University of London
Imperial College of Science, Technology and Medicine
Department of Epidemiology and Biostatistics
School of Public Health

Understanding the COVID-19 pandemic in England

Context, acute symptomatology, and post-acute complications of COVID-19

Matthew Whitaker

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy, February 2023

Declaration of originality

I hereby declare that this thesis represents my original work, and that any work (including data, tables, figures, and text citations) reproduced or adapted from any other sources is duly cited and referenced as such.

Copyright statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

I am extremely grateful to my supervisors Prof Marc Chadeau-Hyam, Prof Paul Elliott and Prof Helen Ward, for their support and encouragement throughout my PhD. Thank you, especially, to Marc, for all your patience, good humour and good ideas.

I'm indebted also to all the other members of the Imperial REACT team – working on the REACT programme was the most intellectually rewarding thing I've ever done and I've learnt a huge amount from collaborating with some extremely smart people, who have been generous with their time and guidance. Thank you to Prof Graham Cooke and Dr Monica Pirani for your advice and suggestions on early drafts of this research, to Josh Elliott for all the stimulating conversation and productive collaborations, and to everyone in Chateau Chadeau for being great people to do research alongside. Thanks also to Alex Menys, for derailing my life in such a productive way.

Finally, I would like to thank my parents, my sister Catherine, and my fiancée Milly, who have all supported me in more ways than I can name. And last but not least, thank you to my dog Billie-Jean for keeping my feet warm during the write-up.

Abstract

The COVID-19 pandemic has represented an unprecedented global health challenge. The Imperial REACT-1 and REACT-2 programmes formed a part of the research response in England, analysing community prevalence of the SARS-CoV-2 virus and antibodies through a series of large cross-sectional samples.

The introductory chapter of this thesis presents a detailed overview of the epidemiological aspects of the COVID-19 pandemic, summarising key research papers and milestones as well as describing the public health response and the development of treatments. The literature on the symptomatology of COVID-19, and of post-COVID syndrome, is reviewed in depth, and the REACT-1 and REACT-2 studies are introduced. Chapter two presents an algorithmic interrogation of the full corpus of COVID-19 literature published between 2020 and 2022, with description of research trends and analysis of different metrics of impact contextualised within all non-COVID-19 literature that has been published since 2019.

Chapters three and four use data from the REACT-1 and REACT-2 programmes to investigate (i) the symptomatology of COVID-19 in the acute phase of this disease and how this has changed over time with different variants and varying levels of population immunity, and (ii) the characteristics, risk factors and sub-types of post-COVID syndrome, where symptoms persist beyond 12 weeks after initial infection.

The thesis concludes with general remarks contextualising the findings and discussing potential and actual extensions of the research.

Contents

Abstract **iii**

List of Figures **ix**

List of Tables **xiv**

Abbreviations and acronyms **xvi**

M Whitaker publications **xvii**

1 Introduction **2**

1.1 The COVID-19 pandemic 2

1.1.1 Public health response 3

1.1.2 COVID-19 pathogenenic mechanisms 4

1.1.3 Risk factors for COVID-19 5

1.1.4 Treatments for COVID-19 7

1.1.5 Vaccines 8

1.1.6 Variants 11

1.2 The COVID-19 Research Landscape 12

1.2.1 Sharing and communicating research 13

1.2.2 Infodemics and bibliometrics . . . 13

1.3 Symptomatology in the acute phase of COVID-19 . . . 15

1.3.1 Using symptoms to identify cases . . . 18

1.4 Post-COVID syndrome ("Long COVID") . . . 20

1.4.1 Causes of Post-COVID syndrome . . . 20

1.4.2 Sub-types of Post-COVID syndrome . . . 22

1.4.3 Risk factors for Post-COVID syndrome . . . 23

1.5 About the REACT studies . . . 25

1.5.1 REACT-1 . . . 25

1.5.2 REACT-2 . . . 26

1.5.3 Real-time reporting from the REACT studies . . . 26

1.6 Analytical tools . . . 29

1.6.1 Classical statistical methods . . . 29

1.6.2 Stability-based approaches . . . 29

1.6.3 Supervised machine learning approaches . . . 30

1.6.4 Unsupervised approaches . . . 32

2 The COVID-19 Research Landscape 33

2.1 Motivation . . . 34

2.2 Web app for data exploration . . . 35

2.3 Methods . . . 36

2.3.1 Data set . . . 36

2.3.2 Statistical analyses . . . 38

2.4 Results . . . 42
2.4.1 Changing publication volumes . . . 44
2.4.2 Evidence of diversion of research output from other areas into COVID-19 46
2.4.3 Citation rates . . . 47
2.4.4 Altmetrics . . . 48
2.4.5 Correlations between Altmetrics and citation metrics . . . 52
2.4.6 Topic modelling . . . 55
2.4.7 Topics as drivers of attention and citation . . . 56
2.4.8 Subsequent publication of preprints . . . 60
2.4.9 Predictors of subsequent high-impact publication of preprints . . . 63
2.5 Discussion . . . 64
2.5.1 Limitations . . . 66
2.6 Conclusion . . . 66

3 Symptomatology in the acute phase of COVID-19 67

3.1 Motivation . . . 67
3.2 Background . . . 68
3.3 Methods . . . 68
3.3.1 Study population . . . 68
3.3.2 Statistical analyses . . . 69
3.3.3 Symptom trajectory . . . 73
3.4 Results . . . 75
3.4.1 Descriptive analysis . . . 75

3.4.2 Logistic regression . . . 76
3.4.3 Multivariable analysis for variable selection . . . 79
3.4.4 Omicron BA.1 vs BA.2 . . . 79
3.4.5 Symptom trajectory . . . 84
3.4.6 Ct values . . . 85
3.5 Discussion . . . 88
3.6 Conclusion . . . 90

4 Persistent symptoms of COVID-19 91

4.1 Motivation . . . 91
4.2 Methods . . . 92
4.2.1 Study participants . . . 92
4.2.2 Data analysis . . . 93
4.3 Results . . . 97
4.3.1 Community prevalence of persistent symptoms . . . 99
4.3.2 Prevalence by symptom . . . 101
4.3.3 Risk factors for persistent symptoms . . . 103
4.3.4 Identifying subtypes of persistent symptoms . . . 108
4.4 Discussion . . . 112
4.5 Conclusion . . . 116

5 Conclusions 117

Bibliography 120

Appendices **145**

A Appendix: The COVID-19 Research Landscape **146**

B Appendix: Symptom profiling **156**

C Appendix: Persistent symptoms **168**

List of Figures

1.1 REACT-1 and REACT-2 timeline of key papers 27

2.1 Dashboard for exploring COVID-19 publications. 35

2.2 Flowchart showing study population and exclusions 42

2.3 Publication rates of COVID-19 papers and non-COVID-19 papers, 2019–2022 . . 45

2.4 Publication rates of COVID-19 papers in the pre-pandemic period (2019) and the pandemic period (2020-2022) 47

2.5 Publication rates, total media attention and average media attention for COVID-19 papers, 2019–2022 50

2.6 Heatmaps showing correlation between media attention, as measured by article Altmetrics score, with article citation metrics and journal impact metrics. 53

2.7 Correlation plots showing correlation between media attention, as measured by article Altmetrics, with article citation metrics and journal impact metrics. . . 54

2.8 Optimisation plot showing average pairwise Jensen-Shannon divergence scores across all topics for a range of values of k. 55

2.9 Line plot showing changes in the relative enrichment of paper abstracts in four key topics, over time. 56

2.10 Univariate analaysis of relationships between article topic and (i) Altmetrics score and (ii) Citation rate 58

2.11 Heatmap showing results of univariable regression of Altmetrics scores onto abstract topics on a rolling 8-week time window throughout the pandemic period 2020–2022, among preprints only. 60

2.12 Kaplan-Meier plots showing time-to-event curves for publication in preprints. . . 61

2.13 Kaplan-Meier plots showing time-to-event curves for publication in preprints from the first wave of COVID-19 (to August 2020). 62

2.14 Panel plot showing multivariable feature importance and model evaluation performance when predicting subsequent publication in a high impact factor (IF $\geq$ 10) journal, in a sample of N=15,336 preprints from 2020. 63

3.1 Flowchart showing study population for symptom profiling, plus exclusions and variant transitions in the UK, based on UK sequencing data from GISAID. . . . 76

3.2 Comparison of ORs for swab positivity based on presence or absence of any of 26 symptoms surveyed in N=1,542,510 participants across five variant-phases of REACT-1. 77

3.3 Results of LASSO stability selection proportions with each of 26 symptoms as predictors and swab positive/negative as the binary outcome variable, for five SARS-CoV-2 variants in England. 79

3.4 Odds ratios for infection with BA.2 vs BA.1 among swab-positive respondents . 80

3.5 Modelled probability of each of eight common symptoms at N days' remove from first symptom onset. 85

3.6 Plots showing: (i) Ct values in N=6,395 swab-positive participants with Omicron BA.2 and BA.1 variants in REACT-1 rounds 17–19, stratified by symptom status (presence of any of 26 surveyed symptoms) 87

4.1 A flowchart of the primary and replication study populations, showing exclusions and proportion of participants experiencing persistent symptoms at least 12 weeks after symptom onset, and study population for estimation of background symptom rates. 98

4.2 Individual symptom prevalence in rounds 3–5 (September 2020 – February 2021) and separately in round 6 (May 2021) 102

4.3 Logistic regression models with one or more symptoms at 12 weeks (y/n) as the binary outcome variable, both adjusted for age-sex and mutually adjusted. . . . 104

4.4 Mean contribution to area under the curve (AUC) that each variable made to a multivariable boosted tree model 105

4.5 Plots showing modelled probability of persistent symptoms at 12 weeks as a function of age and sex, using generalised additive models with splines on age and interactions between age and sex 107

4.6 Results of clustering analysis using the CLARA PAM (Partitioning around medoids) algorithm 109

4.7 Results of Latent Class Analysis (LCA) on symptom profile at 12 weeks in rounds 3–5 110

4.8 Prevalence of symptoms persisting for 11 or more days in 1,879,842 PCR-negative respondents to REACT-1 111

A.1 Treemap plot showing journal fields among non-COVID-19 papers (2019–2022). 147

A.2 Treemap plot showing journal fields among COVID-19 papers (2020–2022). . . . 148

A.3 Accumulation velocity of social media Altmetrics scores for the 20 COVID-19 papers with the highest Altmetrics scores 149

A.4 Publication rates of non-COVID-19 medical papers in the pre-pandemic period (2019) and the pandemic period (2020-2022) 150

A.5 Heatmap showing results of univariable regression of Altmetrics scores onto abstract topics on a rolling 8-week time window throughout the pandemic period 2020–2022, in journal articles only (excluding preprints). 155

B.1 Sensitivity analysis of the univariable symptom profiling among N=1,542,510 participants, additionally adjusted for time since symptom onset 163

B.2 Pooled logistic regression analysis of N=1,542,510 participants across five variant-phases 164

B.3 Scatter plot showing N-gene Ct values among symptomatic swab-positive individuals plotted against time (in days) since symptom onset, for BA.1 and BA.2 . 165

B.4 Heatmaps showing co-occurrence of symptoms across variants among swab-positive individuals 166

B.5 Heatmap showing co-occurrence of first symptoms among swab positive individuals in rounds 17–19 167

C.1 Results of clustering analysis in the replication cohort (round 6) 174

List of Tables

1.1 REACT-1 invitations and response rates 28

1.2 REACT-2 invitations and response rates 28

1.3 Summary of sample extraction and fieldwork dates, and response rates, by round, for REACT-1 (top) and REACT-2 (bottom) 28

2.1 Confusion matrix showing the results of algorithmic detection of COVID-19 papers vs Dimensions COVID-19 inclusion criteria vs manual assessment of paper topic (COVID-19 / not COVID-19) 37

2.2 Summary of all papers indexed in the Dimensions database in 2019, 2020, 2021 and 2022 43

2.3 Top 20 most highly cited COVID-19 papers published between 1 January 2020 and 31 December 2022. 51

2.4 Top 20 COVID-19 papers ranked by Altmetrics score, published between 1 January 2020 and 31 December 2022. 51

2.5 Top 20 COVID-19 papers, ranked by citation count (top) and Altmetrics score (bottom) 51

3.1 Odds ratios for swab positivity based on presence or absence of any of 26 symptoms surveyed in N=1,542,510 participants across selected rounds of REACT-1. 78

3.2 Odds ratios generated by logistic regression of the response to the question on symptom severity 82

3.3 Results from log-linear regression of symptom count as a function of BA.1 / BA.1 infection and covariates 83

4.1 Proportions of respondents in (i) rounds 3–5 and (ii) round 6 who reported one or more (or three or more) symptoms 12 weeks after initial symptom onset. . . . 100

A.1 Data points available for each article, with sources 151

A.2 140 topics identified by topic modelling of abstracts from 486,030 COVID-19 papers published in 2020-2022 154

B.1 Characteristics of study population 159

B.2 Characteristics of swab positives in study population 162

C.1 Numbers and weighted prevalence of current persistent symptoms in the main study population (rounds 3–5) 170

C.2 Numbers and weighted prevalence of current or past persistent symptoms in the main study population (rounds 3–5) 172

C.3 Odds ratios for logistic regression models with persistent symptoms y/n as the binary outcome variable. 173

Abbreviations and acronyms

ARDS Acute Respiratory Distress Syndrome

BMI Body Mass Index

CFR Case Fatality Rate

CCSR Clinical Classifications Software Refined

CFS Chronic Fatigue Syndrome

CKD Chronic Kidney Disease

CNS Central Nervous System

CSF Cerebrospinal Fluid

EHR Electronic Healthcare Record

GAM Generalised Additive Model

IMD Index of Multiple Deprivation

ICD10 International Classification of Diseases

IFR Infection Fatality Rate

JCI Journal Citation Indicator

JCVI Joint Committee on Vaccination and Immunisation

JIF Journal Impact Factor

LASSO Least Absolute Shrinkage and Selection Operator

LFIA Lateral Flow Immunoassay

ONS Office for National Statistics

PCR Polymerase Chain Reaction

REACT Real-time Assessment of Community Transmission

SHAP SHapley Additive exPlanation

Publications

Papers based on analyses presented in this thesis

1. **Whitaker, M.,** Elliott, J., Bodinier, B., Barclay, W., Ward, H., Cooke, G., Donnelly, C. A., Chadeau-Hyam, M. & Elliott, P. Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England. en. *Nat. Commun.* **13,** 6856 (Nov. 2022)

2. **Whitaker, M.,** Elliott, J., Chadeau-Hyam, M., Riley, S., Darzi, A., Cooke, G., Ward, H. & Elliott, P. Persistent COVID-19 symptoms in a community study of 606,434 people in England. en. *Nat. Commun.* **13,** 1957 (Apr. 2022)

3. Elliott, J., **Whitaker, M.,** Bodinier, B., Eales, O., Riley, S., Ward, H., Cooke, G., Darzi, A., Chadeau-Hyam, M. & Elliott, P. Predictive symptoms for COVID-19 in the community: REACT-1 study of over 1 million people. en. *PLoS Med.* **18,** e1003777 (Sept. 2021)*

*J. Elliott and M. Whitaker joint first authors

Other publications

1. Elliott, P., **Whitaker, M.,** Tang, D., Eales, O., Steyn, N., Bodinier, B., Wang, H., Elliott, J., Atchison, C., Ashby, D., Barclay, W., Taylor, G., Darzi, A., Cooke, G. S., Ward, H., Donnelly, C. A., Riley, S. & Chadeau-Hyam, M. Design and Implementation of a National SARS-CoV-2 Monitoring Program in England: REACT-1 Study. en. *Am. J. Public Health* **113,** 545–554 (May 2023)

2. Ward, H., **Whitaker, M.,** Flower, B., Tang, S. N., Atchison, C., Darzi, A., Donnelly, C. A., Cann, A., Diggle, P. J., Ashby, D., Riley, S., Barclay, W. S., Elliott, P. & Cooke, G. S. Population antibody responses following COVID-19 vaccination in 212,102 individuals. en. *Nat. Commun.* **13,** 907 (Feb. 2022)

3. Atchison, C. J., **Whitaker, M.,** Donnelly, C. A., Chadeau-Hyam, M., Riley, S., Darzi, A., Ashby, D., Barclay, W., Cooke, G. S., Elliott, P. & Ward, H. Characteristics and predictors of persistent symptoms post-COVID-19 in children and young people: a large community cross-sectional study in England. en. *Arch. Dis. Child.* (Mar. 2023)

4. Cooper, E., Lound, A., Atchison, C. J., **Whitaker, M.,** Eccles, C., Cooke, G. S., Elliott, P. & Ward, H. Awareness and perceptions of Long COVID among people in the REACT programme: Early insights from a pilot interview study. en. *PLoS One* **18,** e0280943 (Jan. 2023)

5. Ward, H., Atchison, C., **Whitaker, M.,** Ainslie, K. E. C., Elliott, J., Okell, L., Redd, R., Ashby, D., Donnelly, C. A., Barclay, W., Darzi, A., Cooke, G., Riley, S. & Elliott, P. SARS-CoV-2 antibody prevalence in England following the first peak of the pandemic. en. *Nat. Commun.* **12,** 905 (Feb. 2021)

6. Elliott, J., Bodinier, B., **Whitaker, M.,** Delpierre, C., Vermeulen, R., Tzoulaki, I., Elliott, P. & Chadeau-Hyam, M. COVID-19 mortality in the UK Biobank cohort: revisiting and evaluating risk factors. en. *Eur. J. Epidemiol.* **36,** 299–309 (Mar. 2021)

7. Chadeau-Hyam, M., Bodinier, B., Elliott, J., **Whitaker, M.,** Tzoulaki, I., Vermeulen, R., Kelly-Irving, M., Delpierre, C. & Elliott, P. Risk factors for positive and negative COVID-19 tests: a cautious and in-depth analysis of UK biobank data. en. *Int. J. Epidemiol.* **49,** 1454–1467 (Oct. 2020)

8. Elliott, P., Bodinier, B., Eales, O., Wang, H., Haw, D., Elliott, J., **Whitaker, M.,** Jonnerby, J., Tang, D., Walters, C. E., Atchison, C., Diggle, P. J., Page, A. J., Trotter, A. J., Ashby, D., Barclay, W., Taylor, G., Ward, H., Darzi, A., Cooke, G. S., Chadeau-Hyam, M. & Donnelly, C. A. *Rapid increase in Omicron infections in England during December 2021: REACT-1 study* en. Mar. 2022

9. Elliott, P., Eales, O., Steyn, N., Tang, D., Bodinier, B., Wang, H., Elliott, J., **Whitaker, M.,** Atchison, C., Diggle, P. J., Page, A. J., Trotter, A. J., Ashby, D., Barclay, W.,

Taylor, G., Ward, H., Darzi, A., Cooke, G. S., Donnelly, C. A. & Chadeau-Hyam, M. Twin peaks: The Omicron SARS-CoV-2 BA.1 and BA.2 epidemics in England. en. *Science* **376,** eabq4411 (June 2022)

10. Eales, O., de Oliveira Martins, L., Page, A. J., Wang, H., Bodinier, B., Tang, D., Haw, D., Jonnerby, J., Atchison, C., Ashby, D., Barclay, W., Taylor, G., Cooke, G., Ward, H., Darzi, A., Riley, S., Elliott, P., Donnelly, C. A. & Chadeau-Hyam, M. Dynamics of competing SARS-CoV-2 variants during the Omicron epidemic in England. en. *Nat. Commun.* **13,** 4375 (July 2022)

11. Atchison, C. J., Moshe, M., Brown, J. C., **Whitaker, M.,** Wong, N. C. K., Bharath, A. A., McKendry, R. A., Darzi, A., Ashby, D., Donnelly, C. A., Riley, S., Elliott, P., Barclay, W. S., Cooke, G. S. & Ward, H. Validity of Self-testing at Home With Rapid Severe Acute Respiratory Syndrome Coronavirus 2 Antibody Detection by Lateral Flow Immunoassay. en. *Clin. Infect. Dis.* **76,** 658–666 (Feb. 2023)

12. Ward, H., Cooke, G. S., Atchison, C., **Whitaker, M.,** Elliott, J., Moshe, M., Brown, J. C., Flower, B., Daunt, A., Ainslie, K., Ashby, D., Donnelly, C. A., Riley, S., Darzi, A., Barclay, W. & Elliott, P. Prevalence of antibody positivity to SARS-CoV-2 following the first peak of infection in England: Serial cross-sectional studies of 365,000 adults. en. *Lancet Reg Health Eur* **4,** 100098 (May 2021)

13. Chadeau-Hyam, M., Eales, O., Bodinier, B., Wang, H., Haw, D., **Whitaker, M.,** Elliott, J., Walters, C. E., Jonnerby, J., Atchison, C., Diggle, P. J., Page, A. J., Ashby, D., Barclay, W., Taylor, G., Cooke, G., Ward, H., Darzi, A., Donnelly, C. A. & Elliott, P. Breakthrough SARS-CoV-2 infections in double and triple vaccinated adults and single dose vaccine effectiveness among children in Autumn 2021 in England: REACT-1 study. en. *EClinicalMedicine* **48,** 101419 (June 2022)

14. Elliott, P., Eales, O., Bodinier, B., Tang, D., Wang, H., Jonnerby, J., Haw, D., Elliott, J., **Whitaker, M.,** Walters, C. E., Atchison, C., Diggle, P. J., Page, A. J., Trotter, A. J., Ashby, D., Barclay, W., Taylor, G., Ward, H., Darzi, A., Cooke, G. S., Chadeau-Hyam, M. & Donnelly, C. A. Dynamics of a national Omicron SARS-CoV-2 epidemic during January 2022 in England. en. *Nat. Commun.* **13,** 4500 (Aug. 2022)

15. Wong, N. C. K., Meshkinfamfard, S., Turbé, V., **Whitaker, M.,** Moshe, M., Bardanzellu,

A., Dai, T., Pignatelli, E., Barclay, W., Darzi, A., Elliott, P., Ward, H., Tanaka, R. J., Cooke, G. S., McKendry, R. A., Atchison, C. J. & Bharath, A. A. Machine learning to support visual auditing of home-based lateral flow immunoassay self-test results for SARS-CoV-2 antibodies. en. *Commun. Med.* **2,** 78 (July 2022)

16. Chadeau-Hyam, M., Wang, H., Eales, O., Haw, D., Bodinier, B., **Whitaker, M.,** Walters, C. E., Ainslie, K. E. C., Atchison, C., Fronterre, C., Diggle, P. J., Page, A. J., Trotter, A. J., Ashby, D., Barclay, W., Taylor, G., Cooke, G., Ward, H., Darzi, A., Riley, S., Donnelly, C. A., Elliott, P. & COVID-19 Genomics UK consortium. SARS-CoV-2 infection and vaccine effectiveness in England (REACT-1): a series of cross-sectional random community surveys. en. *Lancet Respir Med* **10,** 355–366 (Apr. 2022)

17. Jefferson, E., Cole, C., Mumtaz, S., Cox, S., Giles, T. C., Adejumo, S., Urwin, E., Lea, D., Macdonald, C., Best, J., Masood, E., Milligan, G., Johnston, J., Horban, S., Birced, I., Hall, C., Jackson, A. S., Collins, C., Rising, S., Dodsley, C., Hampton, J., Hadfield, A., Santos, R., Tarr, S., Panagi, V., Lavagna, J., Jackson, T., Chuter, A., Beggs, J., Martinez-Queipo, M., Ward, H., von Ziegenweidt, J., Burns, F., Martin, J., Sebire, N., Morris, C., Bradley, D., Baxter, R., Ahonen-Bishopp, A., Smith, P., Shoemark, A., Valdes, A. M., Ollivere, B., Manisty, C., Eyre, D., Gallant, S., Joy, G., McAuley, A., Connell, D., Northstone, K., Jeffery, K., Di Angelantonio, E., McMahon, A., Walker, M., Semple, M. G., Sims, J. M., Lawrence, E., Davies, B., Baillie, J. K., Tang, M., Leeming, G., Power, L., Breeze, T., Murray, D., Orton, C., Pierce, I., Hall, I., Ladhani, S., Gillson, N., **Whitaker, M.,** Shallcross, L., Seymour, D., Varma, S., Reilly, G., Morris, A., Hopkins, S., Sheikh, A. & Quinlan, P. A Hybrid Architecture (CO-CONNECT) to Facilitate Rapid Discovery and Access to Data Across the United Kingdom in Response to the COVID-19 Pandemic: Development Study. en. *J. Med. Internet Res.* **24,** e40035 (Dec. 2022)

Chapter 1

Introduction

1.1 The COVID-19 pandemic

The coronavirus disease of 2019 (COVID-19) – the disease caused by the SARS-CoV-2 virus – was first detected in Wuhan, China, in late 2019. The World Health Organization (WHO) made a statement about the novel coronavirus on 10 January 2020 [1], and declared COVID-19 a global pandemic on 11 March 2020 [2]. As of April 2023, more than 765 million cases and 6.9 million deaths have been reported worldwide [3], and 24.5 million cases and 224,000 deaths in the UK. Analysis of excess deaths has suggested that the true worldwide death toll may be nearly three times as high [4]. On top of the immediate human costs, the second-order effects of the pandemic – in terms of reduced economic output and huge governmental borrowing [5, 6], over-stretched health services [7], and the burden of people suffering from ongoing poor health as a result of COVID-19 [8] – will continue to be felt for many years.

This chapter introduces some of the key topics relating to the COVID-19 pandemic and to the disease itself, with a focus on the areas that are investigated in following chapters: the response of the research community, symptomatology in the acute phase of COVID-19, and persistent COVID-19 symptoms.

1.1.1 Public health response

The COVID-19 pandemic has elicited an unparalleled coordinated effort in international public health, informed by a similarly unprecedented response from the research community.

Early work on small numbers of patients infected with COVID-19 in Wuhan characterised the genome of the virus, identifying similarities with bat-derived coronaviruses and suggesting that the virus binds to the ACE-2 receptor in humans [9–14]. Subsequent early epidemiological research – largely single-centre studies from China and Italy, which were the first countries to be severely affected by the epidemic – identified clinical characteristics of the disease and putative risk factors [15–17]. Fever, persistent coughing and fatigue were identified as common symptoms [17]. While anosmia and ageusia were mentioned as putative symptoms as early as February 2020 [18], it was several months before they were established as key charceristic manifestations of the disease [18, 19] and subsequently added to official symptom lists in the UK [20] and elsewhere [21].

Research into the contained epidemic on the Princess Louise cruise ship showed that some (estimated proportion of 17.9%) of those infected with SARS-CoV-2 did not present any symptoms at all [22]. Subsequent research revised the estimated asymptomatic proportion upwards to 40–45% [23], and established that transmission from asymptomatic or presymptomatic carriers was likely to be a large contributor to epidemic spread [24, 25].

Measures to combat the pandemic fell into two categories: non-pharmaceutical interventions (NPIs) – including social distancing, restriction of group activities and mixing, masking and hand-washing – aimed to control the spread of the virus; and pharmaceutical interventions – including vaccination and therapeutic treatments such as antivirals and monoclonal antibodies – aimed to reduce the likelihood of transmission and mitigate the effects of the disease.

In the early phase of the pandemic, between the onset of the first wave in early 2020 to the start of vaccine roll-outs at the end of 2020, non-pharmaceutical interventions were the only public health tools available to control the spread of SARS-CoV-2. These were informed by research into transmission mechanisms [26] and dynamics [27]. Covid-19 can be spread through droplets, short and long-range aerosols, and contaminated surfaces,[24, 26, 28, 29], although public health policy was sometimes slow to acknowledge the role of airborne transmission and focused initially on fomite transmission [30]. Most countries implemented some form of lockdown [31], imposing

restrictions on citizens' movements and mandating the closure of shops, restaurants and public gathering places. In England, the first lockdown began on 23 March 2020 [32] and remained in place until a partial relaxation was announced on 13 May, followed by a full relaxation and replacement with social distancing and masking guidance on 23 June 2020. Lockdowns continued in many countries, with varying degrees of restriction, in late 2020 and early 2021; only a small number of countries continued with significant restrictions beyond 2021: China maintained a zero-COVID policy, deploying targeted lockdowns, mandatory mass testing, and quarantines, until abruptly revoking all restrictions in January 2023 [33]. Some countries, as of April 2023, retain some public policy measures such as requirements for masking in public places (eg India), or travel and immigration restrictions on unvaccinated people (eg Niger) [34].

1.1.2 COVID-19 pathogenenic mechanisms

Viral entry

Viral entry generally occurs through the nose, mouth or eyes [35], with subsequent viral replication occurring initially in the upper respiratory tract [14, 36], then subsequently descending into the lower respiratory tract (as evidenced by findings that viral load is highest the upper respiratory tract in the early phase of the disease, and then increases in the lower respiratory tract [37, 38]). The SARS-CoV-2 virus has various mechanisms of spreading into other host systems, after initially infiltrating the respiratory system. It can pass through the epithelial barrier and enter the bloodstream by infecting endothelial cells in the blood-brain barrier or blood-CSF (Cerebrospinal Fluid) barrier [39]. It can infect white blood cells, which then spread to other tissues in the host body and cross the barrier to the central nervous system (CNS) (a process known as a "Trojan horse mechanism" [40]). It is also hypothesised that the SARS-COV-2 virus can spread locally through the epithelial lining, via direct extension (Xia *et al* found SARS-CoV-2 virus in the tears and conjunctival secretions of COVID-19 patients and suggested that the virus had passed through the nasolacrimal duct [41]).

Pathogenesis

COVID-19 causes inflammation and damage across multiple systems including the respiratory [42], cardiovascular [43], nervous [44], musculoskeletal [45], gastrointestinal [46] renal [47] and hematopoietic systems [48]. The multi-systemic nature of the disease has been hypothesised to be related to the fact that the SARS-CoV-2 virus binds to the angiotensin-converting enzyme 2 (ACE2) receptors, which are located on many different human cells [36, 46].

If the virus spreads beyond the respiratory system to other cells and tissues, a systemic immune response can be induced. This does not always indicate severe disease: even in mild cases, manifestations of COVID-19 can be the product of disturbance of multiple biological systems [49].

COVID-19 can also induce a cytokine storm, where the body over-produces pro-inflammatory cytokines, causing systemic inflammation, acute respiratory distress syndrome (ARDS), and, potentially, multi-organ damage or failure and death [50]. The cytokine storm is also implicated in coagulation dysfunction [51], and thrombotic complications, which are commonly experienced by COVID-19 patients and are associated with increased disease severity and poorer clinical outcomes [52–55].

1.1.3 Risk factors for COVID-19

The term 'risk factor' is used to describe two separate types of risk: risk of becoming infected (or susceptibility), and risk of progression to severe disease or death, conditional on infection. The latter could also be subdivided into a second subtype: the risk of progression to intensive care or mortality, conditional on hospitalisation.

The earliest investigations into risk factors were primarily descriptive studies of hospitalised patients in Wuhan, which identified increasing age as a risk factor for mortality among those hospitalised [15]. An over-representation of men, and patients with comorbidities including diabetes, among those hospitalised was also reported [17], although the extent to which this reflected intrinsic predisposition to infection or severe disease (as opposed to behavioural or environmental factors influencing the likelihood of exposure to the virus) could not be determined. Subsequent multi-centre studies in Italy and in New York supported the findings of elevated

risk for older patients and males, and further identified: chronic kidney disease, heart failure and increased body mass index (BMI) [56]; and smoking, diabetes, hypertension, cardiovascular disease and respiratory disease [57] as risk factors for severe disease and mortality.

Population-based studies in England, including the REACT-1 study [58], investigated risk factors for infection and for severe disease and mortality at a population level in the early phase of the pandemic, demonstrating a demographic component to the infection risk with some communities (in England: healthcare and care home workers people of black or Asian ethnicity, as well as those of older age, and men), at greater risk of infection in 2020 [59]. Higher rates of infection naturally lead to higher rates of hospitalisation and death, so disambiguating the infection risk (influenced by environment and behaviour) from risk of severe disease conditional on infection became an increasingly important public health question.

As data began to be merged across populations in sufficient volume to disentangle these effects, age emerged as the most consistent and significant risk factor for severe disease and death, conditional on infection. Two large systemic analyses of seroprevalence data and mortality rates from multiple countries estimated the relationship between age and risk of mortality conditional on infection (the age-specific Infection Fatality Rate (IFR)) [60, 61], and found a strikingly log-linear relationship, with the risk of mortality increasing approximately tenfold for every additional 15 years of age [61]. This finding informed the vaccine rollout in the UK, with age group being the primary variable used to prioritise vaccine recipients, alongside comorbidities.

Genetic predisposition to COVID-19

In the case of genetic predisposition there is likely to be an overlap between susceptibility to infection and susceptibility to severe disease, as the mechanisms for infection (cell entry via the binding of the spike protein to the ACE-2 receptor) and disease progression (subsequent transfer to ACE-2 receptors in different cell tissues in other host systems) are affected by levels of expression of ACE-2. A meta-analysis (SeyedAlinaghi *et al* [62]) analysed multiple reports supporting the existence of a genetic predisposition to both COVID-19 infection and to progression to severe COVID-19, that was related to ACE2 polymorphisms and to expression levels of of type 2 transmembrane serine proteases (TMPRSS2) in the host. Jackson *et al* also

note in a 2022 review [63] that many of the risk factors for severe COVID-19 – age, sex, smoking status, diabetes, hypertension – have been associated with expression levels of ACE-2, although the directions of effect are not always consistent and the evidence for this mechanism of risk is far from conclusive.

