vahidy, F. S. *et al.* Racial and ethnic disparities in SARS-CoV-2 pandemic: analysis of a COVID-19 observational registry for a diverse US metropolitan population. *BMJ Open* **10**, e039849 (2020).
18. Park, C., Ayoubkhani, D. *et al.* Short Report on Long COVID. (2021).
19. Santamaria-Garcia, H. *et al.* Factors associated with healthy aging in Latin American populations. *Nat Med* **29**, 2248–2258 (2023).
20. Lu, J. *et al.* Educational inequalities in mortality and their mediators among generations across four decades: nationwide, population based, prospective cohort study based on the ChinaHEART project. *BMJ* **382**, e073749 (2023) doi:10.1136/bmj-2022-073749.
21. ISARIC Clinical Characterization Group *et al.* ISARIC-COVID-19 dataset: A Prospective, Standardized, Global Dataset of Patients Hospitalized with COVID-19. *Sci Data* **9**, 454 (2022).
22. International Severe Acute Respiratory and emerging infection Consortium. COVID-19 Long term protocol. https://isaric.org/research/covid-19-clinical-research-resources/covid-19-long-term-follow-up-study/

23. Menkir, T. & ISARIC Working Group. The relative influence of clinical and sociodemographic variables on the long-term COVID-19-associated QALYs lost: Analysis Plan. (2022).
24. Euroqol. EQ-5D-5L | About. (2021). https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-5l/
25. Sandmann, F. G. et al. Long-Term Health-Related Quality of Life in Non-Hospitalized Coronavirus Disease 2019 (COVID-19) Cases With Confirmed Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection in England: Longitudinal Analysis and Cross-Sectional Comparison With Controls. *Clin Infect Dis* **75,** e962-e973 (2022).
26. Breiman, L. Random Forests. *Machine Learning* **45** (2001).
27. Hejazi, N. S., Rudolph, K. E., Van Der Laan, M. J. & Díaz, I. Nonparametric causal mediation analysis for stochastic interventional (in)direct effects. *Biostatistics* **24**, 686–707 (2023).
28. Hejazi, N., Rudolph, K. & Díaz, I. medoutcon: Nonparametric efficient causal mediation analysis with machine learning in R. *JOSS* **7**, 3979 (2022).
29. Robins, J.M. & Greenland, S. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology* **3** (1992).
30. Benkeser, D., Díaz, I. & Carone, M. Statistical Learning in Mediation Analysis - Chapter 3: Natural direct and indirect effects. (2021).
31. Qeadan, F. *et al.* Racial disparities in COVID-19 outcomes exist despite comparable Elixhauser comorbidity indices between Blacks, Hispanics, Native Americans, and Whites. *Sci Rep* **11**, 8738 (2021).
32. VanderWeele, T. & Vansteelandt, S. Mediation Analysis with Multiple Mediators. *Epidemiologic Methods* **2** (2014).
33. Mizrahi, B. *et al.* Long covid outcomes at one year after mild SARS-CoV-2 infection: nationwide cohort study. *BMJ* **380**, e072529 (2023).
34. Cai, J. *et al.* A one-year follow-up study of systematic impact of long COVID symptoms among patients post SARS-CoV-2 omicron variants infection in Shanghai, China. *Emerg Microbes & infect* **12**, 2220578 (2023).
35. Smith, L. H. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. *Curr Epidemiol Rep* **7**, 179–189 (2020).
36. US National Center for Health Statistics. Long COVID: Household Pulse Survey. (2022).https://www.cdc.gov/nchs/covid19/pulse/long-covid.htm

Main text figures and tables

Table 1: Summary of demographic variables (excluding SES proxies) and common comorbidities in the final study populations for each cohort, post-missing data imputation.

