Modeling the relative influence of socio-demographic variables on post-acute COVID-19 quality of life: an application to settings in Europe, Asia, Africa, and South America

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]

[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

Abstract

Long-term COVID-19 complications are a globally pervasive threat, but their plausible social drivers are often not prioritized. Here, we use data from a multinational consortium to quantify the relative contributions of social and clinical factors to differences in quality of life among participants experiencing long COVID and measure the extent to which social variables' impacts can be attributed to clinical intermediates, across diverse contexts. In addition to age, neuropsychological and rheumatological comorbidities, educational attainment, employment status, and female sex were identified as important predictors of long COVID-associated quality of life days (long COVID QALDs). Furthermore, a great majority of their impacts on long COVID QALDs could not be tied to key long COVID-predicting comorbidities, such as asthma, diabetes, hypertension, psychological disorder, and obesity. In Norway, 90% (95% CI: 77%, 100%) of the effect of belonging to the highest versus lowest educational attainment quintile was not attributed to intermediate comorbidity impacts. The same was true for 86% (73%, 100%) of the protective effects of full-time employment versus all other employment status categories (excluding retirement) in the UK and 74% (46%,100%) of the protective effects of full-time employment versus all other employment status categories in a cohort of four middle-income countries (MIC). Of the effects of female sex on long COVID QALDs in Norway, UK, and the MIC cohort, 77% (46%,100%), 73% (52%, 94%), and 84% (62%, 100%) were unexplained by the clinical mediators, respectively. Our findings highlight that socio-economic proxies and sex may be as predictive of long COVID QALDs as commonly emphasized comorbidities and that broader structural determinants likely drive their impacts. Importantly, we outline a multi-method, adaptable causal machine learning approach for evaluating the isolated contributions of social disparities to long COVID quality of life experiences.

Introduction

Long-term COVID-19 sequelae have resulted in a pressing public health crisis since early 2020. A diagnosis of long COVID is defined by the World Health Organization as unexplainable symptoms which persist at least three months after an infection, and occur over the span of two or more months.[1] The widespread presence and impacts of this condition 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), and total disability-adjusted life years (DALYs) for hospitalized patients have been estimated to be as high as 642.8 DALYs/1000 individuals.[3,4] Prior work 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 laboratory measures, severe acute infection, vaccination, age, and sex.[5–14] Conditions that have been consistently identified as key correlates of long-term sequelae include obesity, asthma and other pulmonary diseases, chronic cardiac disease, diabetes, tuberculosis (TB), liver disease, lung disease, and reactivated Epstein-Barr infection.[5–8,10]

Beyond clinical factors, social vulnerabilities are often critical determinants of differential disease burden overall, with such inequities attributed to broader challenges in access to disease prevention and management services and an array of health-limiting exposures, including but not limited to food and housing insecurity, financial discrimination, and air pollution.[15–19]

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.[6,8,10,12,14,20–26] For instance, a 2021 study in the United Kingdom (UK) found that living in high-deprivation settings was associated with both higher and lower odds of symptom persistence, depending on the measure of deprivation index used, 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.[8,21] It is also important to highlight that many of these studies rely on self-reported binary measures of COVID recovery or continued symptomatology, some of which may be subjective classifications, with potential between-group differences in the tendency to report such experiences.

Given this context, we aimed to complement existing efforts centered on uncovering disparities in long-term COVID outcomes, leveraging a large dataset from a prospective, observational, multinational study of hospitalized and non-hospitalized COVID patients with post-infection follow-up data, focusing on Norway, the United Kingdom (UK), India, Brazil, Russia, and South Africa. Specifically, we formally assessed the relationship between a diverse group of biological and social exposures, and our long-term long COVID QoL measure, reasoning that factors like socio-economic status (SES) would matter as much or more than commonly considered comorbidities, as has been recently illustrated for related outcomes, such as "healthy aging".[27] We further evaluated the extent to which clinical intermediates contribute to any observed disparities, hypothesizing that they may only partially explain these effects. Our analysis applies

a similar mediation-centered lens to that of Vahidy et al.[25] and Lu et al.[28] However, rather than focusing on evaluating whether various mediators can independently explain the effects of a given social factor on long COVID risk[25] or on how much social factors mediate disparities in all-cause mortality[28] we measured the degree to which the effects of social variables on long COVID-associated QoL cannot be explained by comorbidities.

Methods

This study uses data from the International Severe Acute Respiratory and emerging Infection Consortium's (ISARIC) multi-cohort consortium.[29] This prospective study across 76 countries collected demographic and illness-associated data during acute SARS-CoV-2 infection, with a subset of sites assessing participants 3 months (+/-1) following infection and 3-6 months thereafter. Recruitment was aimed at clinical settings. Complete details on the study design and recruitment procedures can be found in the published follow-up protocol.[30]

We selected representative locations defined by varying pandemic experiences and demographic/socio-cultural contexts, with data available on indicators of socio-economic status (educational attainment and employment status), resulting in the following individual cohorts: Norway (n=1672) and the UK (n=1064). We combined data from India, Brazil, Russia, and South Africa into one cohort representing upper and lower middle-income countries (MICs), as defined by the World Bank[31]. Sample sizes were insufficient (n<1000) for each of these locations individually (n=264, n=125, n=57, for India, South Africa, and Brazil, respectively), with the exception of Russia (n=1155). Our study incorporates information on two outcome measures capturing long COVID experiences, self-reported continued symptomatology and post-illness quality of life, thereby reducing the influence of between-group reporting patterns that may exist within either of these measures.

Long-term health utility values were obtained using standard quality of life-adjustment estimation procedures, based on subjects' responses to the EQ-5D-5L survey included in the ISARIC follow-up case report forms.[32] The purpose of the EQ-5D-5L survey is to elicit self-reported rankings on five dimensions of health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), representing the intensity of problems experienced along that dimension. All possible rankings for each of the five dimensions were assigned weights, which quantify preferences for different health states. Participants were asked to report these five rankings based on their experiences prior to their COVID infection and in the present, yielding two sets of rankings. Weights were sourced from studies conducted in the country in question or in a related country (see Supplementary Appendix Table S1).

We subset each country cohort to only subjects reporting, at any follow-up survey, at least one long COVID-associated symptom that was not present prior to illness. Utility scores were computed following standard practice and then time-transformed (Supplementary Appendix: Methods). Similar to Sandmann et al.[14], we used a measure of quality-adjusted life days or QALDs, as opposed to quality-adjusted life years (QALYs), given the limited variability across

subjects expected for the latter measure, where utility scores would be simply scaled by one year, i.e., the assumed duration, based on other study findings.[33,34] We additionally chose to focus on quality of life at least three months following infection, or quality of life in the 'present' as recorded in each follow-up survey. We did not consider QALD differences post- versus prior to illness, given that issues with recalling experiences several months in the past are likely to result in erroneous measures of pre-COVID QALDs.

We incorporated data on age, socio-demographic variables (female sex at birth and socio-economic status proxies), a set of clinical comorbidities and risk behaviors, COVID-19 severity, antiviral treatment, and vaccination status (Supplementary Appendix: Methods). For all countries included in our analysis, variables directly reflecting socio-economic status, such as income/wealth or job category are not available, so we generated a series of associated indicators depending on the data available. For Norway, we used quintiles of educational attainment (years). For the UK and the combined MIC cohort, we used information on pre-illness employment status (Supplementary Appendix Table S2).

For a subset of variables with incomplete data, we used multiple imputation by chained equations (MICE) to assign values to missing entries (Supplementary Appendix: Methods). We cannot make a definitive conclusion on the missingness mechanisms for sex at birth, i.e., we have no reason to believe that females compared to males are less likely to self-report their sex, which would result in a missing not at random (MNAR) mechanism. Thus, the use of MICE, which assumes a missing (completely) at random mechanism, appears justified. However, vaccination status and socio-economic indicators of educational attainment and employment status are more likely to be MNAR due to social stigmas and fear of judgment. Nonetheless, a relatively low fraction of most of these values were missing, with the exception of the vaccination status indicator in the UK and combined MIC cohorts and antiviral treatment in the Norwegian and combined MIC cohorts (Supplementary Appendix: Methods). Consequently, to avoid imposing significant bias, we chose not to impute these specific variables and excluded them from their respective variable sets. Finally, as the distribution of sex at birth in our sampled populations non-negligibly deviated from that of the underlying population in the individual cohorts, i.e., Norway and the UK, we conducted sensitivity analyses where subjects were assigned post-stratification weights according to the *raking* method (Supplementary Appendix: Methods).[35–37]

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

For RF #1, we trained a simple random forest treating each variable as individual predictors over an ensemble of decision trees. For RF #2, we manually grouped comorbidities depending on similar clinical categorizations or latent mechanisms (Supplementary Appendix Table S3),

applying multiple factor analysis[38] to the groups to provide a summary measure for each prior to implementing the random forest (Supplementary Appendix: Methods).For RF #3, we applied a combined “variable clustering/variable selection” method introduced by Chavent et al.[39], i.e., Clustering of Variables-Variable Selection Using Random Forest (CoV-VSURF), which enables related variables to be grouped prior to being used as inputs in training a random forest.[39]

To estimate the natural direct effects (NDE) and natural indirect effects (NIE) of binary educational attainment and employment status categorizations or female sex on long COVID QALDs in each cohort, we applied a flexible semi-parametric statistical approach.[40,41] We note that the NDE and NIE are, under well-studied assumptions, interpreted causally, as defined by Pearl[42] and Robins and Greenland[43], although our “exposures” of interest are not directly intervenable as they are social constructs.[44] For Norway, our binary proxy of SES was high/low educational attainment, where ‘high’ encompassed quintiles 3-5 and ‘low’ encompassed quintiles 1 and 2. For the UK and the combined MIC cohorts, our binary proxy of SES was high/low employment status. For the UK, ‘high’ encompassed full-time employment and ‘low’ encompassed all other employment status categories (furloughed, part-time employment, student, and the unemployed), with the exception of retirement. For the combined MIC cohort, ‘high’ encompassed full-time employment and ‘low’ encompassed all other employment status categories, where retirees are included due to notable disparities in representation of full-time employees vs all other categories in this cohort .

