Modelling 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

Abstract

Long-term COVID-19 complications are a globally pervasive threat, but their plausible social drivers are often overlooked. Here, using data from a multinational prospective cohort (focusing on Norway, the UK, and Russia), we 1) ranked social and clinical predictors of quality of life (QoL) with long COVID and 2) measured the extent to which clinical intermediates explain any observed relationships between social variables and long COVID QoL. We found that, in addition to age, neuropsychological, and rheumatological comorbidities, educational attainment, employment status, and sex were consistently identified as top predictors of long COVID-associated utility scores, a measure of quality of life. Furthermore, 98.2% (95% CI: 87.3%, 100%) and 89.4% (74.6%, 99.6%) of the adjusted associations between high educational attainment or full-time employment and long COVID utility scores was unexplained by key long COVID-predicting comorbidities in Norway and the UK. The same was true for 84.2% (46%, 100%), 67.6% (45.5%, 89.7%), and 87.2% (71.3%, 100%) of the adjusted associations between female sex and long COVID utility scores in Norway, the UK, and Russia. Thus, socio-economic proxies and sex can be as predictive of long COVID QoL, not fully attributable to commonly emphasized comorbidity pathways, and warrant increased focus in long COVID-targeting efforts.

Introduction

Long-term COVID-19 sequelae, referred to as long COVID, have contributed to a growing public health crisis since 2020. 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 four months after infection (2); further, an estimated 59% of previously infected participants reported a reduced quality of life (QoL) (3,4). 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). Key correlates of long COVID sequelae consistently identified 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 (7,12-20), findings on these relationships have been somewhat mixed (13, 18). 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 and non-significantly correlated 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 how a diverse group of exposures may differentially predict long COVID QoL, reasoning that factors like socio-economic status (SES) would be as much or more critical predictors than important comorbidities, as has been recently illustrated for related outcomes, such as "healthy aging" (21). To further isolate the unique contribution of social determinants of health, we then evaluated the degree by which clinical intermediates may explain social disparities in long COVID QoL, hypothesising that differences in comorbidity risk may not fully account for these disparities. To do so, we apply a similar mediation-centred lens to that of Vahidy et al. (17) and Lu et al. (22). However, rather than independently evaluating the role of various clinical mediators (17) or social factors as mediators (22), we measured how much social variables' associations with long COVID QoL *cannot* be explained by a collective set of comorbidities.

Methods

Data

This study uses data from the International Severe Acute Respiratory and emerging Infection Consortium's (ISARIC) multi-cohort consortium (23). This prospective study across 76 countries collected demographic and medical data during acute SARS-CoV-2 infection from 2020 to 2022, based on hospital admissions case reporting forms, with a subset of sites assessing participants through detailed surveys starting from one-to-three months post-infection and periodically thereafter (23,24). Complete details on the study design and recruitment procedures can be found in the published follow-up protocol (23,24). Within this multi-national cohort, we subset to 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 participants subjects. The countries meeting this criteria were Norway (n=1672), the UK (n=1064), and Russia (n=1155). Our study population consisted of subjects self-reporting one or more long COVID-associated symptoms that were not present prior to illness, at least one month following infection. Health utility values among these subjects were obtained based on their responses to the EQ-5D-5L survey in follow-up forms, eliciting self-reported rankings of the intensity of problems felt for each of five dimensions of health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) (25), based on their experiences prior to their COVID-19 illness and in the present.

Exposures and covariates

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: educational attainment in years for Norway and employment status groups for the UK and Russia. For our descriptive causal mediation analysis, we used estimands based on the natural direct and indirect effects, dichotomizing socioeconomic indicators as follows: high/low educational

attainment (quintiles 3-5 versus quintiles 1 and 2) (Norway) and full-time versus other employment status categories (UK and Russia). Our mediators of interest were selected based on a literature search of studies distinguishing key long COVID predictors (6,7), consistent with the data available for each country. See Supplementary Appendix: Methods for a full list of mediators for each cohort.

Outcomes

We estimated QoL scores among those classified as experiencing long COVID, as defined previously. Composite utility scores for each participant were first computed following standard practice, incorporating relevant country-specific external weights for each EQ-5D-5L response. Utility scores can take any value from -1 to 1, where a value approaching 1 would indicate fewer problems experienced across the EQ-5D-5L survey dimensions, and an overall greater quality of life, while a value approaching 0 would indicate an overall poorer quality of life. Negative utility scores reflect a quality of life state considered “worse-than-dead” (26). Due to the absence of information about subject-specific duration of symptoms, we were unable to compute a quality of life time-adjusted measure without having to make a strict simplifying assumption of equivalent duration across subjects. Furthermore, given that issues with recalling QoL experiences several months in the past are likely to bias measures of pre-COVID QALDs, we did not evaluate any changes in QoL pre- versus post-COVID condition.

Statistical Analysis

For our first aim, we identified social predictors of long COVID utility scores by applying a series of random forest ensemble learners (27) 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 three different approaches to accommodate a range of assumptions about the relationships between variables. For RF#1, we consider each variable as a distinct predictor, whereas for RF #2 and #3, we allow for shared relationships between variables; the former

incorporates subject matter knowledge, while the latter is agnostic to such context. For RF #1 and #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.

For our second aim, probing mechanisms by which social factors relate to long COVID utility scores, we estimated the natural direct effects (NDEs) and natural indirect effects (NIEs) of binary SES proxies or female sex on utility scores in each cohort. Formally, let Z capture the baseline confounders (sex and age for the high SES contrast, and SES indicator for the female sex contrast), A captures the binary exposure category (high SES or female sex), M captures the considered mediators, and Y is the outcome (long COVD utility scores). The NDE is define $E[Y(a,\ M(a'))] - E[Y(a',\ M(a'))]$, while the NIE is defined $E[Y(a,\ M(a))] - E[Y(a,\ M(a'))]$. The NDE and NIE arise from a decomposition of the total effect (TE), or average treatment effect, $E[Y(a)] - E[Y(a')]$, as first described by Robins and Greenland (30). Throughout, we use Y(a) to denote the outcome, possibly contrary to fact, that would have been observed for a given unit, had the exposure taken the value A = a; Y(a') is similarly defined for the alternative exposure contrast A = a'. Analogously, M(a) denotes the mediators under exposure contrast A = a, and M(a') their value under A = a'. Importantly, note that Y(a, M(a)) = Y(a), as the hypothetical intervention A = a connotes that the mediators take value based on the upstream intervention on exposure A. Following Pearl (28), we conceive of these hypothetical interventions for purely descriptive purposes, using counterfactuals to define causal estimands, and the statistical functionals that identify these causal estimands as well-defined measures of association.

Our estimand of interests are the proportion mediated, defined NIE/TE, and the proportion non-mediated, defined NDE/TE. In context, the NDE describes the relationship between a given social variable (exposure) and long COVID utility scores, operating through all pathways excluding the mediators of interest, while the NIE describes this exposure–outcome relationship *through* the mediators. The proportion non-mediated, i.e., NDE/TE, describes the

share of the TE of the social variable on long COVID utility scores that cannot be explained by the clinical intermediates.