1.1.4 Treatments for COVID-19

A wide range of treatments for COVID-19 have been investigated. Early in the pandemic, treatments were limited to in-hospital mechanical ventilation and supplemental oxygen. Multiple drugs and treatments were proposed and entered accelerated trials. One of the earliest breakthroughs came from the RECOVERY Collaborative Group, who demonstrated that administration of oral or intravenous dexamethasone reduced 28-day mortality in hospitalised patients (with an age-adjusted mortality rate ratio vs placebo of 0.83; 95% confidence interval 0.75–0.93) [64]. Monoclonal antibodies – laboratory made proteins that mimic the body's immune response and bind to SARS-CoV-2 spike protein – were also demonstrated to reduce risk of infection and of severe outcomes [65, 66].

In 2021, antivrial treatments Molnupiravir and Paxlovid were shown to reduce the risk of hospitalisation and death [67, 68], and are now widely used across the world. Remdesivir was one of the earliest antivirals to begin trials [69], and early results indicated that it reduced recovery time in patients hospitalised with COVID-19 [70]; however, the WHO Solidarity Trial, a large multi-centre study involving more than 11,000 patients in 30 countries, found no significant reduction in mortality risk versus a placebo control group [71]. Hydroxychloroquine was also initially proposed as a treatment and was granted emergency use authorisation in the United States [72], but repeated studies found no evidence for its effectiveness either as a prophylaxis [73] or to reduce risk of hospitalisation or mortality [71] and the emergency use authorisation was revoked in June 2020. Despite this, Hydroxychloroquine and other drugs such as Ivermectin continue to be promoted as effective treatment for COVID-19 by some, largely through social media [74].

The quantity of low-quality information on the internet about treatments for COVID-19, informed in some cases by small studies that have failed to replicate but are still cited as robust evidence of effectiveness, has presented a public health challenge. It has also motivated and

informed the second chapter of this thesis, in which the relationship between the impact of a report on social media and the academic merit of that report is investigated.

1.1.5 Vaccines

Eight days after the first reported cluster of COVID-19 cases in Wuhan, researchers shared the genetic code for the SARS-CoV-2 virus [75] (later published by Wu et al [13]) and multiple research groups began pre-clinical work on developing a vaccine. In order to maximise the chances of finding a successful candidate, research into multiple different vaccine technologies was initiated, with funding agencies often deliberately funding a spread of different approaches and technologies [76].

Viral Vector Vaccines

Viral vector vaccines use a modified virus as a vehicle to deliver genetic information into host cells, eliciting an immune response against a specific pathogen. The virus is engineered to be non-pathogenic while retaining its capacity to enter cells and deliver genetic instructions, [77] which prompt cells to synthesise a protein associated with the target pathogen, such as SARS-CoV-2's spike protein. The immune system identifies this protein as foreign, generating an immune response that confers protection against future infections by the pathogen. Viral vectors have demonstrated the ability to provoke both innate and adaptive immune responses in mammalian hosts. These are currently being researched as potential vaccine candidates for multiple other conditions including malaria and human immunodeficiency virus (HIV) [77].

On 30 April 2020 AstraZeneca announced a collaboration with Oxford University in the UK to produce a vaccine [78] using viral-vector technology, and in July 2020, interim results from the Phase 1/2 trial of the Oxford-AstraZeneca vaccine showed promising immune responses and safety, with neutralising antibody responses detected in 91% of recipients [79]. Johnson and Johnson also developed an adenovirus-based vaccine [80], which was authorised for use in the USA in April 2021 [81] but was never used in the UK.

Recombinant protein vaccines

Recombinant protein vaccines contain fragments, such as proteins, from the virus, which are chosen to elicit an immune response in the host. The fragments are reproduced at scale in a heterologous expression system (where a gene that codes for the critical protein is inserted into cells from another organism, often bacteria or yeast, which acts as a factory to produce the protein) [82]. After purification, the proteins are used as the principal ingredient in the vaccine and introduced into the host to induce an immune response by producing antibodies. Around a third of the vacines what went into development in the first year of the pandemic used a recombinant protein platform [83], although none were widely authorised for use in the first 18 months of the pandemic. The Novovax vaccine, which used residues of the receptor-binding domain of the SARS-CoV-2 spike protein, produced promising early results in animals [84] and later passed phase 2 and 3 trials and was approved in 40 countries worldwide [85].

mRNA vaccines

mRNA vaccines are created by identifying a protein from the virus that produces an immune response in the host, and designing an mRNA sequence that will be translated into the target protein by host cells [86, 87]. The mRNA sequence is then synthesised in a laboratory through a process of in vitro transcription, and encapsulated within lipid nanoparticles, which protect the mRNA and aid its delivery into host cells.

The two most widely used vaccines in the world have both been based on mRNA technology [88]: the Pfizer-BioNTech vaccine (BNT162b2) and the Moderna vaccine (mRNA-1273). Both are lipid nanoparticle-encapsulated mRNA-based vaccines that encode the spike protein of the SARS-CoV-2 virus, and both showed more than 90% efficacy in preventing COVID-19 illness in phase 2/3 trials [89, 90]. The Pfizer-BioNTech vaccine became the first vaccine to be approved and administered anywhere in the world when the UK approved its use on 2 December 2020 [91] and began the vaccine rollout on 8 December [92]

The vaccine rollout in England

The vaccine rollout in England was informed by advice from the Joint Committee on Vaccination and Immunisation (JCVI) [93]. The population was stratified into 12 risk groups according to age, healthcare/social care worker status and pre-existing health conditions and vaccines were offered to each group sequentially, beginning with care home residents and medical staff in December 2020. The Pfizer-BioNTech vaccine was offered first, with AstraZeneca's vaccine receiving approval on 30 December 2020 [94] and being offered shortly after, and the Spikevax (Moderna) vaccine receiving approval on 8 January 2021 [95]. Despite the fact that the dosing interval between first and second dose was 3–4 weeks in both the Pfizer-BioNTech and Astrazeneca trials, the UK Chief Medical Officers made the decision to extend the dosing interval to 12 weeks, given the limited supply of vaccine doses and some early evidence that a single dose provided substantial protection against severe infection and death [96]. A target to offer first vaccines to all those aged 50 and over, or aged 16+ and with an underlying health condition, by 15 April 2021 was met, as was a target to offer all adults (aged 16+) a first vaccine by 19 July 2021 [96]. After evidence emerged of the risk of blood clots associated with the AstraZeneca vaccine [53], the government accepted the JCVI's recommendation that an alternative vaccine be offered to all under-30s (later extended to all under-40s on 7 May 2021) [96]. By 19 July 2021, a vaccine had been offered to all adults aged 16 and over. As risks from COVID-19 are low for younger people, the cost-benefit calculations for vaccinating children were less clear cut than for adults, and initially vaccinations were only offered to those under 16 who had specific underlying conditions [97]. In February 2022, the offer of vaccines was extended to all children aged 5 and over [98]. From September 2022, only children (aged 5–16) who were in a high-risk group were eligible for a vaccine, and the offer of first and second vaccines for adults ends in June 2023.

Booster programmes offering third, fourth and fifth vaccines, initially to high-risk groups and then to lower risk groups, have been run in autumn 2021, spring 2022, and spring 2023 in England. These have included vaccines tailored to later variants of SARS-CoV-2 [99], and bivalent vaccines, which target both the Omicron spike proteins and the ancestral wild-type SARS-CoV-2 spike proteins in a single vaccine. Early findings from the US have suggested that bivalent vaccines provide increased protection against hospitalisation when compared with monovalent vaccines [100]; however, Wang *et al* found no difference in virus-neutralizing peak

antibody response when boosting with a bivalent vs a monovalent vaccine [101].

1.1.6 Variants

After the initial wave of the 'Wild-type' variant of SARS-CoV-2, the pandemic was characterised by the emergence of successive new variants which competed with and in some cases totally replaced preceding strains. The first variant to achieve near-complete international replacement was the Alpha (B.1.1.7) variant, which was first detected in England in November 2020 [102] and spread rapidly internationally owing to a 90% increased transmissibility over the wild-type variant. Subsequently, the Delta variant, first detected in India in 2021 [103], achieved near-complete replacement of the Alpha variant in the UK by July 2021 [104] and in most of the rest of the world in the following months [105]. Finally, the Omicron variant, which was first detected in South Africa and Botswana in late 2021 [106], achieved near complete worldwide replacements of previous variants in early 2022. As of April 2023, Omicron subvariants continue to be the dominant strains of COVID-19 worldwide [105].

1.2 The COVID-19 Research Landscape

The COVID-19 pandemic has induced an unprecedented response from the scientific research community. It has been the first global pandemic of the internet era, and the first of the 'modern' era of scientific publishing: the influenza pandemic of 1918 happened before the era of formalised peer review [107], when medical journals were circulated by post to a select few libraries and members of the publishing organisation. A PubMed search for articles with "Influenza" in the title between 1918 and 1920 returns 17 results. By contrast the World Health Organization (WHO) COVID-19 literature database lists more than 440,000 papers published since the start of the COVID-19 pandemic [108].

The severity and speed of onset of the pandemic, the novelty of the disease, and the concomitant demand for rapid execution and dissemination of research has represented a stress test for the institutions and incentive structures of scientific publishing and communication. The most immediate pressure point has been the peer review process. Peer review can be a lengthy procedure [109], especially when reviewers are themselves dealing with additional work demands. Some higher impact journals introduced COVID-19 peer review 'fast streams' and were able to accelerate peer review times by 49% [110]. Nonetheless, the sheer volume of COVID-19 submissions – Elsevier estimated an increase of 63% in submissions to their medical journals in February–May 2020 compared with February–May 2019 [111] – caused concerns about the burden on editors and peer reviewers leading to a reduction in the rigorousness of the review process [112] and a degradation of overall research quality [113]. Schonhaut et al [114], for example, found that while submission–acceptance turnaround times were 11 times faster for COVID-19 papers than for generic influenza papers, withdrawal rates were an order of magnitude higher for COVID-19 papers than for influenza papers from the same period (0.26% vs 0.023%).

While research journals were racing to adapt to the increased volume of submissions, the demand for rapid sharing of results (and authors' desire to be first to publish a particular finding) meant that the volume of preprints – papers published immediately without peer review on domain-specific, non-paywalled servers – grew even more dramatically during the pandemic: monthly submissions to medRxiv increased by a factor of ten between January and May 2020 [115], and more than two-thirds of all the papers published on medRxiv in 2020 were

related to COVID-19 [116]. Researchers noted the scale of COVID-19 research output and asked whether the scientific publishing industry might be permanently altered by the rapid adoption of preprints [117]. However, the initial exponential trajectory of COVID-19 preprint publication was not maintained much beyond mid-2020, and preprint publication rates plateaued [118].

1.2.1 Sharing and communicating research

The second key pressure point in the research trajectory comes after publication: communicating the latest research findings to the scientific community, policy-makers, and the public. Here, preprints have occupied a conflicted position. Despite the disclaimers that they are not peer reviewed and should not be reported, the huge public demand for information on COVID-19 meant that many preprints received a lot of attention in the media and on social media: medRxiv was receiving more than 10 million page views per month at the peak of the pandemic, compared with 64,000 in late 2019 [115]. This circumvention of the normal quality control mechanism has undoubtedly led to some spurious findings entering the bloodstream of scientific discourse [119–121], and influencing attitudes, behaviour and public policy. The World Health Organization has called this phenomenon, when the demand for information on a new disease exceeds supply and normal quality control mechanisms become disrupted, an 'infodemic' [122]. It has significant implications for public health policy and messaging [123], and is therefore an important field of research.

1.2.2 Infodemics and bibliometrics

Analysing the impact of research has become an academic field in itself owing to the increasingly rich and varied data streams that are now available for tracking how academic papers are consumed and shared online, and by whom. Altmetrics [124] – measures of article impact and attention in news outlets, blogs and on social media – can be used to track how papers are shared on social media [125–127] and can be combined with other data to better understand the mechanisms by which information enters the discourse. The analysis of this article usage and impact data, in conjunction with other article metadata and, sometimes, rich textual analysis of the paper itself, is generally known as bibliometrics or scientometrics.

The next chapter of this thesis uses bibliometrics to describe and analyse how the scientific information ecosystem functioned under the acute stress of a pandemic. The first goal is to provide a descriptive overview of all published COVID-19 research literature, using the tools and language of epidemiology. Specifically, this involves: reconstruction of 'epidemic curves' of (i) COVID-19 papers (both preprints and journal articles), and (ii) the media and social media attention generated by those papers; summarising the most-cited most-shared papers; and using textual analysis to show how research topics evolved throughout the pandemic, and which ones were driving attention at which times. The second goal is to investigate questions relating to the functioning of the scientific information ecosystem. Specifically: were the 'best' papers (as measured by traditional citation and journal quality metrics) getting the most attention (as measured by Altmetrics)? Was the huge wave of COVID-19 preprints in 2020 of lower average quality than other preprints (using time-to-event analysis to describe publication trends)? And did COVID-19 research 'cannibalise' other research? The chapter uses a large corpus of all research literature (more than 17 million papers) published between the start of 2019 and the end of 2022. Analysis of COVID-19 literature is therefore contextualised within this larger corpus, which provides a benchmark against which to compare trends.

1.3 Symptomatology in the acute phase of COVID-19

Clinical manifestations of COVID-19 are highly heterogenous, reflecting the pathogenic breadth of the disease and the various host systems that it can affect. The majority of infections involve mild (or no) symptoms, most commonly coughing, fatigue, Anosmia/dysnosmia (loss/disturbance of sense of smell) and ageusia/dysgeusia (loss/disturbance of sense of taste) and fever, but studies have identified at least 30 symptoms that are associated with infection [128]. Early research focused on symptomatology in patients admitted to hospital in the countries earliest affected (China [17, 129] and Italy [130]). These studies necessarily focused on hospitalised patients, and therefore the prevalence estimates are weighted towards symptoms that are associated with severe disease, and the later stages of the disease course. A meta-analysis of these early papers by Grant *et al* found that fever, cough, fatigue and hyposmia were the most common symptoms [128]. In a 2020 meta-analysis of patients who tested PCR positive in hospital and experienced only mild-to-moderate COVID-19, Lechien *et al* found the most common symptoms, in order, were headache (Lechien *et al* found a pooled prevalence of 70.3% for headache, compared to only 13% in Grant *et al*), anosmia, nasal obstruction, asthenia, myalgia, rhinorrhea, gustatory dysfunction, sore throat, and fever [131].

Among the first representative studies of symptom prevalence among non-hospitalised individuals was the UK Office for National Statistics' (ONS's) Coronavirus (COVID-19) Infection Survey [132], which tested and surveyed a cohort of 150,000 individuals from a representative sample of households in the UK approximately fortnightly between August 2020 and March 2023 [133]. Regular bulletins were published summarising the results; the first of these to contain symptom data was published in May 2021 [134] and summarised symptom prevalence among 'strong' SARS-CoV-2 positive individuals (with a Ct value of 30 and under) at monthly intervals between December 2020 and April 2021 in the UK – therefore capturing the end of the Wild-Type variant wave and the emergence of the Alpha variant.

In December 2020, 54.69% (95% CI: 52.92,56.44) of SARS-CoV-2-positive participants reported experiencing one or more of the 12 surveyed symptoms within 35 days of their first positive PCR test. The ONS reports did not report test-negative symptom prevalence, but owing to lockdowns and reduced social interactions the prevalence of other circulating respiratory viruses, and therefore background symptom prevalence, was low [135]. The most common symptoms

among SARS-CoV-2 positive participants were cough (34.50% [32.83,36.20]), headache (32.96% [31.31,34.64]) and fatigue/weakness (33.73% [32.07,35.42]).

Fever

Fever is a common immune response to infection and illness, and was generally found to be the most prevalent COVID-19 symptom in hospital-based studies: Islam *et al* found a pooled prevalence of 79.43% [95% CI: 77.05–81.80] [136] in a meta-analysis of studies of hospitalised COVID-19 patients published in the first four months after the start of the COVID-19 outbreak (1 December 2019 – 3 April 2020). The ONS study found a prevalence of 21.73% (20.29,23.22) among SARS-COV-2-positive individuals in December 2020.

Fatigue

As with fever, fatigue is a common consequence of viral infections and is a highly prevalent but low-specificity symptom of COVID-19. It is thought to be caused by neuroinflammation of the brain tissue [137], and functions to conserve energy expenditure to allow the immune system to fight pathogens. In a meta-analysis of hospitalised patients with mild to moderate COVID-19, Lechien *et al* found a pooled prevalence of 63.2% [131]; by contrast, the symptom prevalence in the representative ONS sample of non-hospitalised SARS-CoV-2 positive individuals was 33.73% (32.07,35.42) in December 2020.

Anosmia and ageusia

Anosmia/dysnosmia and ageusia/dysgeusia are highly specific symptoms of COVID-19, and are likely caused by viral entry into cells in the olfactory epithelium (potentially non-neuronal cells such as support cells, stem cells, and perivascular cells [138]), and subsequent inflammation [139]. In hospitalised patients, two 2020 meta-analyses found very similar prevalences: Agyeman *et al* found that anosmia and ageusia are common in up to 41% of COVID-19 cases [140]; Favas *et al* found a prevalence of 35.8%; 95% CI 21.4–50.2) for anosmia/dysnosmia, and 38.5%; 95%CI 24.0–53.0) for ageusia/dysgeusia [141]. In the ONS study, prevalence of anosmia was 17.01% (15.71,18.38) and ageusia was 17.75% (16.42,19.13) in December 2020.

Respiratory symptoms

COVID-19 presents most commonly as an upper respiratory disease, with viral replication in the upper and lower respiratory tract inducing a localised inflammatory response. The most common respiratory symptom of COVID-19 is a cough: Lechien *et al* found a pooled prevalence of 63.2% [131] in hospitalised patients with 'mild to moderate', ie not requiring intensive care, COVID-19, while the ONS study found a prevalence of 34.50% (32.83,36.20) in December 2020. Sore throat is also a common symptom (pooled prevalence in hospitalised patients in Lechien *et al* [131]: 52.9%, ONS representative prevalence 22.40% (20.95,23.91)). Less commonly, patients may report dyspnoea (breathlessness) (ONS prevalence 14.60% [13.38,15.89]), chest pain or tightness, haemoptysis (coughing up blood) or wheezing [128].

Neurological symptoms

While anosmia/dysnosmia and ageusia/dysgeusia are the most COVID-19-specific neurological symptoms, other common neurological symptoms have been identified. A meta-analysis of neurological COVID-19 symptoms, which included non-hospitalised patients, identified several symptoms that were described in multiple papers [141]. The most common were myalgia (muscle aches and pains; pooled prevalence *et al* 19.3%; 95% CI 15.1–23.6), headache (pooled prevalence 14.7%; 95% CI 10.4–18.9), dizziness (pooled prevalence 6.1%; 95% CI 3.1–9.2) and syncope (fainting or passing out; pooled prevalence 1.8%; 95% CI 0.9–4.6). Headache (ONS prevalence 32.96% [31.31,34.64]) has been shown to co-occur frequently with anosmia and ageusia, and to be indicative of a shorter disease course [142].

Gastrointestinal symptoms

Gastrointestinal symptoms that have been reported in association with COVID-19 include nausea, vomiting, abdominal pain, diarrhea and anorexia/appetite loss [143]. A meta-analysis of hospitalised patients in 2020 found that one in every eight (13%) people with COVID-19 reported had at least one gastrointestinal symptom [144]. This included 5.6% with nausea and/or vomiting, 2.7% with abdominal pain, 7.8% with diarrhoea, and as many as 27.1% patients with anorexia/appetite loss. The ONS study found a prevalence of nausea/vomiting of 10.65% (9.59,11.79) in December 2020.

Changing prevalence of symptoms over time

The symptomatology of COVID-19 has been shown to have changed since the wild-type variant: the most significant trends being an increase in the prevalence of 'cold-like' symptoms associated with the Omicron variant [145]. The symptom profile of COVID-19 has also been shown to vary in people with some existing immunity, either from a previous infection or from vaccination (See supplementary material in Antonelli *et al*, which shows changing symptom prevalence in people with one or two vaccinations, versus unvaccinated controls [146]).

1.3.1 Using symptoms to identify cases

Variation in methodology between early studies and meta-analyses resulted in divergent estimates of symptom prevalence, with most study populations involving all or mostly hospitalised patients. Identifying a subset of symptoms with sufficient sensitivity and specificity for accurate diagnosis of COVID-19 in the community has therefore been a significant ongoing challenge, particularly against a backdrop of changing background prevalence of COVID-19 and of symptoms.

Before mass testing began in England in April 2020 [147], symptoms were the only way to diagnose probable infection and to recommend that people self-isolate. After testing became widely available, symptoms were used as a screening device to allocate tests [148], based on early research into COVID-19 manifestations [17]. In early 2020, only a high temperature and a new continuous cough were listed by the NHS as potential COVID-19 symptoms. In May 2020, anosmia and ageusia were added to the symptom list [20], shortly after findings from the Zoe COVID-19 app were published [149] which suggested that loss of sense of smell or taste were highly specific symptoms of COVID-19: 59% of participants who reported a positive PCR test also reported loss of sense of smell or taste (against 18% who tested negative).

Subsequent community-based studies assessed the capacity of symptom data to predict PCR positivity for SARS-CoV-2, in some cases using variable selection and ranking methods to identify the 'best' (set of) symptoms to identify cases [150–152]. These studies included an early analysis based on REACT-1 data from June 2020 to January 2021, when the Wild-type variant was dominant [151]. The primary research goal of the paper was driven by domestic

public health priorities: how could community cases be better detected using symptom data to trigger PCR testing, and what would be the impact of expanding the symptom 'trigger' list beyond loss of smell and taste, fever and new continuous cough. The paper identified 7 symptoms that were jointly positively predictive of PCR positivity and demonstrated that expanding the trigger list to these seven symptoms would increase community detection rates without exceeding testing capacity. The paper also used early cases of B.1.1.7 (Alpha variant) that were detected in the final round of fieldwork from that period to demonstrate differences in symptom profile from the Wild Type variant, which subsequent studies confirmed in different data sets [153, 154]. The NHS has, as of April 2023, widened symptom criteria to include shortness of breath, feeling tired or exhausted, an aching body, headache, sore throat, blocked or runny nose, loss of appetite, diarrhoea and feeling sick or being sick [155].

The third chapter of this thesis builds on the above paper [151], with less emphasis on community detection and a greater emphasis on describing the changing patterns of symptomatology over time, and identifying the underlying factors that influence COVID-19 symptoms.

1.4 Post-COVID syndrome ("Long COVID")

COVID-19 has been associated with symptoms and sequelae that persist for many months after initial infection [156, 157]. Colloquially, this phenomenon is usually referred to as 'long COVID', but the National Institute for Health and Care Excellence (NICE) has settled on the umbrella term 'Post-COVID-19 syndrome', defining it as

> "[...] signs and symptoms that develop during or after COVID-19 and continue for more than 12 weeks and are not explained by an alternative diagnosis" [158].

This broad definition is useful chiefly in that it defines a threshold duration. Studies have variously used four weeks [159, 160] or eight weeks [161] or 12 weeks [162] as the duration threshold to define long COVID, and the international public health community has yet fully to agree on a duration: the Department of Health and Human Services in the US, for instance, still uses four weeks [163]. 'Signs and symptoms' also encompasses a broad range of phenotypes, because such a wide and diverse range of symptoms have been associated with COVID-19. Reported persistent symptoms include tiredness and severe fatigue, shortness of breath, chest pains, palpitations, fever, anosmia, ageusia, cognitive impairment (memory loss, 'brain fog'), skin rashes and joint pain or swelling [157, 160, 164, 165]. Similarly, estimates of symptom prevalence and duration have varied substantially. This may be attributable to heterogeneity in study design, or ongoing lack of agreement about syndrome definition(s) [157, 166–168].

1.4.1 Causes of Post-COVID syndrome

While the biological mechanisms and determinants of long COVID are yet to be fully described, multiple theories have been proposed, some explaining specific symptoms or aspects of long COVID and some more holistic.

Systemic inflammation is one potentially biologically plausible explanation. The PHOSP-COVID Collaborative Group discovered that individuals with more severe physical and mental health impairments one year after being discharged from the hospital exhibited elevated levels of multiple biomarkers associated with systemic inflammation and lung damage [169], and proposed therapeutic treatments to reduce inflammation. These findings were supported by

proteomic and metabolomic analysis in an exploratory study of patients with and without long COVID symptoms, which found stable patterns of anti-inflammatory molecules in the blood plasma of those with long COVID [170].

Bastard *et al* found that approximately 10% of patients with severe COVID-19 pneumonia had high levels of neutralizing autoantibodies against type I interferon-alpha2 (IFN-α2) and interferon-omega (IFN-ω) [171], suggesting that some long COVID may be a manifestation of an autoimmune mechanism [172].

There are clear parallels between many people's experience of long COVID and Chronic Fatigue Syndrome (CFS) [173], a multi-system condition characterised by chronic fatigue and post-exertional malaise that commonly occurs after a bacterial or viral infection, and it has been estimated that approximately half of people with long COVID meet the diagnostic criteria for CFS [174]. CFS is also poorly understood [175] so any commonalities with long COVID unfortunately confer few additional clues as to potential biological mechanisms.

It has also been proposed that long COVID may be a manifestation of a latent chronic COVID-19 infection [176]: O'Donnell *et al* cite studies that detected SARS-CoV-2 N protein and viral RNA in the intestinal epithelia of patients some months after infection [177] and suggests that this failure to fully clear the infection may lead to disturbance of the gut microbiome and facilitate the establishment of a chronic infection. There is some support for this theory in the more recent findings that vaccinations are associated with a reduction in symptoms in people with long COVID: Ayoubkhani *et al* found that a first vaccine dose was associated with a 12.8% decrease in the odds of reporting long COVID, and a second dose with an 8.8% decrease [178].

Yong [179] proposed a hypothesis that persistent brainstem dysfunction could play a role in long COVID, noting that that the brainstem, which regulates essential bodily functions, has a higher expression of ACE2 receptors than other brain regions and has been previously shown to be vulnerable to damage from pathological immune or vascular activation. Yong cites multiple autopsy studies reporting evidence of brainstem involvement in COVID-19, and notes that the symptoms of long COVID, which centre around the respiratory, cardiovascular, gastrointestinal, and neurological systems, overlap substantially with the processes controlled by the brainstem. Finally he notes brainstem abnormalities have been previously implicated in CFS [180, 181], and that neurons are known not to regenerate easily, meaning that brainstem dysfunction might be persistent and long-lasting in a way that matches with patient experiences of long COVID

and CFS.

Finally, it has been acknowledged that there are inherent challenges in studying long COVID while the pandemic has been ongoing [8]. Lockdown measures and other societal disruptions have been shown to have measurable detrimental impacts on psychological health and wellbeing [182], manifesting as elevated levels of fatigue [183], depression and anxiety [184]. Disambiguating these effects from the effects of long COVID, especially the neurological and psychological effects, is complex, particularly in the absence of a pre-existing literature on the psychological impacts of pandemics on this scale.

1.4.2 Sub-types of Post-COVID syndrome

Some research has hypothesised that long COVID encompasses an overlapping set of disparate conditions including post-viral syndromes, ongoing inflammation and long-term organ or tissue damage, and investigated this hypothesis by applying unsupervised learning methods to identify subgroups of individuals with different persistent symptom profiles [185, 186], including some recent studies using electronic health records (EHRs) from large cohorts.

Reese *et al* [187] used an algorithm to characterise and quantify different phenotypes from the EHRs of more than 6,000 patients experiencing persistent COVID-19 symptoms, and applied a k-means [188] clustering algorithm to the quantitative data to identify sub-groups with common phenotypic profiles. Six clusters were identified, which were characterised as a multi-system cluster (with high co-occurrence of symptoms across multiples systems), a pulmonary cluster, a neuropsychiatric cluster, a cardiovascular cluster, a pain/fatigue cluster and a multi-system pain cluster. The multi-system cluster was associated with severe acute COVID-19 and with various comorbidities. It was also associated with higher mortality rates in the period after assessment.

Zhang *et al* [189] used topic modelling to distil textual data from ICD10 (International Classification of Diseases 10th Revision) codes and CCSR (Clinical Classifications Software Refined) categories into continuous variables capturing post-acute COVID conditions in a large (N=20,881) cohort of test-confirmed SARS-COV-2 positives in the INSIGHT cohort [190], with validation in the OneFlorida cohort [191]. Hierarchical agglomerative clustering [192] was then applied to the topic data to identify clusters of subphenotypes. They identified four clusters,

which they characterised as 1: cardiac and renal, 2: respiratory, sleep and anxiety, 3: musculoskeletal and nervous, 4: digestive and respiratory. As with Reese *et al*'s clusters, the severity of the acute COVID-19 phase was related to cluster membership, as were several comorbidities.

1.4.3 Risk factors for Post-COVID syndrome

The literature suggests that the occurrence of post-COVID syndrome is correlated with the severity of COVID-19, with high rates of persistent symptoms observed among individuals who were hospitalised with COVID-19 [193–195]. Additionally, pre-existing poor health, including mental health conditions, obesity and a number of comorbidities, have been identified in multiple studies as risk factors for post-COVID syndrome [157, 164, 196, 197]. A large meta-analysis of EHR-based studies from the UK identified psychological stress, obesity, diabetes and asthma as specific conditions associated with higher risk of symptoms lasting 12 weeks or more [196]. Additionally, a self-rated metric of 'poor overall health' was associated with greater risk of long COVID (Odds ratio: 1.66 [1.16,1.82]).The same meta-analysis found a highly linear relationship between age and the risk of long COVID, conditional on symptomatic infection [196]. When comparing population incidence at different ages using GP records, however, there was an inverted 'U' trend, peaking at age 50. This may partly reflect differential incidence rates by age group. Finally, black and South Asian ethnicity, deprivation and female sex were also identified as risk factors [196], and have been further identified elsewhere [157, 164].

Latterly, as noted above, much research effort has been directed towards investigating the impact of vaccination on risk of persistent symptoms, with a majority of studies finding that vaccination has a protective effect against experiencing persistent symptoms after infection [198], and potentially a therapeutic effect when administered post-infection [199]. The role of the COVID-19 variant in influencing likelihood of persistent symptoms has also been of considerable public health interest, and a meta-analysis of studies into the topic found that infection with wild-type variant was associated with higher risk of persistent symptoms [200]; however, the effect of variant is extremely difficult to disambiguate from factors such as vaccination and prior infection, and the heterogeneity of the studies and the data limit any strong conclusions.

While significant research efforts have been directed at post-COVID syndrome, it remains a poorly understood condition, almost certainly encompassing multiple disparate and overlapping

disorders with distinct biological pathways and risk factors. This deficit in understanding motivated the research presented in chapter four of this thesis, which was conducted in late 2020 and early 2021, when still less was known about post-COVID syndrome. This chapter uses data from four rounds of field work in the REACT-2 programme conducted between September 2020 and February 2021 to estimate population prevalence of post-COVID syndrome in England in 2020/2021, to identify risk factors for persistent symptoms, and to identify sub-types of post-COVID syndrome using unsupervised learning.

1.5 About the REACT studies

The Imperial REACT programme comprised two large community surveillance studies designed to track the prevalence of SARS-CoV-2 virus (REACT-1) and antibodies (REACT-2) using large representative samples from the population of England [58, 201]. In each study, between 100,000 and 160,000 people were recruited each round, approximately monthly for REACT-1 (in 19 rounds of fieldwork between May 2020 and March 2022) and for six rounds of fieldwork in REACT-2 (between June 2020 and May 2021). Participants were asked to complete an extensive survey on health, lifestyle, behaviour and suspected COVID-19 history or symptoms, in addition to taking self-administered swab test for PCR analysis (in REACT-1), or a self-administered lateral flow antibody test (in REACT-2).

1.5.1 REACT-1

REACT-1 ran for 19 rounds between 1 May 2020 and 31 March 2022 [202]. Participants were recruited from the National Health Service patient register to form random, nationally representative samples of individuals aged 5 years and above, with the aim of achieving approximately equal numbers of participants in each of the 315 lower-tier local authorities (LTLAs) in England. In later rounds, the sampling methodology was changed to recruit random samples proportional to the population, thus increasing the density of sampling in urban areas and reducing it in rural areas. The sampling methodology employed was stratified by LTLA, with larger numbers selected in some LTLAs to account for variable response rates at the LTLA level. A total of up to 160,000 valid responses and viable swabs were obtained at each round. Participation in the study involved a self-administered throat and nasal swab (parent/guardian administered for children aged 5 to 12 years), as well as the completion of a short online or telephone questionnaire that included information on demographic variables, household composition, behaviours (such as mask-wearing), and recent symptoms. The questionnaires used in the study are available on the study website (https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/react-1-study-materials/). At the time of survey completion, participants were blind to the results of their swab test. The survey also asked participants about new symptoms that occurred in the week preceding the swab.

1.5.2 REACT-2

The REACT-2 programme ran in six rounds between June 2020 and May 2021. The community prevalence of SARS-CoV-2 anti-spike protein antibody positivity in England was evaluated by taking randomly selected samples of adults every 2-4 months using the National Health Service (NHS) patient list in order to achieve a similar number of participants in each of the 315 lower-tier local authority areas. Participants were invited to register via an online portal or by telephone and were subsequently sent a test kit by post, which included a self-administered point-of-care lateral flow immunoassay (LFIA) test accompanied by instructions and a link to an online video. Upon completion of their self-test, participants completed a survey via either online or telephone means, providing information on demographics, household composition, comorbidities, and whether or not they believed they had had COVID-19.

1.5.3 Real-time reporting from the REACT studies

Throughout two years of the pandemic, the REACT studies produced regular reports on the prevalence of SARS-CoV-2 virus and antibodies [203]. The research presented in this thesis was conducted as part of the REACT research programme, in response to real-time public health questions about the changing symptomatology of COVID-19 in light of evolving strains of the virus and changing levels of population immunity, and the nature of long COVID and the likely burden of disease on society.