Estimates of the NDE and NIE are obtained using targeted maximum likelihood estimation.[40,45] Through this procedure, “nuisance parameters”[40,41,45,46] are first estimated via ensemble machine learning or super learning, then updated in such a way to diminish bias and inform a more correct characterization of uncertainty for the target parameter of interest (Supplementary Appendix: Methods).[40,41] In this context, nuisance parameters include the conditional probability of each of our binary SES proxies or female sex, dependent on confounders; the conditional probability of each of our binary SES proxies or female sex, dependent on confounders and mediators; and the conditional mean of long COVID QALDs dependent on each of our binary SES proxies or female sex, the confounders, and mediators.

To obtain these estimates, we incorporated a multi-model learner consisting of an assortment of parametric and nonparametric approaches (Supplementary Appendix: Methods). The causal structures assumed here are depicted in the Directed Acyclic Graphs (DAGs) presented in Supplementary Appendix Figures S1-2. Mediators were selected following a literature search of studies distinguishing key long COVID predictors[11,12], consistent with the data available for each country, resulting in the following final mediator sets: asthma, chronic cardiac disease (not hypertension), hypertension, chronic pulmonary disease (not asthma), type 1 diabetes (T1D), type 2 diabetes (T2D), diabetes (type not specified), psychological disorder, smoking, and vaccination status (Norway), asthma, chronic cardiac disease (not hypertension), hypertension, chronic pulmonary disease (not asthma), T2D, diabetes (type not specified), psychological disorder, ischemic heart disease, smoking, obesity, and antiviral treatment (UK) and asthma, chronic cardiac disease (not hypertension), hypertension, chronic pulmonary disease (not asthma), T1D, T2D, diabetes (type not specified), smoking and obesity (combined MIC cohort).

Here, the NDE broadly describes the effect of a given social variable on long COVID QALDs not operating through the included mediators while the NIE specifically describes the effect of that social variable on long COVID QALDs through the included mediators. We define the proportion non-mediated as the % contribution of NDE to the total effect (TE), reflecting the share of the TE of the social variable on long COVID QALDs that cannot be explained by the clinical intermediates, consistent with the social disparities conceptual framing provided in Bellavia et al., 2018.[47] For the combined MIC cohort, we additionally conducted country-specific analyses to assess whether the directionality of effects were generally consistent compared to their analogues in the overall analysis.

Data availability statement

The data that underpin this analysis are highly detailed clinical data on individuals hospitalised with COVID-19. Due to the sensitive nature of these data and the associated privacy concerns, they are available via a governed data access mechanism following review of a data access committee. Data can be requested via the IDDO COVID-19 Data Sharing Platform (http://www.iddo.org/covid-19). The Data Access Application, Terms of Access and details of the Data Access Committee are available on the website. Briefly, the requirements for access are a request from a qualified researcher working with a legal entity who have a health and/or research remit; a scientifically valid reason for data access which adheres to appropriate ethical principles.The full terms are at: https://www.iddo.org/document/covid-19-data-access-guidelines. A small subset of sites who contributed data to this analysis have not agreed to pooled data sharing as above. In the case of requiring access to these data, please contact the corresponding author in the first instance who will look to facilitate access.

All code (with the exception of code used to process the individual datasets) is publicly available at: https://github.com/goshgondar2018/social_long_covid.

Results

Combined MIC cohort (India, Brazil, Russia, and South Africa)

The mean age of participants was 58.4 years (SD: 14.4 years). Females outnumbered males, constituting 57% of the cohort. Hypertension was the most frequently reported comorbidity (58%). Estimated long COVID QALDs were greatest in this cohort (median=346; Interquartile range (IQR): 316-365), when compared to the Norway and UK cohorts.

Employment status was markedly skewed towards full-employment (53%) and retirement (35%), with unemployment (7%) and part-time employment (2%) seeing less representation. Carers and students were least represented (2% and 0.3%). Long COVID QALDs were highest among students, full-time employees, and carers, and lowest among those in the retired and unemployed categories. Males reported marginally higher long COVID QALDs than females ($p<0.0001$).

Age dramatically eclipsed all other variables in terms of its predictive power for long COVID QALDs in this cohort, followed by employment status indicators, chronic neurological disorder, and hypertension. Country fell next in the rankings, indicating that underlying cross-country differences in long COVID QALD trends that are not tied to the variables under interest merit consideration, validating our adjustment by country in the subsequent mediation analyses as a source of confounding (Figure 1a). The acute COVID-19 severity indicator was found to be unimportant. RF #2 supported these findings. The cluster containing solely age led the rankings, followed by the principal components (PCs) of the cluster containing socio-demographic variables employment status and sex, and the PCs of the cluster containing psychological disorder and chronic neurological disorder (Figure 1c). Similarly, for RF #3, age, chronic cardiac disease (not hypertension), chronic neurological disorder, as well as dementia, employment status indicators, hypertension, rheumatological disorder, and sex led the set of most frequently selected variables (Figure 1b). However, there were no perceptible differences in variable selection across nearly all predictors, indicating that the CoV-VSURF procedure may not be well-suited for variable rankings in a combined MIC cohort.

While we cannot directly comment on the directionality of associations based on these results, the correlation matrix for all variables in the combined MIC cohort (Supplementary Figure S5) indicates that QALDs were independently negatively associated with age, the acute COVID-19 severity indicator, and all top predicting comorbidities, particularly for chronic neurological disorder, hypertension, and rheumatological disorder. Subjects with those comorbidities present reported lower long COVID QALDs.

Full-time employment was associated with higher long COVID QALDs compared to all other employment status categories. 8.95 (95% CI: 1.82, 16.1) more long COVID QALDs were expected among subjects self-reporting full-time employment compared to all other employment categories due to the direct effect of self-reported full-time employment on long COVID QALDs (the NDE). An additional 4.03 (0.82, 7.24) long COVID QALDs were expected among subjects self-reporting full-time employment compared to all other employment categories, due to the indirect effect of self-reported full-time employment on long COVID QALDs through the intermediates (the NIE). The proportion non-mediated was estimated to be 0.74 (0.46, 1), i.e., 74.2% of the effect of full-time employment versus all other employment status categories on long COVID QALDs could not be explained by the included mediators and must thus be attributed to broader structural factors or other mechanisms. Female sex was associated with lower COVID QALDs, with an estimated NDE of -5.24 (-10.8, 0.33), NIE of -0.85 (-2.59, 0.90), and proportion non-mediated of 0.84 (0.62, 1), indicating that a high fraction of the sex-at-birth effect cannot be tied to mediation by the considered clinical intermediates. We note that the upper and lower bounds of the confidence intervals (CIs) for the proportions mediated for all cohorts are not necessarily exactly aligned with what we expect based on the bounds of the CIs for the corresponding NDEs and NIEs. The relationship between these measures may not hold exactly for computational reasons, because CIs are estimated separately for each of the measures using cross-validation, which may introduce technical noise due to the randomness inherent in sample splitting. Finally, from the country-specific analyses, we observed the same

directionality (positive), albeit at far reduced precision, for the NDE for the binary high SES indicators in all countries as in the combined analysis, but opposing directionality (negative) for the NIE for the binary high SES indicators in India and Brazil (Supplementary Figure S9). Nonetheless, the TE for both these countries was positive overall, suggesting that, as in our main analysis, high SES is associated with increased long COVID QALDs, but that the mediators have an opposing intermediary effect (Supplementary Figure S9). Importantly, for all countries except South Africa, the direct effect comprises the majority of the total effect, although to differing degrees (Supplementary Figure S9). For the female sex 'exposure', we found consistent directionality for all NDEs and NIEs (negative) as in the combined analysis, with the exception of the estimated NDE for South Africa, and NDEs exceeding NIEs in magnitude (Supplementary Figure S9). For Brazil, NDEs and NIEs could not be estimated because of the large discrepancy between males and females (9 versus 84).