To estimate the NDE, NIE, and proportion (non-)mediated, we applied a flexible semi-parametric statistical approach (Supplementary Appendix Methods) (29,30). While our analysis directly provides model-agnostic statistical (i.e., non-causal) targeted estimates, these estimates would admit a causal interpretation if assumptions around consistency, exchangeability, and positivity, which we fully enumerate and detail in context in the Supplementary Appendix Methods for the sake of completeness, could be satisfied (31,32). That is, under these assumptions, the NDE, NIE, and proportion non-mediated act as descriptive measures of how intervening on social factors in the presence of mediators may alter long COVID QoL and how mediators play a role in any observed disparities.

To estimate the NDE, NIE, and proportion mediated, we use targeted maximum likelihood estimation, incorporating an assortment of algorithms to estimate relevant nuisance parameters (see Supplementary Appendix for further details). The super learner ensemble modeling procedure (33), used for nuisance estimation, then pinpoints the combination of weights to assign to each candidate algorithm to minimize the (cross-validated) MSE across all folds in a theoretically optimal manner (29,30,32,33). In all analyses, 95% Wald-style confidence intervals (CIs) were constructed for our measures of interest. Proportions non-mediated are calculated as the complement of estimated proportions mediated. For CIs of the proportion mediated with upper bounds falling below 0 (i.e., upper bound of the CI of the proportion non-mediated exceeding 1), we reported CIs bounded by 1.

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

Results

Norway

The median of long COVID utility scores was 0.944 (interquartile range/IQR: 0.858, 0.986). There was no broadly consistent trend in utility scores across quintiles of educational attainment, although the lowest scores occurred in the bottom two quintiles. The greatest differences in scores occurred between quintiles 4 and 2 and between 4 and 1. Estimated scores 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 utility scores, anxiety/depression ranked first, followed by educational attainment, rheumatological disorder, and age (Figure 1a). When considering pre-grouped variables as predictors (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 clusters identified as important, followed by educational attainment (in years) and a dummy educational attainment indicator for quintile 5 (versus 1) (Figure 1c).

We estimated that falling in the top two quintiles of educational attainment was significantly associated with 0.0324 (0.0186,0.0462) additional units of utility, on average, via the NDE, and 0.000722 (-0.00333, 0.00478) additional units of utility (non-significant), via the NIE, corresponding to a proportion non-mediated of 0.982 (0.873, 1). That is, 98.2% of the adjusted association between high educational attainment on long COVID utility scores 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 numerical reasons, as point estimates and CIs are estimated using cross-validation and separately for each of the measures. We obtained consistent directionality in findings for pairwise comparisons of quintiles 4 and 1 ($p=0.042$), quintiles 4 and 2 ($p<0.001$), quintiles 5 and 1 ($p=0.33$), and quintiles 5 and 2 ($p=0.0067$).

A clear and statistically significant negative association was also observed between female sex and utility scores, with an estimated NDE of -0.0181 (-0.0326, -0.00365), NIE of -0.00411 (-0.0115, 0.0033) and proportion non-mediated of 0.842 (0.463,1).

UK

The median (IQR) of long COVID utility scores was 0.809 (0.637,0.937). Employment status was skewed towards full-employment (51%), retirement (30%), part-time employment (10%) and unemployment (7%). Estimated scores were greatest among participants who reported being furloughed, students, or full-time employees and lowest among those in the unemployed and retired categories. Estimated scores 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 utility scores, followed by anxiety/depression, age, chronic neurological disorder, and rheumatological disorder (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 the rankings, closely followed by the first PC of the group of mental health and neurological disorders (Figure 2b). Age alone was also an influential predictor, even in comparison to grouped factors (Figure 2b). Findings from RF #3 aligned well with these results, with psychological disorder, chronic neurological disorder, employment status indicators being

the most commonly selected variables, followed by age, severity indicator and rheumatological disorder, across key clusters (Figure 2c).

Increased income/job stability, as proxied by employment status, was consistently associated with increased utility scores, irrespective of the binary designation. We found that self-reported full-time employment versus any other employment status category (excluding retirement) was significantly associated with, on average, 0.0908 (0.0434, 0.138) increased units of utility scores, via the NDE, and 0.0139 (0.000804, 0.0271) increased units of utility scores, via the NIE, with a proportion non-mediated of 0.894 (0.746, 0.996). We obtained an even stronger and significant relationship between self-reported full-time employment versus unemployment, with, on average, 0.216 (0.136, 0.297) increased utility scores observed among the full-time employed relative to the unemployed (NDE) and 0.0353 (0.0140, 0.0566) increased utility scores observed among the full-time employed versus unemployed (NIE). The proportion non-mediated was 0.886 (0.804, 0.967), suggesting that 88.6% of the adjusted association between full-time employment versus unemployment on long COVID utility scores does not operate through the considered mediators.

Female sex was associated with lower expected utility scores, with an NDE and NIE of -0.0674 (-0.105, -0.0300) and -0.0247 (-0.0427, -0.00675), respectively. The corresponding proportion non-mediated was the lowest observed among all contrasts, with 67.6% (45.5%, 89.7%) of the adjusted association of female sex and long COVID utility scores being unexplained by the clinical intermediates.

For both the Norway and UK cohorts, we observed consistent directionality and magnitude of estimates of the NDE, NIE, and proportion non-mediated, but with greater variability, in the population correction-sensitivity analysis for both the high SES and female sex contrasts (Supplementary Figures S5B and D and S6B and D). We also obtained a similar ranking of individual variables and grouped variables from all random forest implementations for the population correction-based sensitivity analysis (Supplementary Figures S8 and S9).

Russia

The median (IQR) of long COVID utility scores was 0.966 (0.916,1.0). Full-time employment (55%) and retirement (39%) were the predominant employment status categories. 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%).

Age, followed by employment status indicators, hypertension, and chronic neurological disorder, outranked all other variables in predicting utility scores 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, chronic neurological disorder, employment status indicators, hypertension and dementia led the set of most frequently selected variables (Figure 3c).

Full-time employment was associated with higher long COVID utility scores compared to all other employment status categories via the NDE but lower long COVID utility scores via the NIE, i.e. 0.02085 (0.00344, 0.0383) additional units of utility were expected among subjects self-reporting full-time employment compared to all other employment categories, via the NDE and a decrement of 0.000771 (-0.00384, 0.00229) units of utility were expected among subjects self-reporting full-time employment compared to all other employment categories, via the NIE. The estimated proportion mediated was negative, which corresponds to a proportion non-mediated exceeding 1. This result is intuitive given our positive and negative estimates of the NDE and NIE, indicating that the direct and indirect effects of full-time employment versus all

other employment categories on long COVID utility scores act in opposite directions in this cohort.

Female sex was associated with lower utility scores, with an estimated NDE of -0.0102 (-0.0187, -0.0017) and NIE of -0.00171 (-0.00420, 0.000783), corresponding to a proportion non-mediated of 0.872 (0.713, 1).

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. We addressed this overarching question in two stages, first comparing the predictive role of clinical and social factors on long COVID QoL and subsequently distilling the unique contribution of social factors on long COVID QoL, abstracting away intermediate comorbidity impacts.

In addition to neuropsychological and rheumatological comorbidities and age, we found that educational attainment or employment status, and to a lesser extent sex at birth, were generally as or more predictive of long COVID utility scores. Our mediation analyses further suggest that not only are indicators of social disadvantage notably predictive of lower utility scores, i.e., poorer QoL, but also that the connection between these variables and QoL 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 COVID-19 pandemic (12,17,34).