	Norway (N=1672)	Russia (N=1155)	UK (N=1064)	Overall (N=3891)
AGE				
Mean (SD)	51.8 (13.6)	59.6 (14.4)	59.0 (12.6)	56.1 (14.1)
Median [Min, Max]	52.0 [17.0, 96.0]	59.0 [20.0, 100]	59.0 [19.0, 91.0]	56.0 [17.0, 100]
SEX				
Female	1139 (68.1%)	688 (59.6%)	434 (40.8%)	2261 (58.1%)
Male	533 (31.9%)	467 (40.4%)	630 (59.2%)	1630 (41.9%)
ASTHMA				
ABSENT	1305 (78.1%)	1076 (93.2%)	818 (76.9%)	3199 (82.2%)
PRESENT	367 (21.9%)	79 (6.8%)	246 (23.1%)	692 (17.8%)
CHRONIC CARDIAC DISEASE (NOT HYPERTENSION)				
ABSENT	1547 (92.5%)	877 (75.9%)	952 (89.5%)	3376 (86.8%)
PRESENT	125 (7.5%)	278 (24.1%)	112 (10.5%)	515 (13.2%)
CHRONIC HAEMATOLOGICAL DISEASE				
ABSENT	1642 (98.2%)	1137 (98.4%)	1032 (97.0%)	3811 (97.9%)
PRESENT	30 (1.8%)	18 (1.6%)	32 (3.0%)	80 (2.1%)
CHRONIC KIDNEY DISEASE				
ABSENT	1635 (97.8%)	1081 (93.6%)	1004 (94.4%)	3720 (95.6%)
PRESENT	37 (2.2%)	74 (6.4%)	60 (5.6%)	171 (4.4%)
CHRONIC NEUROLOGICAL DISORDER				
ABSENT	1556 (93.1%)	1087 (94.1%)	1013 (95.2%)	3656 (94.0%)
PRESENT	116 (6.9%)	68 (5.9%)	51 (4.8%)	235 (6.0%)
CHRONIC PULMONARY DISEASE (NOT ASTHMA)				
ABSENT	1602 (95.8%)	1037 (89.8%)	875 (82.2%)	3514 (90.3%)
PRESENT	70 (4.2%)	118 (10.2%)	189 (17.8%)	377 (9.7%)
DIABETES MELLITUS (TYPE 2)				
ABSENT	1542 (92.2%)	942 (81.6%)	846 (79.5%)	3330 (85.6%)
PRESENT	130 (7.8%)	213 (18.4%)	218 (20.5%)	561 (14.4%)
DIABETES MELLITUS (TYPE NOT SPECIFIED)				
ABSENT	1646 (98.4%)	934 (80.9%)	828 (77.8%)	3408 (87.6%)
PRESENT	26 (1.6%)	221 (19.1%)	236 (22.2%)	483 (12.4%)
HYPERTENSION				
ABSENT	1322 (79.1%)	475 (41.1%)	685 (64.4%)	2482 (63.8%)
PRESENT	350 (20.9%)	680 (58.9%)	379 (35.6%)	1409 (36.2%)
MALIGNANT NEOPLASM				
ABSENT	1626 (97.2%)	1099 (95.2%)	1034 (97.2%)	3759 (96.6%)
PRESENT	46 (2.8%)	56 (4.8%)	30 (2.8%)	132 (3.4%)
OBESITY				
ABSENT	1638 (98.0%)	878 (76.0%)	1006 (94.5%)	3522 (90.5%)
PRESENT	34 (2.0%)	277 (24.0%)	58 (5.5%)	369 (9.5%)
RHEUMATOLOGICAL DISORDER				
ABSENT	1356 (81.1%)	1090 (94.4%)	884 (83.1%)	3330 (85.6%)
PRESENT	316 (18.9%)	65 (5.6%)	180 (16.9%)	561 (14.4%)
SMOKING				
ABSENT	1627 (97.3%)	1011 (87.5%)	958 (90.0%)	3596 (92.4%)
PRESENT	45 (2.7%)	144 (12.5%)	106 (10.0%)	295 (7.6%)
Long COVID QALDs				
Mean (SD)	322 (59.2)	338 (48.9)	269 (94.3)	312 (73.4)

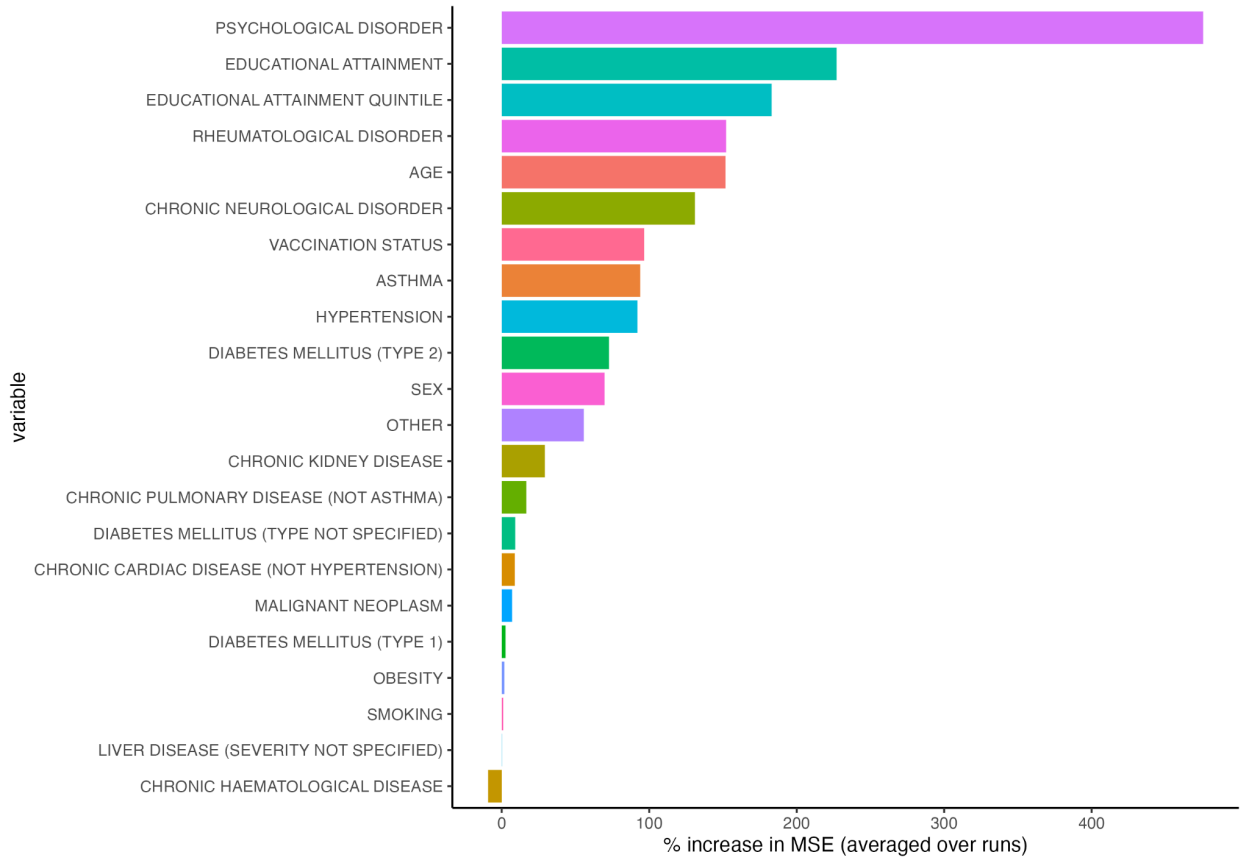

Figure 1a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for Norway. Variables with negative % MSE values are considered unimportant.