Norway

The Norwegian cohort was the youngest, with a mean age of 51.8 years (SD: 13.6 years). Most participants self-identified as white (93%), with non-white race/ethnic minorities notably underrepresented when compared to the approximate race/ethnicity distribution of Norway.[48] As a consequence of this imbalance and sparse numbers in some ethnic groups, we concluded that any assessment of the role of race/ethnicity in shaping long COVID QALDs would not be justified, omitting it from our social variables under consideration. Females constituted a majority of this cohort (68%) compared to males and are thus overrepresented in relation to Norway's underlying population.[49]

The most commonly reported comorbidity was asthma (22%). From the total cohort, 50% reported receiving at least one dose of any COVID-19 vaccine. The median (IQR) value for estimated long COVID QALDs was 345 (313-360).

There was no broadly consistent trend in long COVID QALDs across increasing quintiles of educational attainment, although the lowest mean QALDs occurred in the bottom two quintiles. We observed the greatest differences in long COVID QALDs between, in order of increasing magnitude, quintiles 3 and 1, quintiles 5 and 1, and quintiles 4 and 1. Estimated long COVID QALDs in males slightly exceeded those in females ($p<0.0001$).

Among the leading individual predictors of long COVID QALDs in Norway, anxiety/depression dominated the rankings, followed by educational attainment, rheumatological disorder, and age (Figure 2a). Sex fell lower in the rankings. For RF #2, we observed that 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 2c). 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 2b). Our sensitivity analyses applying the sex-based population corrections

produced nearly identical findings for RF #1 and RF #3 (Supplementary Figure S12a and c). As expected, due to its greater intrinsic variability across model runs, the model-grouped CoV-VSURF procedure provided somewhat contrasting rankings, with variables like T2D and diabetes (type not specified) more frequently selected, and age and sex less frequently selected, than in the main analysis (Supplementary Figure S12b). However, the top-ranking variables – psychological disorder, rheumatological disorder, and chronic neurological disorder – were preserved, and educational attainment indicators retained a similarly high ranking (Supplementary Figure S12b).

QALDs were found to negatively correlate with all top comorbidities and moderately positively correlate with age and the vaccination status indicator (Supplementary Figure S6).

We estimated that falling in the top two quintiles of educational attainment was associated with 12.3 (6.49,18.2) additional long COVID QALDs, on average, via the NDE and 0.67 (-0.98, 2.32) additional long COVID QALDs, on average, via the NIE. The corresponding proportion non-mediated was estimated to be 0.90 (0.77, 1). 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 observed for the quintile 5 vs 1 comparison. However, we note that such pairwise comparisons warrant multiple testing corrections before any inferential claims can be made. A clear negative association was also observed between female sex and long COVID QALDs, with an estimated NDE of -6.79 (-12.8, -0.72), NIE of -3.05 (-5.89, -0.22) and proportion non-mediated of 0.77 (0.46,1). Effect estimates for the NDE, NIE, and proportion non-mediated were nearly equivalent, but with reduced stability, in the sensitivity analysis applying population weights (Supplementary Figure 10).

UK

The UK cohort was skewed towards older adults (mean (SD): 59.0 (12.6) years). Nearly all participants self-identified as white (96%), representing an even greater deviation from the underlying race/ethnicity distribution in the UK.[50] Again, for this reason, we exclude race/ethnicity from our “exposure” variables/predictors for long COVID QALDs. Most were male (59%), which is not reflective of the underlying UK population, as females are the slight majority (51%).[51]

The most commonly reported comorbidity was hypertension (36%). The median (IQR) value for estimated long COVID QALDs was 295 (233, 342), the lowest of all three cohorts.

Employment status was markedly skewed towards full-employment (51%), retirement (30%), part-time employment (10%) and unemployment (7%). The least represented categories were students (0.6%) and the furloughed (0.5%). 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.

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 (Figure 3a). Sex followed in the rankings, which, along with the acute COVID-19 severity indicator, fell among the top ten predictors (Figure 3a). RF #2 also revealed the prime predictive role of employment status and sex as a group, with the PCs of the socio-demographic variables in aggregate dominating other grouped variables, closely followed by the PCs of the group of mental health and neurological disorders (Figure 3c). Age alone ranked highly, even in comparison to grouped factors (Figure 3c). 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 3b). Finally, the population correction sensitivity analysis reported very similar rankings of variables for both RF #1 and RF #3 implementations (Supplementary Figure S13a and c). The same was true for RF #2, where all top ten most selected variables were the same in the main and sensitivity analysis, with the exception of obesity (Supplementary Figure S13b).

QALDs were negatively associated with all major predictive comorbidities, particularly for psychological disorder and chronic neurological disorder, the acute COVID-19 severity indicator, and acute treatment with antivirals (Supplementary Figure S7). In contrast, QALDs were modestly positively associated with age (Supplementary Figure S7).

Increased income/job stability, as proxied by employment status, was consistently associated with increased long COVID QALDs, irrespective of the binary designation used. We found that self-reported full-time employment compared to any other employment status category (excluding retirement) was associated with, on average, 31.7 (14.2, 49.3) higher expected long COVID QALDs, via the NDE, and 4.90 (-0.065, 9.86) higher expected long COVID QALDs, via the NIE. These results yield a proportion non-mediated of 0.86 (0.73, 0.996), i.e., 86% of the effects of full-time employment versus all other employment status categories on long COVID QALDs were unexplained by the mediators in question. We obtained an even stronger relationship between self-reported full-time employment versus unemployment status, with, on average, 79.5 (50.0, 109.0) 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 is thus slightly higher at 0.91 (0.83, 0.98), suggesting that around 91% of the effect of full-time employment versus unemployment on long COVID QALDs cannot be attributed as operating 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: only 73% (52%, 94%) of the effect of female sex on long COVID QALDs were explained by the clinical intermediates. As in Norway, the population correction sensitivity analysis reproduced these findings, but with greater uncertainty (Supplementary Figure S11).

Discussion

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

The data provided compelling evidence for specific categories of pre-existing comorbidities, namely neurological, psychological, and rheumatological, being major predictors of long COVID-associated quality of life. Age also consistently ranked highly as a predictor across the multiple middle and high income countries represented in Europe, Asia, Africa, and South America, although the direction of this association differed by context. However, of note, our evaluated bivariate associations reflect the *unadjusted* role of a given factor of interest on long COVID QALDs and cannot be used to draw any meaningful conclusions about causality. For instance, the negative association observed between antiviral treatment and long COVID QALDs in the UK is likely to be an artifact of confounding by acute disease severity.

Importantly, we observed that socio-demographic variables educational attainment or employment status and sex at birth were generally as or more predictive of long COVID QALDs than the aforementioned comorbidities. Our mediation analyses further suggest that not only are indicators of societal disadvantage highly predictive of lower long COVID QALDs, but also that the impacts of these variables on 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.[20,25,52]

Our study benefited from the use of a sizable and diverse international cohort with data on long-term quality of life following SARS-CoV-2 infection, providing more exhaustive information on post-acute COVID-19 experiences beyond simply whether patients experience long COVID symptoms. Given the sufficiently large sample sizes for our selected cohorts, we were able to apply adaptable machine learning tools, including recent developments in causal machine learning.[40,46] The variable selection methods we used avoid strict and potentially biased modeling assumptions required for commonly used parametric regressions and accommodate inherent variable groupings. Additionally, the causal mediation approaches we applied extend traditional causal inference frameworks[53] through the integration of flexible, but simultaneously relatively precise, model-free regression algorithms [40,46], providing a promising alternative, as aligned with previously outlined frameworks[54], to parametric, model-based approaches.

There are several important limitations of our analysis. First, as our outcome of interest was post-acute COVID-19 long-term quality of life, we were unable to examine the varying roles of the different social and clinical factors on *changes* in quality of life. For future work, it is thus imperative to collect information on quality of life measures at all stages of illness, not simply ex post facto, to avoid issues with recall by positioning readily implementable epidemic study protocols at the outset of an outbreak. Relatedly, we did not have information on subject-specific

duration of long COVID, and instead assumed uniform duration for all participants consistent with examples in the literature on the duration of long COVID.[33,34] Additionally, while participant recruitment was extensive for each of the cohorts considered in this analysis, filtering each dataset to subjects with available demographic, comorbidity, and quality of life data led to sharply reduced sample sizes, especially for MICs. This led to reduced power and the need to combine multiple, highly varying, MIC countries into one cohort, despite potentially profoundly different relationships between each of the social factors and long COVID QALDs in each country. While we found evidence for generally comparable directions of association for the country-specific analyses compared to those of the full combined cohort, which would indicate that there may not be meaningful effect modification by country, some measures did vary in sign and magnitude, notably lower estimated proportions non-mediated, and all results were generally less stable. Consequently, country-specific analyses where sample sizes are sufficient are necessary for disentangling whether these represent real differences by country within the combined cohort, or possible artifacts of small samples. Finally, variables like vaccination status and antiviral treatment were sparsely recorded in some cohorts mainly due to early implementation of the follow-up studies before vaccination was available.