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 flexible data-adaptive machine learning tools, including recent developments in causal machine learning (28,29). The variable

selection methods we used allow us to avoid relying on restrictive modelling assumptions and further accommodate inherent variable groupings. The causal mediation approaches we applied integrate flexible algorithms to estimate nuisance parameters and semi-parametric theory for best-in-class estimation (29,30), providing an alternative to strictly parametric approaches (35) that place restrictive assumptions about the underlying and unknown data-generating process.

There are several important limitations of our analysis. First, as we noted, we did not have information on subject-specific duration of long COVID, so we could not evaluate a measure of long COVID-adjusted quality of life over the full time span of the condition. 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 QoL between socioeconomic groups and self-reported sex were simply artefacts of baseline QoL differences, we compared pre-COVID QoL scores across groups. We found that for Norway and Russia, cross-group variability in estimated responses by EQ-5D-5L dimensions was generally higher post- versus pre-COVID, based on variability of medians by group by dimension, suggesting that pre-COVID heterogeneity in QoL did not fully drive the differences we observed across groups (Supplementary Appendix). The UK cohort had pre-COVID QoL measures for only 21% of subjects, so we could not make a standardised 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, and tabulating information about duration of symptoms.

It can also be concluded that survivorship bias (36) 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. Furthermore, our study may be limited by the fact that

it self-selects for subjects who were able to contact the healthcare system while they were ill with COVID-19, irrespective of disease severity, and thus misses those who were not seen in a hospital setting. For instance, we may not be effectively accounting for populations who have reduced overall healthcare access who would not have been able to visit a hospital setting during their illness. If such populations generally experienced lower QoL, then the estimates we provided here would provide a lower bound of the relationship between socioeconomic disadvantage and QoL with long COVID.

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 emphasise 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 (37).

Conclusion

We provide a robust statistical framework for highlighting the role of social disparities to chronic ill-health. 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 programmes 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 marginalised communities.

References

1. World Health Organization. Post COVID-19 condition (Long COVID) [Internet]. 2022. Available from: https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition#:~:text=Definition,months%20with%20no%20other%20explanation.
2. O'Mahoney LL, Routen A, Gillies C 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*. 2023;55:101762.
3. Malik P, Patel K, Pinto C et al. Post‐acute COVID‐19 syndrome (PCS) and health‐related quality of life (HRQoL)—A systematic review and meta‐analysis. *J Med Virol.* 2022;94(1):253–62.
4. Sudre CH, Murray B, Varsavsky T et al. Attributes and predictors of long COVID. *Nat Med.* 2021;27(4):626–31.
5. Thompson EJ, Williams DM, Walker AJ et al. Long COVID burden and risk factors in 10 UK longitudinal studies and electronic health records. *Nat Commun*. 2022;13(1):3528.
6. Tsampasian V, Elghazaly H, Chattopadhyay R et al. Risk Factors Associated With Post−COVID-19 Condition: A Systematic Review and Meta-analysis. *JAMA Intern Med*. 2023 Jun;183(6):566.
7. Subramanian A, Nirantharakumar K, Hughes S et al. Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat Med*. 2022;28(8):1706–14.
8. Office of Disease Prevention and Health Promotion. Discrimination [Internet]. Available from: https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-health/interventions-resources/discrimination
9. Jones CP, Jones CY, Perry GS, Barclay G, Jones CA. Addressing the Social determinants of children's Health: A cliff Analogy. *JHCPU*. 2009;20:1–12.

10. Berger Z, Altiery De Jesus V, Assoumou SA, Greenhalgh T. Long COVID and Health Inequities: The Role of Primary Care. *Milbank Quarterly*. 2021;99(2):519–41.
11. Bibbins-Domingo K. Integrating Social Care Into the Delivery of Health Care. *JAMA*. 2019;322(18).
12. Shabnam S, Razieh C, Dambha-Miller H et al. Socioeconomic inequalities of Long COVID: a retrospective population-based cohort study in the United Kingdom. *J R Soc Med*. 2023;116(8):263–73.
13. Hirschtick JL, Titus AR, Slocum E, Power LE, Hirschtick RE, Elliott MR, et al. Population-based estimates of post-acute sequelae of SARS-CoV-2 infection (PASC) prevalence and characteristics. *Clin Infect Dis*. 2021;73(11).
14. Müller SA, Isaaka L, Mumm R et al. Prevalence and risk factors for long COVID and post-COVID-19 condition in Africa: a systematic review. *Lancet Glob Health*. 2023; 11(11):e1713–24.
15. Robinson-Lane SG, Sutton NR, Chubb H et al. Race, Ethnicity, and 60-Day Outcomes After Hospitalization With COVID-19. *JAMDA*. 2021;22(11):2245–50.
16. Naidu S, Shah A, Saigal A 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*. 2021;76:A141.
17. Vahidy FS, Nicolas JC, Meeks JR 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*. 2020;10(8):e039849.
18. Park C, Chaturvedi N, Sterne J et al. Short Report on Long COVID [Internet]. 2021. Available from: https://www.gov.uk/government/publications/ons-short-report-on-long-covid-22-july-2021
19. Xiang, J, Zheng, H, Cai, Y, Chen, S, Wang, Y, Chen, R. Cumulative social disadvantage and its impact on long COVID: insights from a U.S. national survey. *BMC Med*. 2025;23(207).

20. Cheetham, N. et al. Social determinants of recovery from ongoing symptoms following COVID- 19 in two UK longitudinal studies: a prospective cohort study. *BMJ Public Health*. 2025;3:e001166.
21. Santamaria-Garcia H, Sainz-Ballesteros A, Hernandez H et al. Factors associated with healthy aging in Latin American populations. *Nat Med*. 2023;29(9):2248–58.
22. Lu J, Wu C, Zhang X 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*. 2023;19;e073749.
23. ISARIC Clinical Characterization Group, Abbas A, Abdukahil SA et al. ISARIC-COVID-19 dataset: A Prospective, Standardized, Global Dataset of Patients Hospitalized with COVID-19. *Sci Data*. 2022;9(1):454.
24. COVID-19 Long term protocol [Internet]. Available from: https://isaric.org/research/covid-19-clinical-research-resources/covid-19-long-term-follow-up-study/
25. Euroqol. EQ-5D-5L | About [Internet]. 2025. Available from: https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-5l/
26. Liao, M, Yang, Z, Karimi, M, Rand, K, Luo, N. How Do Individuals Value Worse-Than-Dead EQ-5D-5L Health States in Composite Time Trade-Off Tasks? A Qualitative Study. *Value in Health*. 2025;28(4).
27. Breiman L. Random Forests. *Machine Learning*. 2001;45(5–32).
28. Pearl J. Direct and Indirect Effects. *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM Books; 2022.
29. Hejazi NS, Rudolph KE, Van Der Laan MJ, Díaz I. Nonparametric causal mediation analysis for stochastic interventional (in)direct effects. *Biostatistics.* 2023;24(3):686–707.
30. Hejazi N, Rudolph K, Díaz I. medoutcon: Nonparametric efficient causal mediation analysis with machine learning in R. *JOSS*. 2022;7(69):3979.