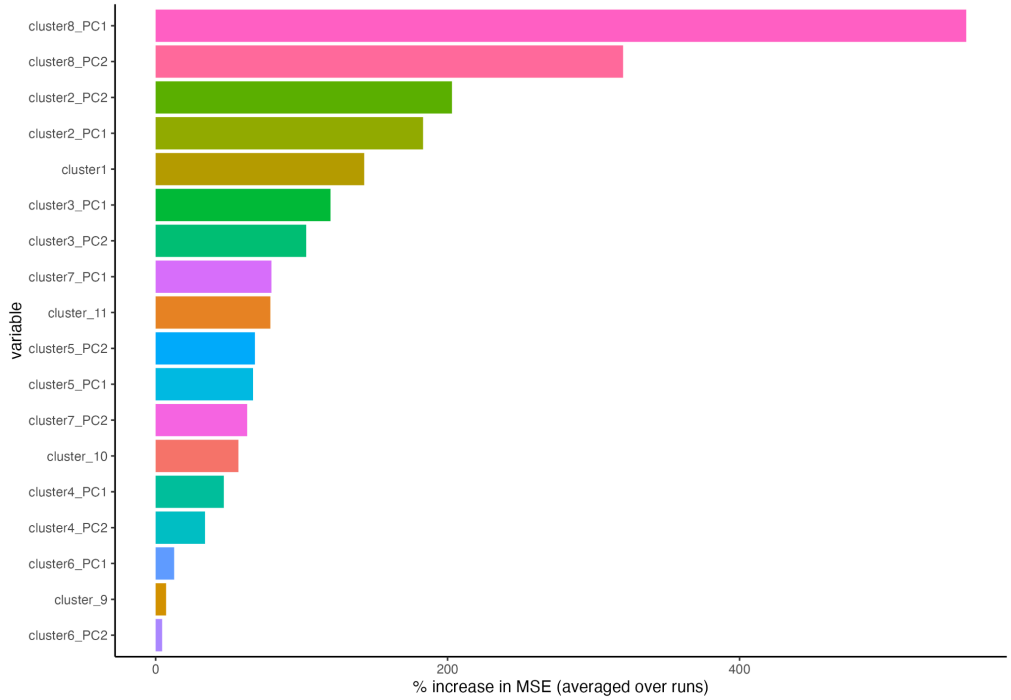

Figure 1b. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for Norway. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding principal components, if the cluster consists of multiple variables. PC1 denotes principal component 1 and PC2 denotes principal component 2.

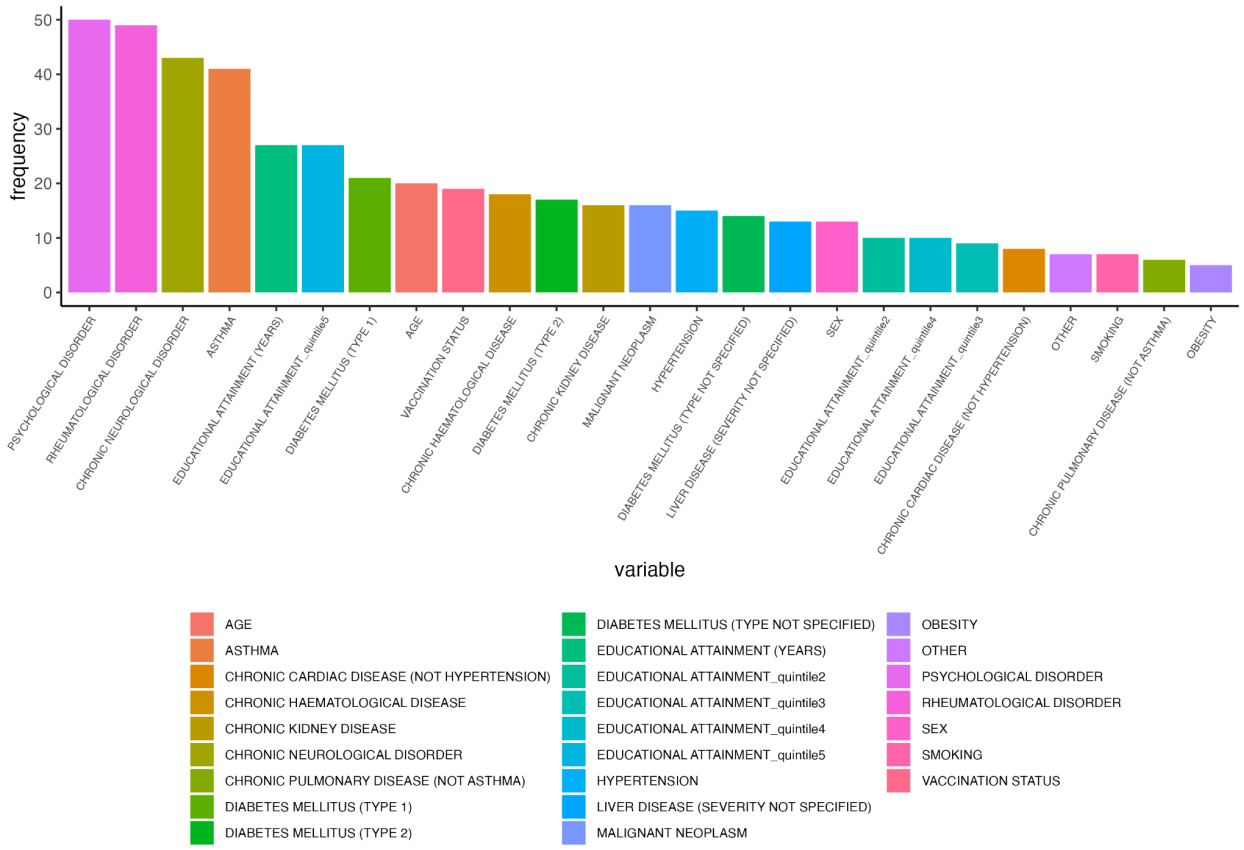

Figure 1c. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for Norway.