In the Norway cohort, we observed that vaccination status was a very weak predictor of long COVID QALDs, but we note that there could be possible type-specific, dose-specific, or duration-specific effects that we were unable to evaluate here given that the relevant data was nonexistent or insufficient. Thus, the prioritization of both data unification procedures that enable the synchronous collection of the range of variables considered, and the collection of data in less resource-rich/high-attention settings, is crucial.

Finally, it can be concluded that survivorship bias might affect our results, as only subjects who completed a follow-up survey at any follow-up interval can have their quality of life measures recorded. Those lost to follow up due to death from early myocardial infarction, vascular strokes, etc. are also likely to have a reduced quality of life prior to this event.[55] However, no participants in the Norway and UK cohorts died at any point during follow up and only 1.3% of subjects in the combined MIC cohort with follow-up information died by the conclusion of the study.

There were additional limitations related to the type of information we had at our disposal. First, the marked underrepresentation of race/ethnic minorities in the UK and Norway cohorts, and the lack of race/ethnicity data or relevance of this construct in some contexts, such as the Russian and Indian cohorts, prevented us from investigating the impacts of race/ethnicity. It is crucial that future studies seek to recruit more balanced cohorts in this respect, and, where applicable, address constraints with documenting such information. Importantly, in order for a meaningful analysis of race/ethnic risk factors to be conducted in settings where race/ethnicity is a relevant determinant of health, controlled definitions of minority groups must be tailored to and assessed independently in each context. Additionally, while we did have data available for socio-economic proxies, we were limited to specific variables that may not fully reflect participants' levels of socio-economic deprivation and had to apply somewhat arbitrary aggregations of variables due to data and methodological constraints. For future work, it would be useful to both emphasize the collection of more proximate indicators of socio-economic status and consider extensions of

the mediation approaches that allow for categorical exposures. Finally, we were limited to data on sex at birth, which does not capture important gender-based disparities that exist beyond this binary.[56]

Despite these challenges, the central aim of this analysis was to outline a robust statistical and causal-analytic framework, applied to example case studies, for highlighting the contribution of social disparities to chronic ill-health. Our framework can be used both as a standardized comparison of a collective of diverse variables, grouping related factors when necessary, as determinants of worsened post-acute COVID-19 health and well-being and a way to distill the unique impacts of any social variable of interest on these outcomes. Our data highlights the multifactorial relationship between pre-existing risk factors and socio-economic factors and recovery from SARS-CoV-2 infection. As such, we demonstrate that accounting for social vulnerabilities when evaluating determinants of post-acute COVID-19 trajectories is essential and that studies focusing solely on clinical targets may not be sufficient. Consequently, approaches aiming to alleviate social disparities in long COVID recovery by exclusively targeting comorbidity prevention and management are likely to be restricted in their impact. Conversely, transformational societal interventions, that can address disease exposures and access to care, educational, and employment, and other social determinants of health, 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. *Journal of Medical Virology* **94**, 253–262 (2022).
4. Bowe, B., Xie, Y. & Al-Aly, Z. Postacute sequelae of COVID-19 at 2 years. *Nat Med* **29**, 2347–2357 (2023).
5. Cervia, C. *et al.* Immunoglobulin signature predicts risk of post-acute COVID-19 syndrome. *Nat Commun* **13**, 446 (2022).
6. Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nat Med* **27**, 626–631 (2021).
7. Su, Y. *et al.* Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell* **185**, 881-895.e20 (2022).
8. Park, C. *et al.* Short Report on Long COVID. (2021).
9. 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).
10. 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).
11. Tsampasian, V. *et al.* Risk Factors Associated With Post−COVID-19 Condition: A Systematic Review and Meta-analysis. *JAMA Intern Med* **183**, 566 (2023).
12. Subramanian, A. *et al.* Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat Med* **28**, 1706–1714 (2022).
13. Bai, F. *et al.* Female gender is associated with long COVID syndrome: a prospective cohort

study. *Clinical Microbiology and Infection* **28**, 611.e9-611.e16 (2022).

14. 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. *Clinical Infectious Diseases* ciac151 (2022) doi:10.1093/cid/ciac151.
15. 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.
16. 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).
17. Berger, Z., Altiery De Jesus, V., Assoumou, S. A. & Greenhalgh, T. Long COVID and Health Inequities: The Role of Primary Care. *Milbank Quarterly* **99**, 519–541 (2021).
18. Bibbins-Domingo, K. Integrating Social Care Into the Delivery of Health Care. *JAMA* **322**, (2019).
19. Heard-Garris, N. *et al.* Structuring Poverty: How Racism Shapes Child Poverty and Child and Adolescent Health. *Academic Pediatrics* **21**, S108–S116 (2021).
20. 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).
21. 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).
22. Müller, S. A. *et al.* Prevalence and risk factors for long COVID and post-COVID-19 condition in Africa: a systematic review. *The Lancet Global Health* **11**, e1713–e1724 (2023).
23. 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).

24. 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).
25. 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).
26. Nolen, L. T., Mukerji, S. S. & Mejia, N. I. Post-acute neurological consequences of COVID-19: an unequal burden. *Nat Med* **28**, 20–23 (2022).
27. Santamaria-Garcia, H. *et al.* Factors associated with healthy aging in Latin American populations. *Nat Med* **29**, 2248–2258 (2023).
28. 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* e073749 (2023) doi:10.1136/bmj-2022-073749.
29. 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).
30. International Severe Acute Respiratory and emerging infection Consortium. COVID-19 Long term protocol.
31. The World Bank. World Bank Country and Lending Groups.
32. Euroqol. EQ-5D-5L | About. (2021).
33. Mizrahi, B. *et al.* Long covid outcomes at one year after mild SARS-CoV-2 infection: nationwide cohort study. *BMJ* e072529 (2023) doi:10.1136/bmj-2022-072529.
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. *Emerging Microbes & Infections* **12**, 2220578 (2023).
35. Matthew DeBell & Jon A. Krosnick. *Computing Weights for American National Election Study Survey Data*. https://electionstudies.org/wp-content/uploads/2018/04/nes012427.pdf.

36. ReStore National Centre for Research Methods. 5. Adjusting for non-response by weighting. https://www.restore.ac.uk/PEAS/nonresponse.php.
37. Pew Research Center. 1. How different weighting methods work. https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work/#:~:text=With%20raking%2C%20a%20researcher%20chooses,the%20population%20for%20those%20variables (2018).
38. Pages, J. Multiple Factor Analysis: Main Features and Application to Sensory Data. *Revista Colombiana de Estadística* **27**, 1–26 (2004).
39. 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).
40. 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).
41. Hejazi, N., Díaz, I. & Rudolph, K. Efficient causal mediation analysis with the natural and interventional effects. https://code.nimahejazi.org/medoutcon/articles/intro_medoutcon.html (2022).
42. Pearl, J. Direct and Indirect Effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence* (2001).
43. James M. Robins & Greenland, S. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology* **3**, (1992).
44. Díaz, I. Non-agency interventions for causal mediation in the presence of intermediate confounding. *Journal of Royal Statistical Society Series B: Statistical Methodology* **qkad130**,.
45. Díaz, I., Hejazi, N. S., Rudolph, K. E. & Van Der Laan, M. J. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika* **108**, 627–641 (2021).

46. Hejazi, N., Rudolph, K. & Díaz, I. medoutcon: Nonparametric efficient causal mediation analysis with machine learning in R. *JOSS* **7**, 3979 (2022).
47. 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).
48. Statistics Norway. Immigrants and Norwegian-born to immigrant parents. (2023).
49. Statista Research Department. Population in Norway from 2013 to 2023, by gender. (2023).
50. UK Office for National Statistics. Population of England and Wales. (2023).
51. UK Office for National Statistics. Male and female populations. (2023).
52. 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).
53. VanderWeele, T. & Vansteelandt, S. Mediation Analysis with Multiple Mediators. *Epidemiologic Methods* **2**, (2014).
54. Tchetgen Tchetgen, E. J. & Shpitser, I. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Ann. Statist.* **40**, (2012).
55. Smith, L. H. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. *Curr Epidemiol Rep* **7**, 179–189 (2020).
56. US National Center for Health Statistics. Long COVID. (2023).