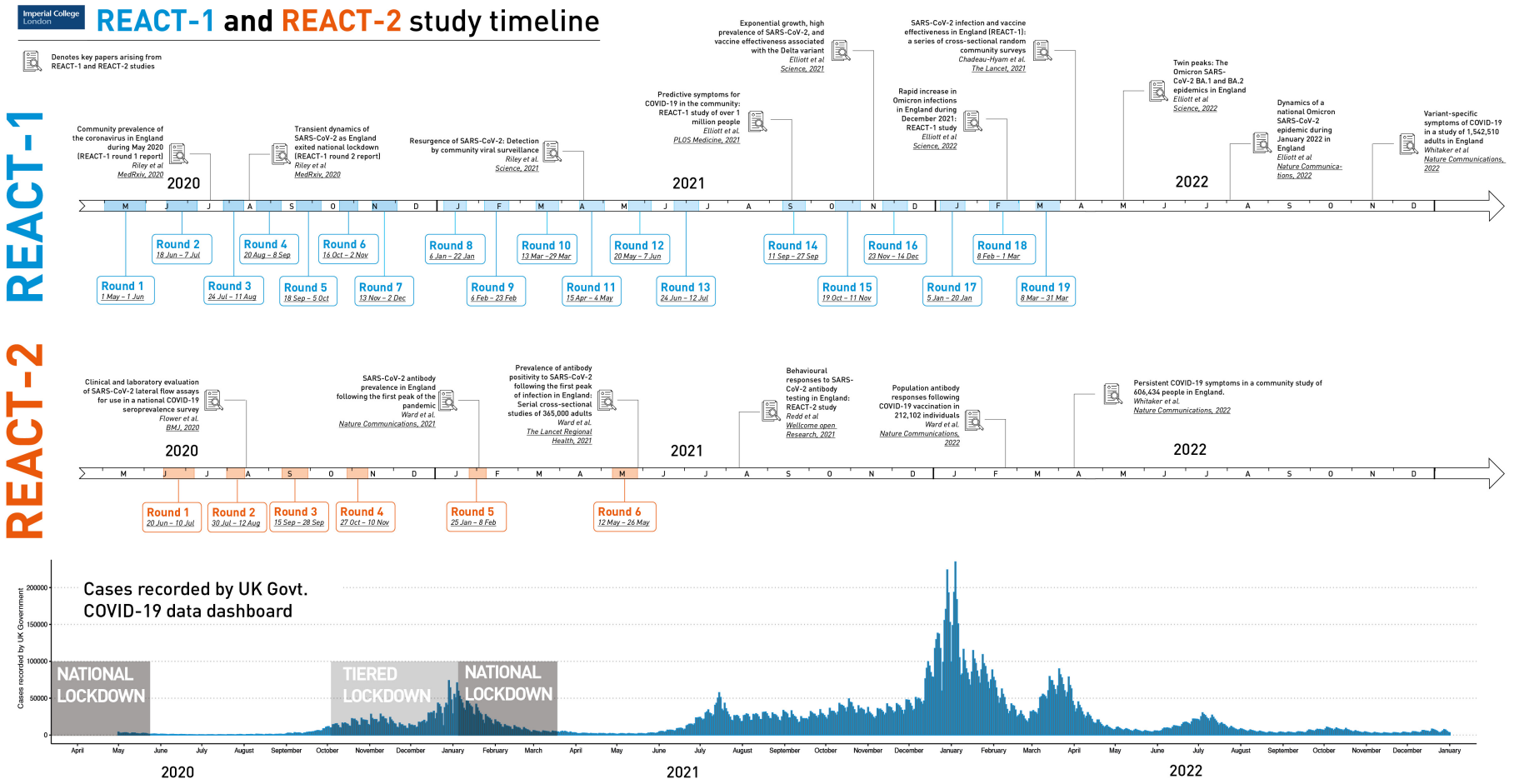


Figure 1.1: Timeline of REACT-1 and REACT-2 studies. Key research papers are annotated, in addition to fieldwork dates. Bottom plot shows daily cases as recorded by UK Government [204]

Round	Month	Date sample drawn	Date invitation letters sent	Fieldwork dates	Number of registration letters sent out	Number of PCR kits sent out	Number of PCR tests returned	Response rate (tests/letters)	PCR test return rate (tests/kits)
1	May 2020	29/04/2020	30/07/2020	01/05/20 – 01/06/20	395,020	161,497	120,620	30.5%	74.7%
2	June 2020	29/05/2020	11/06/2020	19/06/20 – 07/07/20	600,000	219,633	159,199	26.5%	72.5%
3	July 2020	10/07/2020	14/07/2020	24/07/20 – 11/08/20	710,000	225,615	162,821	22.9%	72.2%
4	August 2020	06/08/2020	11/08/2020	20/08/20 – 08/09/20	710,000	211,291	154,325	21.7%	73.0%
5	September 2020	03/09/2020	08/09/2020	18/09/20 – 05/10/20	727,312	232,158	174,949	24.1%	75.4%
6	October 2020	21/09/2020	06/10/2020	16/10/20 – 02/11/20	725,158	211,438	160,175	22.1%	75.8%
7	November 2020	29/10/2020	05/11/2020	13/11/20 – 03/12/20	750,140	213,192	168,181	22.4%	78.9%
8	January 2021	07/12/2020	10/12/2020	06/01/21 – 22/01/21	761,007	213,746	167,642	22.0%	78.4%
9	February 2021	20/01/2021	26/01/2021	06/02/21 – 23/02/21	760,000	210,046	165,456	21.8%	78.8%
10	March 2021	19/02/2021	02/03/2021	13/03/21 – 29/03/21	760,000	180,106	140,844	18.5%	78.2%
11	April 2021	19/03/2021	01/04/2021	15/04/21 – 04/05/21	823,812	170,503	127,408	15.5%	74.7%
12	May 2021 *	23/04/2021	04/05/2021	20/05/21 – 07/06/21	814,633	161,037	108,911	13.4%*	67.6%*
13	June 2021 *	26/05/2021	07/06/2021	24/06/21 – 12/07/21	841,227	146,084	98,233	11.7%*	67.2%*
14	September 2021 *	20/08/2021	23/08/2021	11/09/21 – 27/09/21	822,176	146,140	100,527	12.2%*	68.8%*
15	October 2021 *	21/09/2021	04/10/2021	19/10/21 – 05/11/21	859,184	141,892	100,112	11.7%*	70.6%*
16	December 2021 *	07/11/2021	11/11/2021	23/11/21 – 14/12/21	803,864	129,534	97,089	12.1%*	75.0%*
17	January 2022 *	26/11/2021	06/12/2021	05/01/22 – 20/01/22	840,529	138,925	102,174	12.2%*	73.5%*
18	February 2022 *	17/01/2022	24/01/2022	08/02/22 – 01/03/22	635,000	133,118	94,950	15.0%*	71.3%*
19	March 2022 *	16/02/2022	21/02/2022	08/03/22 – 31/03/22	697,055	147,640	109,181	15.7%*	74.0%*
					14,036,117	**3,393,595**	**2,512,797**	**17.9%**	**74.0%**

Round	Month	Date sample drawn	Date invitation letters sent	Fieldwork dates	Number of registration letters sent out	Number of LFT kits sent out	Number of LFT tests completed	Response rate (tests/letters)	LFT completion rate (tests/kits)
1	July 2020	10/06/2020	15/06/2020	20/06/20 - 10/07/20	314,998	121,976	105,651	33.5%	86.6%
2	August 2020	16/07/2020	21/07/2020	30/07/20 - 12/08/20	344,737	126,960	107,382	31.1%	84.6%
3	September 2020	27/08/2020	02/09/2020	15/09/20 - 28/09/20	560,375	194,484	161,565	28.8%	83.1%
4	November 2020	07/10/2020	13/10/2020	27/10/20 - 10/11/20	565,000	193,611	164,213	29.1%	84.8%
5	February 2021	04/01/2021	11/01/2021	25/01/21 - 08/02/21	600,018	194,762	157,698	26.3%	81.0%
6	May 2021	19/04/2021	26/04/2021	12/05/21 - 25/05/21	749,225	255,750	209,482	28.0%	81.9%

Table 1.3: Summary of sample extraction and fieldwork dates, and response rates, by round, for REACT-1 (top) and REACT-2 (bottom)

1.6 Analytical tools

In this thesis the tools of classical statistics and epidemiology are used alongside more advanced computational approaches designed for high dimensional data, and, where appropriate, machine learning approaches to mine large data sets for patterns.

1.6.1 Classical statistical methods

All of the analyses in this thesis make use of 'classical' biostatistics: descriptive and inferential statistics, parametric and non-parametric tests, univariable and multivariable analysis using eg linear and logistic regression, survival analysis.

Partly as a byproduct of working on the REACT studies and the requirement to produce repeated analyses of large data sets in near real-time the author created an R package, 'OverReact' [205], to automate some of the more common tasks, including:

- Creation of cross-tabs and table-one in consistent format, with integrated confidence intervals, row-wise or column-wise percentages, and statistical tests to compare group differences
- Creation of Odds Ratio tables based on a binary outcome modelled using logistic regression
- Creation of prevalence tables
- Formatting of figures in consistent style using Imperial College colour palette [206]
- Utility functions for saving and exporting figures and tables
- Data-cleaning functions

1.6.2 Stability-based approaches

Where variable selection is required, stability-based approaches have been shown to be effective in identifying sparse sets of variables that contribute the maximum orthogonal information to

predicting a given outcome. Briefly, the principle is to run multiple iterations of a predictive model on subsamples of a dataset, extracting a metric of variable importance each time, then averaging the variable importance across all subsamples and selecting all variables whose average importance exceeds a given threshold, which may be calibrated or chosen arbitrarily. Bodinier *et al* have implemented an extension of stability selection as applied to the LASSO (Least Absolute Shrinkage and Selection Operator), which uses an internally calibrated threshold to select variables based on the proportion of subsampled iterations in which a variable is selected by a LASSO model [207]. This model is used as the main variable selection engine in chapter three to identify sparse sets of symptoms that are maximally predictive of specific COVID-19 variants.

1.6.3 Supervised machine learning approaches

For supervised machine learning, the algorithm most commonly used in this thesis is boosted tree modelling as implemented in Catboost [208]. Boosted tree models work on the principle of weak learners (ie decision trees) built sequentially, where each subsequent tree is being fitted on the errors of the previous tree [209]. They have been shown to perform extremely well on prediction tasks [210, 211]. Catboost is similar to other popular Gradient Boosting Machines (GBMs) such as XGBoost [212] and LightGBM [213], but with some advantageous features: it can handle missing data without exclusion or imputation; it has been shown to give excellent performance with default parameters and requires less hyperparameter tuning [214]; it can be trained across multiple CPU cores or on an GPU; it handles categorical features natively, without the need for one-hot encoding (expanding individual categorical features into multiple binary features); and it has been shown to train more quickly than other GBMs [211]. Finally, it features some integrated explanatory tools for feature importance and attribution, including integrated calculation of SHAP (SHapley Additive exPlanation) values [215].

The author has compiled an R package, stabilityshap [216], which uses Catboost as an underlying algorithm to automate the application of some more complex supervised learning tasks, and some black-box model explanation methods. Some of these are novel implementations of experimental methods; some are implementations of more established ML tasks. These include:

- **Cross-validation:** Automated K-fold cross validation to estimate model generalisability,

and to gain more stable estimates of feature importance based on multiple model trains.

- **Boruta feature selection:** An implementation of the Boruta [217] feature ranking and selection method, previously only applied to random forest models. Briefly, the principle is to create a permuted 'shadow' feature for every feature, train a model on the full matrix of features and shadow features, deriving feature importance scores for each feature and shadow feature. The feature importance scores for the shadow features are taken as a null score. Features whose importance exceeds the maximum shadow importance score are 'hits'. This process is repeated N times with new permutations each time. The proportion of hits per feature across all runs is taken as a more robust feature importance measure; a selection threshold can be set at 95% hits.

- **SHAP analysis with stability extension:** An experimental implementation of an idea discussed by Scott Lundberg, who developed SHAP values, here: (https://github.com/slundberg/shap/issues/13), to apply a stability analysis to SHAP values derived from a predictive model. This is designed to overcome an innate problem with SHAP values, which is the possibility of spurious attributions arising from poor predictive models.

- **Permutation feature importance:** An implementation of a permutation-based method for computing feature importance in a model [218]. Briefly, a model is trained on all data in a training set, and performance evaluated on a holdout data set using an appropriate loss measure. Next, each feature in turn is permuted, and performance re-evaluated on the holdout data. The degradation in performance when each feature is permuted can be interpreted as a measure of the marginal contribution of that feature to the predictive accuracy of the model. To obtain a distribution and smooth out the effects of randomness, the permutation process is repeated 100 or 1,000 times (depending on data size and computational feasibility), and the average loss function change calculated. If required, feature selection can be performed by setting a threshold for the proportion of permuted runs that have a lower performance than the unpermuted run: eg a feature could be selected with $p<0.05$ if 950 or more of the 1,000 permuted runs have degraded performance.

All methods use the principle of a 'watchlist' to calibrate the complexity of the underlying

model and avoid overfitting. A watchlist is a subset of the training data (15% is a suitable size) that is withheld from the model. After each additional tree is added to the model in training, model performance is evaluated on the watchlist. When performance on the watchlist stops improving, the model ceases training.

1.6.4 Unsupervised approaches

Elsewhere in this thesis, unsupervised machine learning algorithms are used to mine for clusters in post-COVID-19 symptomatology data, and to identify latent topics in the abstracts of COVID-19 research papers. The specific methods are described in detail in the relevant chapters.

Chapter 2

The COVID-19 Research Landscape: publication Patterns, Media Coverage, and Evolving Topics

This chapter uses data on more than 17 million research papers to describe and analyse the scientific information ecosystem throughout the COVID-19 pandemic, using the tools of big data, statistics and bibliometrics.

There are two research goals. The first is primarily descriptive, and amounts to an algorithmic literature review of the full corpus of COVID-19 research output: quantifying the volume of research output over time, identifying highly cited papers and papers that generated the most media and social media attention, using textual analysis to identify key research topics and how they evolve over time, and comparing the characteristics of COVID-19 papers and preprints to non-COVID-19 papers. The second goal is to investigate questions relating to the functioning of the scientific information ecosystem, with a particular eye on the role of preprints in that ecosystem. Did COVID-19 research 'cannibalise' other research? Were the 'best' papers (as measured by traditional citation and journal quality metrics) getting the most attention (as measured by Altmetrics)? Was the huge wave of COVID-19 preprints in 2020 of lower average quality than other preprints (using time-to-event analysis to describe publication trends)? And what were the characteristics of the preprints that were most likely to be published in high-impact journals?

2.1 Motivation

The COVID-19 pandemic has induced an unprecedented volume of research output, far more than any researcher can keep abreast of. Therefore fully describing and characterising the corpus of COVID-19 research helps create a means to navigate the huge volume of information and identify important themes, papers and trends.

Secondarily, a functioning research economy is an essential for the scientific community and for society as a whole. Therefore an investigation of how the publishing ecosystem (encompassing the researchers, publishers, communicators, general public, and policymakers) functioned under the acute stress of a pandemic may yield actionable insights that could help improve the research response to future public health crises.

2.2 Web app for data exploration

A web dashboard has been built by the author, to search and interrogate the data used in this chapter:

https://lookerstudio.google.com/reporting/f1f0a990-1591-4975-acfe-81bcfeac9e36

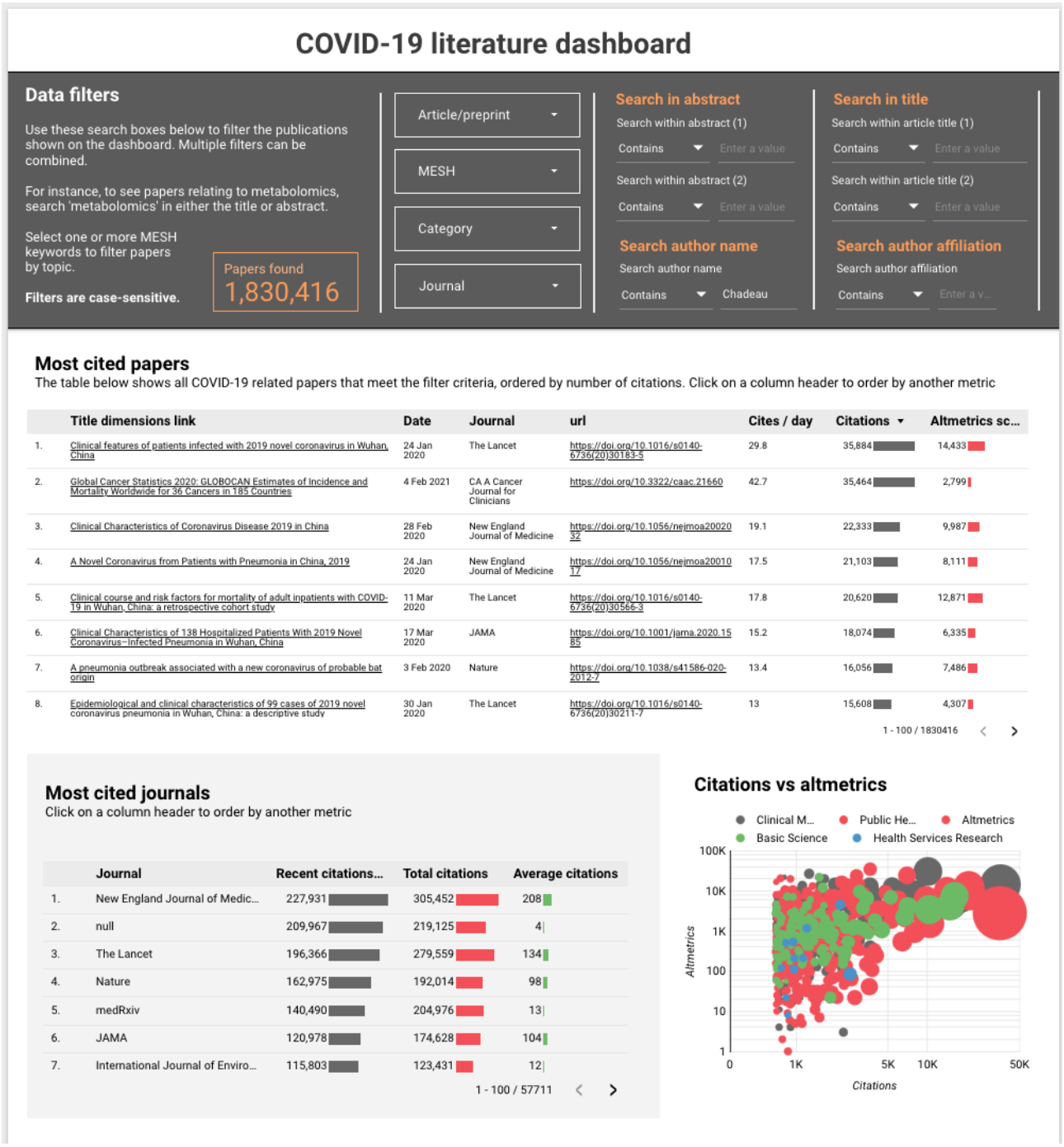


Figure 2.1: Dashboard for exploring COVID-19 publications.

2.3 Methods

2.3.1 Data set

The Dimensions research database is a 'real-time' bibliographic data set that collates multiple different types of research object — referenced journal articles, preprints, patents, clinical trials, grants and policy documents [219] — from across all fields of science, arts and humanities [1]. Rich metadata is available for each article, including article title and abstract text, author details, publication dates, DOI and Altmetrics ID (which permits linking with more extensive Altmetric data through their API).

A slice of data from 1 January 2019 to 31 December 2022 was extracted from the Dimensions research database. Non-English language papers were filtered using Google's Compact Language Detector [220]. Book chapters, monographs, conference proceedings, and edited book types were excluded, leaving only research articles and preprints (See Figure 2.2 for data flow chart and exclusions).

Criteria for designation as a COVID-19 article

Articles were classified as COVID-19-related if the title text returned a positive result for the following Boolean search: "2019-nCoV" OR "COVID-19" OR "SARS-CoV-2" OR "HCoV-2019" OR "hcov" OR "NCOVID-19" OR "severe acute respiratory syndrome coronavirus 2" OR "severe acute respiratory syndrome corona virus 2" OR (("coronavirus" OR "corona virus") AND (Wuhan OR China OR novel))

Dimensions curate an open access COVID-19-specific publications database [221], which has been used for some already published research into the COVID-19 literature [222]. The inclusion criteria are broader than those applied in this analysis: they include any paper that returns a positive in a Boolean search for the same set of COVID-19 keywords above in the title or abstract or full text [223]. This results in the inclusion of a large number of papers that are not primarily about COVID-19 but merely reference it obliquely. Therefore, for the purposes

[1] Articles from the sciences are more heavily represented owing to the greater volume of research output in scientific fields. Figure A.2 and Figure A.1 in Appendix A show the distribution of COVID-19 and non-COVID-19 articles by field

		Chapter two COVID-19 data set		Dimensions COVID-19 data set	
		Not included	Included	Not included	Included
Manual evaluation	Not included	TN= 96	FP= 0	93	3
	Included	FN= 0	TP= 4	0	4
		Accuracy: 1 Precision: 1 Recall: 1 F1 score: 1		Accuracy: 0.97 Precision: 0.57 Recall: 1 F1 score: 0.73	

Table 2.1: Confusion matrix showing the results of algorithmic detection of COVID-19 papers vs Dimensions COVID-19 inclusion criteria vs manual assessment of paper topic (COVID-19 / not COVID-19)

of this analysis, it is desirable to apply a stricter filter.

To compare the papers included in this database with the filtered data set used here and justify a stricter filter, a random sample of 100 papers was taken from the Dimensions COVID-19 data set, and read to manually judge whether COVID-19 was a central element of the paper. The human judgements were then compared with the Boolean search results for inclusion in the data set used here, and a confusion matrix and associated statistics were created.

The Boolean search criteria applied in this chapter performed better at identifying papers with COVID-19 as a central element. As expected, the Dimensions COVID-19 data set included some papers (three of 100) which are on unrelated topics but mention COVID-19: *On neighborhood Zagreb index of product graphs* by Mondal *et al* [224], *Undergraduates' Sense of Belonging, Community, and Support at a Title V Center* by Lucas and Roberts [225], and *Indoor air quality analysis in naturally ventilated university training laboratories: a health risk assessment* by Seseña *et al* [226].

Additional data harvesting

Additional data on citations was gathered using the CrossRef API as implemented in the R package *rcrossref* [17]. Data on non-citation-related impact metrics was gathered through the Altmetrics API as implemented in the R package rAltmetric [18]. Altmetrics include a range of metrics measuring article shares and mentions on different social media platforms, on Wikipedia and Youtube, in the news media, on blogs, on reference managers and in policy documents. They also generate an aggregated Altmetrics Score, which represents a weighted sum of attention across all tracked digital platforms. The exact algorithm is not in the public

domain but a summary of weightings is available online[2]. Journal-level impact and citation data were gathered from the Thompson Reuters journal citation reports service [19]. See Table A.1 in Appendix A for a list of all data points and sources.

Linking preprints to subsequent publications

Where preprints have later been published in a journal, the preprint is linked to this subsequent publication in the Dimensions data set. Several mechanisms exist to detect subsequent publications. First, Crossref monitors preprints and journal articles and, when it detects and article that was previously preprinted, the preprint server is notified [227]. Second, some of the larger preprint servers (eg MedRxiv and BioRxiv) monitor their own preprints for subsequent publication and link to the journal article when the preprint authors confirm a link [227]. Finally, authors may notify the preprint server when a preprint has been published in a journal and ask for it to be linked [227].

2.3.2 Statistical analyses

Descriptive analyses

Primary descriptive analyses were conducted to characterise the full data set, including summary tables and plots of the 'epidemic curve' of published literature on COVID-19. Correlation analysis was conducted to examine relationships between metrics of academic article impact (citation metrics and publishing journal impact metrics) and non-academic article impact (Altmetrics), and to see how these relationships vary between COVID-19 and non-COVID-19 papers, and between journal articles and preprints.

To investigate changing rates of publishing output, linear regression models were fit to the monthly publication data, with the specification:

$$y[PapersPerMonth] \sim \alpha + \beta[Months] \tag{2.1}$$

[2]https://tinyurl.com/4jhvzvs3

To test whether the pandemic era was associated with a change in the rate of publication of non-COVID-19 research, a further regression model was fit to the monthly publication data including a binary independent variable denoting pre-pandemic (2019) or pandemic era (2020-2022), including an interaction between the independent variables:

$$y[PapersPerMonth] \sim \alpha + \beta[Months] + \beta[PandemicEra] + \beta[Months] * \beta[PandemicEra] \tag{2.2}$$

Topic Modelling

Topic modelling is a method for making large corpuses of text computationally tractable by revealing underlying structure, and it has been shown to be effective for conducting exploratory literature reviews at scale [228]. In this analysis, topic modelling was applied to the abstracts[3] of the COVID-19 papers using Latent Dirichlet Allocation (LDA), as implemented in the R package topicmodels [20]. LDA is a generative probabilistic model that aims to infer the underlying topic structure that best explains the observed document-word co-occurrence patterns in a corpus of text, by assuming that each document is a mixture of topics, and each topic is a mixture of words.

To prepare the data set, the abstract text corpus was cleaned: papers without abstracts or with incomplete abstracts were excluded altogether; html and xml artefacts, 'stop words' (words that do not add significant meaning to a sentence) and common conjunctions (eg the, and, with, etc etc) were removed. Text data were summarised into a bag of words [21] matrix containing any word or bigram (combination of two words) that occurred 1,000 times or more in the corpus (the threshold of 1,000 was set as a tradeoff between fidelity and computational feasibility). The final matrix dimensions were 348,934 (papers/rows) * 5,497 (ngrams/columns).

After initialising the LDA model with random topic assignments for each word in each document, variational Gibbs sampling was used to iteratively update the topic assignments for each word in each document based on the probabilities of each word belonging to each topic, and the probabilities of each topic occurring in each document. To choose k, the number of topics

[3]In an ideal world, topic modelling would be applied to the full text of the papers. However, given the size of this data set, this was not computationally feasible

in the final model, a number of models were fit with values of k ranging between between 5 and 500. The final model was selected with the value of k that maximised the average pairwise Jensen–Shannon divergence (a metric that measures the distance between two probability distributions) (D) between probability distributions of topics [22].

The final output from the fitted model with optimal k is two matrices: one matrix of β values, with topics as rows and words as columns, where β represents the probability of a word appearing in a topic; and one matrix of γ values with documents as rows and topics as columns, where γ represents the distribution of topics within each document.

Within each topic:

$$\sum_{i=1}^{n} \beta = 1$$

And within each document:

$$\sum_{i=1}^{t} \gamma = 1$$

The mix of topics in a given document is therefore zero sum – if a document is enriched in one topic, that topic must necessarily be 'taking up space' that other topics could have occupied. This intuition has implications for the association analysis presented in this chapter.

Investigating drivers of attention and citation

To investigate the relationship between article topics and (i) media attention and (ii) citation rates, univariable linear regression models were constructed as:

$$y[citationrate/altmetricscore] \sim \ \beta[Topicgammascore] \tag{2.3}$$

Where the Gamma score represents the 'quantity' of a given topic in a given paper. All topics were modelled in this way for each outcome, and the results visualised as volcano plots. Regression coefficients were compared across outcomes to see whether the same associations

existed for citations as for Altmetrics, and to identify topics that had different effects for the two outcomes. To examine how drivers of attention evolved over the course of the pandemic period, this association analysis was repeated in rolling time windows of 8 weeks between 1 January 2020 and 31 December 2022.

Rates of subsequent publication in preprints

Rates of publication within one year and two years were calculated for all preprints. Additionally, time-to-event analysis was conducted (i) on all preprints (COVID-19 related and non-COVID-19 related) from 2019 to 2022, and (ii) on preprints (COVID-19 related and non-COVID-19 related) published between 1 January 2020 and 1 August 2020. The 'event' was publication, and results were visualised as Kaplan-Meier plots.

Predictors of subsequent high-impact publication in preprints

To investigate predictors of high-impact publication, a subset of articles that were preprinted in 2020 was selected. Later articles were excluded to ensure that all articles under analysis would have enough time to be peer reviewed for publication. The binary outcome variable of subsequent publication y/n was modelled against the citation and Altmetrics variables of the preprint, in addition to the topic β data. Logistic regression models were constructed to test univariable associations, then multivariable boosted tree models, as implemented in Catboost [229], were constructed, and the predictive performance of different families of predictor variables (citation metrics, Altmetrics, abstract topics) was evaluated using the Area Under the Received Operating Characteristic (AUC). Variable importance within multivariable models was calculated using a combination of SHAP feature attribution [230] and computation of the change in loss (measured in AUC) incurred when each variable was removed from the model.

2.4 Results

After excluding (N=4,612,729) non-English language papers and (N=3,796,841) document types other than journal articles and preprints, the final 2019–2022 data set comprised 17,736,043 documents. 3,901,322 of these were published in 2019 (the pre-pandemic period) and 13,834,721 in the pandemic period of 1 January 2020 – 31 December 2022. Throughout the pandemic period, 486,030 papers were classified as COVID-19 papers (420,860 journal articles and 65,170 preprints). This constitutes 3.4% of English-language journal articles, and 5.0% of preprints published during that time (Figure 1, Table 2.2).

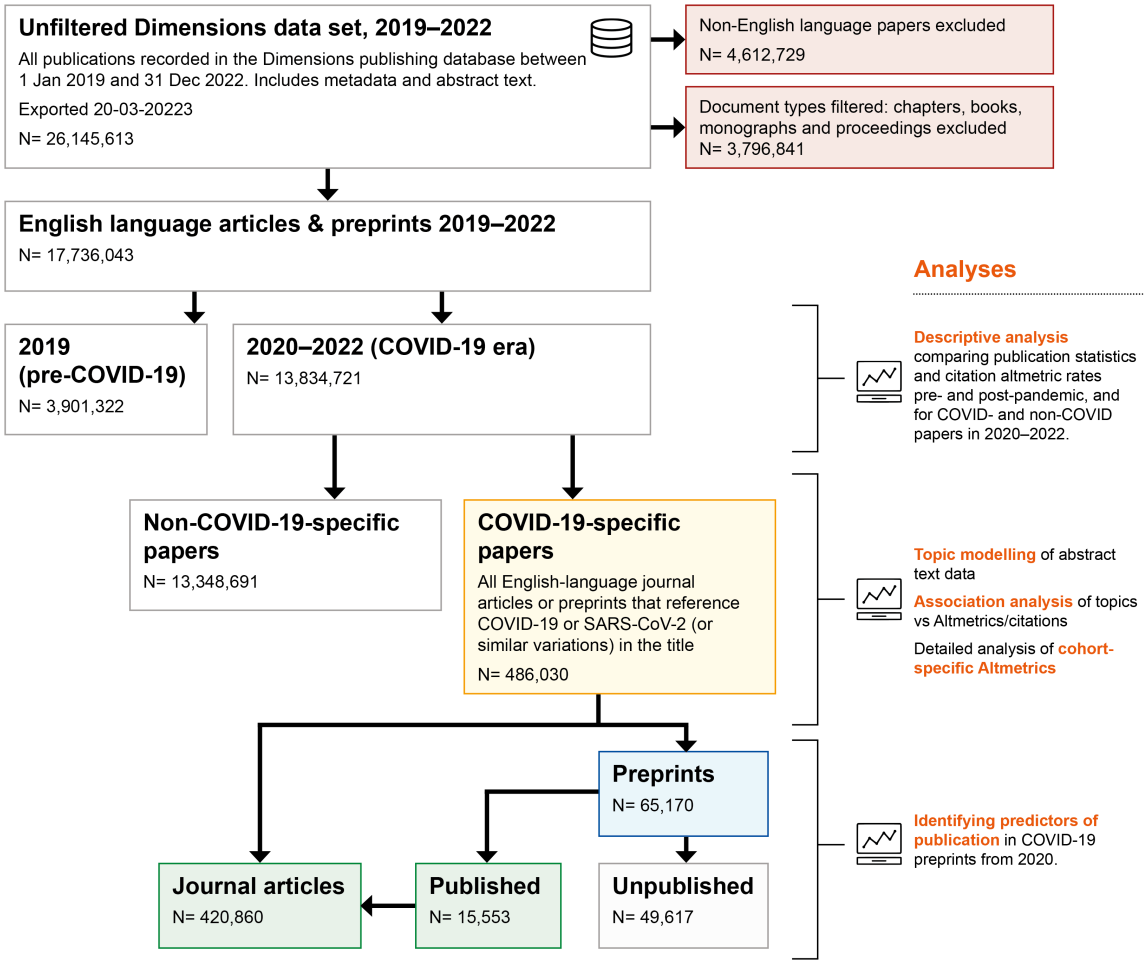


Figure 2.2: Flowchart showing study population and exclusions.