31. James M. Robins, Greenland S. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*. 1992;3(2).
32. Benkeser D, Díaz I, Carone M. Statistical Learning in Mediation Analysis - Chapter 3: Natural direct and indirect effects [Internet]. 2021. Available from: https://si.biostat.washington.edu/sites/default/files/modules/siscer21mod13_ch3_0.pdf
33. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *sagmb*. 2007;6(1).
34. Qeadan F, VanSant-Webb E, Tingey B, Rogers TN, Brooks E, Mensah NA, et al. Racial disparities in COVID-19 outcomes exist despite comparable Elixhauser comorbidity indices between Blacks, Hispanics, Native Americans, and Whites. Sci Rep. 2021;11(1):8738.
35. VanderWeele T, Vansteelandt S. Mediation Analysis with Multiple Mediators. *Epidemiol Methods.* 2014;2(1):95-115.
36. Smith LH. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. *Curr Epidemiol Rep*. 2020;7(4):179–89.
37. US National Center for Health Statistics. Long COVID [Internet]. 2023. Available from: https://www.cdc.gov/nchs/covid19/pulse/long-covid.htm

Main text figures and tables

	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	1138 (68.1%)	688 (59.6%)	434 (40.8%)	2260 (58.1%)
Male	534 (31.9%)	467 (40.4%)	630 (59.2%)	1631 (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 utility scores				
Mean (SD)	0.882 (0.162)	0.926 (0.134)	0.738 (0.258)	0.855 (0.201)
Median [Min, Max]	0.944 [-0.167, 1.00]	0.966 [-0.590, 1.00]	0.809 [-0.179, 1.00]	0.930 [-0.590, 1.00]

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

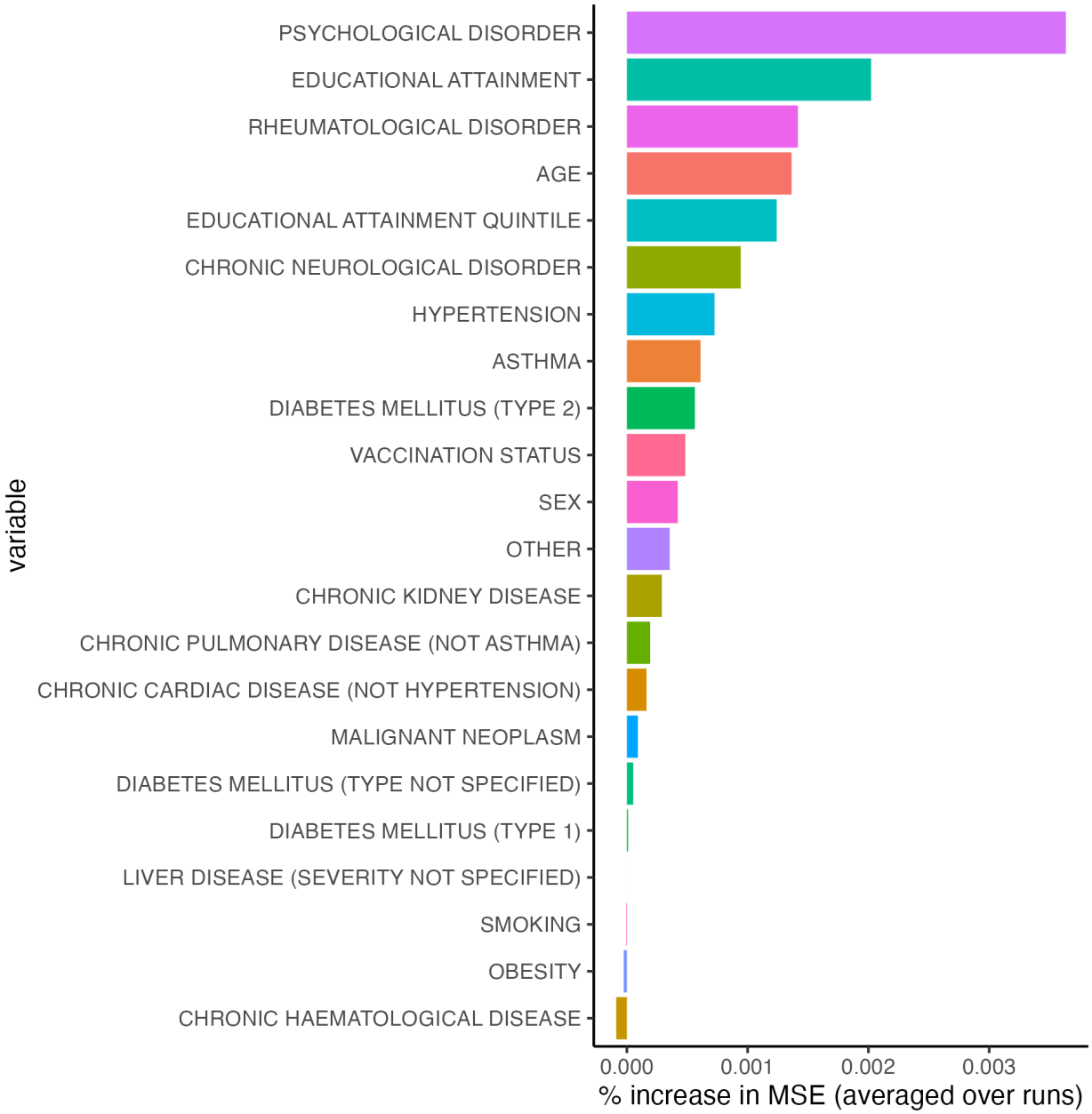

Figure 1a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for Norway. Variables with negative % MSE values are considered unimportant.