Main text figures and tables

	Combined MIC cohort (N=1601)	Norway (N=1672)	UK (N=1064)	Overall (N=4337)
AGE				
Mean (SD)	58.4 (14.4)	51.8 (13.6)	59.0 (12.6)	56.0 (14.1)
Median [Min, Max]	58.0 [1.00, 106]	52.0 [17.0, 96.0]	59.0 [19.0, 91.0]	56.0 [1.00, 106]
SEX				
Female	907 (56.7%)	1139 (68.1%)	434 (40.8%)	2480 (57.2%)
Male	694 (43.3%)	533 (31.9%)	630 (59.2%)	1857 (42.8%)
ASTHMA				
ABSENT	1493 (93.3%)	1305 (78.1%)	818 (76.9%)	3616 (83.4%)
PRESENT	108 (6.7%)	367 (21.9%)	246 (23.1%)	721 (16.6%)
CHRONIC CARDIAC DISEASE (NOT HYPERTENSION)				
ABSENT	1300 (81.2%)	1547 (92.5%)	952 (89.5%)	3799 (87.6%)
PRESENT	301 (18.8%)	125 (7.5%)	112 (10.5%)	538 (12.4%)
CHRONIC HAEMATOLOGICAL DISEASE				
ABSENT	1582 (98.8%)	1642 (98.2%)	1032 (97.0%)	4256 (98.1%)
PRESENT	19 (1.2%)	30 (1.8%)	32 (3.0%)	81 (1.9%)
CHRONIC KIDNEY DISEASE				
ABSENT	1520 (94.9%)	1635 (97.8%)	1004 (94.4%)	4159 (95.9%)
PRESENT	81 (5.1%)	37 (2.2%)	60 (5.6%)	178 (4.1%)
CHRONIC NEUROLOGICAL DISORDER				
ABSENT	1530 (95.6%)	1556 (93.1%)	1013 (95.2%)	4099 (94.5%)
PRESENT	71 (4.4%)	116 (6.9%)	51 (4.8%)	238 (5.5%)
CHRONIC PULMONARY DISEASE (NOT ASTHMA)				
ABSENT	1480 (92.4%)	1602 (95.8%)	875 (82.2%)	3957 (91.2%)
PRESENT	121 (7.6%)	70 (4.2%)	189 (17.8%)	380 (8.8%)
DIABETES MELLITUS (TYPE 1)				
ABSENT	1559 (97.4%)	1634 (97.7%)	1050 (98.7%)	4243 (97.8%)
PRESENT	42 (2.6%)	38 (2.3%)	14 (1.3%)	94 (2.2%)
DIABETES MELLITUS (TYPE 2)				
ABSENT	1258 (78.6%)	1542 (92.2%)	846 (79.5%)	3646 (84.1%)
PRESENT	343 (21.4%)	130 (7.8%)	218 (20.5%)	691 (15.9%)
DIABETES MELLITUS (TYPE NOT SPECIFIED)				
ABSENT	1356 (84.7%)	1646 (98.4%)	828 (77.8%)	3830 (88.3%)
PRESENT	245 (15.3%)	26 (1.6%)	236 (22.2%)	507 (11.7%)
HYPERTENSION				
ABSENT	675 (42.2%)	1322 (79.1%)	685 (64.4%)	2682 (61.8%)
PRESENT	926 (57.8%)	350 (20.9%)	379 (35.6%)	1655 (38.2%)
MALIGNANT NEOPLASM				
ABSENT	1539 (96.1%)	1626 (97.2%)	1034 (97.2%)	4199 (96.8%)
PRESENT	62 (3.9%)	46 (2.8%)	30 (2.8%)	138 (3.2%)
PSYCHOLOGICAL DISORDER				
ABSENT	1591 (99.4%)	1484 (88.8%)	897 (84.3%)	3972 (91.6%)
PRESENT	10 (0.6%)	188 (11.2%)	167 (15.7%)	365 (8.4%)
OBESITY				
ABSENT	1291 (80.6%)	1638 (98.0%)	1006 (94.5%)	3935 (90.7%)
PRESENT	310 (19.4%)	34 (2.0%)	58 (5.5%)	402 (9.3%)
RHEUMATOLOGICAL DISORDER				

ABSENT	1528 (95.4%)	1356 (81.1%)	884 (83.1%)	3768 (86.9%)
PRESENT	73 (4.6%)	316 (18.9%)	180 (16.9%)	569 (13.1%)
SMOKING				
ABSENT	1421 (88.8%)	1627 (97.3%)	958 (90.0%)	4006 (92.4%)
PRESENT	180 (11.2%)	45 (2.7%)	106 (10.0%)	331 (7.6%)
Long COVID QALDs				
Mean (SD)	330 (46.4)	322 (59.2)	269 (94.3)	312 (70.3)
Median [Min, Max]	346 [27.0, 365]	345 [-61.0, 365]	295 [-65.3, 365]	336 [-65.3, 365]

Table 1: Summary of demographic variables (excluding SES proxies) and common comorbidities in the final study populations for each cohort, post-missing data imputation. Note the combined middle-income country cohort consists of the following countries: India, Brazil, Russia, and South Africa

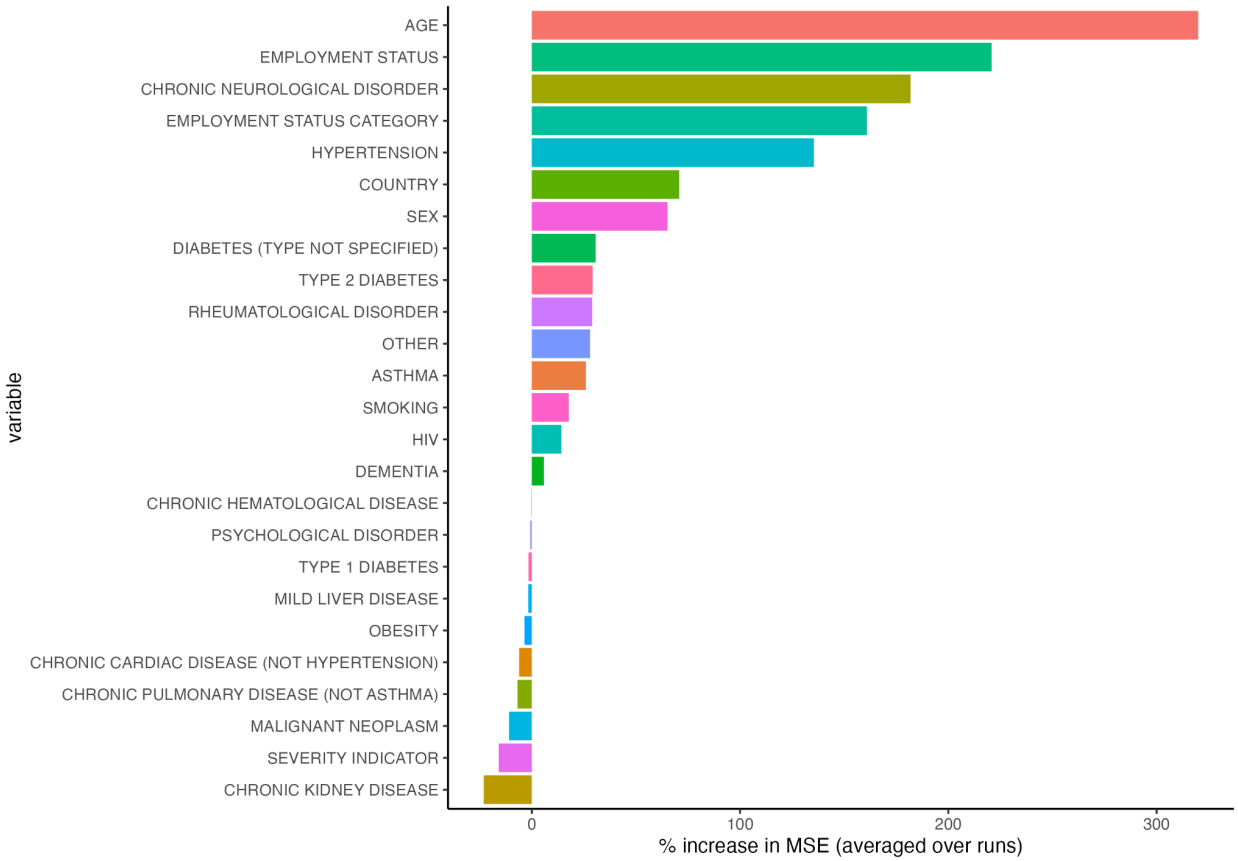

Figure 1a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for the combined middle-income country cohort.