		2019		*2020-2022*				
Variable	**Category**	**Journal article (2019)**	**Preprint (2019)**	**Journal article (2020-2022)**	**Preprint (2020-2022)**	**COVID-19 journal article**	**COVID-19 preprint**	**Total**
Papers published	nobs	3,637,951 (100%)	263,371 (100%)	12,117,213 (100%)	1,231,478 (100%)	420,860 (100%)	65,170 (100%)	17,736,043
Year	2019	3,637,951 (100%)	263,371 (100%)	-	-	-	-	3,901,322
	2020	-	-	3,906,106 (32.2%)	370,085 (30.1%)	112,964 (26.8%)	30,880 (47.4%)	4,420,035
	2021	-	-	4,099,681 (33.8%)	396,010 (32.2%)	164,070 (39%)	21,154 (32.5%)	4,680,915
	2022	-	-	4,111,426 (33.9%)	465,383 (37.8%)	143,826 (34.2%)	13,136 (20.2%)	4,733,771
Published in journal	No	-	155,671 (59.1%)	-	899,943 (73.1%)	-	49,617 (76.1%)	17,281,255
	Yes	-	107,700 (40.9%)	-	331,535 (26.9%)	-	15,553 (23.9%)	454,788
Published within one year	No	-	178,681 (67.8%)	-	953,813 (77.5%)	-	50,700 (77.8%)	17,359,218
	Yes	-	84,690 (32.2%)	-	277,665 (22.5%)	-	14,470 (22.2%)	376,825
Published within two years	No	-	167,036 (63.4%)	-	933,622 (75.8%)	-	50,004 (76.7%)	17,326,686
	Yes	-	96,335 (36.6%)	-	297,856 (24.2%)	-	15,166 (23.3%)	409,357
Number of citations	Mean (SD)	9.87 (31.23)	0.58 (3.46)	4.06 (19.76)	0.31 (2.66)	16.22 (131.32)	5.32 (22.74)	5.23 (29.81)
Relative citation ratio	Mean (SD)	1.38 (3.16)	-	1.55 (6.58)	1.09 (2.21)	4.63 (21.2)	1.65 (3.1)	1.64 (7.17)
Field citation ratio	Mean (SD)	3.42 (10.46)	0.24 (1.55)	2.87 (12.19)	0.25 (2.01)	13.68 (76.15)	3.36 (12.22)	3.02 (15.72)
Altmetrics score	Mean (SD)	4.29 (44.07)	5.49 (42.23)	3.98 (49.95)	4.23 (41.78)	36.99 (378.39)	44.72 (360.89)	5.02 (78.47)
Time from preprint to article publication	Mean (SD)	-	194.93 (194.72)	-	140.42 (158.21)	-	141.88 (121.3)	153.52 (168.15)

Table 2.2: Summary of all papers indexed in the Dimensions database in 2019, 2020, 2021 and 2022. Papers are separated into preprints and journal articles, and into COVID-19 and non-COVID-19 papers. COVID-19 papers are so designated if COVID-19/SARS-CoV-2 (or similar variations) appear in the article title.

2.4.1 Changing publication volumes

The monthly number of COVID-19 journal articles being published rose from zero at the end of 2019 to more than 10,000 a month by May 2020 and remained stable throughout the rest of the pandemic period, with no obvious sign of decline (Figure 2). COVID-19 preprints spiked to a high of more than 4,000/month in May 2020 but have been on a steadily declining trend since then, to fewer than 1,000/month in the last quarter of 2022.

Non-COVID-19 journal article publication rates have remained in a range between 250,000 and 400,000 per month throughout 2019–2022. Regression modelling showed a small average increase of 1,400 papers per month (p=0.000008). The rate of pre-printing of non-COVID-19 papers increased steadily from around 20,000/month at the start of 2020 to more than 40,000/month by the end of 2021, although they seem to have plateaued throughout 2022.

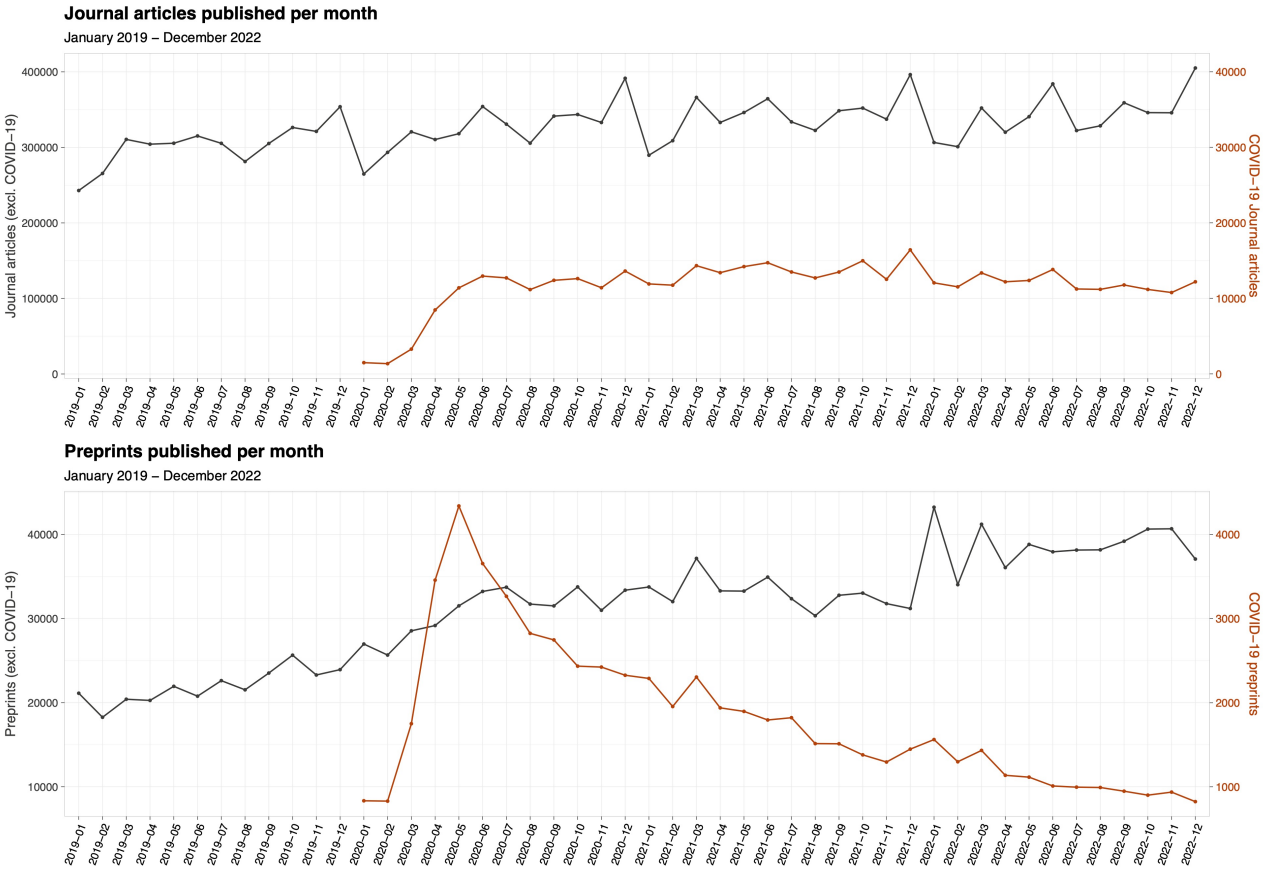


Figure 2.3: Comparison of monthly publication numbers of journal articles and preprints between 2019 and 2022. Grey line (left axis) shows all non-COVID-19 related papers; red line (right axis) shows all COVID-19 related papers. COVID-19 journal articles rise to 12000/month in March 2020 and plateau, while non-COVID-19 articles remain consistent at around 300,000–400,000/month throughout 2019–2022. COVID-19 preprints spike in March 2020, to a high of 4,000/month, and then steadily decline over the remainder of the study period, while numbers of non-COVID-19 preprints are on a steady upward trajectory from 20,000/month the start of 2019 to 400,000/month at the end of 2021, then plateau for the remainder of the study period. NOTE: left and right axes are on different (1:10) scales.

2.4.2 Evidence of diversion of research output from other areas into COVID-19

Monthly non-COVID-19 publication rates are higher in the COVID-19 era (2020–2022) than the pre-COVID-19 era (2019): an average of 336,576 non-COVID-19 journal articles and 34,201 preprints were published each month in 2020–2022, compared with 303,082 and 21,941 respectively in 2019. However, the rate of scientific output increases each year (approximately 5% annually, according to Larsen *et al* [231]), so an increase is to be expected. The data presented here offer some mixed evidence of a decline in the rate of growth of non-COVID-19-related research output during the pandemic period, compared with 2019. As shown in Figure 2.4, in journal articles, the slope of the regression line is shallower in the pandemic era compared with the pre-pandemic era (p=0.03), suggesting a decline in growth rate during the pandemic period. In preprints, there is no change in the slope (p=0.44), but the overall level of output is increased in the pandemic era relative to the pre-pandemic era. The patterns were consistent within medical journal articles and bioRxiv/medRxiv preprints (Appendix A, Figure A.4).

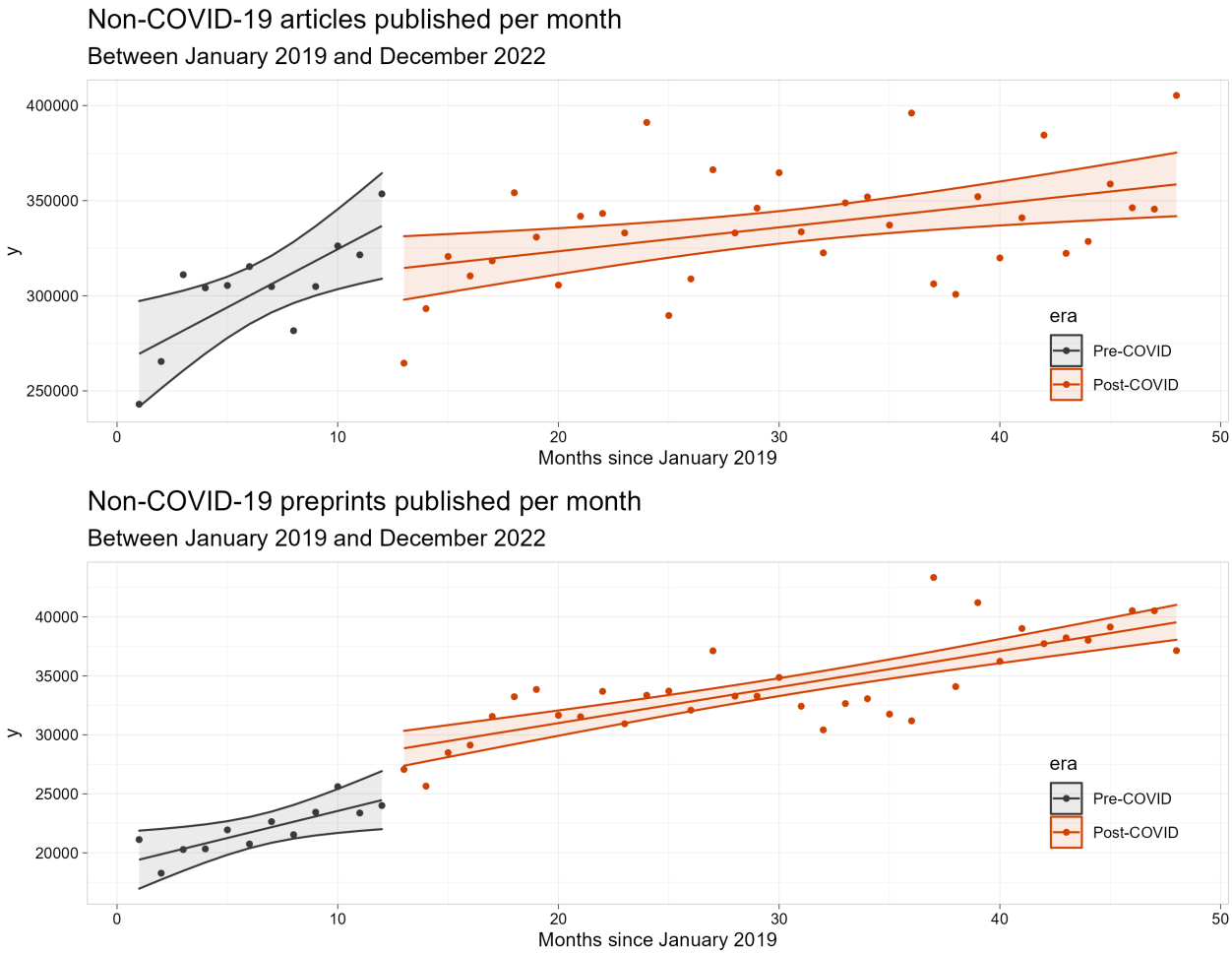


Figure 2.4: Publication rates of COVID-19 papers in the pre-pandemic period (2019) and the pandemic period (2020-2022). Regression lines and 95% CIs are shown with articles/month regressed onto time (in months since January 2019) and a binary pre-pandemic/post pandemic variable, plus an interaction effect between these two independent variables. In journal articles, this interaction effect is significant (p=0.03), suggesting a decline in growth rate during the pandemic period. In preprints, the interaction effect is not significant (p=0.44), while the main effect for the pre/post-pandemic binary variable is significant (p=0.003), with a positive effect, suggesting a boost in preprint publishing and no decline in growth rate.

2.4.3 Citation rates

As of March 2023 (latest data export date), the average COVID-19 journal article had received 16.2 citations and the average COVID-19 preprint 5.32 citations, as compared with 4.06 and 0.31 respectively for non-COVID-19 journal articles and preprints from the same period, and 9.87 and 0.58 for journal articles and preprints from 2019 (Table 2.2). Using the relative citation ratio (which is normalised to the time of publication and the co-citation network of the paper [232]), COVID-19 journal articles scored 3 times as highly as either non-COVID-19 papers or

pre-COVID-19 papers.

The most highly cited COVID-19 papers are, as might be expected, the early research into hospitalised patients in Wuhan from the first three months of 2020 [9, 15–17, 129, 233] (Huang *et al* having received 33,357 citations at the time of export [17]) (Table S3). Also highly cited are the Polack et al, reporting the results of the phase II/III trials of the Pfizer vaccine [89] (which had received 8,574 citations at the time of data export) and the RECOVERY group's paper reporting the results of a Dexamethasone trial [64] (7,172 citations at the time of export).

2.4.4 Altmetrics

COVID-19 research generated nearly 10 times the media and social media attention as non-COVID-19-related research either from 2019 or from 2020–2022: the average Altmetrics score for a COVID-19 journal article was 36.99, compared to 3.98 for a non-COVID-19 article (Table 2.2). Preprints also generated more attention than journal articles, on average (the average Altmetrics score was 44.72 vs 36.99 for COVID-19 preprints vs journal articles). This pattern was consistent across COVID-19 papers and non-COVID-19 papers (the average Altmetrics score was 4.23 vs 3.98 for non-COVID-19 preprints vs journal articles).

Over the course of 2020–2022, the total media and social media attention paid to COVID-19 journal articles peaked in the first wave of COVID-19 (March–May 2020) and gradually declined from there to around one third of the peak level by the end of 2022. For preprints, there were secondary 'waves' in July/August 2021 and December 2021, coinciding with the emergence of the Delta and Omicron waves in the US (Figure 2.5).

The per-paper average media attention shows the disparity between journal article trends and preprint trends more starkly. Per-paper attention for journal articles has declined to one-tenth of the first-wave peak, with no significant peaks in between. Per-paper attention for preprints has seen two significant peaks of equal size to the first wave peak.

These results show that preprints are, on average, greater drivers of public attention than journal articles, and they also point to one of the likely reasons: preprints are by their nature topical in a way that journal articles, by consequence of peer review and production times, cannot be. Therefore when an acute event such as the emergence of a new variant occurs,

preprints naturally fill the information vacuum and become sources and drivers of attention. While this effect can be detected in the COVID-19 papers because it is simple to correlate infection rates with publication rates, the fact that average Altmetrics scores are higher in preprints than journal articles in *non*-COVID-19 papers, as well as COVID-19 papers (Table 2.2), suggests that this is not a pandemic-specific effect.

Accumulation velocity of Altmetrics scores

Previous research has found that the half-life (ie the length of time it takes for Altmetrics to reach half their total value) of Altmetrics scores ranges from 7 and 13 days for Reddit and Twitter respectively, to 22 days for news media and 30 days for Facebook [234]. Analysis here of the accumulation velocity of Altmetrics news and social media scores for the 20 highest-scoring COVID-19 papers showed similar results: the half-life of social media Altmetrics scores was 14 days, with 59.7% of all social media mentions occurring within 30 days of publication (Appendix A, Figure A.3). The half life for news coverage was longer (73 days, with 38.2% of news coverage occurring within the first 30 days)

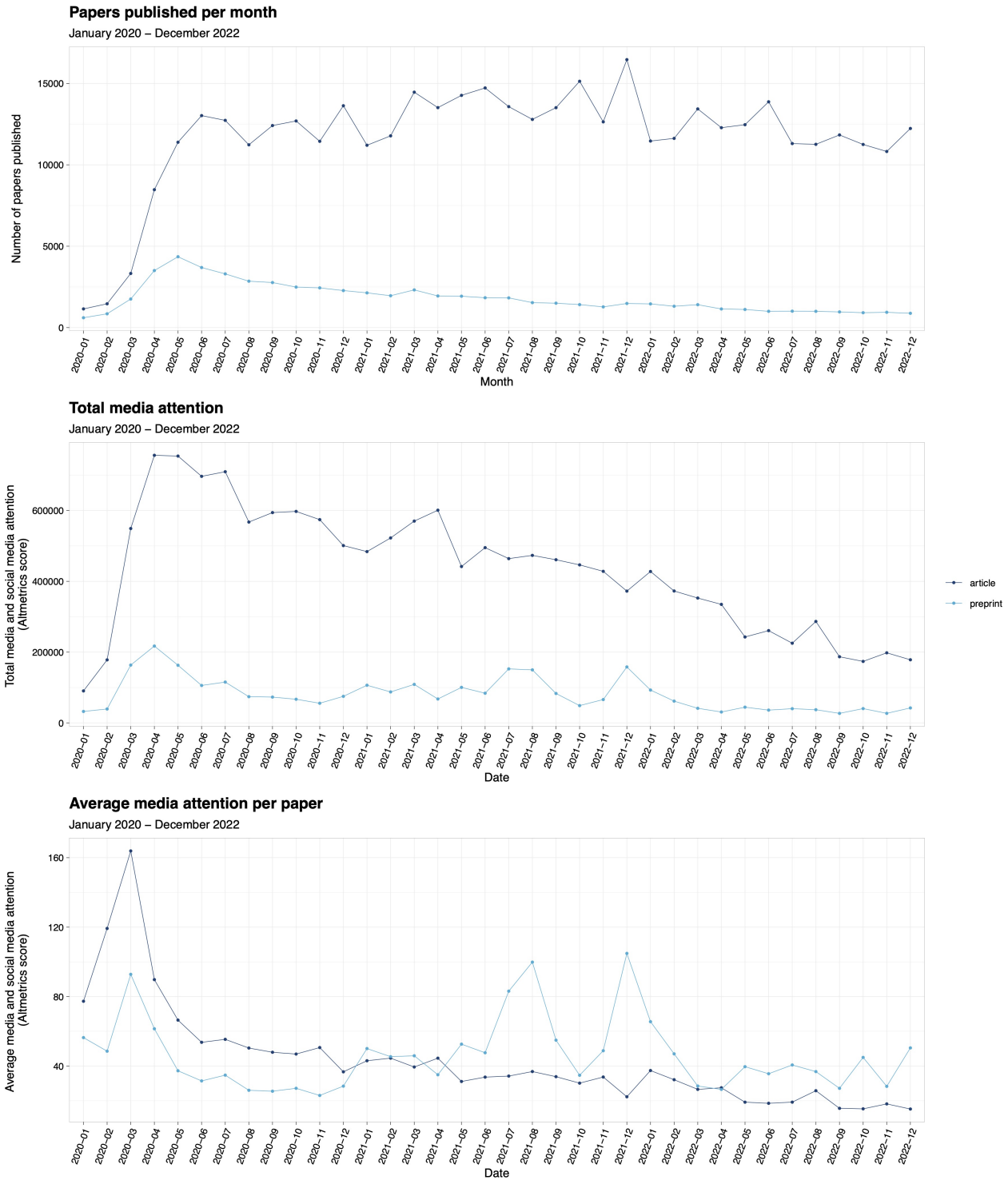


Figure 2.5: Publication rates, total media attention and average media attention for COVID-19 papers, 2019–2022. Top panel: lines show the number of COVID-19 papers published per month for journal articles (dark blue line) and preprints (light blue line). Middle panel: lines show the monthly sum of media attention for journal articles (dark blue line) and preprints (light blue line); bottom: lines show the monthly average of media attention per paper (sum of media attention divided by number of papers) for journal articles (dark blue line) and preprints (light blue line).

Article title	Journal	Altmetrics score	Citations	Publication date	DOI
Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China	The Lancet	14260	33357	2020-01-24	10.1016/s0140-6736(20)30183-5
Clinical Characteristics of Coronavirus Disease 2019 in China	New England Journal of Medicine	10021	20910	2020-02-28	10.1056/nejmoa2002032
A Novel Coronavirus from Patients with Pneumonia in China, 2019	New England Journal of Medicine	7887	19602	2020-01-24	10.1056/nejmoa2001017
Clinical course and risk factors for mortality of adult inpatients with COVID19 in Wuhan, China: a retrospective cohort study	The Lancet	13035	19208	2020-03-11	10.1016/s0140-6736(20)30566-3
Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China	JAMA	6362	17052	2020-03-17	10.1001/jama.2020.1585
Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study	The Lancet	4237	14677	2020-01-30	10.1016/s0140-6736(20)30211-7
SarsCoV2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor	Cell	4183	13515	2020-03-05	10.1016/j.cell.2020.02.052
Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID19) Outbreak in China	JAMA	10934	13042	2020-04-07	10.1001/jama.2020.2648
Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia	New England Journal of Medicine	6787	11767	2020-01-29	10.1056/nejmoa2001316
Safety and Efficacy of the BNT162b2 mRNA COVID19 Vaccine	New England Journal of Medicine	30089	8574	2020-12-10	10.1056/nejmoa2034577
Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding	The Lancet	4167	8371	2020-01-30	10.1016/s0140-6736(20)30251-8
An interactive web-based dashboard to track COVID19 in real time	The Lancet Infectious Diseases	925	7305	2020-02-19	10.1016/s1473-3099(20)30120-1
Clinical course and outcomes of critically ill patients with SarsCoV2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study	The Lancet Respiratory Medicine	3434	7287	2020-02-24	10.1016/s2213-2600(20)30079-5
Dexamethasone in Hospitalized Patients with COVID19	New England Journal of Medicine	9372	7172	2020-07-17	10.1056/nejmoa2021436
Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID19 in the New York City Area	JAMA	7665	6699	2020-05-26	10.1001/jama.2020.6775
Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation	Science	4146	6578	2020-02-19	10.1126/science.abb2507
Aerosol and Surface Stability of SarsCoV2 as Compared with SARS-CoV-1	New England Journal of Medicine	24684	6533	2020-03-17	10.1056/nejmc2004973
A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster	The Lancet	4468	6508	2020-01-24	10.1016/s0140-6736(20)30154-9
Pathological findings of COVID19 associated with acute respiratory distress syndrome	The Lancet Respiratory Medicine	2036	6295	2020-02-18	10.1016/s2213-2600(20)30076-x
Structure, Function, and Antigenicity of the SarsCoV2 Spike Glycoprotein	Cell	2207	6207	2020-03-09	10.1016/j.cell.2020.02.058

Article title	Journal	Altmetrics score	Citations	Publication date	DOI
COVID19: Researcher blows the whistle on data integrity issues in Pfizer's vaccine trial	The BMJ	43374	22	2021-11-02	10.1136/bmj.n2635
The proximal origin of SarsCoV2	Nature Medicine	35225	3418	2020-03-17	10.1038/s41591-020-0820-9
Comparing SarsCoV2 natural immunity to vaccine-induced immunity: reinfections versus breakthrough infections	medRxiv	32733	140	2021-08-25	10.1101/2021.08.24.21262415
Effectiveness of Adding a Mask Recommendation to Other Public Health Measures to Prevent SarsCoV2 Infection in Danish Mask Wearers	Annals of Internal Medicine	32008	218	2020-11-18	10.7326/m20-6817
Ivermectin for Prevention and Treatment of COVID19 Infection: A Systematic Review, Meta-analysis, and Trial Sequential Analysis to Inform Clinical Guidelines	American Journal of Therapeutics	30837	123	2021-06-21	10.1097/mjt.0000000000001402
Safety and Efficacy of the BNT162b2 mRNA COVID19 Vaccine	New England Journal of Medicine	30089	8574	2020-12-10	10.1056/nejmoa2034577
Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID19 vaccine: an interim analysis of a randomised controlled phase 3 trial in Russia	The Lancet	27327	1080	2021-02-02	10.1016/s0140-6736(21)00234-8
Intracellular Reverse Transcription of Pfizer BioNTech COVID19 mRNA Vaccine BNT162b2 In Vitro in Human Liver Cell Line	Current Issues in Molecular Biology	26725	23	2022-02-25	10.3390/cimb44030073
Aerosol and Surface Stability of SarsCoV2 as Compared with SARS-CoV-1	New England Journal of Medicine	24684	6533	2020-03-17	10.1056/nejmc2004973
Increases in COVID19 are unrelated to levels of vaccination across 68 countries and 2947 counties in the United States	European Journal of Epidemiology	24258	49	2021-09-30	10.1007/s10654-021-00808-7
Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SarsCoV2 and COVID19: a systematic review and meta-analysis	The Lancet	23975	2602	2020-06-01	10.1016/s0140-6736(20)31142-9
The epidemiological relevance of the COVID19-vaccinated population is increasing	The Lancet Regional Health - Europe	22917	16	2021-11-20	10.1016/j.lanepe.2021.100272
Post-lockdown SarsCoV2 nucleic acid screening in nearly ten million residents of Wuhan, China	Nature Communications	22079	122	2020-11-20	10.1038/s41467-020-19802-w
RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID19: a multinational registry analysis	The Lancet	22035	766	2020-05-22	10.1016/s0140-6736(20)31180-6
Ten scientific reasons in support of airborne transmission of SarsCoV2	The Lancet	21691	489	2021-04-15	10.1016/s0140-6736(21)00869-2
Identifying airborne transmission as the dominant route for the spread of COVID19	PNAS	20395	819	2020-06-11	10.1073/pnas.2009637117
Detection of Messenger RNA COVID19 Vaccines in Human Breast Milk	JAMA Pediatrics	20075	2	2022-09-26	10.1001/jamapediatrics.2022.3581
Long-term cardiovascular outcomes of COVID19	Nature Medicine	19812	333	2022-02-07	10.1038/s41591-022-01689-3
Preliminary Findings of mRNA COVID19 Vaccine Safety in Pregnant Persons	New England Journal of Medicine	19789	488	2021-04-21	10.1056/nejmoa2104983
Outbreak of SarsCoV2 Infections, Including COVID19 Vaccine Breakthrough Infections, Associated with Large Public Gatherings - Barnstable County, Massachusetts, July 2021	MMWR Morbidity and Mortality Weekly Report	19735	389	2021-08-06	10.15585/mmwr.mm7031e2

Table 2.5: Top 20 COVID-19 papers, ranked by citation count (top) and Altmetrics score (bottom)

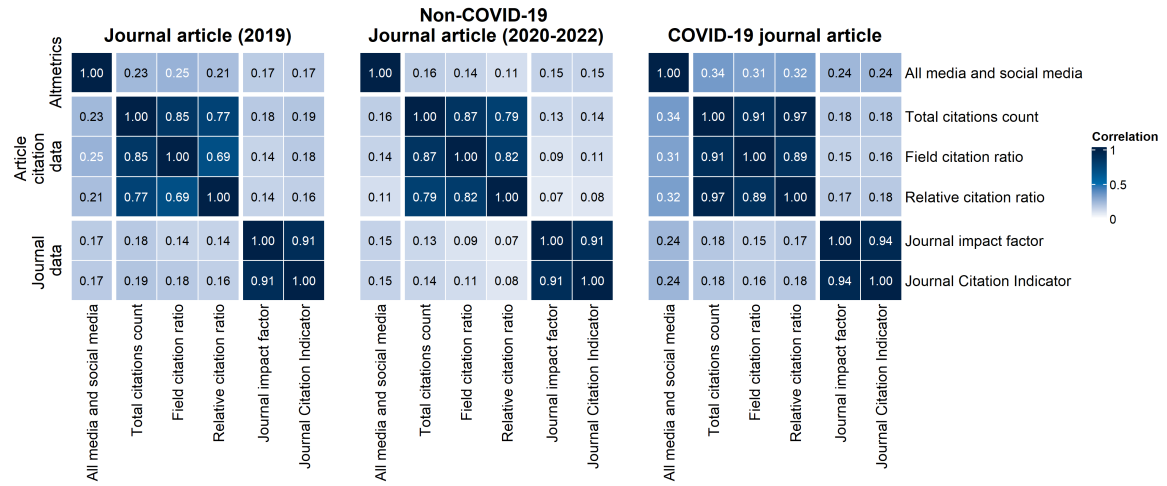
2.4.5 Correlations between Altmetrics and citation metrics

Do Altmetrics scores correlate with citation-based metrics?

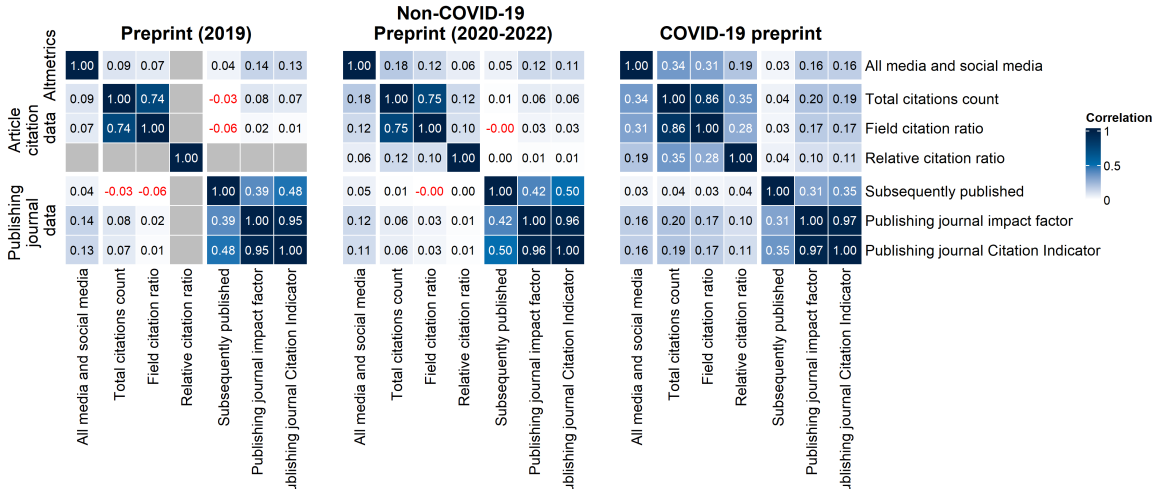
The correlation coefficient between Altmetrics score and field citation ratio indicates the strength of the relationship between which papers are getting recognised and valued by the scientific community (in the form of citations and publication in indexed journals), and what is being paid attention by the news media and by people on social media.

For COVID-19 journal articles, there is a correlation coefficient of 0.34 between the aggregated Altmetrics score and the citation count of the paper (Figure 2.6). This is stronger than the equivalent correlation in either pre-COVID-19 journal articles (from 2019) or non-COVID-19 journal articles from 2020–2022. The same holds true for the normalised citation metrics (citation count, relative citation ratio, field citation ratio), and for journal impact metrics (Journal Impact Factor and Journal Citation Indicator), in both journal articles (Figure 2.6, top row) and preprints (Figure 2.6, bottom row).

These results indicate that the scientific import of an article and the degree of public attention it receives are more closely aligned in COVID-19 papers than in non-COVID-19 papers, although the direction of effect cannot be discerned.



(a) Journal articles



(b) Preprints

Figure 2.6: Heatmaps showing correlation between media attention, as measured by article Altmetrics score, with article citation metrics and journal impact metrics. (a) shows the correlations within journal articles, stratified by pre-pandemic/Non-COVID-19/COVID-19 status. (b) shows correlations within preprints, stratified by pre-pandemic/ Non-COVID-19/ COVID-19 status. Correlations are calculated with Pearson's R, using complete cases only.

Analysis of platform- and cohort-specific Altmetrics in COVID-19 papers

Figure 2.7 explores the correlations between granular Altmetrics (media coverage, mentions on specific social media sites, Wikipedia, Policy document citations, and social media cohorts – general public / scientists / medics / science communications) and measures of article and journal quality (citation rate, total citations, Journal Impact Factor, and Journal Citation Indicator), in COVID-19 journal articles and preprints.

The weakest correlation between article quality metrics (as measured by article citations and journal impact) and Altmetrics is with the social media platform Twitter. Facebook and YouTube mentions correlate more strongly with citation metrics than either Twitter or mentions in the mainstream media. Readers on reference managers, and specifically on Mendeley, have the highest correlations with citation rate (0.86 for both).

These results show that of all the Altmetrics data points available, Twitter mentions are the least closely aligned with conventional, citation-based metrics of article quality. This may be indicative of a lower standard of scientific discourse on Twitter which results in lower-quality papers getting attention, and/or the effect may go in the other direction: Twitter discussion may be less of a driver of subsequent citation than other media platforms.