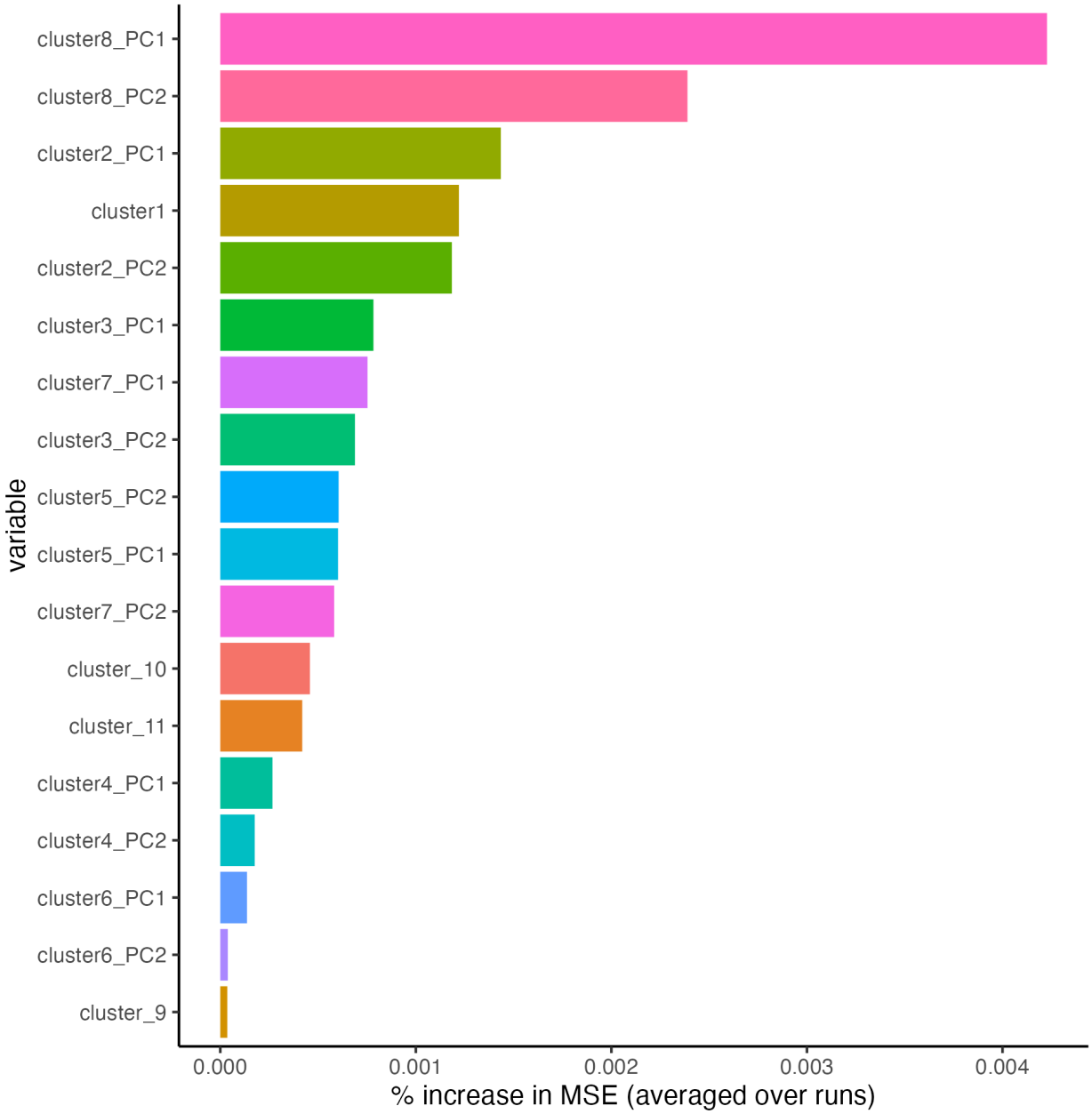

Figure 1b. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for Norway. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding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

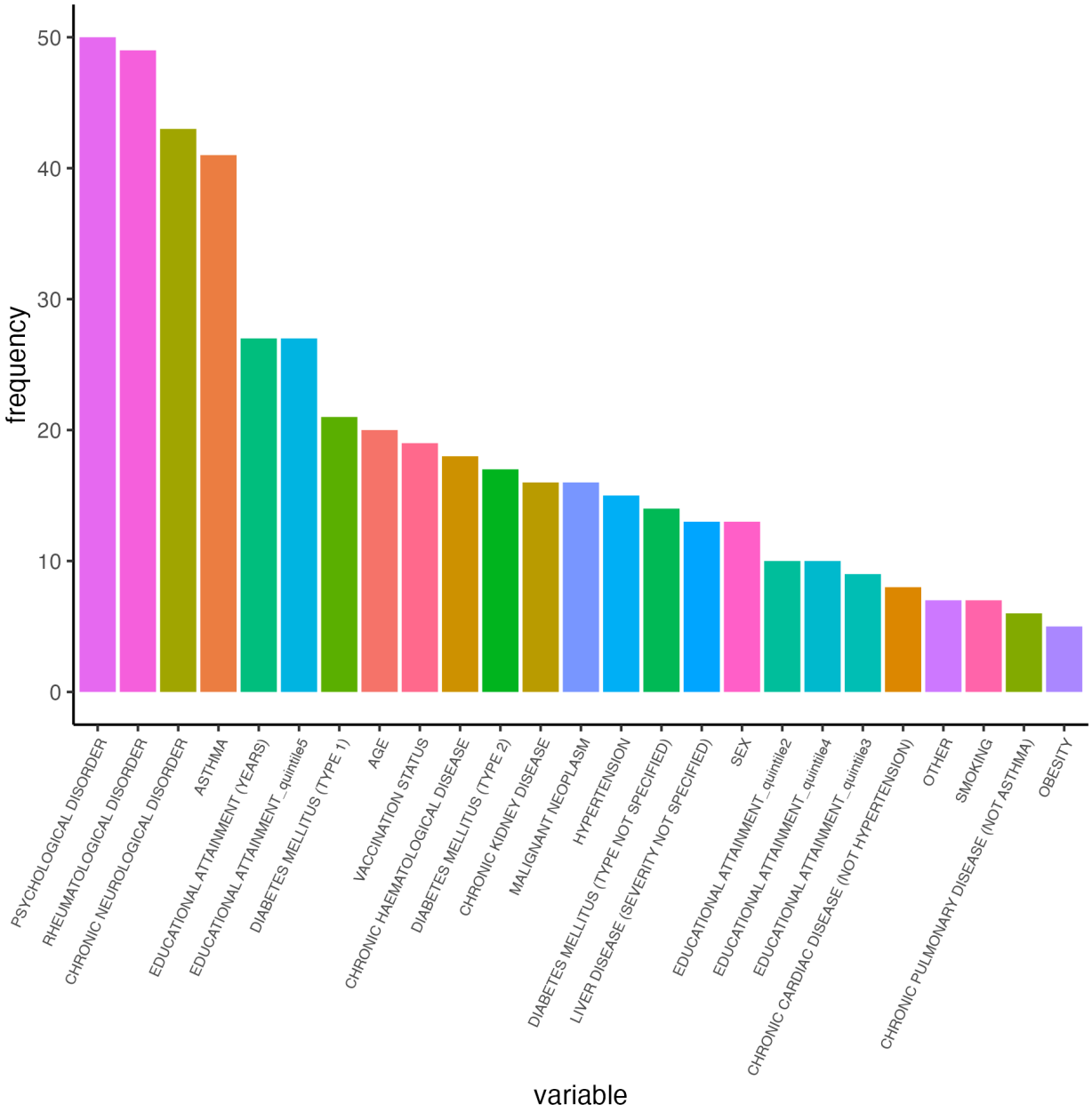

Figure 1c. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for Norway.