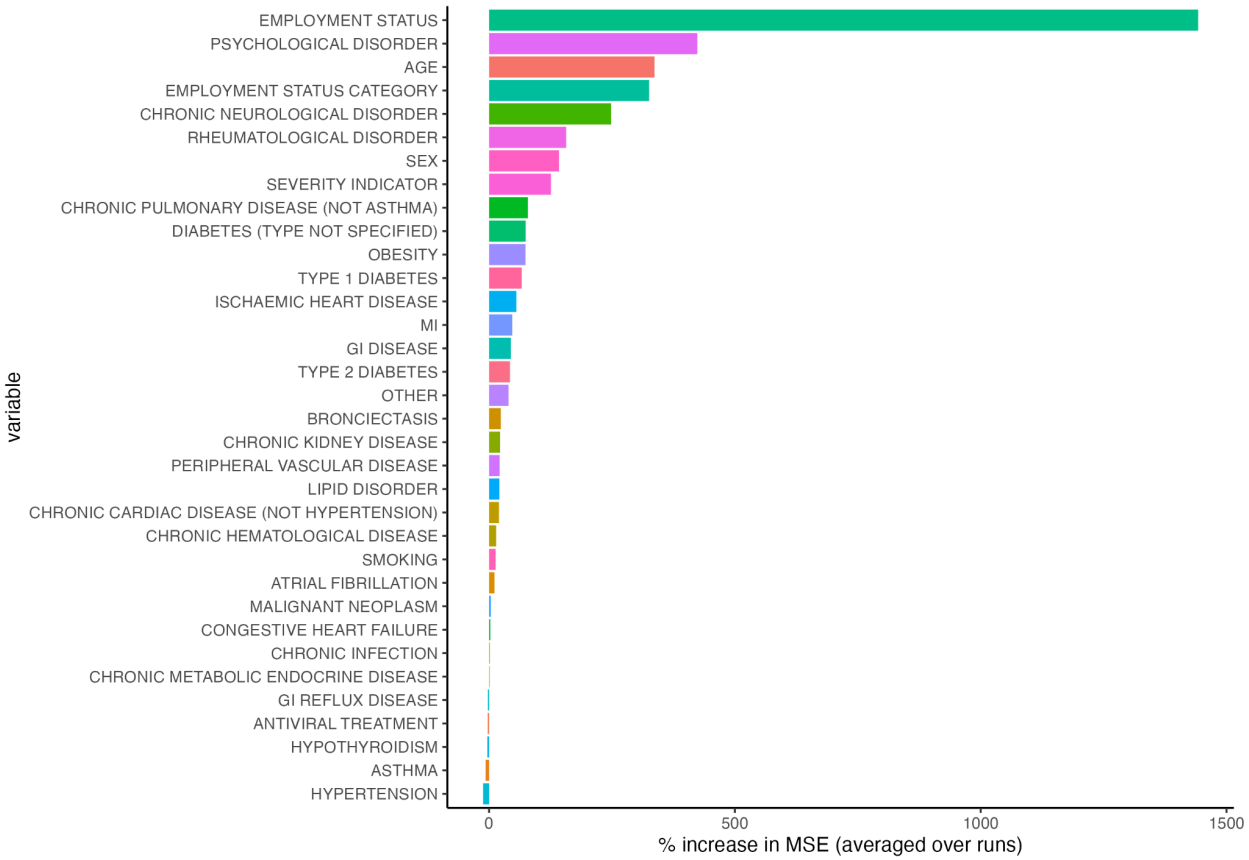

Figure 2a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for the UK. Variables with negative % MSE values are considered unimportant.