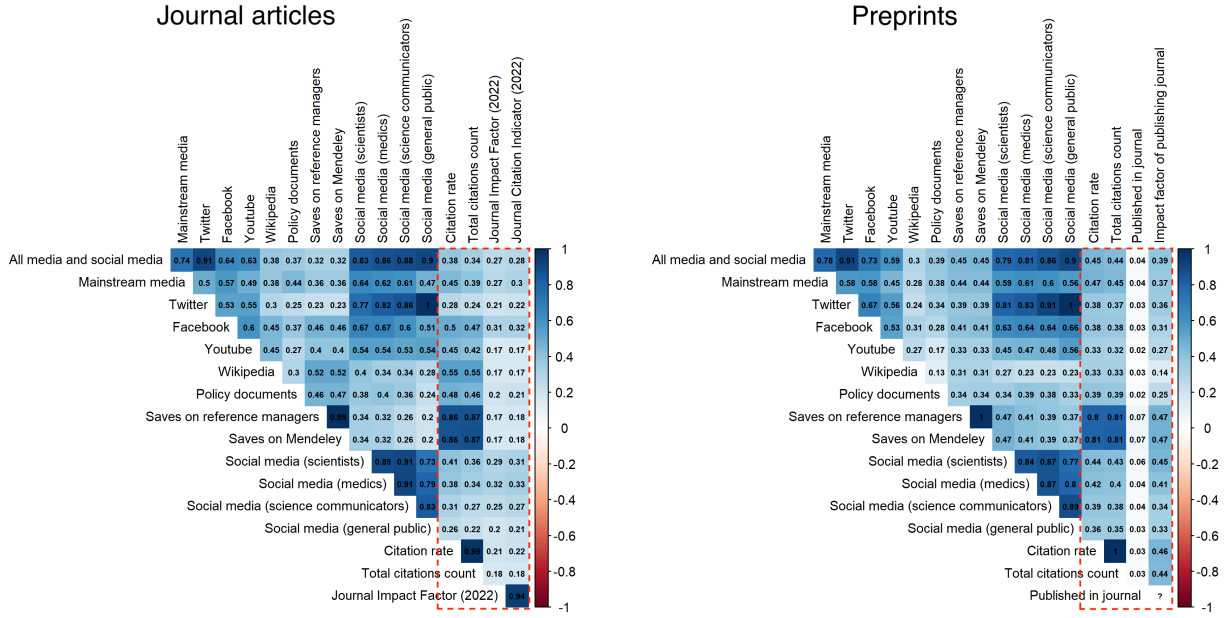


Figure 2.7: Correlation plots showing correlation between media attention, as measured by article Altmetrics, with article citation metrics and journal impact metrics. Left: journal articles; right: preprints. Correlations are calculated with Pearson's R, using complete cases only.

2.4.6 Topic modelling

The calibration process of the LDA topic model showed that 140 topics maximised the Jenson-Shannon divergence. Figure 2.8 shows the optimisation for Jenson-Shannon divergence.

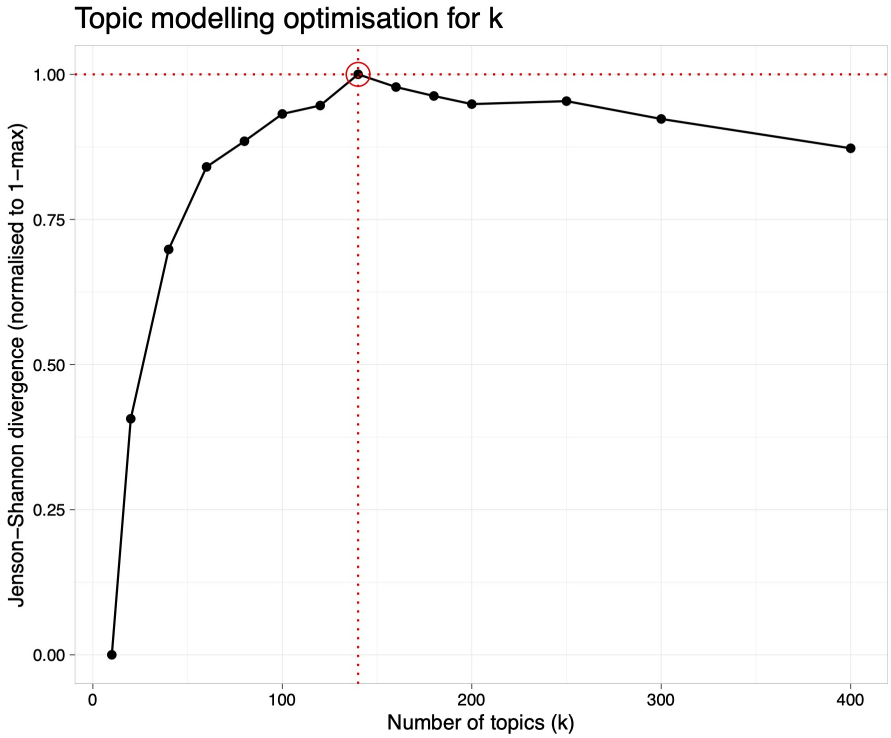


Figure 2.8: Optimisation plot showing average pairwise Jensen-Shannon divergence scores across all topics for a range of values of k. Scores are normalised to a range from zero to one.

The Gamma scores (measuring how much a document is enriched in a particular topic) can be averaged across all papers over time to show evolving areas of research interest and output: Figure 2.9 picks out four COVID-19 specific topics and shows the how the average enrichment of the research literature evolves by month across the study period: topic 13 (epidemic control) descends from an initial peak in January 2020; topic 26 (variants) is close to zero for most of 2020, then increases steadily from December 2020 when the Alpha variant was first detected; topic 7 (post-COVID recovery) increases steadily across the whole study period; and topic 98: (vaccines) gradually increases throughout 2020 then spikes in January 2021 and becomes the dominant research topic for the remainder of the study period.

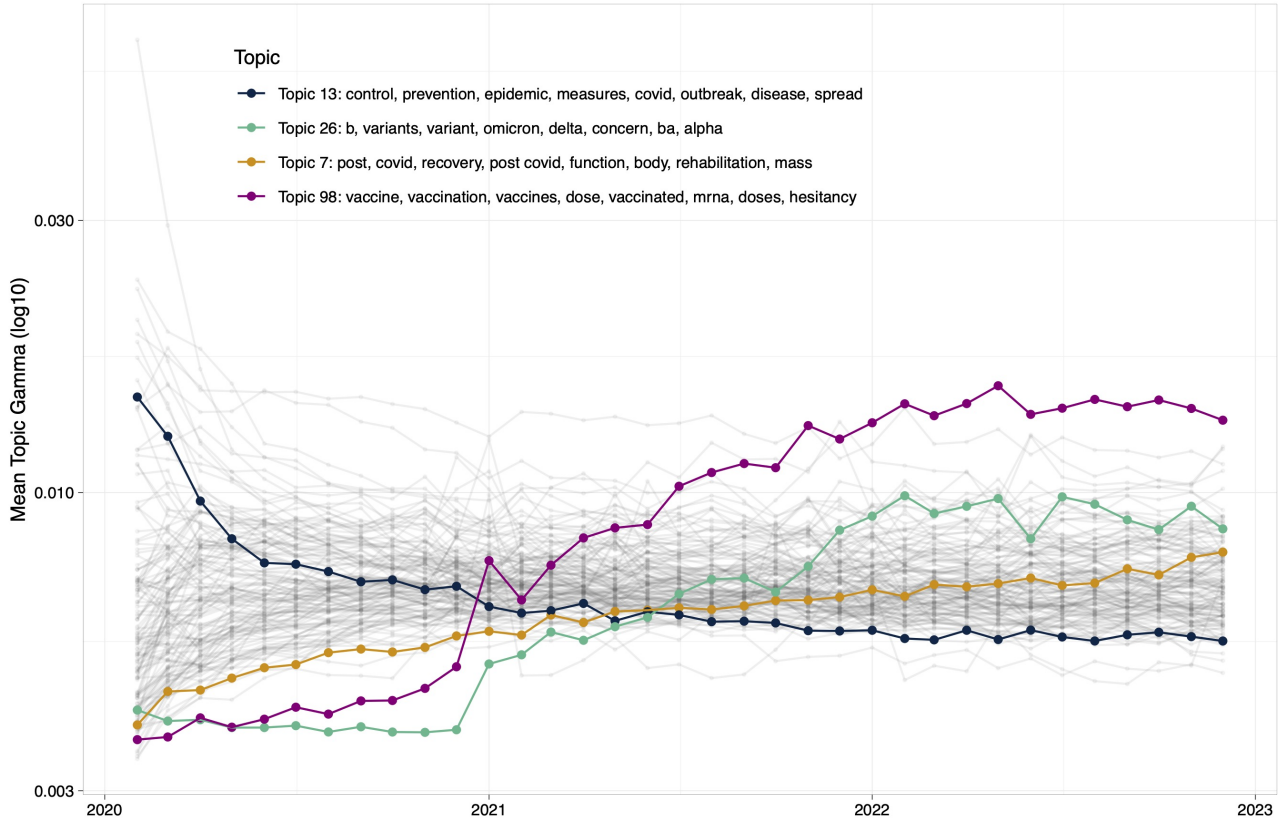


Figure 2.9: Line plot showing changes in the relative enrichment of paper abstracts in four key topics, over time. Enrichment is quantified as the mean γ value for each topic across the abstracts published in a given month. Topic 13, which captures words and word-pairs relating to the acute phase of an epidemic (eg *control, prevention, epidemic*), declines sharply after a peak in early 2020. Topics 98 and 26, which relate to vaccines and variants respectively, are almost absent from the literature in 2020 and then climb steadily throughout 2021 and 2022. Topic 7, which relates to post-COVID conditions and recovery, increases steadily throughout the three year period.

2.4.7 Topics as drivers of attention and citation

In linear regression models associating topics with Altmetrics scores and citation rates, the strongest positive associations with Altmetrics score were topic 98 (vaccines), topic 65 (clinical trials) and topic 26 (variants) (Figure 2.10). The strongest associations with citations were topic 110 (which captures the initial outbreak in China) and topic 137 (a generic 'coronavirus' topic) (Figure 2.10). These are clearly influenced by being the topics that dominated the early period of the pandemic and therefore appear in the most highly cited papers.

There was a high degree of consistency between the drivers of attention and the drivers of citation: the effect direction of the Altmetrics–topic and citations–topic associations were con-

sistent in 76 of 83 (91.6%) of the significant associations (Figure 2.10). Notably topic 15, which captures face masking, was a significant positive driver of Altmetrics but a negative driver of citations: the effectiveness of masking in reducing transmission was a topic that generated (and continues to generate) a huge amount of online debate, as evidenced by the fact that two of the 20 COVID-related papers with the highest Altmetrics scores were on the topic of masking (one of these was the DANMASK-19 study [235], which found no statistically significant association between mask-wearing and infection rates and continues to be a topic of debate [236]) (Table 2.4).

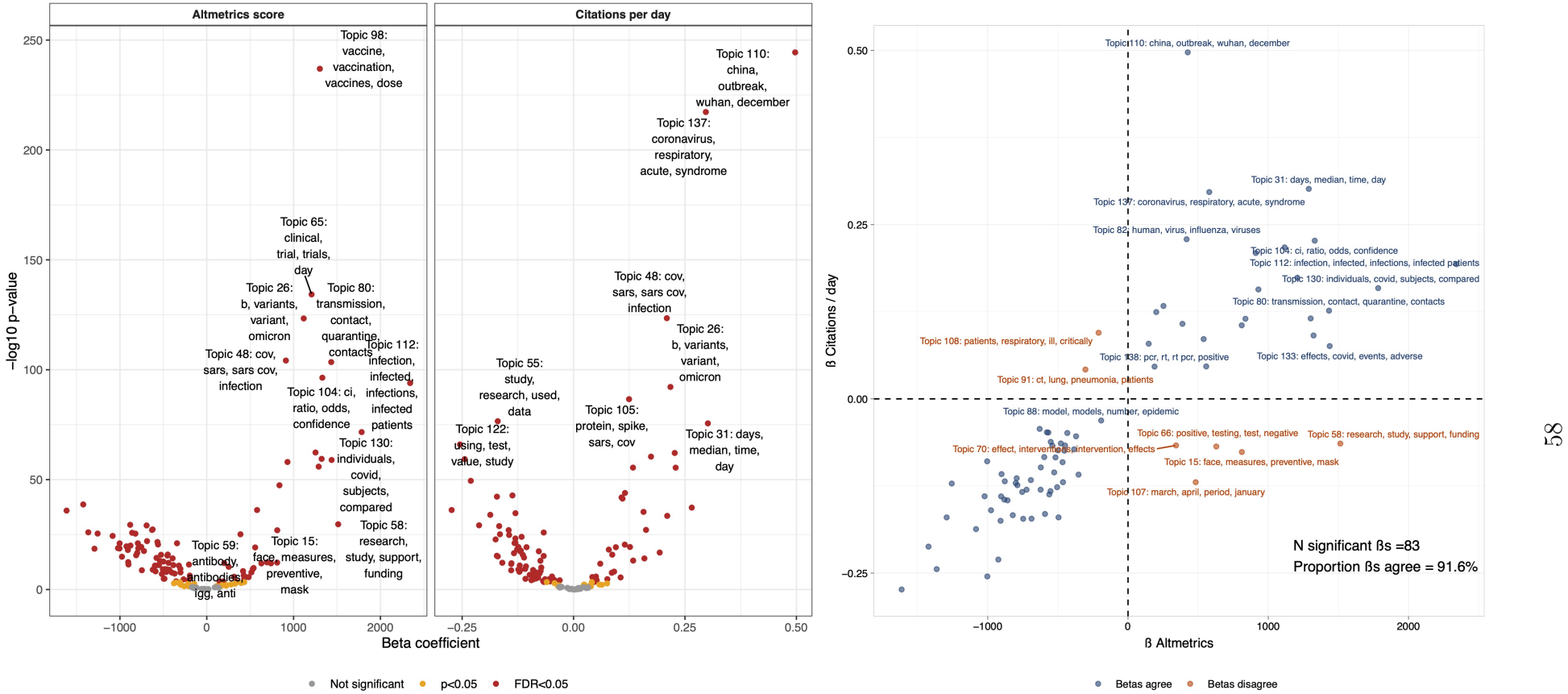


Figure 2.10: Panel plot showing volcano plots of the results of univariable regression of (left) Altmetrics and (centre) citation rates onto abstract topics, and (right) scatter plot comparing beta coefficients in topics with significant associations with citation rates and Altmetrics. In volcano plots, the x-axis shows regression β coefficient, y-axis shows log-10 p-value. In 91.6% of topics, the sign on the (significant) regression coefficient is the same for both citation rate and Altmetrics score, indicating that there is consistency between the drivers of Academic attention (as measured by citations) and public/media attention (as measured by the aggregated Altmetrics score).

Time-varying drivers of attention in preprints and journal articles

In the time-varying association analysis among preprints, some topics are positively associated with media attention throughout the whole pandemic period (for example topic 112 [infection] and topic 48 [a generic SARS-CoV-2/infection topic]), some are strongly associated for a single short period, potentially driven by a small number of papers that generated a lot of attention (for example topic 89 [older adults]), and some have recurring periods of high attention: Topic 112 [infection], Topic 130 [a methods topic that relates to cases and controls], and Topic 26 [variants], for instance, can be seen to be most strongly positively associated with Altmetrics scores during spikes of media attention in August 2021 and December 2021 (Figure 2.11). The same effect is less apparent in journal articles (Appendix A, Figure A.5), where the stronger associations are predominantly concentrated in the first wave (pre August 2020) [4]. This finding reinforces the earlier analysis suggesting that the topicality of preprints means that they are stronger drivers of public attention than journal articles.

[4]The journal article heatmap (Appendix A, Figure A.5) is notably symmetrical, with a similar number of topics negatively associated with media attention as positively, particularly in early 2020. This is partly a by-product of the zero-sum nature of the topic mixture in each document.

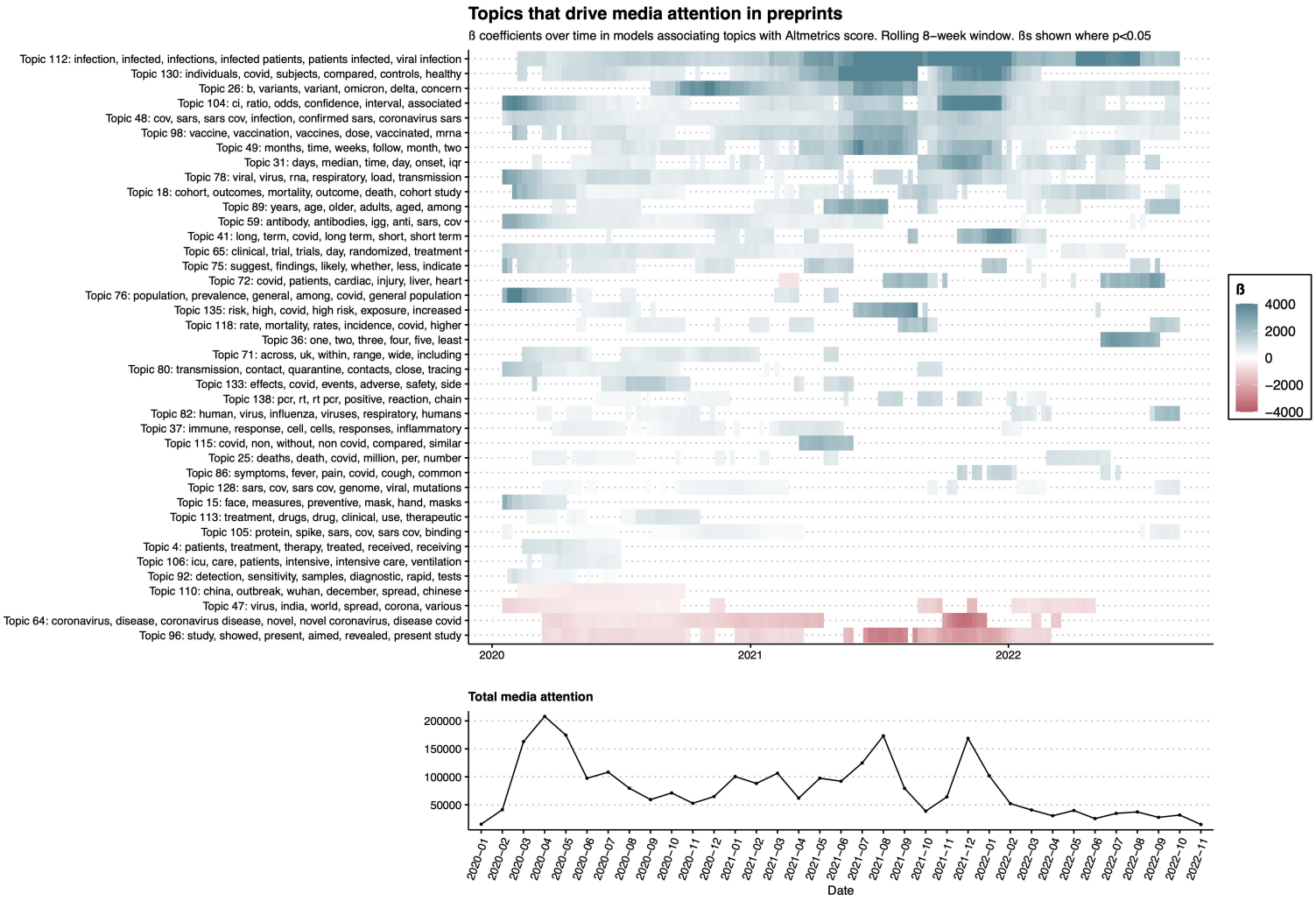


Figure 2.11: Heatmap showing results of univariable regression of Altmetrics scores onto abstract topics on a rolling 8-week time window throughout the pandemic period 2020–2022, among preprints only. Colour represents beta values: blue = positive betas, red = negative betas and colour darkness represents beta value magnitude (darker hue = larger beta). Topics are ordered by average beta value across the three years, and only topics with at least one period of significant association with Altmetrics scores are shown

2.4.8 Subsequent publication of preprints

Fewer COVID-19 preprints had been published in a referenced journal within one year of preprinting than either non-COVID-19 preprints or pre-COVID-19 preprints (22.2% vs 22.5% and 32.2% respectively, Table 2.2). Time-to-event analysis shows that COVID-19 preprints have a lower probability of publication at all time points after preprinting than non-COVID-19 preprints, from either 2019 or 2020–2022: in Cox proportional hazards models, preprints from 2019 had a 60.6% greater 'hazard' of publication than COVID-19 preprints (HR: 1.606, 95%

CI: 1.579, 1.634), and non-COVID-19 preprints from 2020-2022 had a 17.9% higher hazard of publication (HR: 1.179 [1.16,1.198]). Among bioRxiv and medRxiv preprints only, preprints from 2019 had a 94.2% greater 'hazard' of publication than COVID-19 preprints (HR: 1.942 [1.890, 1.995]), and non-COVID-19 preprints from 2020-2022 had a 42.6% higher hazard of publication (HR: 1.426 [1.391,1.461]). Kaplan Meier plots suggest that COVID-19 preprints may actually have a higher probability of publication for the first 150 days after preprinting, and thereafter a lower probability (Figure 2.12).

Constraining the analysis to only preprints from 1 January to 1 August 2020, when the large spike of COVID-19 preprints appeared, shows a similar pattern (Figure 2.13), with non-COVID-19 preprints having a 73.9% higher hazard of publication (HR 1.739 [1.692,1.788]) and non-COVID-19 medRxiv/bioRxiv preprints having a 88.4% higher hazard of publication (HR 1.884 [1.806,1.965]) . Interestingly, the publication rate among non-COVID-19 preprints from the 'first wave' (January–August 2020) was slightly higher than the rate for non-COVID-19 preprints across the whole pandemic period.

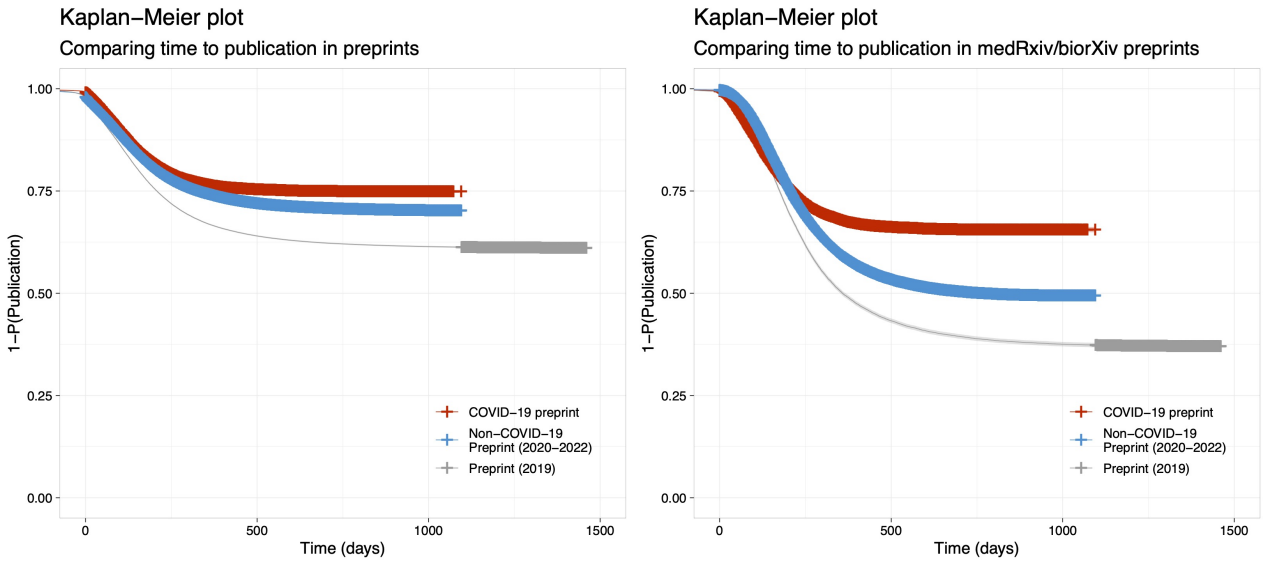


Figure 2.12: Kaplan-Meier plots showing time-to-event curves for publication in preprints. Left plot shows time to publication in all preprints; right plot shows time to publication in medRxiv and bioRxiv preprints only. Compared to all other preprints (left plot), COVID-19 preprints had a lower probability of publication at all time points; among bioRxiv and medRxiv preprints only (right plot), COVID-19 preprints had a higher probability of publication in the first ~200 days after being preprinted, then a substantially lower probability of publication at all subsequent time points.

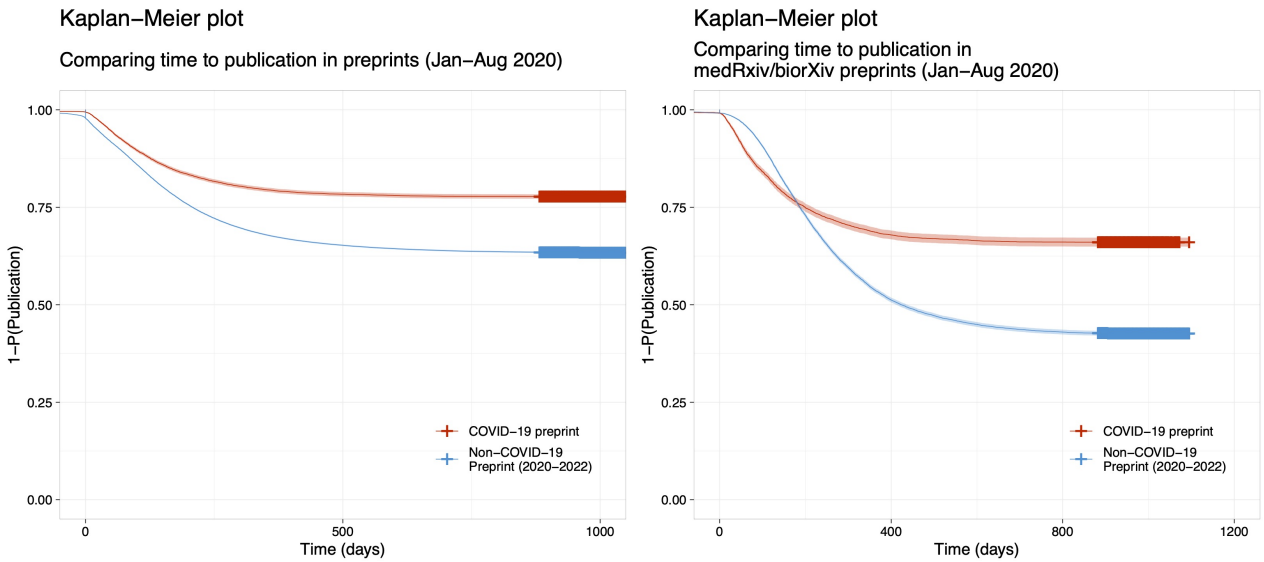


Figure 2.13: Kaplan-Meier plots showing time-to-event curves for publication in preprints from the first wave of COVID-19 (to 1 August 2020). Left plot shows time to publication in all preprints; right plot shows time to publication in medRxiv and bioRxiv preprints only. Compared to all other preprints (left plot), COVID-19 preprints had a lower probability of publication at all time points; among bioRxiv and medRxiv preprints only (right plot), COVID-19 preprints had a higher probability of publication in the first ~200 days after being preprinted, then a substantially lower probability of publication at all subsequent time points.

2.4.9 Predictors of subsequent high-impact publication of preprints

In multivariable boosted tree models predicting subsequent high-impact publication, Altmetrics were shown to be more powerful predictors of subsequent publication than either the preprint's citation metrics, or the article topics (Figure 2.14). In feature importance analysis, which quantifies the degradation in model performance (as measured by in AUC) that results when a variable is excluded, of a model including all Altmetrics data points as predictors alongside preprint citation data and abstract topic data, mentions from the scientific social media cohort, counts of readers on reference managers (and Mendeley specifically), and Altmetric scores (normalised by journal and article age) were the strongest predictors of subsequent publication. This model, with all preidctors included, achieved an AUC of 0.764 (0.748–0.781). A model trained only on Altmetrics data achieved an AUC of 0.747 (0.731–0.763), compared with 0.708 (0.689–0.726) for a model trained on abstract topic data only and 0.68 (0.661–0.699) for a model trained on preprint citation data only (Figure 2.14).

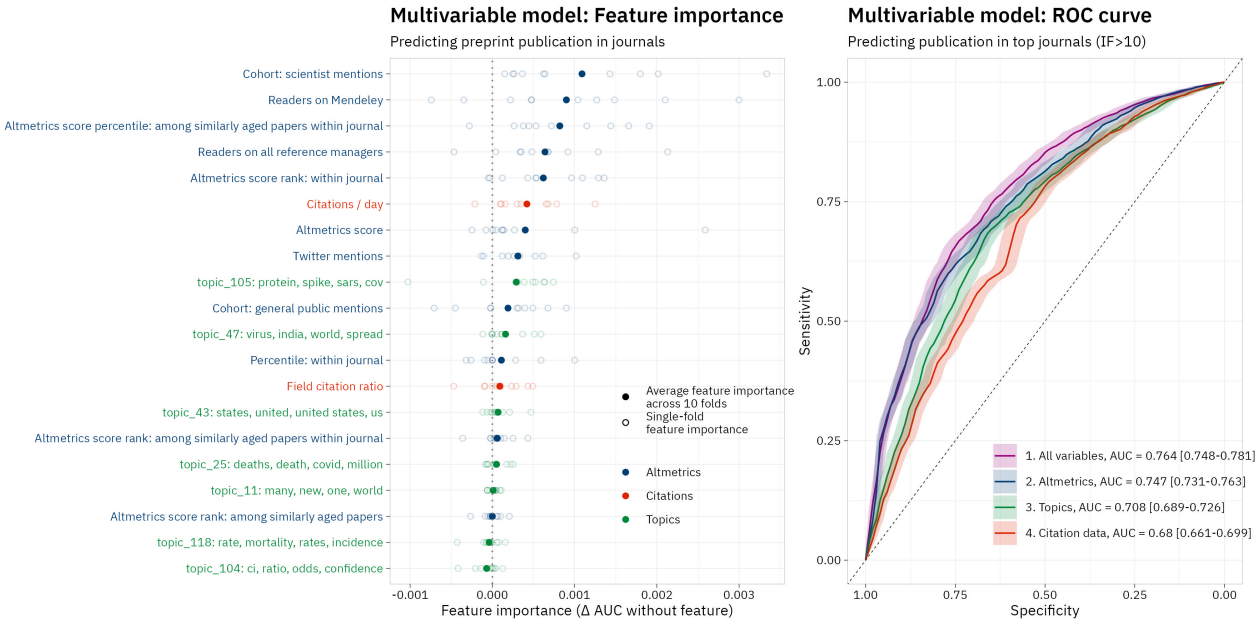


Figure 2.14: Panel plot showing multivariable feature importance and model evaluation performance when predicting subsequent publication in a high impact (IF $\geq$ 10) journal, in a sample of N=15,336 preprints from 2020. Left panel: feature importances from 10 folds of cross-validation, (average score and individual fold scores shown), measured as the change in loss function resulting when each feature is removed from the model. Right panel: Receiver Operating Characteristic (ROC) curve showing the predictive performance of different families of information about preprints. Area Under the ROC Curve (AUC) metrics are shown. Among the variable families, Altmetrics have the highest AUC (0.747 [0.731,0.763]), followed by article abstract topics (0.708 [0.689,0.726]) and citation data (0.680 [0.661, 0.699]). The model with all variables as predictors achieves an AUC of 0.764 [0.748,0.781].

2.5 Discussion

This analysis describes an ongoing surge of COVID-19 research output consisting of more than 420,000 research papers, constituting 3.4% of the English language research output in the world between 2020 and 2022. It also describes a spike of COVID-19 related preprints during the first wave of the pandemic, which has largely subsided. Analysis of article topics revealed evolving areas of research interest over the course of the pandemic, with vaccines, variants, and discussion of mental health issues growing from low baselines in 2020 to become dominant themes by the end of 2020.

Previous research has described the rapid redirection of research efforts into COVID-19 [237] and suggesting that the pandemic may adversely impact scientific output in non-COVID-19 fields [238, 239], this analysis finds little evidence of an overall reduction in the volume of non-COVID-19 related scientific output. Given the lag time for the initiation, execution and publication of new research projects, however, it may be that these effects are yet to be observed.

The results presented here offer mixed support for the hypothesis that the volume of COVID-19 articles and the demand for rapid turnaround caused lower-quality preprints and journal articles to be published [240]. In the time-to-event analysis, COVID-19 preprints were found to be less likely to get published than preprints on other subjects, from either before or during the pandemic, and this effect was more pronounced in preprints from the 'first wave' (up to 1 August 2020). This does suggest a possible quality dip in COVID-19 preprints, and especially so in the first wave.

However, the Altmetrics:citation metrics correlations were higher in COVID-19 articles and preprints than in non-COVID-19 articles and preprints. Previous work has established correlations between Altmetrics and conventional (citation-based) measures of article impact [241, 242], and found a correlation of 0.19 (95% C.I 0.12 to 0.25) in pre-COVID-19 papers in the field of health sciences, which is in line with the analyses presented here of pre-COVID-19 papers and non-COVID-19 papers from 2020–2022. For COVID-19 papers and preprints, the correlation between Altmetrics score and article citations was substantially higher, at 0.34. As discussed previously, this suggests a higher degree of alignment, for COVID-19 articles, between the scientific import of an article and the degree of public attention it receives. There are two ways of interpreting this elevated correlation. Firstly, it may be that papers with intrinsic scientific

merit were gaining more attention and also going on to be more highly cited, and more likely to be published in a higher impact factor journal. Conversely, it could be that the attention a paper received in the media and social media increased its chances of being cited or being published in a higher impact journal, and that this effect was heightened for COVID-19 papers. Previous research has framed Altmetrics impact as being 'predictive' of higher citations [243] but causality has not been investigated and cannot be inferred from this analysis.

The analysis of the citation-attention association on specific platforms found a weak correlation between Twitter attention and metrics of article quality (article citations and journal impact factor), and stronger correlations between Facebook and YouTube mentions and citation-manager readership counts and metrics of article quality. This supports previous research indicating that the Twitter attention an article receives is not strongly related to the quality of that article, but that citation-manager readership is [244].

Preprints are shown to have generated more media and social media attention per paper than journal articles, to have spikes in public attention that tracked spikes in COVID-19 prevalence, and to be more closely aligned to waxing and waning public attention paid to COVID-19 research. This alignment is likely to be bidirectional: research both drives the public discussion and is driven by it, and again these results cannot disambiguate the direction of effect. Whichever the effect direction, the fact that preprints are more responsive to and instigative of public interest around research topics is a natural byproduct of the speed with which research can be made available via preprints, when the peer review process is deferred. This effect is naturally heightened during a pandemic when novel information is at a premium, and concern that non-peer-reviewed work may be driving public opinion and discussion (as, for example, in the case of two studies from early in the pandemic that claimed a protective effect of tobacco and caused panic buying of nicotine in France and cigarettes in Iran [245]) has led some to call for a more robust set of guidelines for discussion of preprint research by the media [119–121].