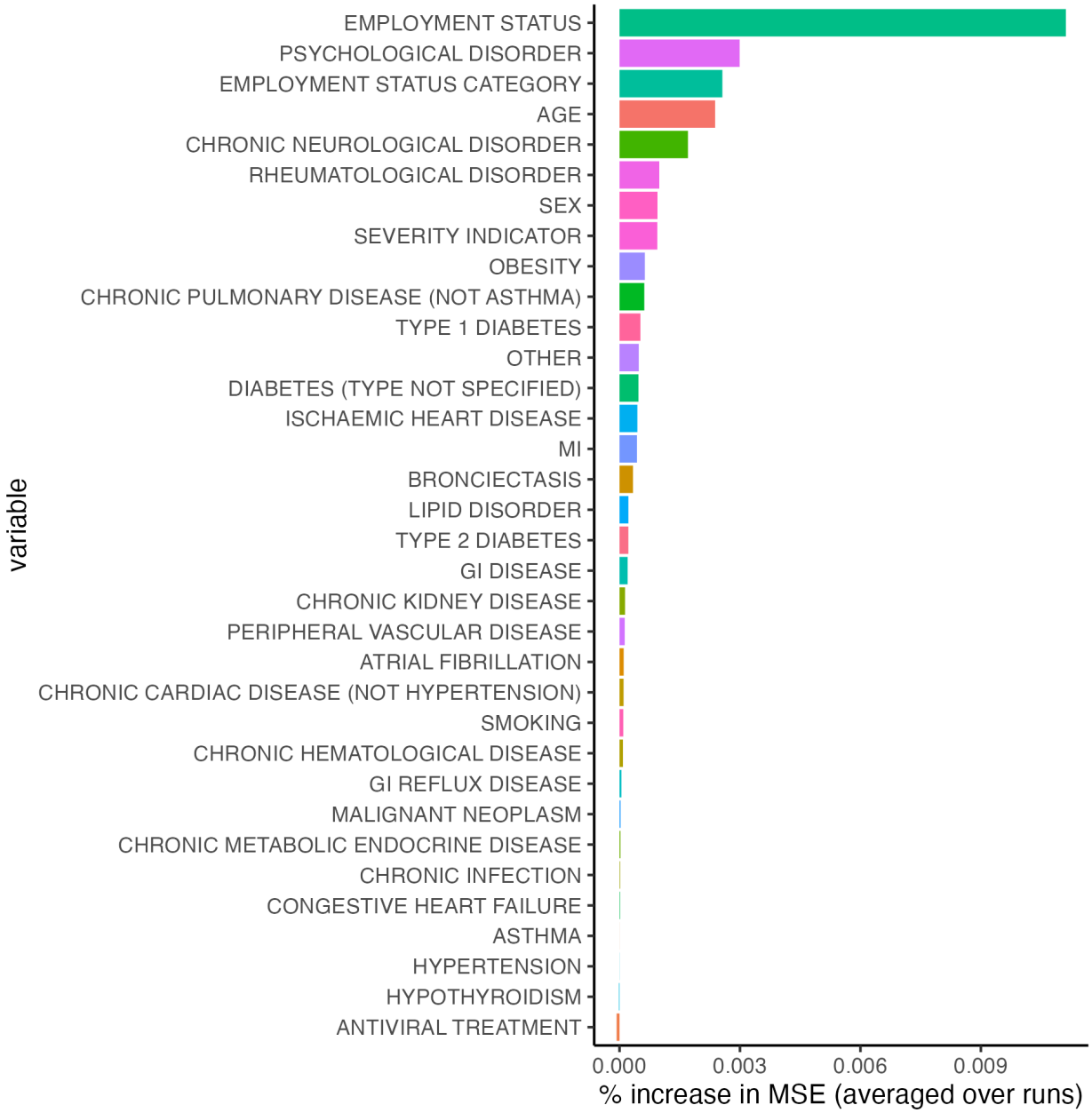

Figure 2a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for the UK. Variables with negative % MSE values are considered unimportant.

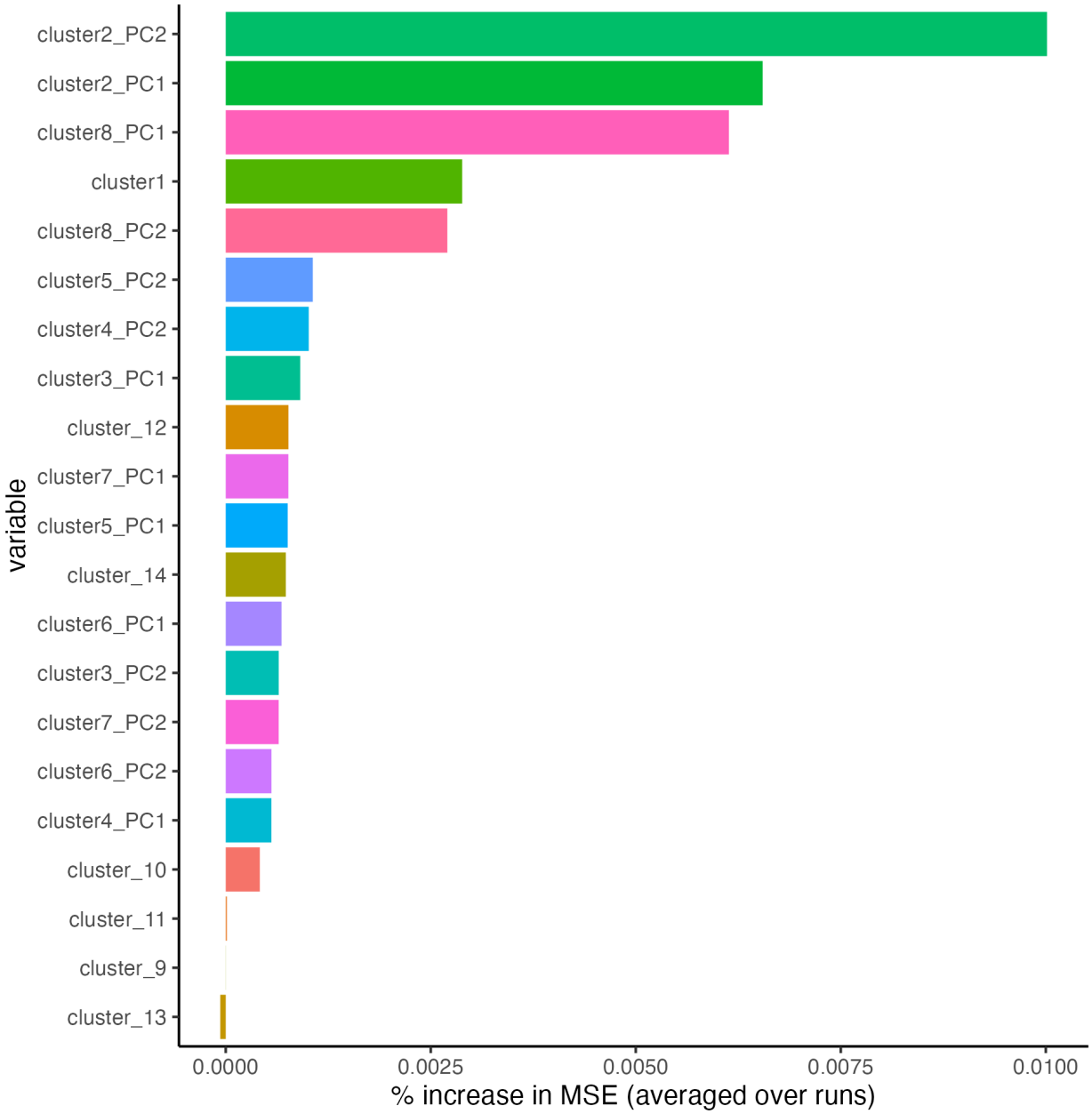

Figure 2b. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for the UK. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding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