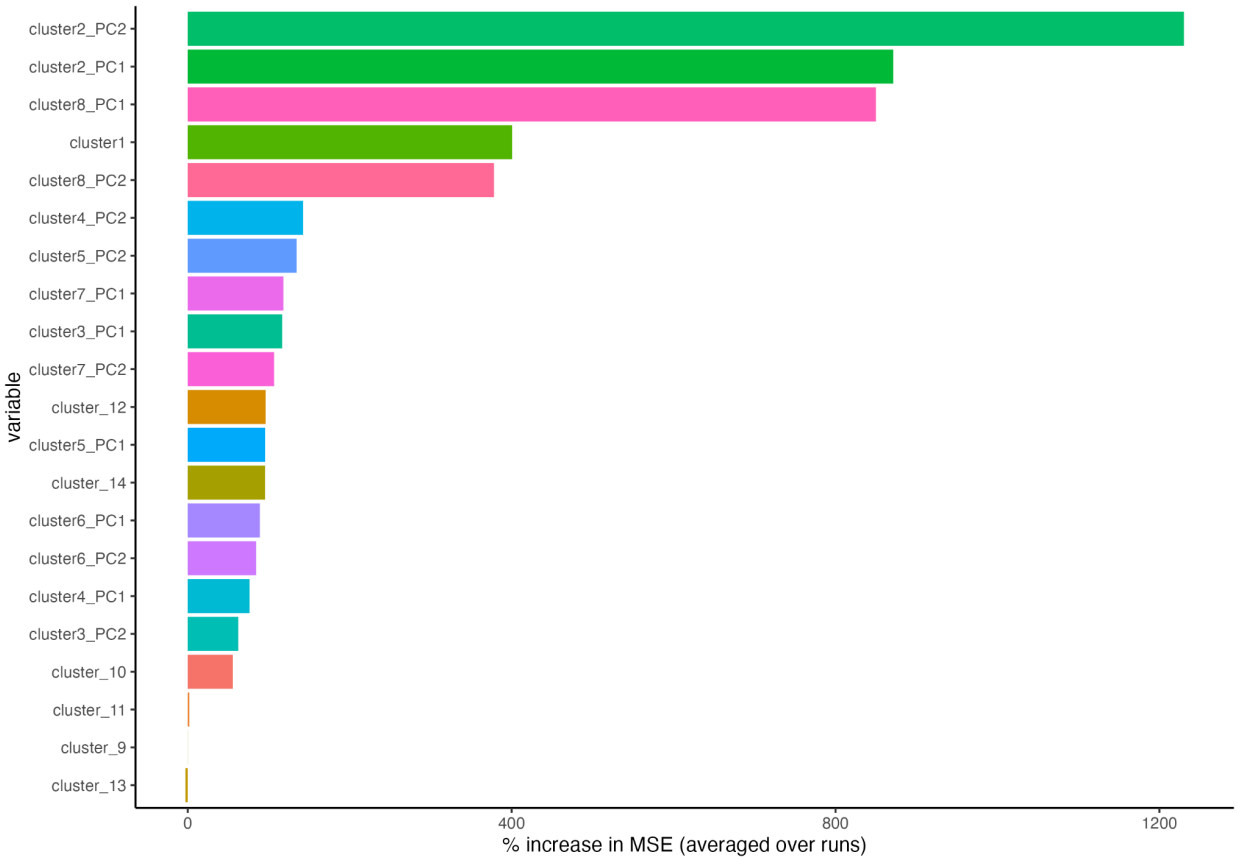

Figure 2b. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for the UK. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding principal components, if the cluster consists of multiple variables. PC1 denotes principal component 1 and PC2 denotes principal component 2.

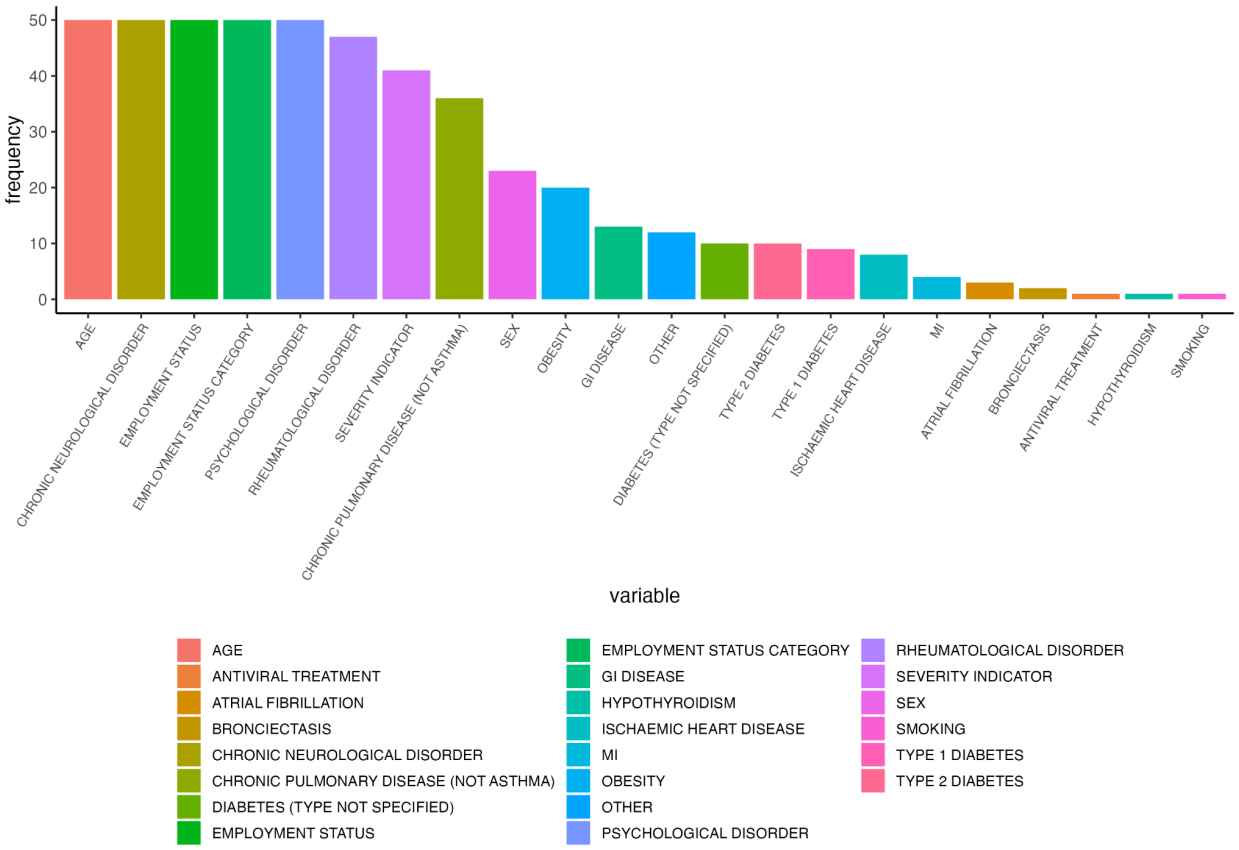

Figure 2c. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for the UK.