The analysis of predictors of high-impact publication of preprints suggests that attention from the scientific community on social media, as measured by the Altmetrics scientist cohort metric, is the best predictor of subsequent publication, alongside numbers of readers on Mendeley and other reference managers. This finding builds on previous findings that higher Altmetrics scores are associated with subsequent publication in preprints [246].

2.5.1 Limitations

Analysis of subsequent publication rates of preprints was contingent on the accurate linking of the preprint to a subsequent publication, which is an ongoing problem in scholarly publishing [227]: linkage is dependent either on algorithmic detection or on authors manually updating their preprint records, both of which are liable to omission. This may result in an underestimate of preprint publication rates, and there is the possibility of bias in the analysis of predictors of subsequent publication if, for example, certain types of paper are more likely to be detected by linking algorithms or logged by their authors. Owing to the limit on calls to the Altmetrics API, it was not possible to pull granular Altmetrics data for for the full data set of 17 million papers. A continuation of this research could use a distributed computing set-up to extend the analysis of granular Altmetrics scores to non-COVID-19 papers.

2.6 Conclusion

In conclusion, this chapter describes a large and ongoing response from the scientific community to the COVID-19 pandemic, with evolving research topics that have aligned with evolving areas of public interest, and may partly be driven by them. While COVID-19 journal articles are being published at the same rate as in 2020, a wave of preprints, potentially including some lower-quality research, peaked in June 2020 and subsequently declined. The findings contribute to an understanding of how scientific discourse functions during a pandemic, particularly on the relationship between the media and the scientific discourse and on the role of preprints in the research ecosystem.

Chapter 3

Symptomatology in the acute phase of COVID-19

This chapter uses data from 15 rounds of the REACT-1 study to investigate the symptomatology of COVID-19, how symptoms vary by age, sex and other biological variables, and how symptom profiles have changed with different variants and in response to natural or vaccine-acquired immunity. Classical statistical methods are used to identify associations between symptoms and PCR positivity for SARS-CoV-2 virus, complemented with machine-learning methods i) to identify sparse subsets of symptoms that most accurately diagnose COVID-19 and different COVID-19 variants, ii) to construct and evaluate multivariable models to predict PCR positivity using symptom data, and iii) to explain the predictions of these models using feature attribution methods.

Note: Findings from this chapter have been published in Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

3.1 Motivation

Identifying potential cases on the basis of symptomatology has been an important component of the public health response to COVID-19. Understanding variations in symptoms and identifying sets of symptoms with the highest sensitivity and specificity to COVID-19 has therefore been

an important research question throughout the pandemic. Further, manifestations of a disease provide information about its pathogenesis and potentially to its mechanisms of entry and transmission. This information may inform public health policy to control disease spread, as well as contributing to the overall understanding of a disease and to the development of new treatments.

3.2 Background

Community-based studies have assessed the capacity of symptom data to predict PCR positivity for SARS-CoV-2, in some cases using variable selection and ranking methods to identify the 'best' (set of) symptoms to identify cases [150–152]. These studies included an early analysis based on REACT-1 data from rounds 2 to 8, when the Wild Type variant was dominant [151]. The primary research goal of the paper was driven by domestic public health priorities: how could community cases be better detected using symptom data to trigger PCR testing, and what would be the impact of expanding the symptom 'trigger' list beyond loss of smell and taste, fever and new continuous cough. The paper identified 7 symptoms that were jointly positively predictive of PCR positivity and demonstrated that expanding the trigger list to these seven symptoms would increase community detection rates without exceeding testing capacity. The paper also used early cases of B.1.1.7 (Alpha variant) from round 8 to demonstrate differences in symptom profile from the Wild Type variant, which subsequent studies confirmed in different data sets[153, 154].

This chapter builds on the analysis published in that paper, with less emphasis on community detection and a greater emphasis on describing the changing patterns of symptomatology over time, and identifying the underlying factors that influence COVID-19 symptoms.

3.3 Methods

3.3.1 Study population

The study design and recruitment methods are covered in the Introduction.

Data from 15 rounds of REACT-1 between 19 June 2020 and 31 March 2022 were split into distinct phases that corresponded with the dominance of different SARS-CoV-2 variants in England: rounds 2–7 (between 19 June and 3 December 2020), when the wild-type variant was dominant; rounds 8–10 (between 6 January and 29 March 2021), when Alpha (B.1.1.7) was dominant; rounds 13–15 (between 24 June and 5 November 2021), when Delta (B.1.617.2) was dominant; and rounds 17–19 (between 5 January and 31 March 2022), when Omicron (B.1.1.529) was dominant. In rounds 17–19, sequencing data were used to identify those participants who were infected with BA.1 or BA.2. Round 1 was excluded because the questions asked relating to symptoms were not consistent with subsequent rounds. Rounds 11, 12 and 16 were excluded from analysis because two variants were competing for dominance in the population while the fieldwork was being conducted. Figure 3.1 shows the study population grouped by these phases and how it aligns with the changing dominance of different variants in the UK, based on GISAID sequencing data [248].

REACT-1 participants reported whether they had experienced any of a list of 26 potential COVID-19 symptoms in the week prior to taking the swab test. In later rounds, the symptom list was expanded to 30 and then again to 32 symptoms. Here, only the 26 symptoms that are surveyed in all rounds are used. These included respiratory and cardiac symptoms (shortness of breath, new persistent cough, chest pain, tight chest), loss or change of sense of taste or smell, influenza-like symptoms (fever, muscle aches, chills, headache), cold-like symptoms (blocked nose, runny nose, sneezing, hoarse voice, sore throat, sore eyes), gastrointestinal symptoms (abdominal pain/belly ache, nausea/vomiting, diarrhoea, appetite loss), fatigue-related symptoms (severe fatigue, tiredness, difficulty sleeping), and others (heavy arms or legs, dizziness, numbness/tingling).

3.3.2 Statistical analyses

Logistic regression

As a primary analysis to look for associations between individual symptoms and risk of PCR positivity, logistic regression models were constructed with PCR positvity (y/n) as the binary outcome variable, and each of the 26 symptoms as predictors. These models adjusted for age group (18–25, then ten-year brackets up to 75+), sex, and vaccination status (quantified as the

number of vaccines received). Equation 3.1 shows the model specification.

$$y(PCRpositive - y/n) \sim \alpha + \beta(age) + \beta(sex) + \beta(vaccines) + \beta(sex) + \beta(symptom) \quad (3.1)$$

Odds ratios and 95% confidence intervals for each of the symptoms, after adjustment, are reported. Separate regressions are conducted within each variant phase so that an odds ratio is generated for each symptom for each variant, showing the risk of PCR positivity for that variant conditional on each of the 26 symptoms.

This analysis shows the changing strength of association between individual symptoms and different SARS-CoV-2 variants. However, the odds of PCR positivity are conditional on experiencing a symptom are also contingent on symptom prevalence and COVID-19 prevalence at the time of data collection. As a complement to this analysis, therefore, to account for some of the effect of changing background prevalence of symptoms and COVID-19, a pooled analysis using data from all 15 rounds, including adjustments for calendar time, tested for the interaction effect between the variant and the symptom in relation to risk of positivity. The specification is show in 3.2.

$$\begin{aligned} y(PCRpositive - y/n) \sim \alpha + \beta(age) + \beta(sex) + \beta(vaccines) + \beta(time) + \\ \beta(symptom) + \beta(variant) + \beta(symptom * variant) \end{aligned} \quad (3.2)$$

In this analysis, the interaction effect captures the change in odds of PCR positivity over time that is attributable to the change in variant (as opposed to changing background symptom/COVID-19 prevalence). Therefore, the odds ratio for a given symptom and a given variant is calculated as the product of the odds ratio for the symptom and the odds ratio for the symptom:variant interaction effect.

Penalised logistic regression with stability selection

The application of stability selection to penalised models for improved variable selection was proposed by Meinshausen and Bühlmann in 2010 [249]. The principle is to run a large number (K) of penalised models on subsamples of data, and to count the number of models in which each variable was selected. This count, taken as a proportion of K, functions as a stability score

indicating the relative importance of a given variable to the final model. To perform variable selection, two parameters must be estimated: first, the penalty parameter lambda (which must be calibrated for a single model), and a threshold proportion (pi), above which a variable is considered to be stably included in the model.

To avoid the use of an arbitrary selection threshold, Bodinier *et al* [250] proposed an internal stability score to derive the optimal selection threshold and lambda simultaneously. Briefly: the variables are assumed to have equal probability of inclusion in a hypothetical, maximally unstable model. Under this model, the distribution of selection counts across all variables follows a binomial distribution, with known parameters n (the number of subsampled models), and p (the probability of inclusion of any single variable in a given model with penalty lambda). The stability score is defined as the negative log likelihood under these assumptions, and increases with stability. A grid search of lambda and pi is conducted and the values that maximise the negative log likelihood are used to give the final selection proportions and the threshold for variable selection.

In this analysis, stability selection was used to identify a parsimonious set of symptoms that were most highly predictive of each variant. The data set was split into 70% training data and 30% model evaluation data. Stability selection was applied to the least absolute shrinkage and selection operator (LASSO) penalised logistic regression, with swab positivity as the binary outcome variable, and the 26 symptoms as predictors. Age, sex and vaccination status (as a count variable showing 0/1/2/3 vaccines) were forced into the model as unpenalised predictors.

$$y(PCRpositive - y/n) \sim \alpha + \vec{\beta}[covariates(agegroup, sex, vaccination)] + \vec{\beta}[symptoms] \tag{3.3}$$

In an early analysis after the first wave of the pandemic, the REACT team conducted variable selection using LASSO stability selection and identified seven symptoms that were jointly positively predictive of PCR positivity (loss or change of sense of smell, loss or change of sense of taste, cough, fever, chills, appetite loss, muscle aches) and additionally one symptom (numbness/tingling) that was negatively predictive of PCR positively when included in a final model with the 7 positive predictors [151]. The selection of a negative predictor presented an interpretability problem as the over-arching public health goal was to identify a sparse set of

symptoms that could be used as triggers for testing, and which were therefore jointly *positively* predictive of PCR positivity. In light of this, in this analysis, the regression coefficients in the LASSO models were constrained to be non-negative.

LASSO models were fit on 1,000 random 50% subsamples of the training data, and a symptom-specific stability score was calculated as the proportion of those 1,000 models in which each symptom was selected. The internal stability score was used to calculate a selection threshold [250], calibrated in conjunction with the LASSO penalty parameter.

Omicron BA.1 vs BA.2

Early findings that Omicron BA.1 was likely to lead to less severe disease than Delta [251] led to some suggestions that this was part of a natural trajectory that would lead to successively milder variants, and eventually to a steady state of COVID-19 endemicity with people getting only mild disease [252, 253]. To investigate the relative severity of Omicron BA.1 vs its immediate successor, BA.2, a series of analyses were conducted specifically comparing symptom profile, burden and severity in individuals with BA.1 vs BA.2.

Viral genome sequencing was conducted on positive swabs from rounds 17–19 to determine Omicron BA.1 and BA.2 lineages. Logistic regression was used to directly compare the symptoms of BA.1 or BA.2 swab-positive individuals. Equation 3.4 shows the specification.

$$y(BA.2 - y/n) \sim \alpha + \beta(age) + \beta(sex) + \beta(vaccines) + \beta(sex) + \beta(symptom) \tag{3.4}$$

This regression analysis was compared with a conditional logistic regression analysis which was conducted on 1:1 matched swab-positive individuals with BA.1 or BA.2. Matching was done on age (+/- 5 years), sex, vaccination status and round in rounds 17–19.

To compare symptom burden in BA.2 and BA.1 in terms of number of symptoms experienced by double- and triple-vaccinated individuals, a log-linear regression model with Poisson error was was constructed, with symptom count as the outcome variable and BA.1/BA.2 as the independent variable of interest. Equation 3.5 shows the specification.

$$y[symptomcount] \sim \alpha + \vec{\beta}[covariates(age, sex, vaccination, \\ PriorCOVID-19, weeks\text{–}since\text{–}symptom\text{–}onset, calendartime)] + \beta[variant] \tag{3.5}$$

Finally, to assess any differences in symptom severity between Omicron BA.1 and BA.2, independent of vaccination history, an analysis of the question relating to symptom severity ("How much, if at all, do the symptoms your/your child have/has had in the last 7 days reduce your/their ability to carry out day-to-day activities?") was conducted. A logistic regression model was constructed with a binary outcome variable where 1 represented a response of "a lot" or "a little" vs "not at all" and the binary independent variable BA.2/BA.1. To isolate the effects of variant vs vaccination or other variables as much as possible, the population was restricted to those who had had two ore more vaccinations, with their second vaccination at least 14 days before their swab test. The models were adjusted for age, sex, vaccine booster (y/n), days since most recent vaccination, prior COVID-19 (at least 28 days before swab test date), calendar time and time since onset of symptoms. Covariates were added to models sequentially, building from a crude model to a fully adjusted model, and odds ratio reported for each model. Additionally, the same subset of individuals was used to model symptom count in log-linear regression models with Poisson error, using a similar specification to 3.5.

3.3.3 Symptom trajectory

The symptoms of COVID-19 are known to change over the disease course [254]. The question of which symptoms manifest at which point in the disease course is an important component of understanding the COVID-19 phenotype, providing clues as to potential underlying pathophysiological mechanisms, as well as being useful in early case detection.

In an initial descriptive analysis, the relationship between days since symptom onset and symptom burden (in terms of number of symptoms reported) was described, stratified by variant. Second, the probability of experiencing each of the 26 symptoms was modelled as a function of days since symptom onset, and variant.

The change in symptom profile over the disease course also has a potential impact on the symptom profiling analysis: sampling for REACT-1 was non-continuous (ie conducted in discrete rounds with some weeks in between each round of fieldwork). As a result, the samples might

have captured different variants at different stages of symptom onset, on average. For instance, the first couple of rounds captured the downslope of the first wave of the Wild Type variant (Figure 3.1), while round 18 captured the transition phase where BA.2 was rapidly replacing BA.1. Those participants captured in round 18 may therefore be more likely to be in the early stages of infection, on average, than those captured in round 2 or 3, and this may influence the symptom profiles detected at different stages of the pandemic.

To investigate the impact that this effect might have on the main symptom profiling findings, the main analysis was repeated, with further adjustment for time since symptom onset.

Ct values

Cycle threshold (Ct) values – ie the number of PCR assay cycles that are required for the fluorescent signal to exceed background levels – are an established proxy for viral load [255, 256], which in turn is a correlate of infectiousness [257, 258]. Identifying symptoms that are associated with particularly high (or low) Ct values is therefore a question of clinical interest, as it would help to assess the likely infectiousness of an individual on the basis of their symptoms. This kind of risk assessment has become more valuable in the latter stages of the pandemic, as governments have stopped mass testing for COVID-19 and largely moved away from mandatory isolation measures.

To this end, an analysis was conducted modelling the relationship between symptoms and Ct values in PCR positive individuals, using a standard linear regression framework with Ct value as the outcome variable (equation 3.6).

$$y(Ct\text{–}value) \sim \alpha + \vec{\beta}[covariates(age, sex, vaccination)] + \beta[symptoms] \tag{3.6}$$

3.4 Results

3.4.1 Descriptive analysis

The study population is summarised in Figure 3.1 and in Table B.1 in Appendix A. There were a total of 1,542,510 adults aged 18 and over, of whom 17,448 tested positive on their PCR test: 2,971 for Wild Type (0.4%, 95% CI:[0.4,0.4]), 2,275 (0.6% [0.6,0.7]) for Alpha, 1,493 (0.6% [0.6,0.7]) for Delta, and 10,709 (4.4% [4.3,4.5]) for Omicron variants (Appendix A, Table B.2).

The proportion of PCR-positive individuals with one or more of the 26 surveyed symptoms increased with each successive variant. Only 45.0% [43.3,46.8] of those with Wild Type reported experiencing any of the surveyed symptoms. This increased to 54.7%, [52.7,56.8] in those with Alpha, 63.8% [61.3,66.2] in those with Delta, 70.0% [68.3,71.6] in those with BA.1 and 75.9% [74.4,77.2] in those with BA.2 (Appendix A, Table B.2).

Symptom burden (in terms of the number of symptoms experienced, and the reported severity of symptoms) also increased with each successive variant. PCR-positive individuals reported an average of 2.70 (95% CI 2.6,2.8) symptoms for Wild Type, 3.4 (3.2,3.6) for Alpha, 4.6 (4.4,4.9) for Delta, 4.6 (4.5,4.8) for BA.1 and 6.0 (5.8,6.2) for BA.2. The proportion of respondents reporting that their symptoms affected their day-to-day activities "a lot" increased from 10.5% in Delta to 17.6% in BA.2 (the question on symptom severity was only included in the post-Alpha rounds).

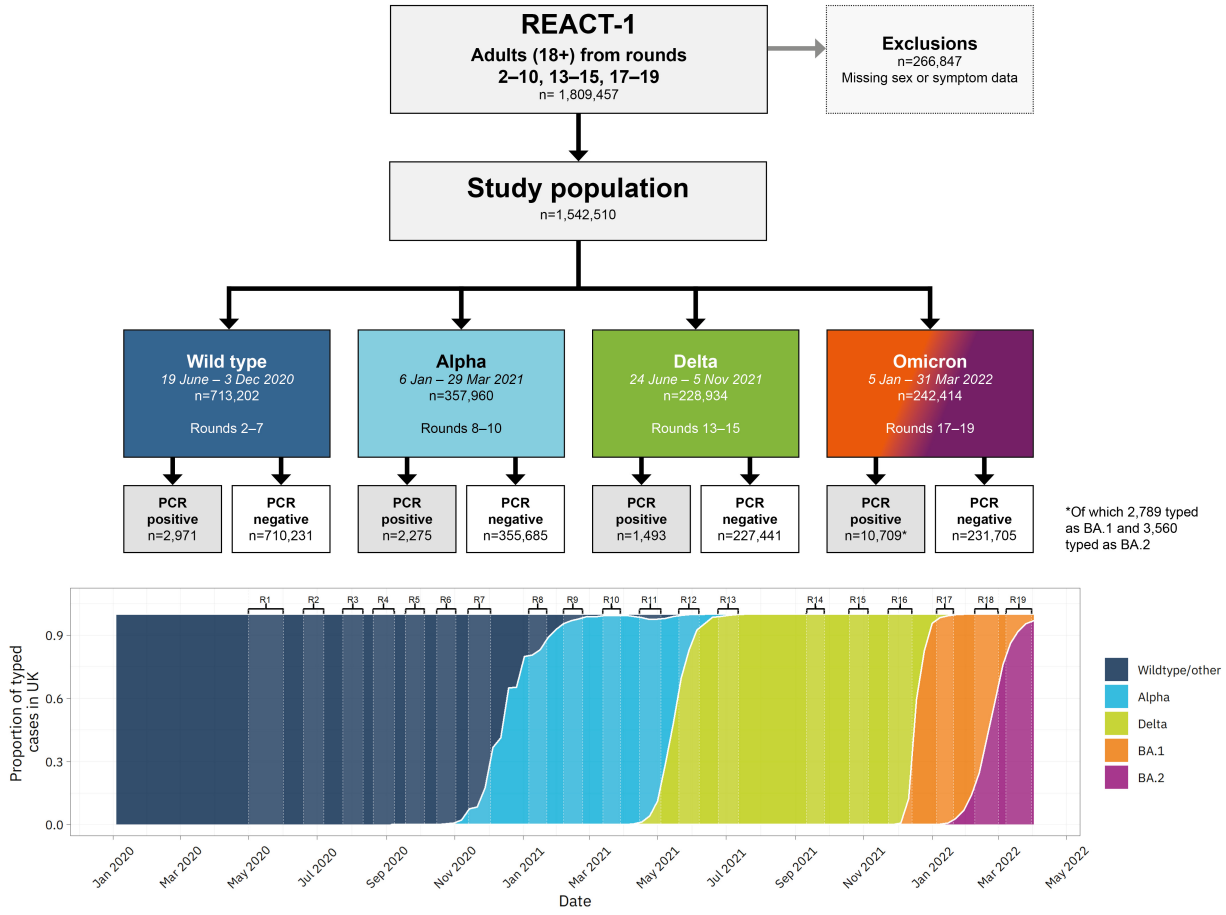


Figure 3.1: Flowchart showing study population for symptom profiling, plus exclusions and variant transitions in the UK, based on UK sequencing data from GISAID [248]. Reproduced from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247]

3.4.2 Logistic regression

All symptoms were positively associated with PCR positivity for all variants (Figure 3.2, Table 3.1). The odds of experiencing any of the 26 symptoms, after adjustment for age group, sex and vaccination status, increased in each successive variant, from an odds ratio of 5.7 [4.8,5.6] for Wild Type to 12.9 [11.9,14.0] for Omicron BA.2. The odds of experiencing coryzal (cold-like) symptoms increased the most significantly across variants: the odds ratio for sore throat, for instance, increased from 3.76 [3.38,4.19] in Wild Type, to 6.56 (5.24,8.20) in Alpha, 7.22 (6.38,8.18) in Delta, 11.6 (10.7,12.6) in BA.1 and finally 18.0 (16.8,19.4) in BA.2 (Figure 3.2, Table 3.1).

By contrast, the odds ratios for loss or change of sense of smell and loss or change of sense

of taste peaked in Delta, with ORs of 73.4 (64.2,83.9) and 68.1 (59.4,78.0) respectively, before dropping to 12.9 (11.1,15.1) and 16.0 (13.9,18.5) in BA.1 and then 17.2 (15.1,19.5) and 21.3 (18.9,24.0) in BA.2. In a sensitivity analysis in which models were further adjusted for days since symptom onset, the results were largely consistent (Appendix A, Figure B.1).

In a pooled analysis designed to account for changing background prevalence of SARS-CoV-2 and non-COVID-related symptoms, the results were broadly consistent with the primary univariable analysis: the Alpha and Delta variants were associated with elevated symptom-specific odds-ratios across most symptoms; the Omicron BA.1 and BA.2 variants were associated with lower odds rations for loss or change of sense of smell and loss or change of sense of taste; omicron BA.1 and BA.2 were associated with increased odds ratios for coryzal symptoms, most notably sore throat and hoarse voice; and BA.2 was associated with higher odds ratios than BA.1, most notably for coryzal symptoms (Appendix A, Figure B.2)

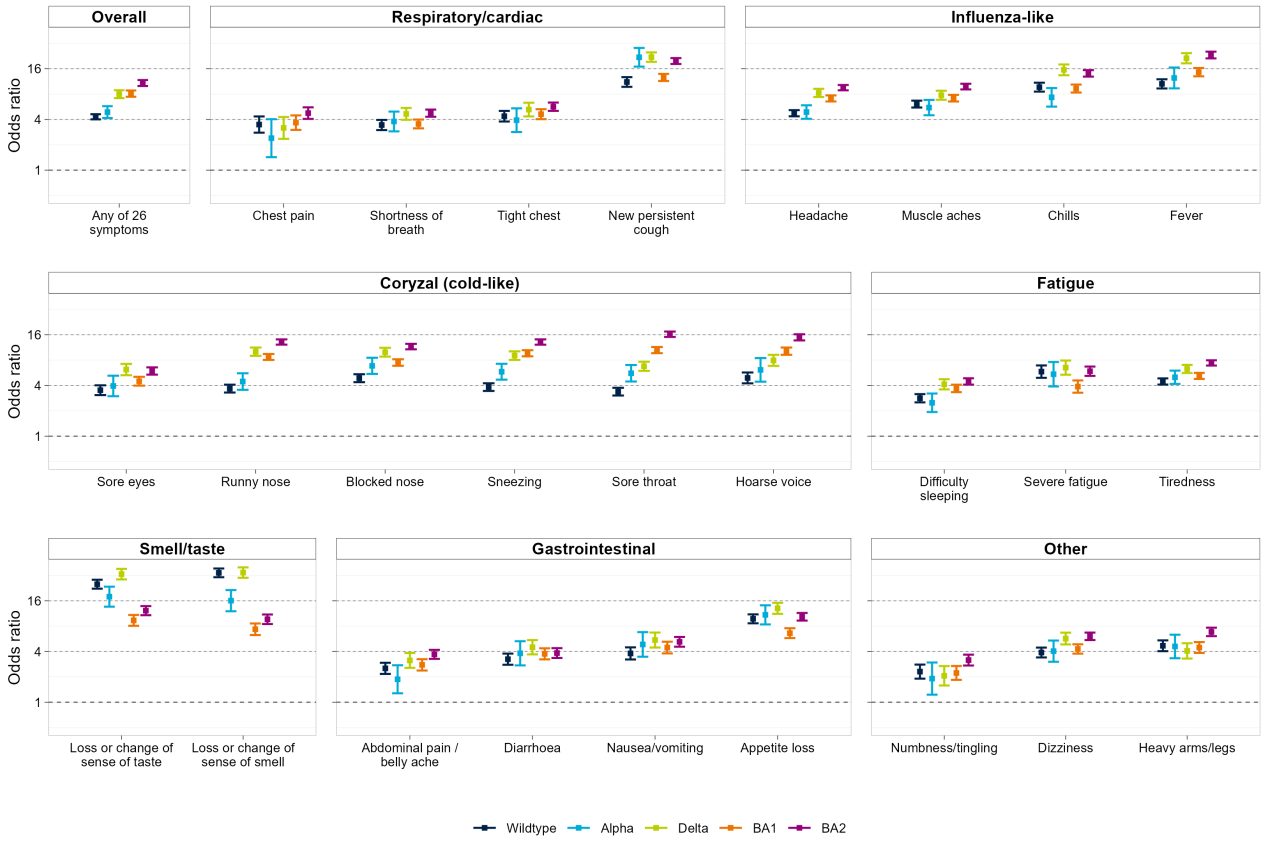


Figure 3.2: Comparison of ORs for swab positivity based on presence or absence of any of 26 symptoms surveyed in N=1,542,510 participants across five variant-phases of REACT-1. Reproduced from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247]

Category	Wildtype	Alpha	Delta	BA1	BA2
Any of 26 symptoms	4.30 (4.01,4.62)	4.89 (4.16,5.75)	7.98 (7.16,8.90)	8.10 (7.44,8.83)	10.8 (9.98,11.8)
Loss or change of sense of smell	34.3 (30.5,38.7)	16.1 (12.0,21.5)	34.6 (30.0,39.9)	7.35 (6.27,8.61)	9.66 (8.48,11.0)
Loss or change of sense of taste	25.2 (22.3,28.4)	17.9 (13.6,23.5)	33.2 (28.7,38.3)	9.36 (8.07,10.9)	12.2 (10.8,13.8)
New persistent cough	11.1 (9.76,12.7)	21.9 (17.0,28.3)	22.0 (19.3,25.0)	12.6 (11.4,13.9)	19.8 (18.2,21.4)
Fever	10.6 (9.33,12.0)	12.4 (9.36,16.6)	21.3 (18.6,24.5)	14.6 (13.0,16.3)	23.3 (21.2,25.5)
Appetite loss	9.77 (8.65,11.0)	10.9 (8.39,14.1)	13.0 (11.2,15.1)	6.60 (5.77,7.56)	10.3 (9.31,11.5)
Chills	9.67 (8.56,10.9)	7.32 (5.68,9.44)	15.5 (13.4,18.0)	9.31 (8.35,10.4)	14.1 (12.9,15.4)
Muscle aches	6.09 (5.54,6.69)	5.53 (4.49,6.82)	7.75 (6.85,8.77)	7.16 (6.53,7.84)	9.82 (9.09,10.6)
Severe fatigue	5.84 (4.93,6.92)	5.44 (3.90,7.60)	6.52 (5.36,7.92)	3.88 (3.28,4.59)	5.90 (5.20,6.71)
Hoarse voice	4.92 (4.24,5.70)	6.12 (4.44,8.46)	7.94 (6.83,9.24)	10.2 (9.24,11.3)	14.9 (13.7,16.2)
Blocked nose	4.88 (4.38,5.44)	6.84 (5.49,8.54)	9.94 (8.81,11.2)	7.49 (6.85,8.19)	11.6 (10.7,12.5)
Headache	4.74 (4.36,5.16)	4.90 (4.07,5.89)	8.26 (7.39,9.24)	7.10 (6.54,7.72)	9.54 (8.87,10.3)
Heavy arms/legs	4.68 (4.04,5.42)	4.59 (3.33,6.33)	4.07 (3.30,5.02)	4.46 (3.85,5.17)	6.83 (6.09,7.67)
Tiredness	4.46 (4.10,4.86)	5.00 (4.17,6.00)	6.29 (5.62,7.02)	5.20 (4.78,5.65)	7.41 (6.90,7.96)
Tight chest	4.37 (3.78,5.05)	3.92 (2.83,5.42)	5.24 (4.35,6.32)	4.62 (4.03,5.29)	5.67 (5.06,6.35)
Dizziness	3.89 (3.40,4.45)	4.04 (3.02,5.40)	5.70 (4.85,6.69)	4.29 (3.77,4.87)	6.06 (5.47,6.71)
Sneezing	3.82 (3.44,4.25)	5.82 (4.69,7.23)	9.04 (8.04,10.2)	9.65 (8.87,10.5)	13.1 (12.2,14.1)
Nausea/vomiting	3.80 (3.22,4.48)	4.85 (3.47,6.79)	5.47 (4.46,6.72)	4.45 (3.80,5.22)	5.21 (4.56,5.95)
Runny nose	3.67 (3.30,4.09)	4.46 (3.55,5.58)	10.1 (9.00,11.3)	8.72 (8.04,9.46)	13.1 (12.2,14.1)
Sore eyes	3.52 (3.08,4.02)	3.95 (2.98,5.22)	6.18 (5.29,7.21)	4.49 (3.98,5.07)	5.94 (5.38,6.56)
Chest pain	3.48 (2.79,4.34)	2.40 (1.43,4.03)	3.17 (2.35,4.28)	3.66 (3.00,4.47)	4.76 (4.06,5.57)
Shortness of breath	3.44 (2.99,3.95)	3.78 (2.89,4.96)	4.66 (3.96,5.48)	3.54 (3.14,3.99)	4.75 (4.30,5.24)
Sore throat	3.38 (3.04,3.76)	5.59 (4.46,7.00)	6.76 (5.96,7.66)	10.5 (9.68,11.5)	16.2 (15.1,17.5)
Diarrhoea	3.24 (2.78,3.78)	3.81 (2.74,5.30)	4.49 (3.70,5.46)	3.75 (3.22,4.36)	3.83 (3.35,4.38)
Difficulty sleeping	2.82 (2.52,3.16)	2.49 (1.94,3.21)	4.13 (3.59,4.74)	3.69 (3.33,4.09)	4.47 (4.10,4.87)
Abdominal pain / belly ache	2.52 (2.17,2.94)	1.87 (1.28,2.75)	3.13 (2.56,3.84)	2.78 (2.38,3.24)	3.70 (3.27,4.18)
Numbness/tingling	2.31 (1.90,2.80)	1.91 (1.23,2.96)	2.06 (1.58,2.68)	2.23 (1.84,2.69)	3.16 (2.73,3.67)

Table 3.1: Odds ratios for swab positivity in N=1,542,510 participants across selected rounds of REACT-1 based on the presence or absence of any of 26 symptoms surveyed. Odds ratios are derived from logistic regression models with swab positive (1/0) as the binary outcome variable, adjusted for age group, sex and vaccination status. 95% confidence intervals are shown. Reproduced from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247]

3.4.3 Multivariable analysis for variable selection

In LASSO penalised logistic regression with stability analysis, loss or change of sense of taste, new persistent cough, and fever were selected as jointly and positively predictive of all variants (Figure 3.3). Notably, coryzal symptoms (sore throat, sneezing and hoarse voice) were only selected for the Omicron BA.1 and BA.2 variants. Chills and muscle aches were selected in all variants except Alpha, where they were marginally selected.