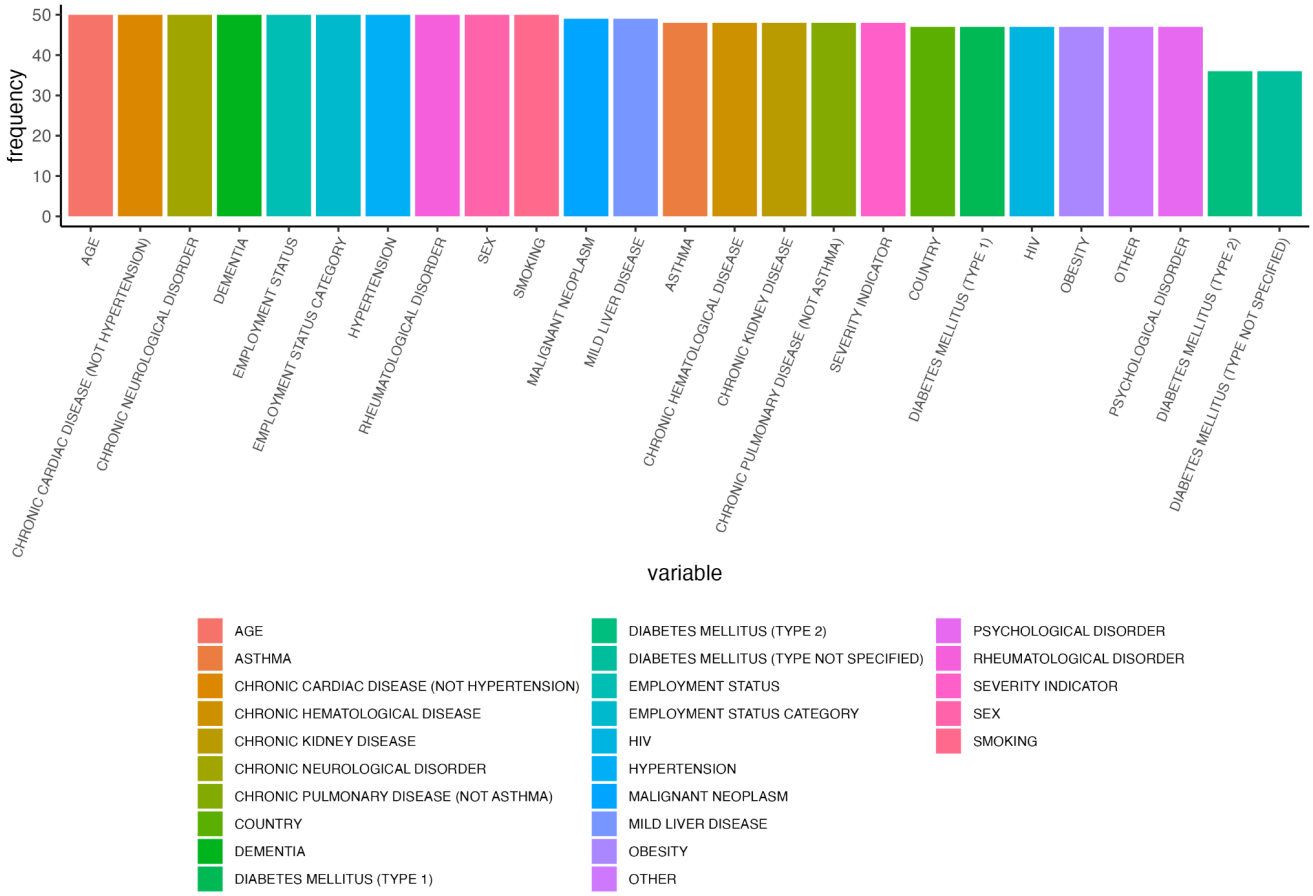

Figure 1b. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for the combined middle-income country cohort.

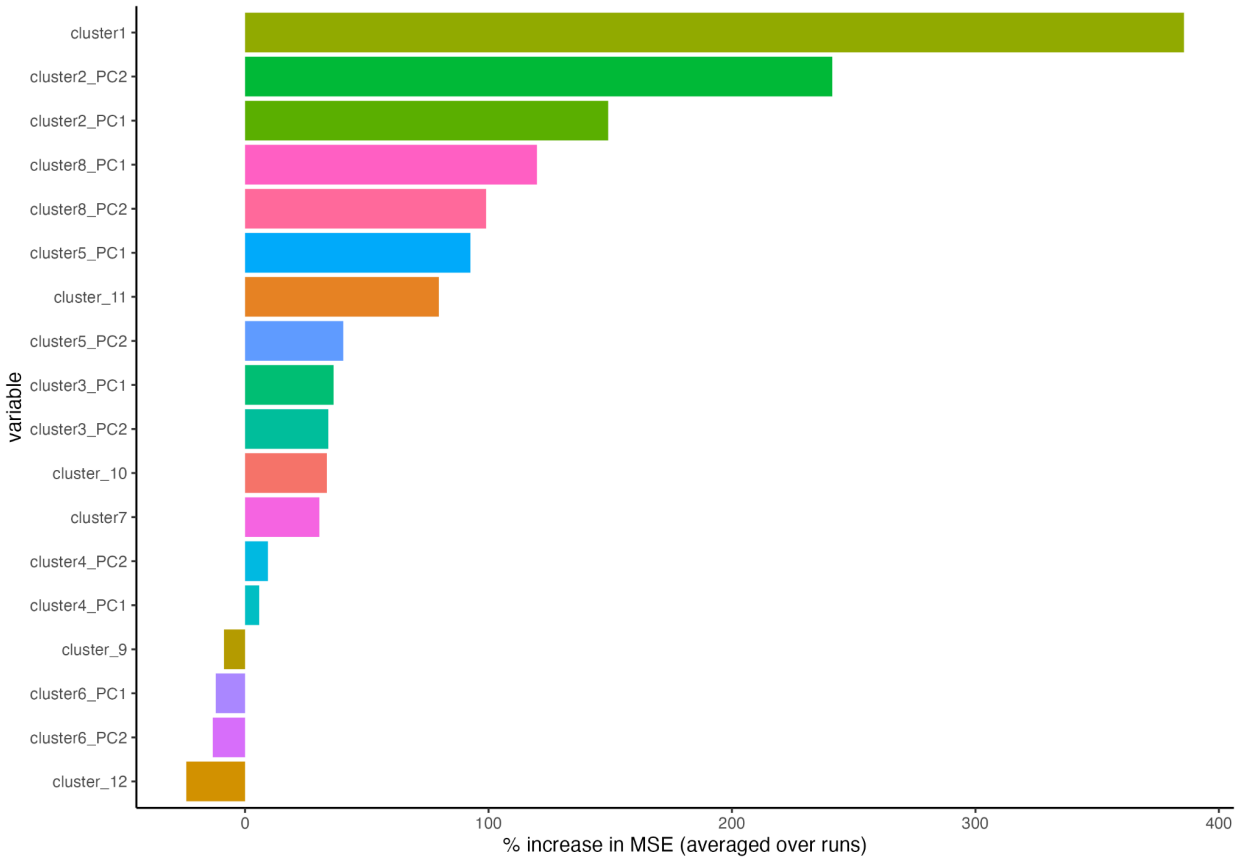

Figure 1c. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for the combined middle-income country cohort. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary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

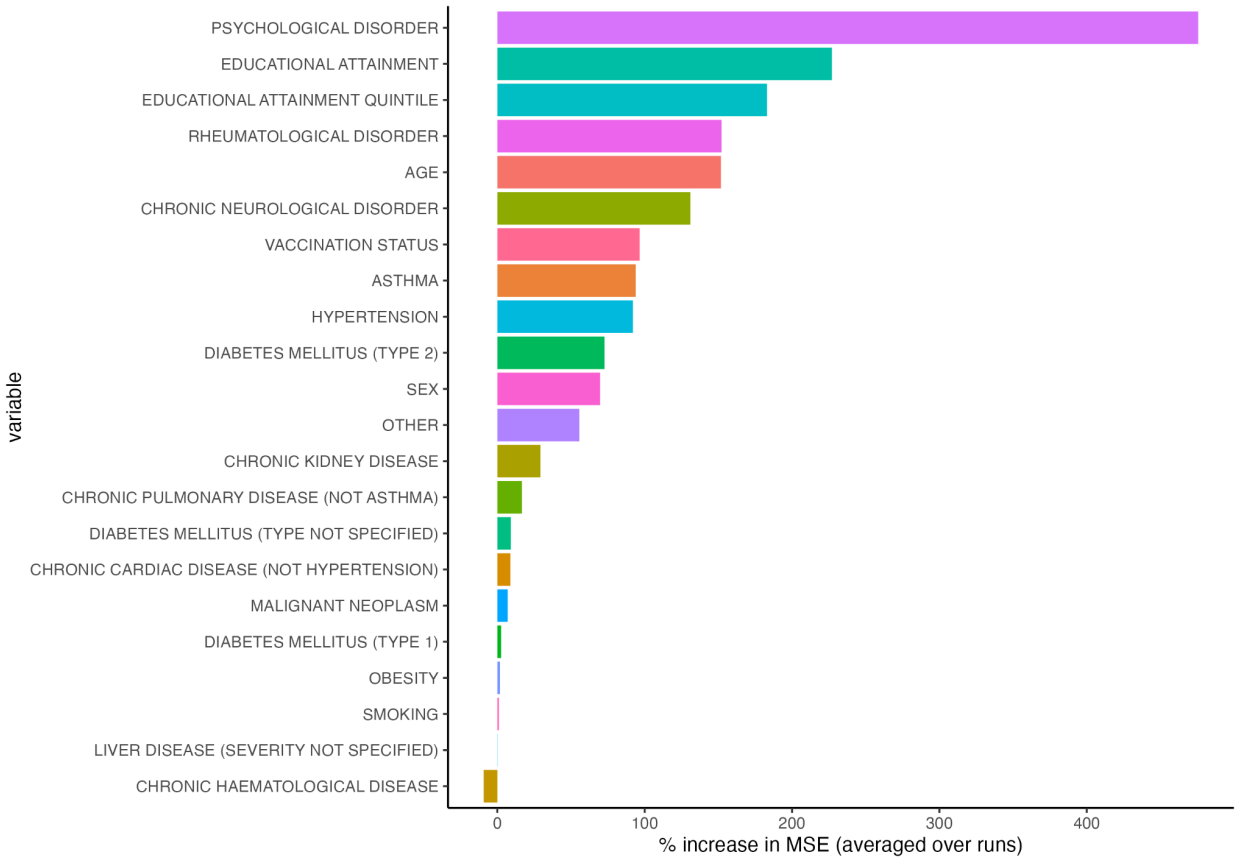

Figure 2a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for Norway.