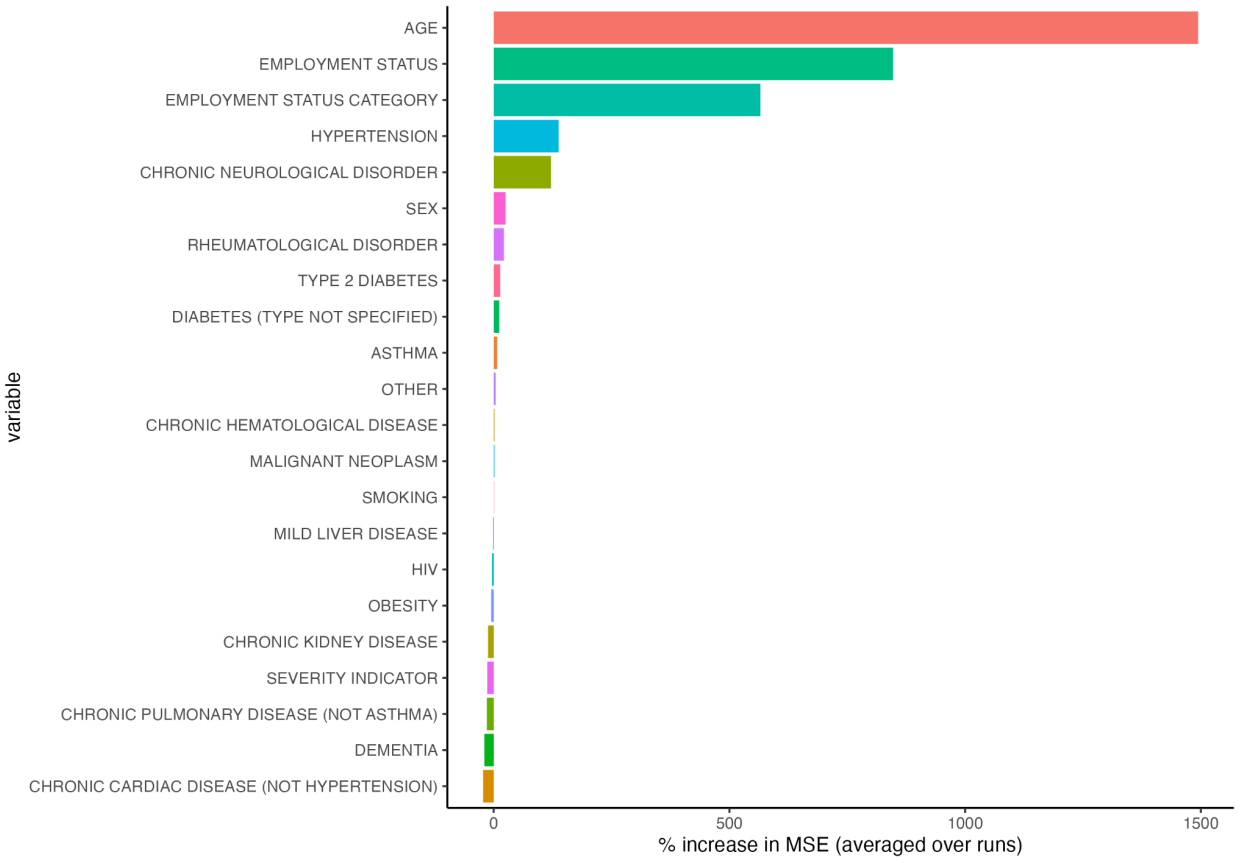

Figure 3a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for Russia. Variables with negative % MSE values are considered unimportant.