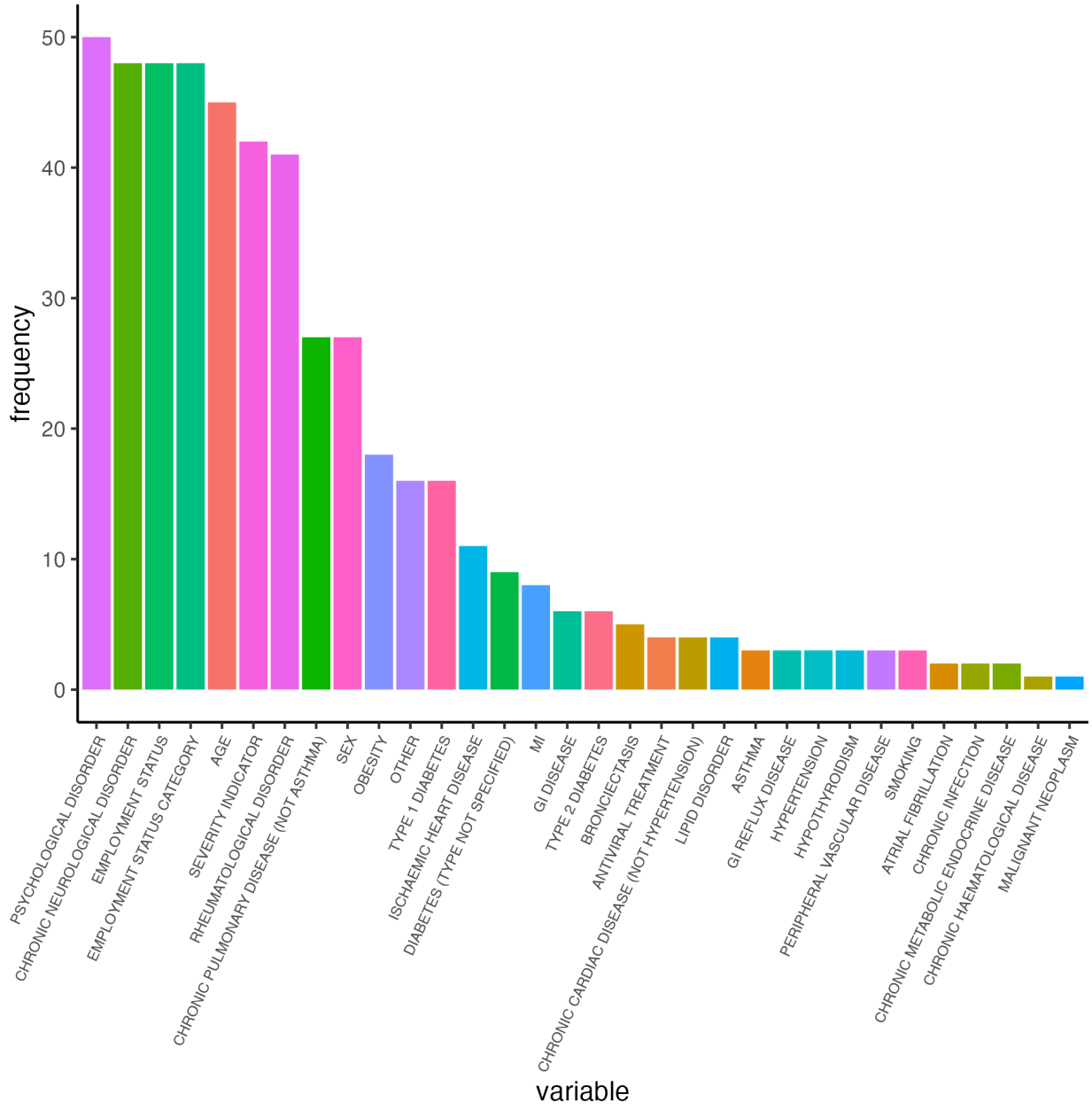

Figure 2c. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for the UK.

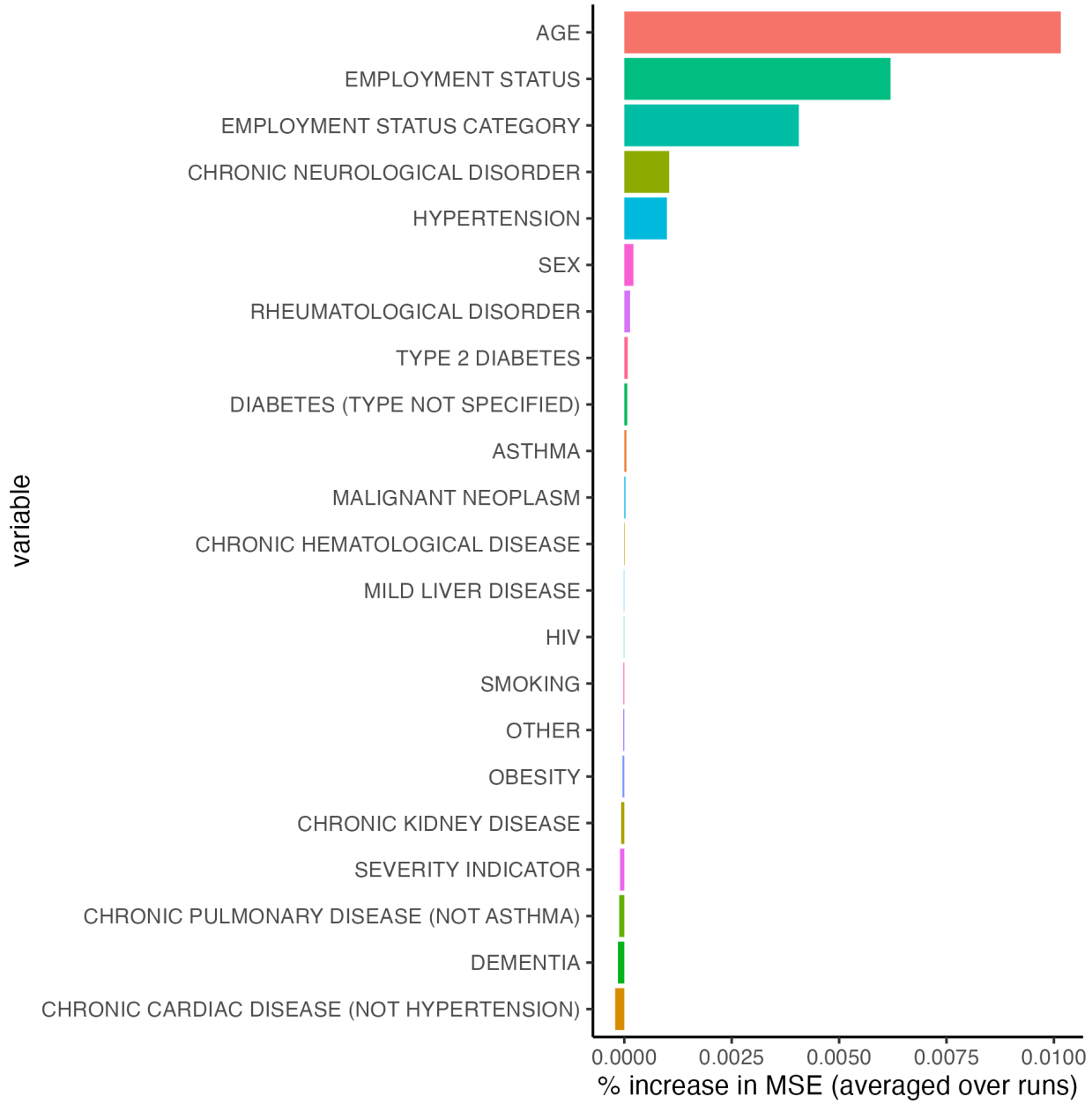

Figure 3a. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from individual random forest implementation (RF #1) for Russia. Variables with negative % MSE values are considered unimportant.