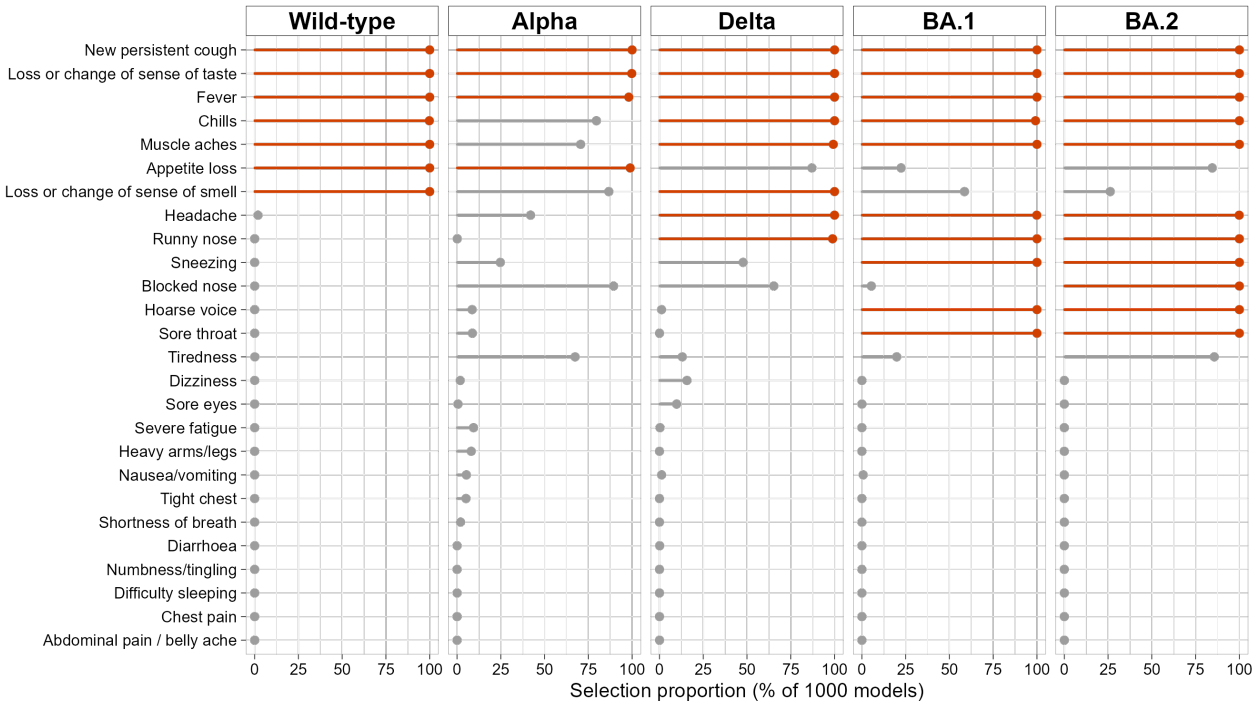


Figure 3.3: Results of LASSO stability selection proportions with each of 26 symptoms as predictors and swab positive/negative as the binary outcome variable, for five SARS-CoV-2 variants in England. Reproduced from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247]

3.4.4 Omicron BA.1 vs BA.2

Figure 3.4 shows the ORs for BA.2 (vs BA.1) in SARS-COV-2 positive individuals from rounds 17-19, in two different analyses: one matched conditional logistic regression and one unmatched logistic regression. In both model types, infection with BA.2 (vs BA.1) is positively associated with severe fatigue, chest pain, sneezing, fever, runny nose, muscle aches, chills, tiredness, blocked nose and headache; in the unmatched logistic regression, infection with BA.2 is further associated with sore eyes, appetite loss and new persistent cough.

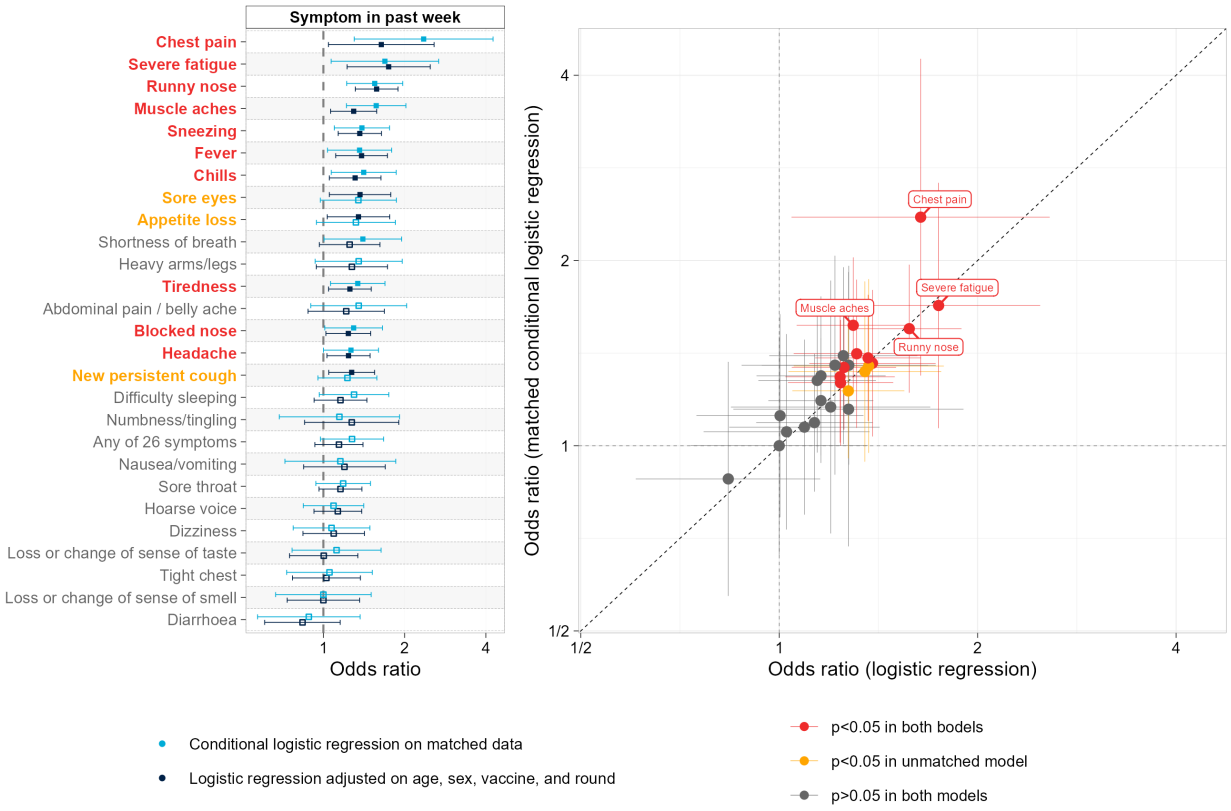


Figure 3.4: ORs for infection with BA.2 vs BA.1 among swab-positive respondents. Left panel shows ORs from logistic regression models among N=5,598 swab-positive individuals with either BA.2 or BA.1 in rounds 17–19 (5 January to 31 March 2022), with each of 26 symptoms as explanatory variables and BA.2 vs BA.1 as the binary outcome variable, adjusted for age group, sex, round and vaccination status. Bars show 95% confidence intervals. Symptoms are ordered by mean OR across BA.1 and BA.2. Right panel shows ORs from conditional logistic regression models among 1,510 swab-positive individuals with either the BA.2 or BA.1 variant in rounds 17–19, with BA.2 vs BA.1 as the outcome variable, matched 1:1 on age (±5 years), sex, vaccination status and round. Bars show 95% confidence intervals and the top 4 symptoms by mean OR across BA.2 and BA.1 are highlighted (named). Right panel directly plots the odds ratios from the two models for comparison. In both model types, infection with BA.2 (vs BA.1) is positively associated with severe fatigue, chest pain, sneezing, fever, runny nose, muscle aches, chills, tiredness, blocked nose and headache; in the unmatched logistic regression, infection with BA.2 is further associated with appetite loss, sore eyes and new persistent cough. Reproduced from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247]

In a subgroup of 5,598 individuals who were double- or triple-vaccinated and tested positive for SARS-CoV-2 via swab, a statistically significant association was observed between infection with the BA.2 variant and increased likelihood of reporting symptoms that significantly impacted daily activities, after adjusting for various demographic and disease-related factors. Specifically, individuals infected with BA.2 were found to be 54% more likely to report severe

symptoms than those infected with BA.1 (OR 1.54 [1.16, 2.06]). Additionally, male individuals were found to be 38% less likely than female individuals to report severe symptoms (0.62 [0.52,0.73]). No association was observed between vaccine booster status or time since vaccination and ability to carry out daily activities. Furthermore, a log-linear regression analysis revealed that individuals infected with BA.2 reported 14% more symptoms, on average, than those infected with BA.1 (OR=1.14 [1.10,1.19]), after controlling for the same covariates as above. (Table 3.2 and Table 3.3)

Variable	Category	crude_mod_OR	plus_age_group_named	plus_sex	plus_boosted	plus_time_since_last_vax	plus_prior_covid_28plus	plus_days_since_symptom_onset	plus_calendar_time
Omicron variant	BA.1 [reference]	-	-	-	-	-	-	-	-
	BA.2	1.86 (1.59,2.18)****	1.94 (1.65,2.28) ****	1.95 (1.66,2.29) ****	1.97 (1.67,2.32) ****	1.92 (1.6,2.29) ****	1.93 (1.61,2.31) ****	1.71 (1.42,2.07) ****	1.56 (1.18,2.07) **
Age	5-17 [reference]		-	-	-	-	-	-	-
	25-34		1.22 (0.8,1.84)	1.26 (0.83,1.91)	1.26 (0.83,1.92)	1.26 (0.83,1.91)	1.27 (0.83,1.93)	1.13 (0.73,1.74)	1.13 (0.74,1.75)
	35-44		1.64 (1.09,2.46) *	1.71 (1.14,2.56) **	1.73 (1.15,2.6) **	1.71 (1.14,2.58) **	1.72 (1.14,2.58) **	1.63 (1.07,2.48) *	1.65 (1.08,2.51) *
	45-54		2.08 (1.39,3.1) ***	2.2 (1.47,3.3) ***	2.25 (1.49,3.38) ***	2.21 (1.47,3.33) ***	2.21 (1.47,3.32) ***	2.09 (1.37,3.19) ***	2.12 (1.39,3.24) ***
	55-64		2.12 (1.42,3.16) ***	2.3 (1.54,3.44) ****	2.36 (1.58,3.55) ****	2.32 (1.54,3.49) ****	2.32 (1.54,3.49) ****	2.24 (1.47,3.41) ***	2.28 (1.49,3.48) ***
	65-74		1.24 (0.81,1.88)	1.39 (0.91,2.12)	1.43 (0.93,2.2)	1.39 (0.9,2.15)	1.37 (0.89,2.12)	1.33 (0.85,2.08)	1.35 (0.86,2.13)
	75+			1.52 (0.93,2.49)	1.57 (0.95,2.58)	1.51 (0.91,2.5)	1.48 (0.89,2.46)	1.74 (1.03,2.96) *	1.79 (1.05,3.05) *
Sex	Female [reference]			-	-	-	-	-	-
	Male				0.56 (0.48,0.66) ****	0.56 (0.48,0.66) ****	0.56 (0.48,0.66) ****	0.61 (0.52,0.73) ****	0.61 (0.52,0.73) ****
Boosted (Yes)	No [reference]				-	-	-	-	-
	Yes					0.97 (0.72,1.31)	0.95 (0.71,1.29)	0.92 (0.67,1.27)	0.88 (0.63,1.23)
Weeks since last vaccination							1 (0.99,1.01)	1 (0.99,1.02)	1 (0.99,1.01)
Prior COVID-19 (28+ days ago)	No [reference]						-	-	-
	Yes							0.92 (0.65,1.29)	0.92 (0.65,1.29)
Weeks since symptom onset									0.81 (0.68,0.98) *
Calendar time									1.02 (0.98,1.05)

Table 3.2: Odds ratios generated by logistic regression of the response to the question "How much, if at all, do the symptoms you have had in the last 7 days reduce your/their ability to carry out day-to-day activities?" as a function of BA.2 / BA.1 infection, age group, sex, booster vaccine received (y/n), weeks since most recent vaccine, prior COVID-19 (28 days or more before testing), weeks since symptom onset, and calendar time (since 1 Jan 2021) among 5,637 PCR-positive individuals with either BA.2 or BA.1 infection who had received two or more vaccines. Reproduced from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

Variable	Category	crude_mod_OR	plus_age_group_named	plus_sex	plus_boosted	plus_time_since_last_vax	plus_prior_covid_28plus	plus_days_since_symptom_onset	plus_calendar_time
Omicron variant	BA.1 [reference]	-	-	-	-	-	-	-	-
	BA.2	1.3 (1.27,1.33)****	1.29 (1.26,1.32) ****	1.29 (1.26,1.32) ****	1.29 (1.26,1.32) ****	1.28 (1.25,1.31) ****	1.29 (1.25,1.32) ****	1.19 (1.16,1.22) ****	1.14 (1.1,1.19) ****
Age	5-17 [reference]		-	-	-	-	-	-	-
	25-34		1.06 (1,1.12) *	1.07 (1.02,1.13) **	1.07 (1.02,1.13) *	1.07 (1.01,1.13) *	1.08 (1.03,1.14) **	1.02 (0.96,1.07)	1.02 (0.97,1.08)
	35-44		1.09 (1.04,1.15) ***	1.11 (1.05,1.17) ***	1.11 (1.05,1.17) ***	1.1 (1.05,1.16) ***	1.11 (1.05,1.17) ***	1.06 (1,1.12) *	1.06 (1.01,1.12) *
	45-54		1.12 (1.06,1.18) ****	1.15 (1.09,1.21) ****	1.14 (1.08,1.2) ****	1.14 (1.08,1.2) ****	1.13 (1.07,1.19) ****	1.07 (1.02,1.13) *	1.08 (1.02,1.14) **
	55-64		1.01 (0.95,1.06)	1.04 (0.99,1.1)	1.03 (0.98,1.09)	1.03 (0.98,1.09)	1.03 (0.97,1.09)	1 (0.95,1.06)	1.01 (0.96,1.07)
	65-74		0.83 (0.79,0.88) ****	0.87 (0.82,0.92) ****	0.87 (0.82,0.92) ****	0.86 (0.81,0.91) ****	0.85 (0.8,0.9) ****	0.85 (0.8,0.9) ****	0.86 (0.81,0.91) ****
	75+			0.73 (0.68,0.78) ****	0.72 (0.67,0.78) ****	0.72 (0.66,0.77) ****	0.7 (0.65,0.76) ****	0.77 (0.72,0.84) ****	0.78 (0.72,0.84) ****
Sex	Female [reference]			-	-	-	-	-	-
	Male				0.79 (0.77,0.81) ****	0.79 (0.77,0.81) ****	0.79 (0.77,0.81) ****	0.85 (0.83,0.87) ****	0.85 (0.83,0.87) ****
Boosted (Yes)	No [reference]				-	-	-	-	-
	Yes					1.04 (1,1.09)	1.02 (0.98,1.07)	0.95 (0.91,0.99) *	0.93 (0.89,0.98) **
Weeks since last vaccination							1 (1,1)	1 (1,1)	1 (1,1)
Prior COVID-19 (28+ days ago)	No [reference]						-	-	-
	Yes							0.86 (0.82,0.91) ****	0.86 (0.82,0.91) ****
Weeks since symptom onset									1.04 (1.01,1.06) *
Calendar time									1.01 (1,1.01) *

Table 3.3: Results from log-linear regression of symptom count as a function of BA.1 / BA.1 infection, age group, sex, booster vaccine received (y/n), weeks since most recent vaccine, prior COVID-19 (28 days or more before testing), weeks since symptom onset, and calendar time (since 1 January 2021) among 5,637 PCR-positive individuals with either BA.2 or BA.1 lineage infection who had received two or more vaccines. Errors are modelled as Poisson distributed. Odd Ratios and 95% confidence intervals are shown. One covariate to the model in each column. BA.2 infection (vs BA.1) is associated with reporting of more symptoms and the effect is robust to adjustment. Reproduced from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

3.4.5 Symptom trajectory

Modelling the probability of experiencing each of the symptoms at N days post-symptom-onset revealed that headache and sore throat were consistently the most likely symptoms to manifest early in the disease course (0-10 days after first symptoms) across all variants, with sore throat being particularly common early in the disease course (Figure 3.5). These symptoms generally resolved within the first 14 days. By contrast, prevalence of loss or change of sense of smell and loss or change of sense of taste was low early in the disease course and gradually increased until around day 10. This pattern was common to all variants, with particularly high prevalence for the Delta variant, and considerably lower overall prevalence for the BA.1 and BA.2 variants. Relative to other symptoms, headache was more prevalent in the first few five days of the disease course in Omicron BA.1 and BA.2 variants than in preceding variants.

It should be noted that this is cross sectional data and that the question about symptom onset was retrospective (ie "How many days ago did these symptoms/ this symptom start?). Therefore, as the analysis is conducted among PCR-positive individuals only, the results are biased by the necessity for the individual to still be PCR positive at N days' remove from symptom onset.

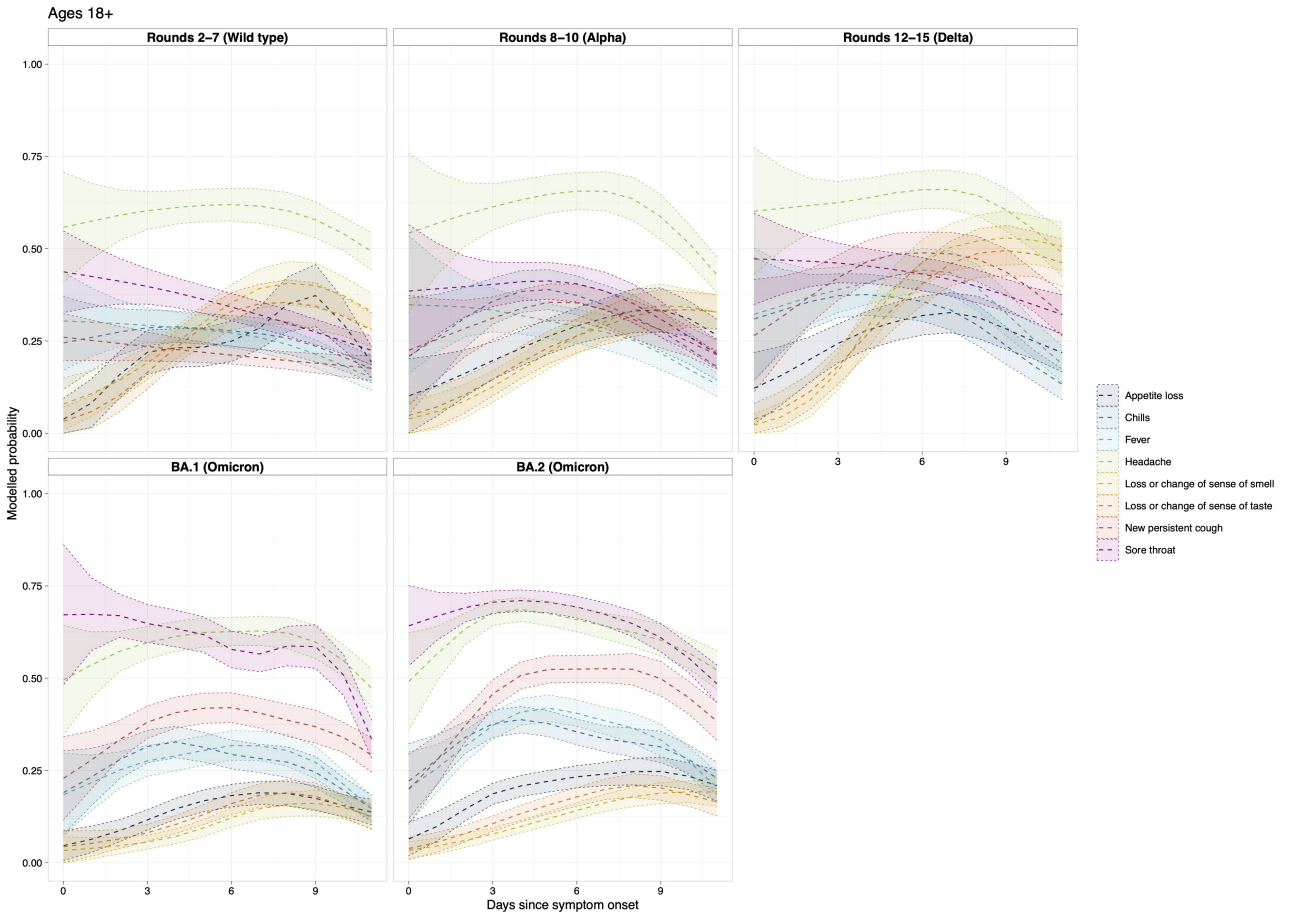
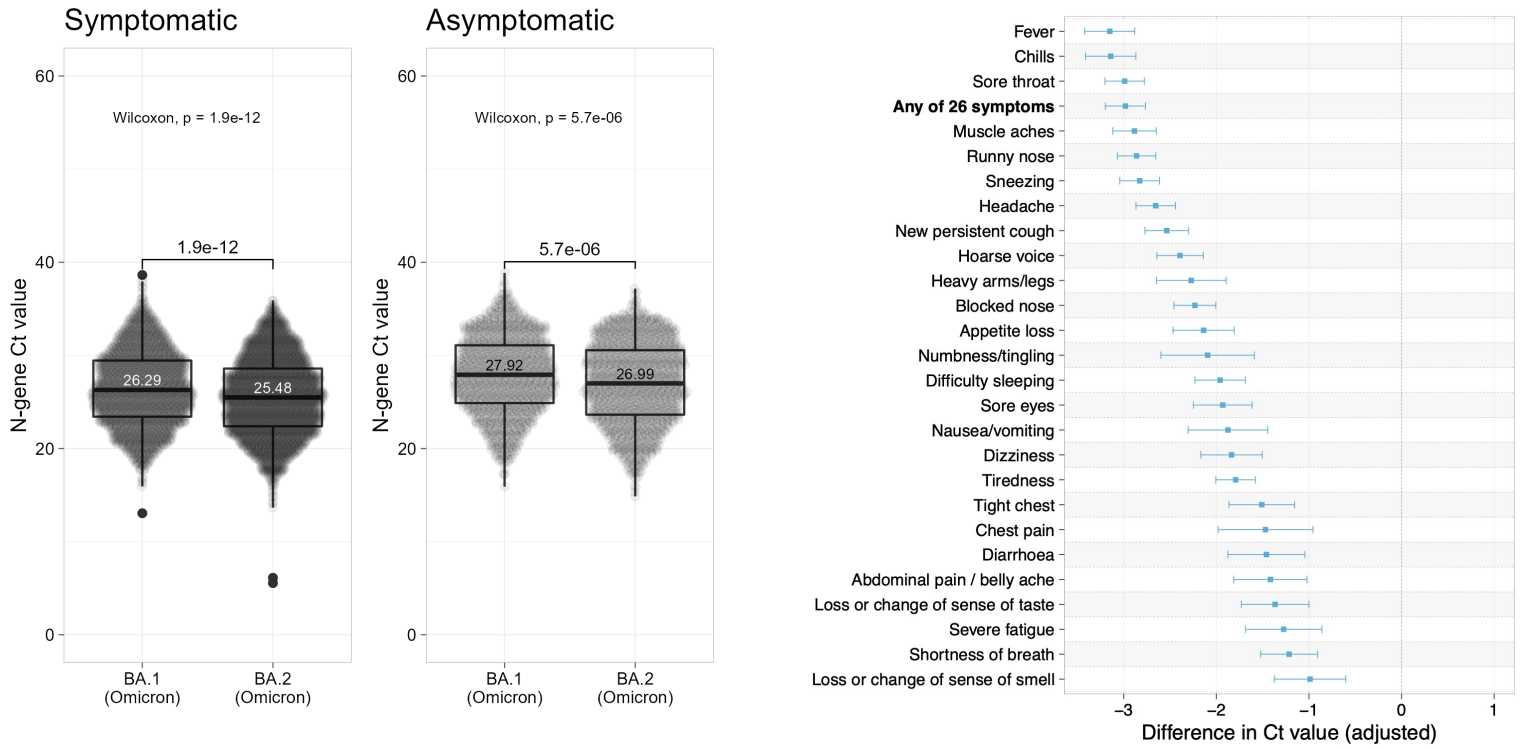


Figure 3.5: Modelled probability of each of eight common symptoms at N days' remove from first symptom onset.

3.4.6 Ct values

Cycle threshold (Ct) values were lower for the BA.2 variant than BA.1 (Figure 3.6a). This may be due to the timing of the sampling in relation to the growth of the variant as more recent infections tend to have lower Ct values, as shown in figure B.3 in Appendix B, which demonstrates a positive correlation between time-since-symptom-onset and Ct values and that the mean time since symptom onset was lower for BA.2 than for BA.1. Additionally, symptomatic individuals had lower Ct values (indicating higher viral loads) than asymptomatic individuals. Linear regression models of swab-positive individuals in rounds 17-19 (from 5 January to 31 March 2022) revealed that symptom reporting was associated with a lower Ct value for each of the 26 surveyed symptoms. The lowest adjusted Ct values were for influenza-like or cold-like symptoms such as fever, chills, sore throat, muscle aches, runny nose, sneezing, and headache (Figure 3.6b), which frequently co-occurred (Appendix B, figure B.4). With the exception of

fever, these symptoms were also commonly reported as the first symptom among symptomatic swab-positive individuals (Appendix B, figure B.5).



(a) Ct values boxplot

(b) Ct values regression model results

Figure 3.6: Plots showing: (i) Ct values in N=6,395 swab-positive participants with Omicron BA.2 and BA.1 variants from REACT-1 rounds 17–19, stratified by symptom status (presence of any of 26 surveyed symptoms). A two-tailed Wilcoxon rank sum test was used to compare the Ct value distributions in BA.2 vs BA.1, within symptomatic and asymptomatic individuals. Ct values are lower in BA.2, in both asymptomatic and symptomatic respondents. (ii) Results of regression models with N-gene Ct values as the outcome and symptoms as predictors, adjusted for age, sex and, where appropriate, vaccination status, in N=10,709 swab-positive individuals in rounds 17–19 (5 January to 31 March 2022). Reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

3.5 Discussion

In this chapter, analysis of more than 1.5 million adults in the REACT-1 programme in England reveals differences in symptom reporting associated with each variant of COVID-19 that has been dominant in England. This could indicate alterations in the pathophysiology underlying the different variants, such as changes in receptor binding, cell entry, or host response, and/or in the presence of varying levels of population immunity resulting from both natural infection and vaccination [259–261].

The results show that symptoms of loss or change of sense of smell or taste were less predictive of a positive swab result for Omicron than for other variants. Conversely, cold-like symptoms were more indicative of Omicron infections than previous variants. These findings align with previous studies[262–264]. In particular, Omicron infections do not exhibit as strong an association with anosmia as earlier variants. The loss or change of the sense of smell or taste following infection with previous variants of SARS-CoV-2 is attributed to the decreased expression of olfactory receptors[265]. Changes in the viral genes responsible for regulating host responses in Omicron may reduce this effect. Further research through transcriptomic studies in both animal and human subjects may assist in uncovering the underlying mechanisms.

It is difficult to compare the intrinsic severity of SARS-CoV-2 variants, owing to changing levels of population immunity from vaccination and prior infection, and potentiating factors such as seasonality. In these data, however, the rapid replacement of BA.1 by BA.2, and the large number of PCR positive individuals with these variants, presented an opportunity to compare the symptom burden and severity of two variants within a narrow time frame, in a population with similar characteristics (in terms of vaccination rates and prior infection rates), against a similar background of non-COVID related illness and symptoms. This comparison revealed that those with BA.2 were more likely to be symptomatic, more likely to report a range of influenza-like and cold-like symptoms, and, in adjusted models, likely to report more symptoms, and to report that these symptoms affected their day-to-day activities “a lot”. The transition from BA.1 to BA.2 was happening during a vaccine booster campaign in the UK [99]; nonetheless, the findings relating to severity and symptom burden were robust to adjustment for both vaccine booster status and for time since most recent vaccine dose and are therefore unlikely to be explained by vaccination status or waning vaccine immunity. The

effects were however attenuated by the addition of calendar time to the model. This implies that the elevated symptom burden and symptom severity of BA.2 relative to BA.1 may partly reflect the detection of swab positivity earlier in the disease course for BA.2 than BA.1. This is a plausible hypothesis given the timing of the REACT-1 fieldwork in rounds 17–19 relative to the variant transitions: round 17 took place at a point where BA.1 had already achieved dominance, while round 18 was conducted just as BA.1 was yielding to BA.2, and round 19 just at the tail end of the transition. It is also likely that the inclusion of calendar time in the models does not fully account for this transition effect; nonetheless, after adjustment for calendar time, BA.2 was associated with 54% higher odds of symptoms that affect day-to-day life "a lot" than BA.1, and also with the reporting of one additional symptom, on average, than BA.1.

To date, five other studies have explicitly set out to compare the severity of Omicron BA.2 vs BA.1, in different geographical populations. An early analysis of hospitalisation data in South Africa, where BA.2 was first reported, found similar rates of hospitalisation in BA.1 and BA.2, and similar rates of progression to severe illness, conditional on hospitalisation [266]. Sievers *et al* analysed hospital admission data in Germany and found similar rates for Omicron BA.1 and BA.2 [267]. Fonager *et al* found similar results in Denmark [268], with no increased risk of hospitalisation or mortality. Strasser *et al* analysed data on COVID-19 cases in New England between March 2020 and June 2022, and found that BA.2 was intrinsically less severe than previous variants (BA.1 and Delta), taking into account vaccination status, prior infection and therapeutic treatment. Lewnard *et al* [269] also compared the severity of BA.2 vs BA.1 in terms of hospitalisations and death in Southern California, and found no difference. These findings are not necessarily incompatible with the results presented here: all these studies quantify severity in terms of hospitalisation or death rates, while severity here is measured in terms of symptom burden and severity. While mortality and hospitalisation rates are clearly the greatest public health concern for an infectious disease, particularly in the acute phase of a pandemic, the impact on quality of life and ability to work is still an important consideration, and the greater symptom burden and severity of BA.2 may still be associated with greater disruption to daily life, and have broader societal and economic impact.

As of early 2023, few countries worldwide have any remaining stay-at-home restrictions and most have moved away from mandatory or voluntary mass testing programmes [34]. In this

post-acute phase of the pandemic, the spread of SARS-CoV-2 may be limited by using symptoms to identify individuals who are particularly likely to be infectious. The analysis of symptom trajectory and Ct values therefore complements the broader symptom profiling by helping determine whether a person is at an early point in their disease course (as infectiousness is shown to be higher in the presymptomatic and early symptomatic phase[27]), and by inferring their likely level of infectiousness on the basis of symptoms. In the Omicron period, sore throat and headache were shown to be the most common symptoms in the early symptomatic phase, and reporting fever, chills, sore throat, muscle aches, runny nose, sneezing and headache was associated with the lowest adjusted Ct values, and therefore likely to be indicative of higher viral load and increased infectiousness.

3.6 Conclusion

Data from REACT-1 show differences in symptom reporting during two years of the COVID-19 epidemic in England. This reflects the emergence and subsequent dominance of different variants over that period, against a background of changing levels of immunity from vaccination and prior infection. Omicron variants BA.1 and BA.2 were associated with lower overall reporting of loss or change of sense of smell and taste, higher reporting of cold-like and influenza-like symptoms, and relatively higher reporting of headache at symptom onset. BA.2 was associated with a greater symptom burden and severity than BA.1, running counter to the prevailing conception that successive variants are intrinsically milder.

Chapter 4

Persistent symptoms of COVID-19

This chapter uses data from four rounds of the REACT-2 study to investigate persistence of symptoms in people who self-report prior COVID-19. Classical statistical approaches are used to (i) estimate the prevalence of persistent COVID-19 symptoms in the community, (ii) estimate the likelihood of experiencing persistent symptoms conditional on infection, and (iii) identify putative risk factors for persistent COVID-19 symptoms. Machine learning methods are deployed to rank predictors of persistent symptoms in a multivariable context, and to cluster individuals by their COVID-19 symptom profile at 12 weeks after symptom onset, with a view to identifying and characterising 'sub-types' of the condition known as long COVID or post-COVID syndrome.

Note: Some findings from this chapter have been published in Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270].

4.1 Motivation

COVID-19 poses not only an acute risk of hospitalisation and mortality, but also a significant burden of prolonged and debilitating illness for some individuals. The frequency, nature, and duration of persistent symptoms associated with COVID-19 remain inadequately understood, constituting a significant gap in knowledge that must be addressed in order to develop effective treatments and management strategies.

In this chapter, self-reported data from more than 600,000 participants in four rounds of REACT-2 are used to describe, characterise, and estimate the prevalence of post-COVID syndrome, using the broad definition of 'experiencing one or more [from a list of 30] symptoms, 12 weeks or more after the date of onset of the first symptom'. A range of phenotypic and socio-demographic variables are investigated as putative risk factors, and unsupervised learning methods are used to identify potential sub-types of post-COVID syndromes.

4.2 Methods

4.2.1 Study participants

The study design and recruitment methods are covered in the Introduction.

The primary study population in this chapter is rounds 3–5 from REACT-2, carried out from 15 to 28 September2020 (round 3), 27 October to 10 November 2020 (round 4) and 25 January to 8 February 2021 (round 5). Round 6 (conducted in May 2021) is used as a replication population, with separate estimates of prevalence calculated, and predictive modelling and clustering analysis repeated. This split of the study population arose as a consequence of this analysis being conducted predominantly in real-time, as the data were gathered, and there being a larger gap between rounds 5 and 6 than in other rounds. Nonetheless the separation is useful as a framework for replication and validation of the results.

While the primary research goal of the REACT-2 study was to track population seroprevalence using LFIA antibody tests, LFIA results were not used in this analysis. All analyses were conducted using survey response data, principally questions around confirmed or suspected prior COVID-19 and the symptoms associated with it. Participants were asked to state if they thought they had previously had COVID-19, and to state the date of symptom onset. As most of the infections occurred before mass testing was available, a confirmatory test was not a requirement. Participants were then asked to select which, if any, symptoms they had experienced as part of their suspected COVID-19, from a list of 29 symptoms, and how long they had experienced the symptom for [1].

[1]The symptom-specific durations were provided either as a count of days or of weeks: a start and end date were not requested for each symptom. This necessitated an assumption – that each symptom had begun on

In addition to the 29 symptoms that were included on the questionnaire in rounds 3–5, 8,370 respondents gave free-text descriptions of other symptoms. Free-text analysis of word occurrence and co-occurrence identified further common symptoms that were not in the round 3–5 survey, including raised blood-pressure, hair-loss, brain-fog, heart-palpitations and severe joint pain. These free text responses informed updates to the questionnaire in round 6 to include additional symptoms: 35 symptoms were included in the survey in round 6, of which 27 were in common with those asked in rounds 3–5.

In order to study risk factors for persistent symptoms conditional on infection, a sub-group was selected of 76,155 participants from rounds 3-5 who self-reported having had COVID-19—either PCR confirmed or suspected—with one or more of the 29 symptoms 12 weeks or more before the survey date. The main analyses were repeated in an independent data set of 13,170 participants from round 6 who reported one or more of an expanded list of 35 symptoms 12 weeks or more before the survey date. To maintain comparability with symptom reporting in rounds 3–5, some analyses in round 6 were restricted to the 27 symptoms in common.