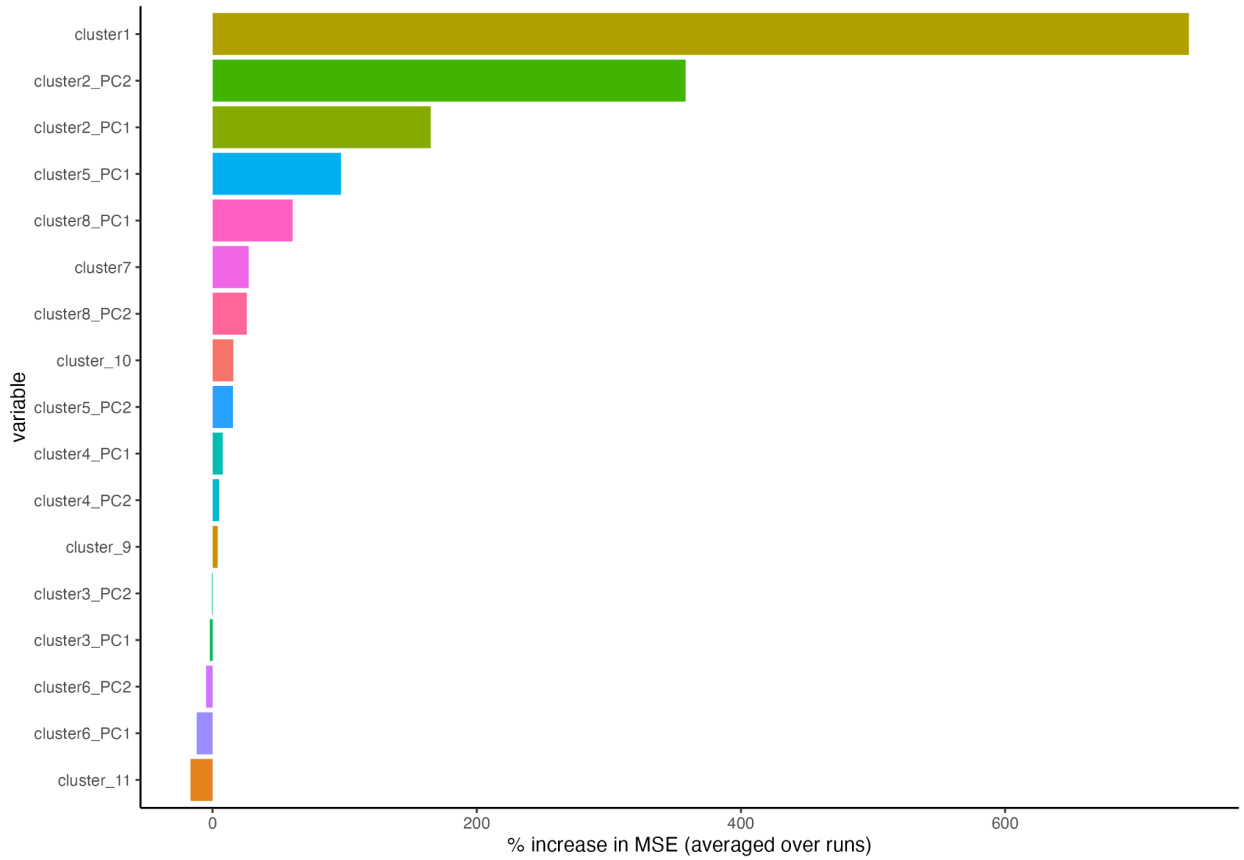

Figure 3b. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for Russia. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding principal components, if the cluster consists of multiple variables. PC1 denotes principal component 1 and PC2 denotes principal component 2.

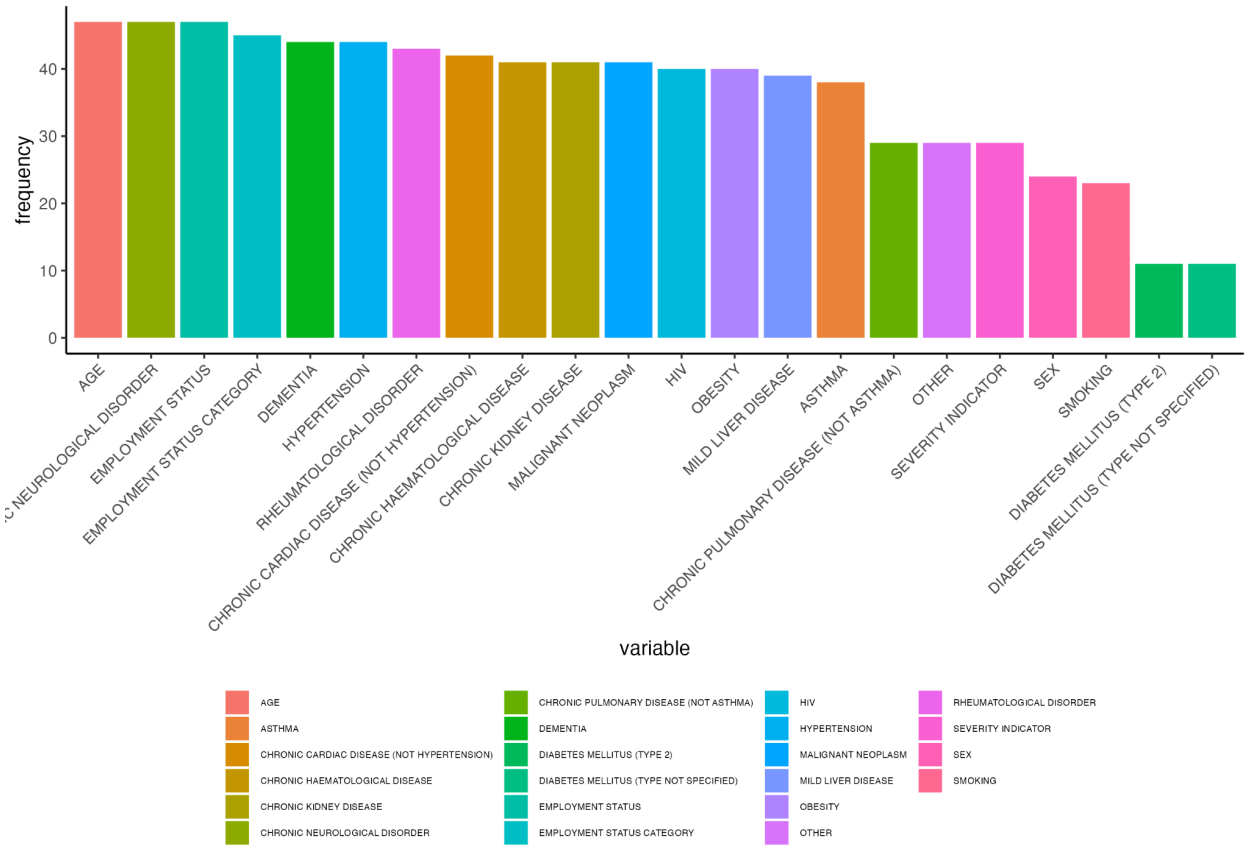

Figure 3c. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for Russia.

Funding Statement

TFM acknowledges support from NIH Training Grant 2T32AI007535. LFR was funded by Universidad de La Sabana (MED-309-2021). MS has been funded (in part) by contracts 200-2016-91779 and cooperative agreement CDC-RFA-FT-23-0069 with the Centers for Disease Control and Prevention (CDC). The findings, conclusions, and views expressed are those of the author(s) and do not necessarily represent the official position of the CDC. MS was also partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM130668.This work was made possible by the UK Foreign, Commonwealth and Development Office and Wellcome [215091/Z/18/Z, 222410/Z/21/Z, 225288/Z/22/Z and 220757/Z/20/Z], the Bill & Melinda Gates Foundation [OPP1209135], the philanthropic support of the donors to the University of Oxford's COVID-19 Research Response Fund (0009109), grants from the National Institute for Health and Care Research (NIHR award CO-CIN-01/DH_/Department of Health/United Kingdom), the Medical Research Council (MRC grant MC_PC_19059), and by the NIHR Health Protection Research Unit (HPRU) in Emerging and Zoonotic Infections at University of Liverpool in partnership with Public Health England (PHE), (award 200907), NIHR HPRU in Respiratory Infections at Imperial College London with PHE (award 200927), Liverpool Experimental Cancer Medicine Centre (grant C18616/A25153), NIHR Biomedical Research Centre at Imperial College London (award ISBRC-1215-20013), and NIHR Clinical Research Network providing infrastructure support, the

Comprehensive Local Research Networks (CLRNs), Cambridge NIHR Biomedical Research Centre (award NIHR203312), the Research Council of Norway grant no 312780, and a philanthropic donation from Vivaldi Invest A/S owned by Jon Stephenson von Tetzchner to the Norwegian SARS-CoV-2 study, the South Eastern Norway Health Authority and the Research Council of Norway.