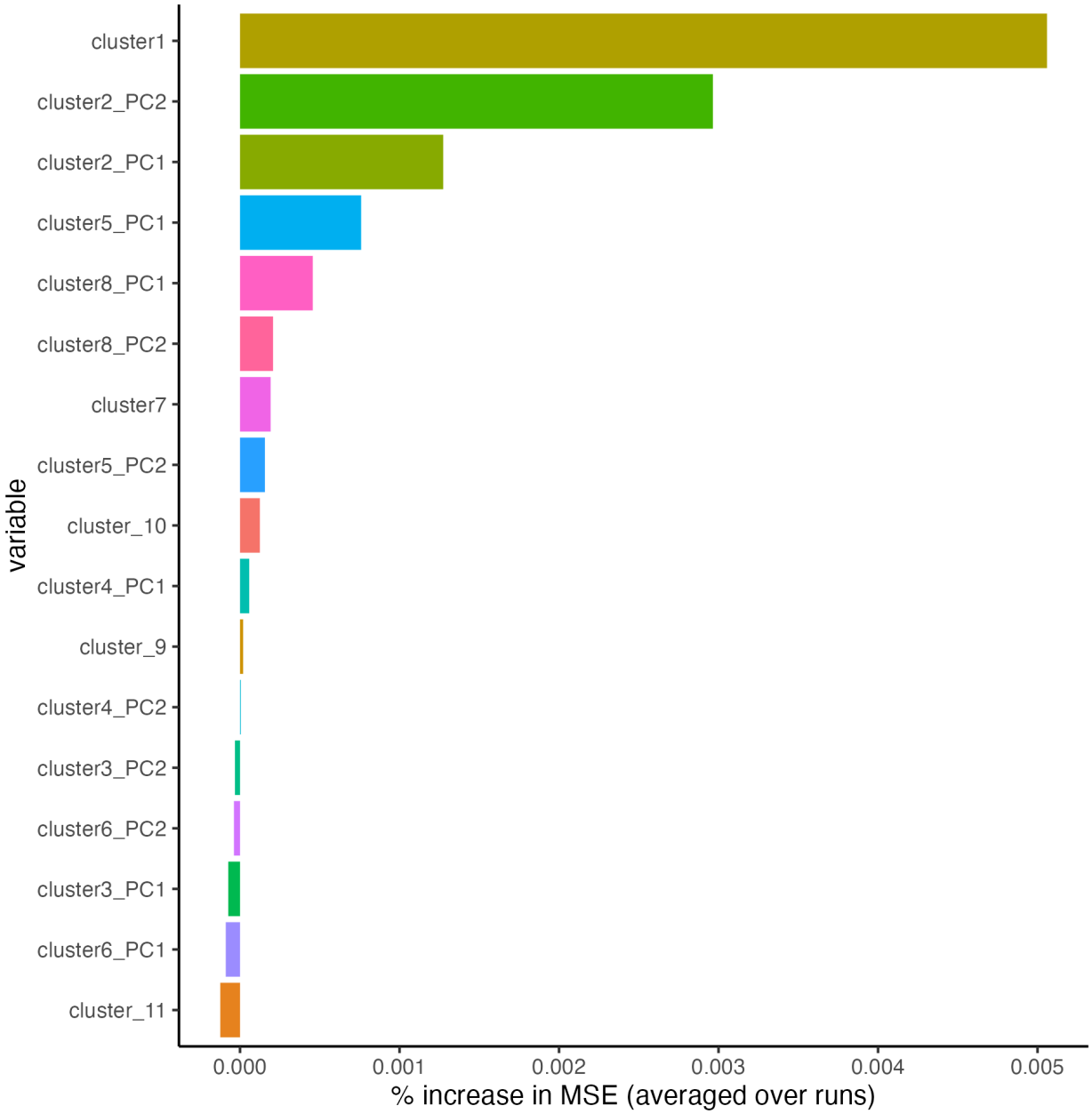

Figure 3b. Estimated variable importance measures, i.e. % increase in mean squared error or MSE, from pre-grouped random forest implementation (RF #2) for Russia. Rows indicate cluster names (a full list of variables belonging to each cluster can be found in Supplementary Table S3) and corresponding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

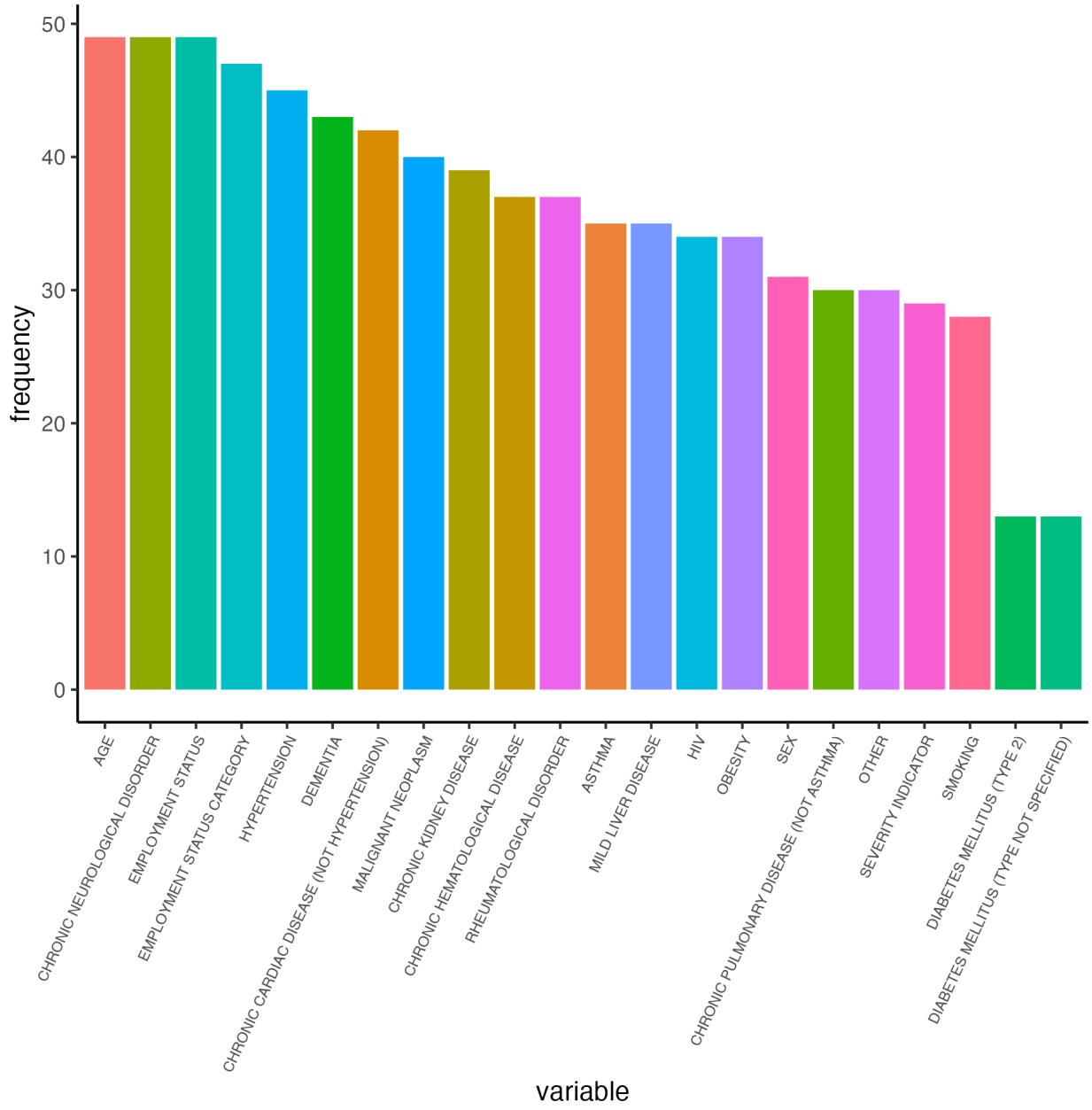

Figure 3c. Number of times (frequency) each variable appears in clusters selected for each CoV-VSURF run (RF #3) for Russia.

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.