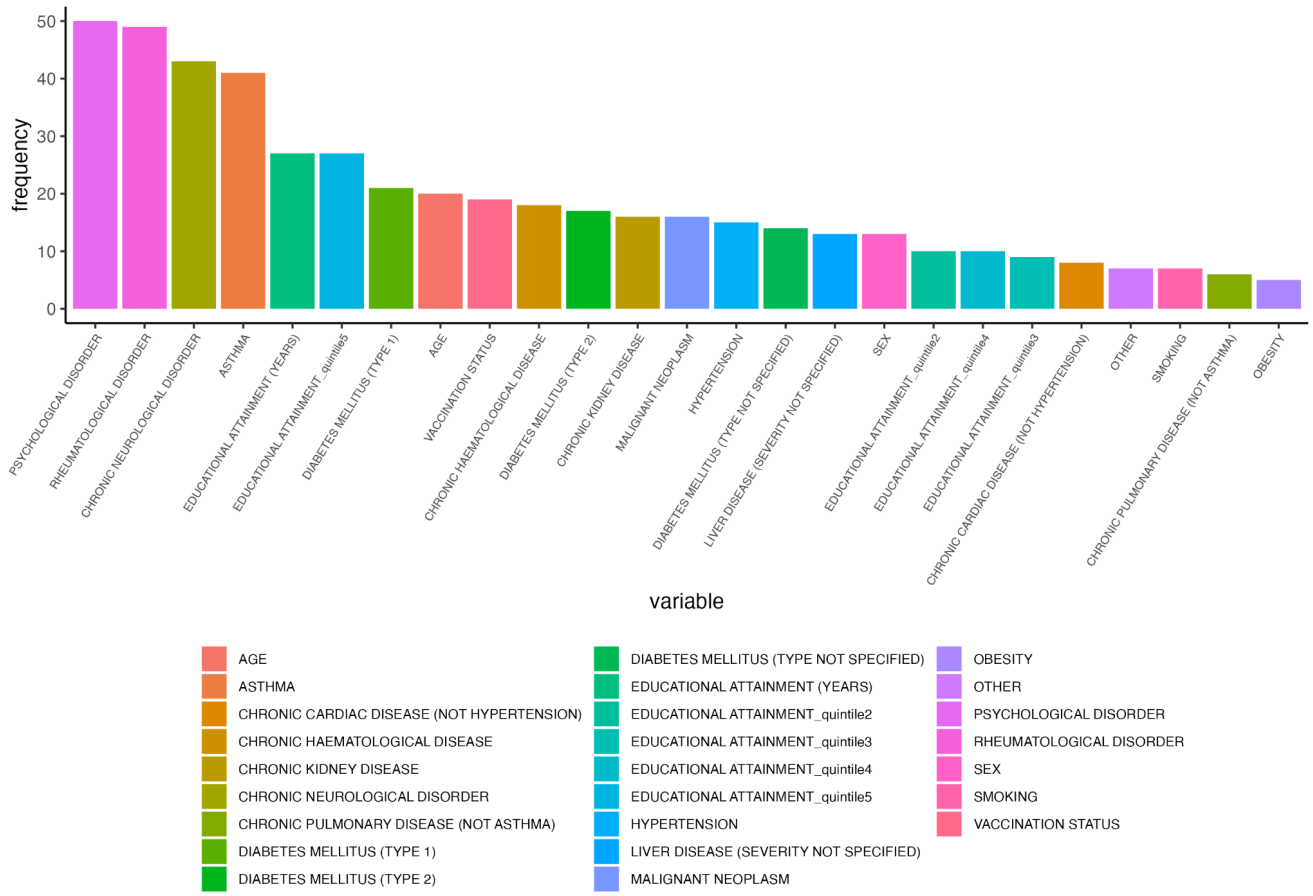

Figure 2b. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for Norway.

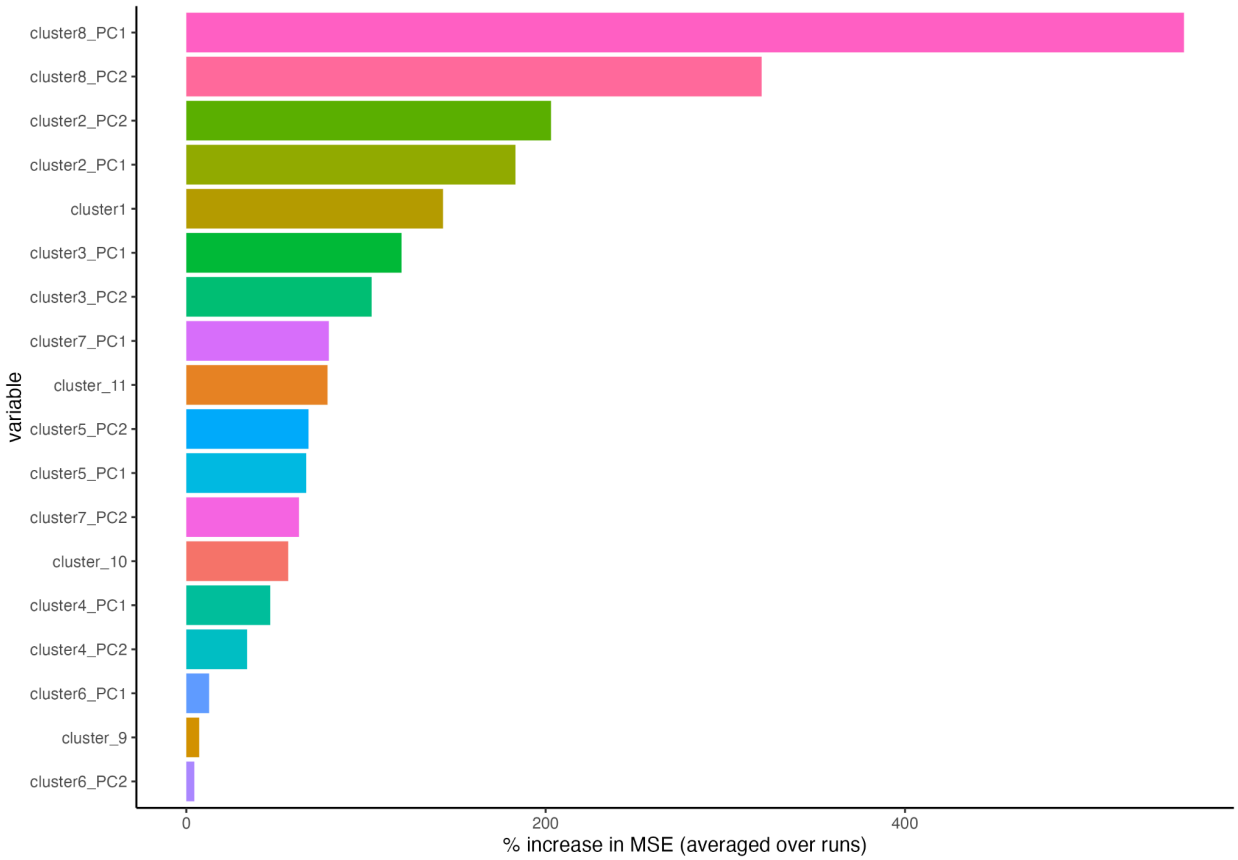

Figure 2c. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for Norway. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding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