Acknowledgments

The investigators thank all the clinical and research staff, who performed the follow-up assessments and collected this data, and the participants for their individual contributions in these difficult times. We would also like to thank the Long Covid Support group and ISARIC's Global Support Centre for their invaluable support.

We also acknowledge the support of the COVID clinical management team, AIIMS, Rishikesh, India; the Liverpool School of Tropical Medicine and the University of Oxford; Imperial NIHR Biomedical Research Centre; the dedication and hard work of the Norwegian SARS-CoV-2 study team; and preparedness work conducted by the Short Period Incidence Study of Severe Acute Respiratory Infection.

This work uses data provided by patients and collected by the NHS as part of their care and support #DataSavesLives. The data used for this research were obtained from ISARIC4C. We are extremely grateful to the 2648 frontline NHS clinical and research staff and volunteer medical students who collected these data in challenging circumstances; and the generosity of the patients and their families for their individual contributions in these difficult times. The COVID-19 Clinical Information Network (CO-CIN) data was collated by ISARIC4C Investigators. We also acknowledge the support of Jeremy J Farrar and Nahoko Shindo.

The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University

Competing interests

MS has received institutional research funds from the Johnson and Johnson foundation and from Janssen global public health. MS also received institutional research funding from Pfizer.

ISARIC Clinical Characterisation Group

Beatrice Alex, Eyvind W. Axelsen, Benjamin Bach, John Kenneth Baillie, Wendy S. Barclay, Joaquín Baruch, Husna Begum, Lucille Blumberg, Debby Bogaert, Fernando Augusto Bozza, Sonja Hjellegjerde Brunvoll, Polina Bugaeva, Aidan Burrell, Denis Butnaru, Roar Bævre-Jensen, Gail Carson, Meera Chand, Barbara Wanjiru Citarella, Sara Clohisey, Marie Connor, Graham S. Cooke, Andrew Dagens, John Arne Dahl, Jo Dalton, Ana da Silva Filipe, Emmanuelle Denis, Thushan de Silva, Pathik Dhangar, Annemarie B. Docherty, Christl A. Donnelly, Thomas Drake, Murray Dryden, Susanne Dudman, Jake Dunning, Anne Margarita Dyrhol-Riise, Linn Margrete Eggesbø, Merete Ellingjord-Dale, Cameron J. Fairfield, Tom Fletcher, Victor Fomin, Robert A. Fowler, Christophe Fraser, Linda Gail Skeie, Carrol Gamble, Michelle Girvan, Petr Glybochko, Christopher A. Green, William Greenhalf, Fiona Griffiths, Matthew Hall, Sophie Halpin, Bato Hammarström, Hayley Hardwick, Ewen M. Harrison, Janet Harrison, Lars Heggelund, Ross Hendry, Rupert Higgins, Antonia Ho, Jan Cato Holter, Peter Horby, Samreen Ijaz, Mette Stausland Istre, Clare Jackson, Waasila Jassat, Synne Jenum, Silje Bakken Jørgensen, Karl Trygve Kalleberg, Christiana Kartsonaki, Seán Keating, Sadie Kelly, Kalynn Kennon, Saye

Khoo, Beathe Kiland Granerud, Anders Benjamin Kildal, Eyrun Floerecke Kjetland, Paul Klenerman, Gry Kloumann Bekken, Stephen R Knight, Andy Law, Jennifer Lee, Gary Leeming, Wei Shen Lim, Andreas Lind, Miles Lunn, Laura Marsh, John Marshall, Colin McArthur, Sarah E. McDonald, Kenneth A. McLean, Alexander J. Mentzer, Laura Merson, Alison M. Meynert, Sarah Moore, Shona C. Moore, Caroline Mudara, Daniel Munblit, Srinivas Murthy, Fredrik Müller, Karl Erik Müller, Nikita Nekliudov, Alistair D Nichol, Mahdad Noursadeghi, Anders Benteson Nygaard, Piero L. Olliaro, Wilna Oosthuyzen, Peter Openshaw, Massimo Palmarini, Carlo Palmieri, Prasan Kumar Panda, Rachael Parke, William A. Paxton, Frank Olav Pettersen, Riinu Pius, Georgios Pollakis, Mark G. Pritchard, Else Quist-Paulsen, Dag Henrik Reikvam, David L. Robertson, Amanda Rojek, Clark D. Russell, Aleksander Rygh Holten, Vanessa Sancho-Shimizu, Egle Saviciute, Janet T. Scott, Malcolm G. Semple, Catherine A. Shaw, Victoria Shaw, Louise Sigfrid, Mahendra Singh, Vegard Skogen, Sue Smith, Lene Bergendal Solberg, Tom Solomon, Shiranee Sriskandan, Trude Steinsvik, Birgitte Stiksrud, David Stuart, Charlotte Summers, Andrey Svistunov, Arne Søraas, Emma C. Thomson, Mathew Thorpe, Ryan S. Thwaites, Peter S Timashev, Kristian Tonby, Lance C.W. Turtle, Anders Tveita, Timothy M. Uyeki, Steve Webb, Jia Wei, Murray Wham, Maria Zambon.