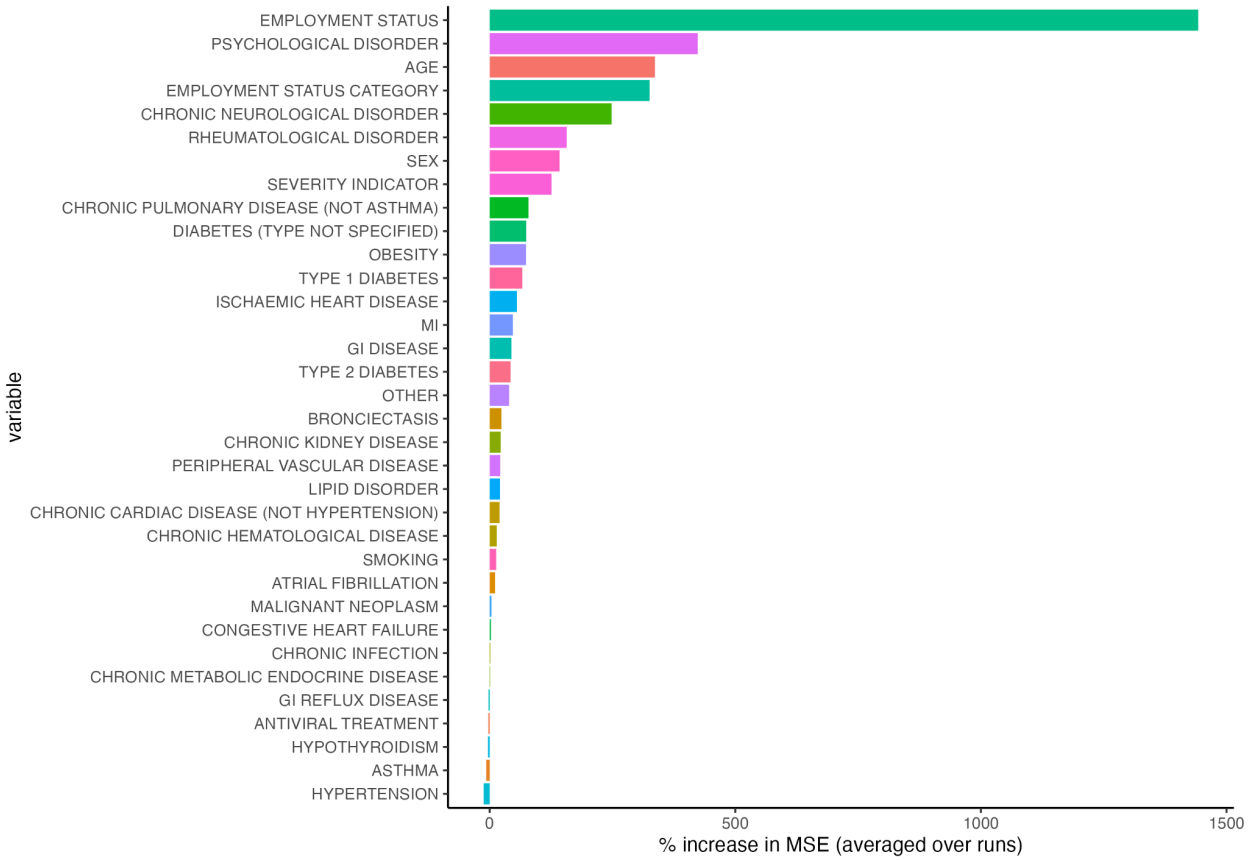

Figure 3a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for the UK.

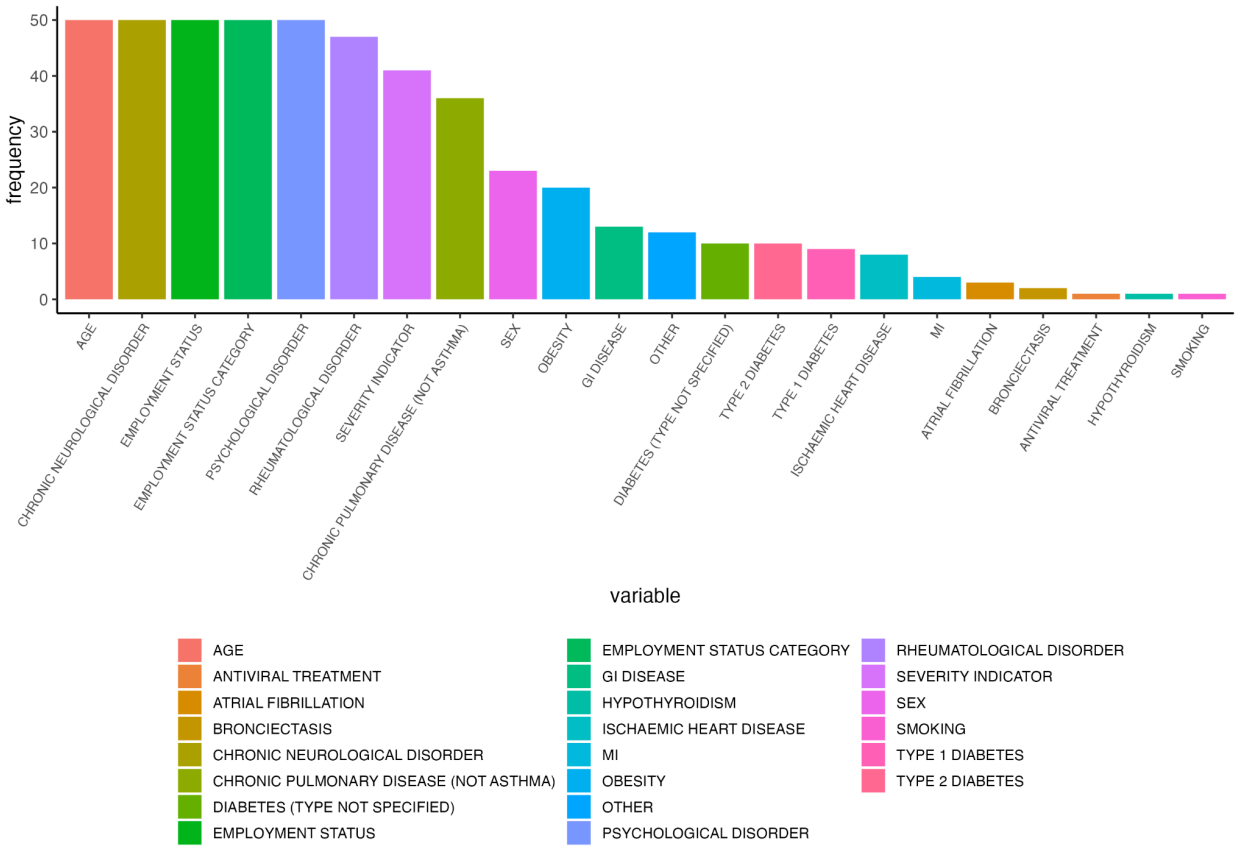

Figure 3b. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for the UK.

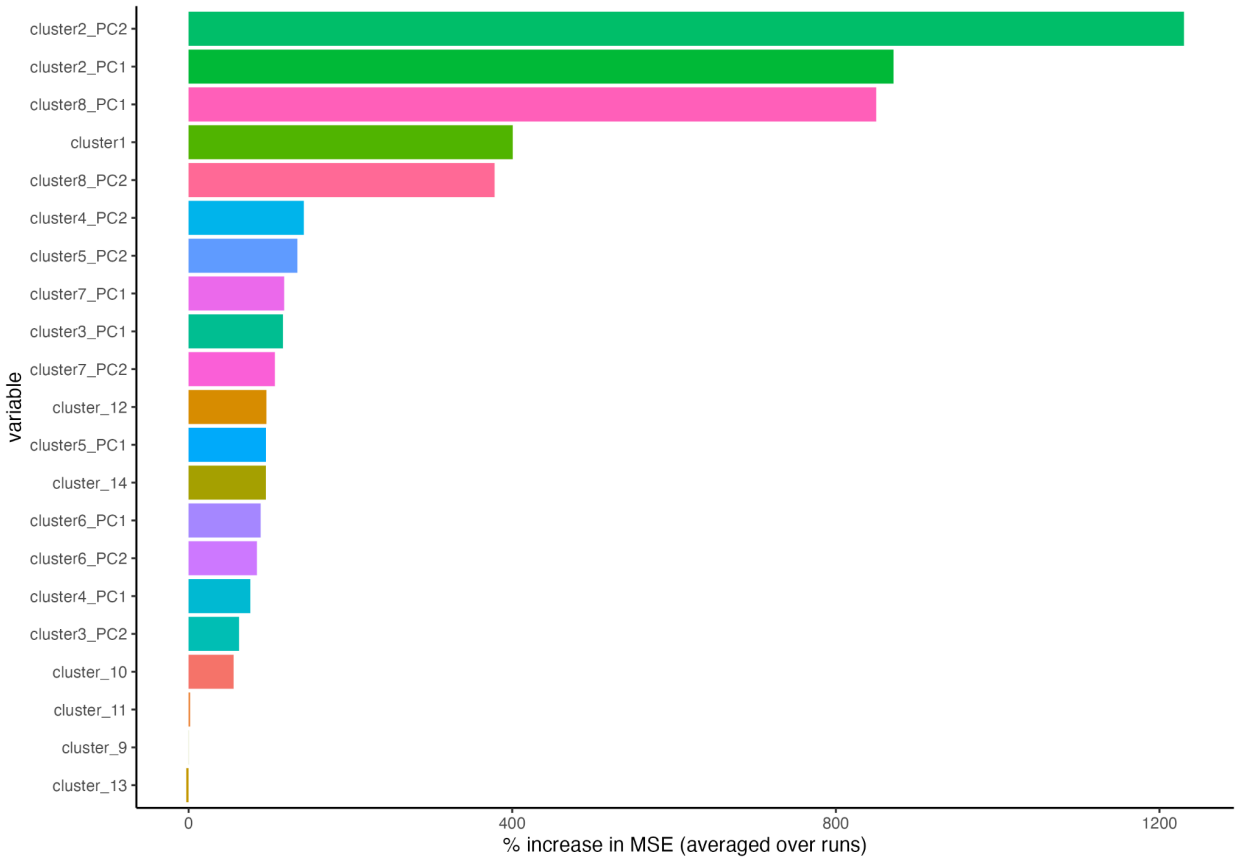

Figure 3c. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for the UK. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding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

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 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) of which PJMO is an NIHR Senior Investigator (NIHR201385); 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 Characterization 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.