In order to determine the background prevalence of symptoms in the population, data from the REACT-1 study was used. Unlike REACT-2, the REACT-1 study includes children aged 5-17 years: these were excluded from the estimation of the background symptom reporting rate. The way symptoms were surveyed in REACT-1 was slightly different to REACT-2: the study collected information on the presence of any of 26 symptoms that persisted for a duration of 11 or more days. Nonetheless, estimating the prevalence of 11-day persistent symptoms using this data provides an upper bound for the background prevalence of persistent symptoms in the population.

4.2.2 Data analysis

Estimating prevalence

Community prevalence of persistent symptoms, defined as one or more symptoms still present 12 or more weeks after symptom onset, was estimated separately in rounds 3–5 and in round

the same date – in order to conduct the analysis. As has been demonstrated in Chapter 1, symptoms tend to manifest at different points in the disease course; however, the majority of symptoms manifest within the first 2–3 weeks.

6. Random Iterative Weighting [271] was applied to the data set to allow prevalence to be re-weighted to be more representative of the population of England in terms of age, sex, ethnicity, deprivation and LTLA population. Within-group prevalence estimates were calculated across a range of characteristics (eg age, sex, ethnic group).

Logistic regression

Logistic regression was used to investigate the associations between a range of demographic and lifestyle factors with persistence of symptoms beyond 12 weeks. Two model specifications were tested: age and sex adjusted, and fully adjusted (age, sex, ethnicity, adiposity, smoking, vaping, household income, deprivation, healthcare worker status, and prior hospitalisation with COVID-19 [y/n]).

Variable ranking and selection

To investigate the predictive capacity of each factor in a multivariable context, gradient boosted models (as implemented in CatBoost [229]) were constructed, with predictive performance with and without each variable compared and quantified as a measure of variable importance.

First, a model was trained on 80% of the data. Overfitting was controlled by further splitting these data into a training pool and a validation pool. The model was fitted on the training pool, and after each round (each additional tree) performance was evaluated on the validation pool. Model training ceased once performance stopped improving on the validation pool. The objective function was logloss. Default settings were used for all other model hyperparameters. Next, the model was used to predict the outcome variable in the 20% holdout data, and the AUC recorded. Next, each variable was permuted in turn in the holdout data, and predictions and AUC calculated each time. This process was repeated 1000x to obtain a distribution of AUCs for the model with each of the variables permuted. The proportion of permuted runs where the AUC was greater than the unpermuted prediction run was taken as a p-value that the variable was contributing to predictive accuracy. Variables whose proportion was below 5% ($p<0.05$) were selected.

Generalised additive models

To investigate how risk of experiencing persistent symptoms changes over the life course, generalised additive models (GAMs) were constructed with symptom persistence (12+ weeks) as the binary outcome variable and age and sex as predictors. A default thin plate spline was appied with an interaction between age and sex, and the smoothed functions were plotted to show the relationship between persistent symptom risk and age. Three versions of the model were constructed:

1. First, to show the risk of experiencing symptoms in the population at any time, a model was constructed with 'any symptom for any duration y/n' as the outcome variable, and age and sex as predictors. This model shows the likelihood of self-reported prior symptomatic COVID-19, and how it varies by age, in the full study cohort (N=508,707).

2. Next, to show the risk of experiencing *persistent* symptoms in the population at any time, a model was constructed with 'any symptom for 12 weeks or more y/n' as the outcome variable, and age and sex as predictors. This model shows the likelihood of self-reported persistent symptoms associated with COVID-19, and how it varies by age, in the full study cohort (N=508,707).

3. Finally, to show the risk of experiencing *persistent* symptoms *conditional on having symptomatic COVID-19*, a model was constructed with 'any symptom for 12 weeks or more y/n' as the outcome variable, and age and sex as predictors, among *only* individuals who reported having experienced symptomatic COVID-19 at least 150 days before the REACT survey date (thus giving 150 days of 'observation time', including the minimum 12 weeks to meet the definition of 'persistent symptoms', and additional time, within the constrictions of the data, to register recovery). This model shows the likelihood of going on to report persistent symptoms associated with COVID-19, and how this varies by age, if you have symptomatic COVID-19.

Identifying persistent symptom clusters

In order to segment participants in rounds 3-5 based on symptom clusters, a binary matrix was generated to represent the presence or absence of each of the 29 surveyed symptoms at 12

weeks post-symptom onset for each individual participant. The Clustering LARge Applications (CLARA) extension of the Partitioning Around Medoids (PAM) algorithm, as implemented in the R package *fpc* [272], was used for this clustering process. PAM aims to identify the most representative data points as cluster centroids by minimising the sum of dissimilarities between data points and their assigned centroids, while CLARA employs a sampling approach to alleviate computational demands for large data sets. The Hamming distance was employed as a measure of dissimilarity between participants; as a sensitivity analysis, clustering was re-run with Dice [273] and Jaccard distance. The optimal number of clusters was determined using the average silhouette width. Two methods were employed to evaluate cluster stability. The first involved bootstrapping and re-clustering 100 times, and quantifying the difference between bootstrapped and non-bootstrapped clusters using the Jaccard coefficient, with a range of 0 indicating no overlap and 1 indicating perfect overlap. The second method involved removing each symptom in turn, re-clustering, and calculating the average proportion of non-overlap (APN) between these clusters and the whole-dataset clusters as a proxy for individual variable importance and contribution to population segmentation. Clusters were characterised and visualised with heatmaps showing within-cluster symptom prevalence.

Latent class analysis

To interrogate the validity of any structure identified in clustering by using a different approach to partitioning, Latent Class Analysis was applied to the symptom data at 12 weeks. Observations were assigned to the class with the highest posterior probability, and the classes were characterised in the same way as the clusters (heatmaps showing symptom prevalence by class).

Estimating background symptom prevalence

The REACT-1 study asked participants what symptoms they had experienced in the past 7 days, and how long before that each symptom had started, up to a maximum of 11+ days. Taking the proportion of PCR-negative people who said they were experiencing one or more symptoms that had started 11 or more days previously gives an upper bound estimate of the prevalence of non-COVID-related persistent symptoms in the community during this period.

4.3 Results

Figure 4.1 shows the study design and population. A total of 508,707 people took part in REACT-2 rounds 3–5, and a further 97,727 in REACT-2 round 6 (excluding a 'booster' sample of additional people recruited at ages 55 years and over), with response rates of 29.4% and 29.9% respectively. Non-responders were disproportionately male, younger (18–24 years) or older (>75 years) adults and from more deprived areas.

In total, 92,116 respondents reported having had COVID-19, either test-confirmed or suspected, in rounds 3–5 (weighted prevalence 19.2% [19.1,19.3]), and 14,562 in round 6 (weighted prevalence 17.9% [17.7,18.0]).[2]

[2]The (somewhat surprising) decrease in prevalence between rounds 3–5 and round 6 is potentially attributable to a (i) the increased availability of testing during round 6, meaning that individuals with suspected COVID-19 were able to test, and some portion of those would be negative, whereas before testing they would simply have been recorded as 'suspected or confirmed', or (ii) greater level of awareness of the symptoms of COVID-19 and a lower likelihood of retrospectively attributing any cold-like illness from early 2020 as COVID-19

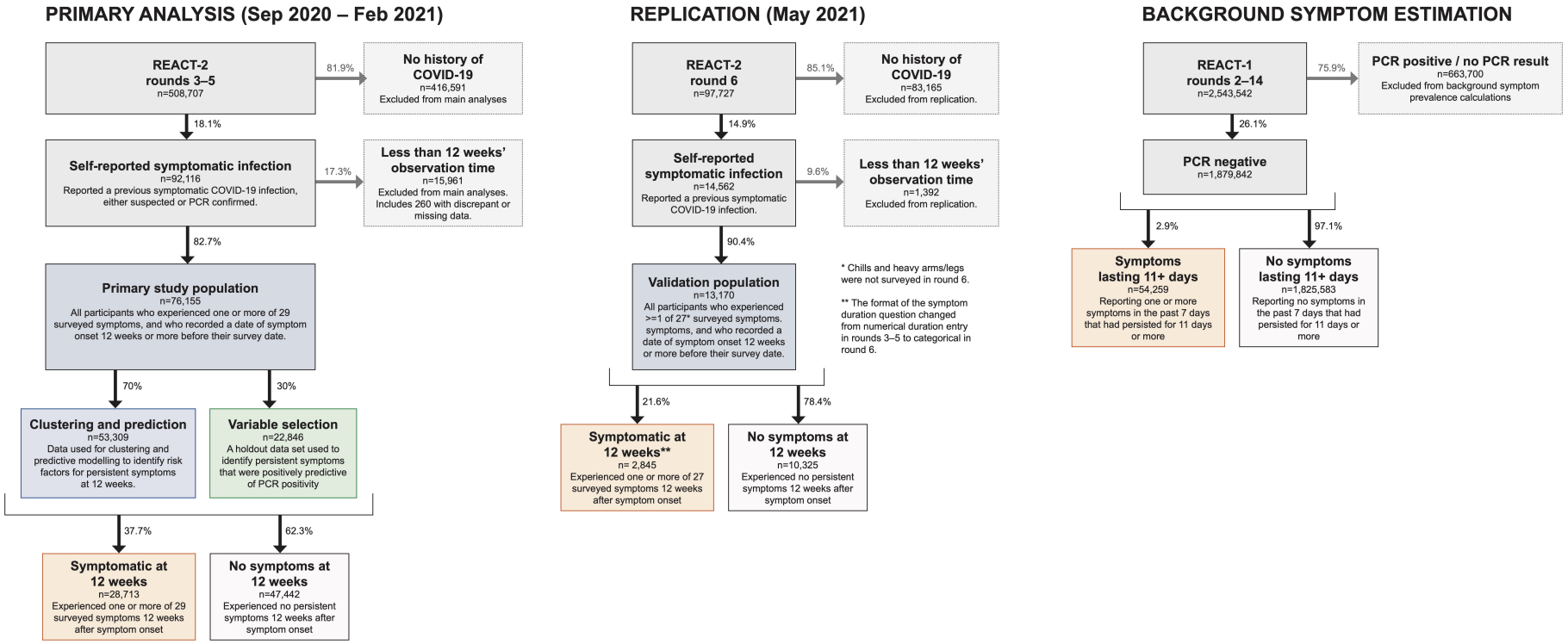


Figure 4.1: A flowchart of the primary and replication study populations, showing exclusions and proportion of participants experiencing persistent symptoms at least 12 weeks after symptom onset, and study population for estimation of background symptom rates. Reproduced from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270].

4.3.1 Community prevalence of persistent symptoms

Community prevalence of current persistent symptoms

Table C.1 (Appendix C) presents the data on still-active persistent symptoms in the study population. A total of 26,943 out of the full population of 508,707 in rounds 3–5 reported they were *currently* experiencing one or more COVID-19 symptoms that had begun at least 12 weeks previously, giving a weighted point prevalence of 5.41% (5.35,5.47). For two or more symptoms, the numbers were 16,368 (3.27% [3.22,3.31]), and for three or more symptoms 10,156 (2.03% [1.99,2.07]). Prevalence of one or more symptoms was higher in the 45–54 age group (6.38% [6.22,6.55]), in women (6.19% [6.09,6.28]), in obese individuals (6.94% [6.78,7.11]) and in those of mixed ethnicity (7.23% [6.59,7.93]). In round 6, the corresponding figures were 3.06% (2.98-3.14) and 1.61 (1.56-1.67) for the 27 symptoms in common with rounds 3-5, and 3.26% (3.18-3.34) and 1.86% (1.80-1.92) for one and three symptoms respectively when all 35 symptoms surveyed in round 6 are taken into consideration.

Community prevalence of current or historical persistent symptoms

Table C.2 (Appendix C) presents the data on current or resolved persistent symptoms in the study population. A total of 28,973 out of the full population of 508,707 in rounds 3–5 reported they currently had, or had had, one or more COVID-19 symptoms that had lasted for at least 12 weeks, giving a weighted prevalence of 5.80% (5.73,5.86). For two or more symptoms, the numbers were 17,966 (3.56% [3.51,3.61]), and for three or more symptoms 11,287 (2.23% (2.19,2.27)). The trends by population characteristic were similar to those with current persistent symptoms: higher prevalence in those in the 45–54 age group, in women, in obese individuals and in those of mixed ethnicity.

Prevalence of persistent symptoms in those with symptomatic infection

Table 4.1 shows the proportion of individuals who reported symptomatic COVID-19 who reported the presence of one or more, or three or more, of the 29 symptoms surveyed at 12 weeks post symptom onset. In rounds 3-5, 37.7% (37.4-38.1) of participants reported one or more persistent symptoms, and 17.5% (17.2-17.7) reported three or more persistent symptoms;

in round 6, 21.6% (20.9-22.3) of participants reported one or more persistent symptoms, and 11.9% (11.4-12.5) reported three or more persistent symptoms.

	n	n with prior symptomatic COVID-19 and 12 weeks' observation time	% with one or more symptoms at 12 weeks	% with three or more symptoms at 12 weeks
Rounds 3–5	508,707	76,155	37.7 [37.4–38.1]	17.47 [17.2–17.7]
Round 6	97,727	13,170	21.6 [20.9–22.3]	11.94 [11.4–12.5]
Round 6 (extended symptom list)			22.8 [22.1–23.5]	13.82 [13.2–14.4]

Table 4.1: Proportions of respondents in (i) rounds 3–5 and (ii) round 6 who reported one or more (or three or more) symptoms 12 weeks after initial symptom onset.

4.3.2 Prevalence by symptom

Figure 4.2 compares the prevalence of each symptom across the main study period (rounds 3–5) and separately in round 6, among the N=76,155 (in rounds 3–5) and N=13,170 participants who reported symptomatic COVID-19 and for whom 12 weeks' observation time was available (ie their date of symptom onset was 12 weeks or more before the date of survey response in REACT-2). Tiredness was the most prevalent symptom in both rounds 3–5 and in round 6; however, the prevalence of persistent tiredness decreased substantially, from 16.8%, (16.5-17.1) in rounds 3–5 to 8% (7.5-8.6) in round 6. The decrease in overall symptom prevalence between rounds 3–5 and round 6 was driven primarily by this decrease in reported persistent tiredness, although smaller declines in the prevalence of 16 of the 26 other symptoms were observed.

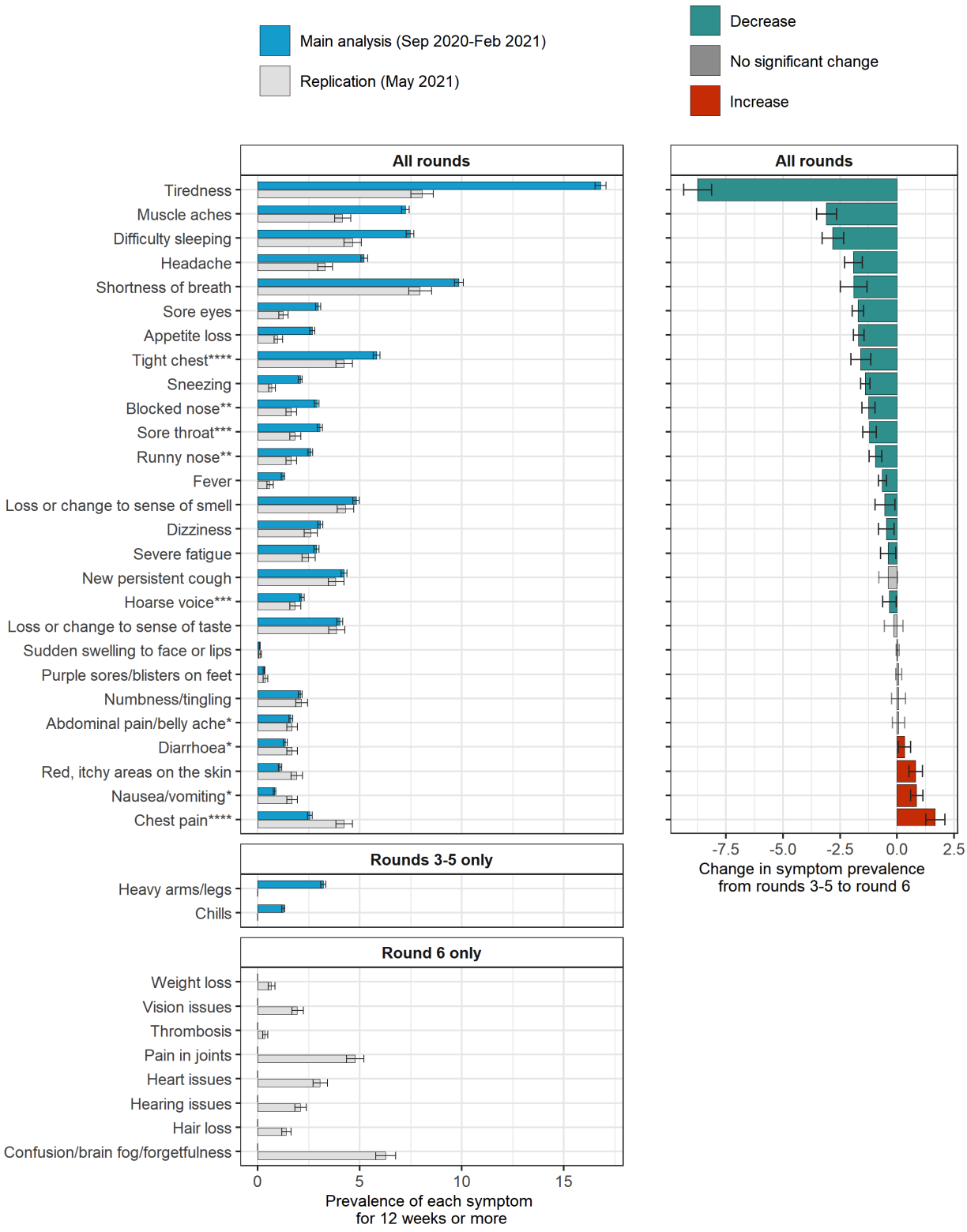
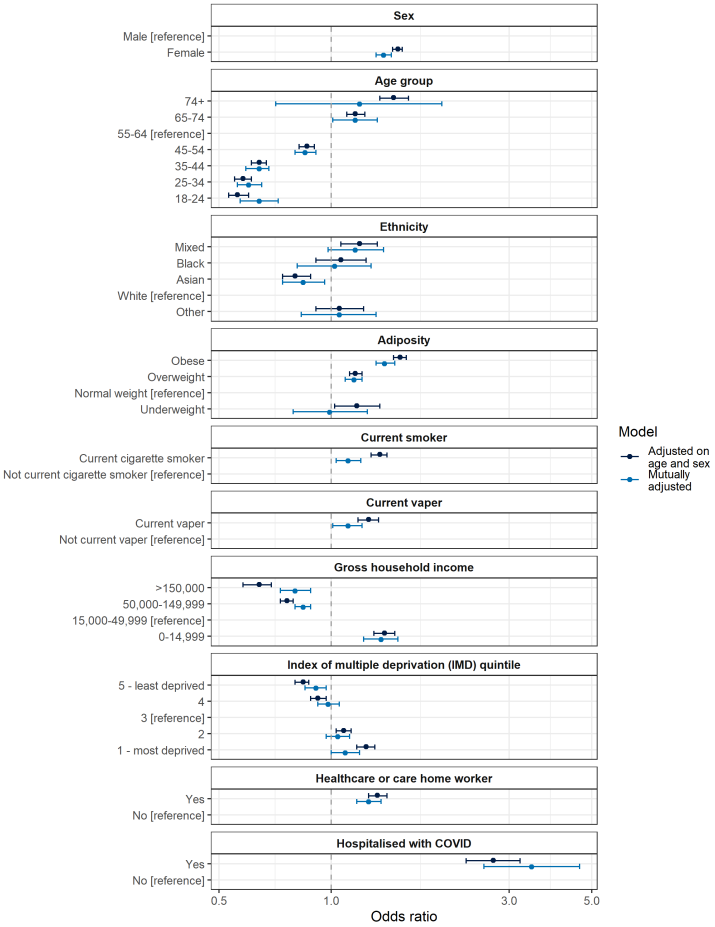


Figure 4.2: Individual symptom prevalence in rounds 3–5 (September 2020 – February 2021) and separately in round 6 (May 2021). Additional symptoms only surveyed in round 6 are shown in the bottom left panel. Right panel shows the change in symptom prevalence between rounds 3–5 and round 6. Asterisks indicate symptoms that were grouped in round 6. Reproduced from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270]

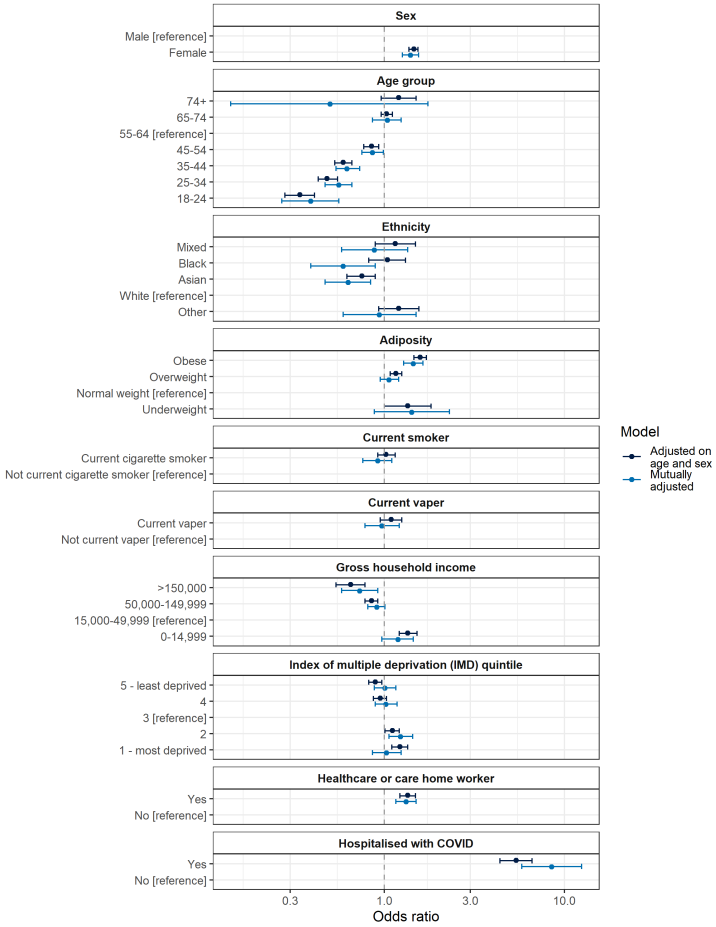
4.3.3 Risk factors for persistent symptoms

Logistic regression

In logistic regression models with the binary outcome variable of one-or-more symptoms at 12 weeks, after mutual adjustment, persistent symptoms were positively associated with female sex, increasing age, being overweight or obese, having been hospitalised with COVID-19, smoking, vaping, low household income and being a care home or healthcare worker. Odds ratios ranged from 1.38 (1.32,1.45) for female sex to 3.45 (2.57,4.64) for hospitalisation with COVID-19. Asian ethnicity was associated with a lower risk of persistent symptoms, after mutual adjustment (figure 4.3). Patterns of association were consistent in round 6.



(a) Rounds 3–5



(b) Round 6

Figure 4.3: Logistic regression models with one or more symptoms at 12 weeks (y/n) as the binary outcome variable, both adjusted for age-sex and mutually adjusted. Reproduced from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270]

Variable ranking and selection

In multivariable analysis for variable ranking and selection, the strongest predictors of persistent symptoms, in order, were age, sex, body mass index (BMI), household income, healthcare/care home worker, deprivation, smoking status, prior hospitalisation with COVID-19 and vaping status (in that order) (figure 4.4). Applying a frequentist approach to the goal of variable selection (ie classifying predictors as selected when permutation of the predictor results in a reduced or unchanged AUC at least 95% of the time over 1,000 runs), all putative risk factors were selected as predictors except for the ethnicities, suggesting that the associations with ethnicity identified in the logistic regression analysis may be attributable to misspecification or model assumptions in the linear model framework that do not hold in this data (eg non-linearity, interactions).

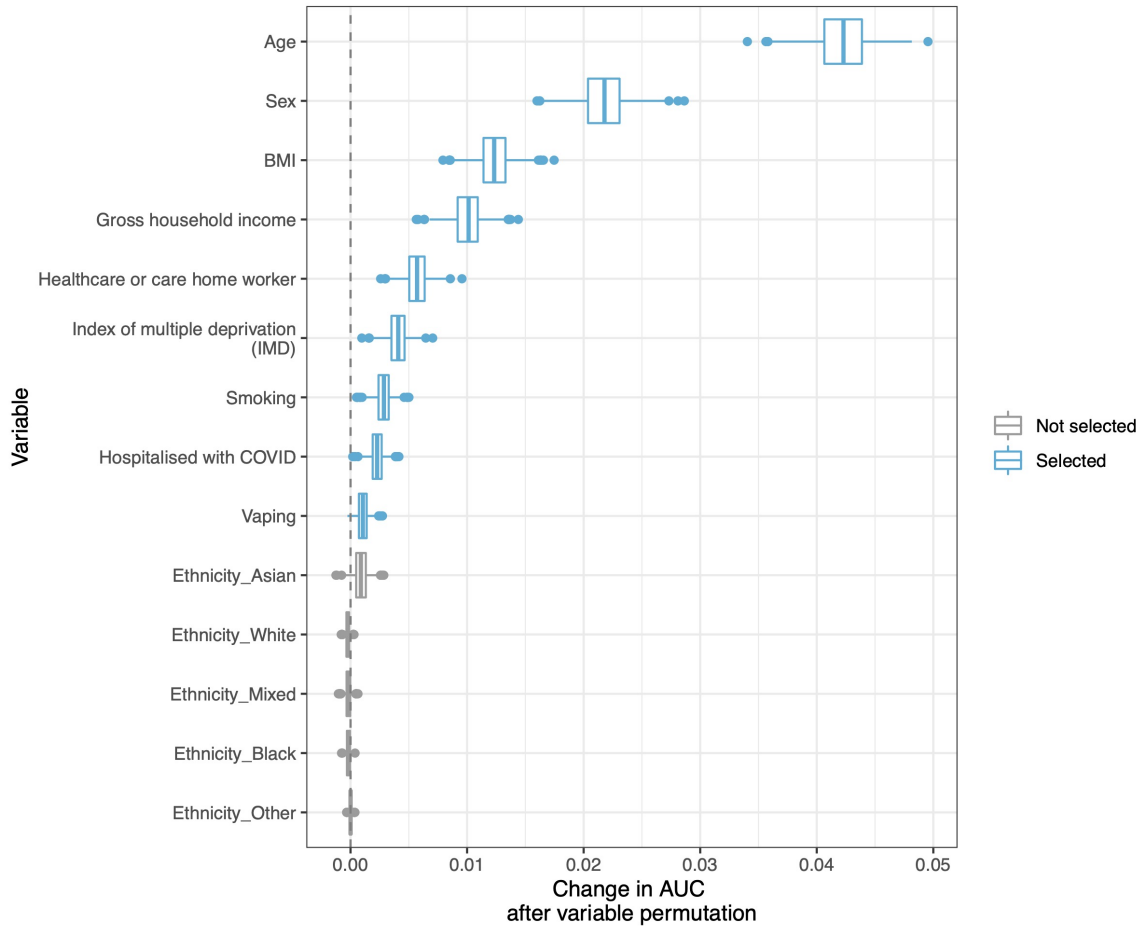
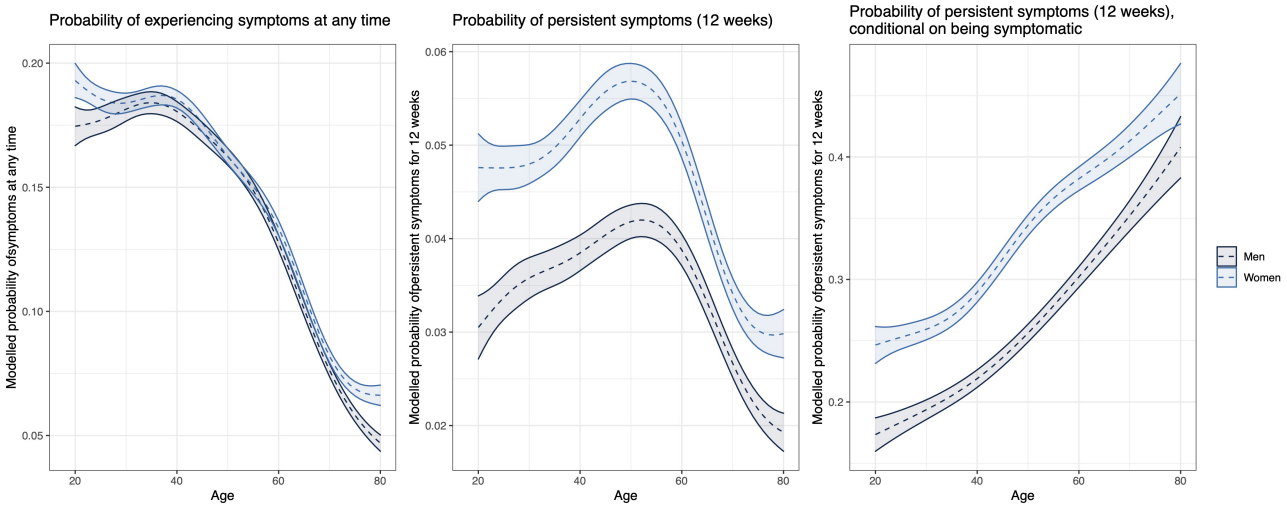


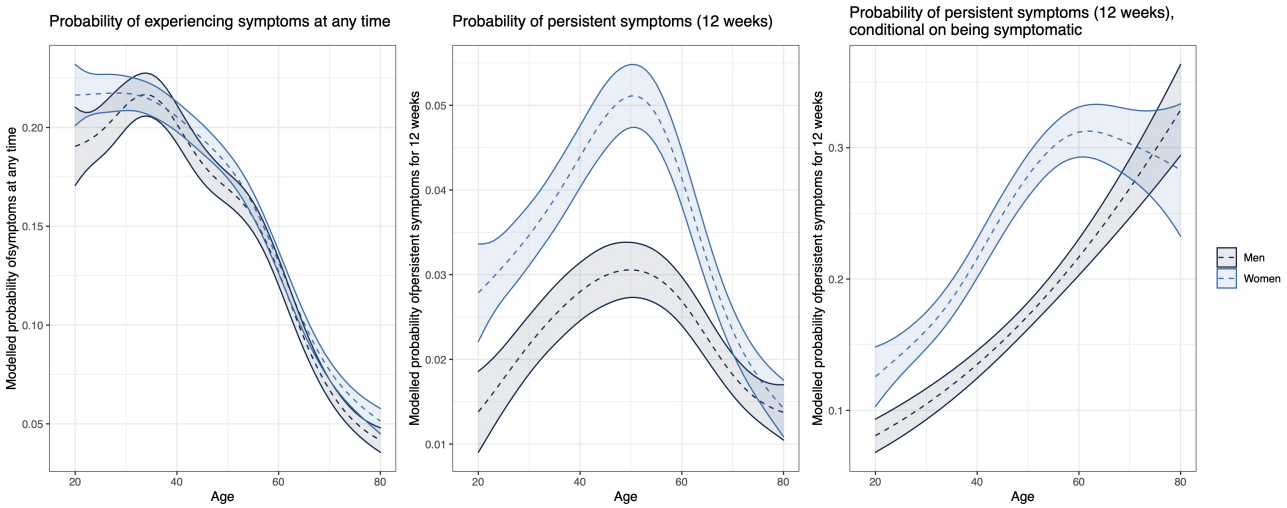
Figure 4.4: Mean contribution to area under the curve (AUC) that each variable made in a multivariable boosted tree model, derived by permuting each variable in turn (1000 times to obtain a distribution) and quantifying the change in model performance. Reproduced from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270]

Generalised additive modelling to investigate the likelihood of persistent symptoms in relation to age

In generalised additive models (figure 4.5), three trends were apparent. First, the probability of experiencing COVID-19 symptoms in the population followed a declining trend after approximately the age of 40 (left panels in figure 4.5). This trend is partly a function of the variation in prevalence of prior infection across ages in the population, with younger people more likely to have been infected in the first wave, and partly a function of the changing likelihood of experiencing asymptomatic COVID-19 across ages. Second, the probability of experiencing *persistent* COVID-19 symptoms in the general population follows an inverted V shape, with a peak at approximately aged 50 (central panels in figure 4.5). Finally, the probability of experiencing persistent symptoms, conditional on having had symptomatic COVID-19, increased linearly with age, with approximately an additional 3.5 percentage points of risk per decade of life (right panels in figure 4.5). Across all ages, women had approximately 8 percentage points higher risk of experiencing persistent symptoms, conditional on an initial symptomatic infection, than men. All these trends were highly consistent across the main study population (rounds 3–5) and the replication cohort (round 6).



(a) Rounds 3–5



(b) Round 6

Figure 4.5: Plots showing modelled probability of persistent symptoms at 12 weeks as a function of age and sex, using GAMs (generalised additive models) with splines on age and interactions between age and sex. From left to right, plots show (i) probability of experiencing COVID-19 symptoms in the population (n=508,707 observations), (ii) probability of experiencing persistent symptoms at 12 weeks in the population, and (iii) probability of experiencing persistent symptoms at 12 weeks, conditional on symptomatic infection (n=71,642 observations for whom we had 150 days' observation time). Conditional on symptomatic infection (likelihood of which decreases after age 50, as shown in the central panels), the risk of experiencing persistent symptoms increases linearly with age (right panels). Reproduced with modification from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270].

4.3.4 Identifying subtypes of persistent symptoms

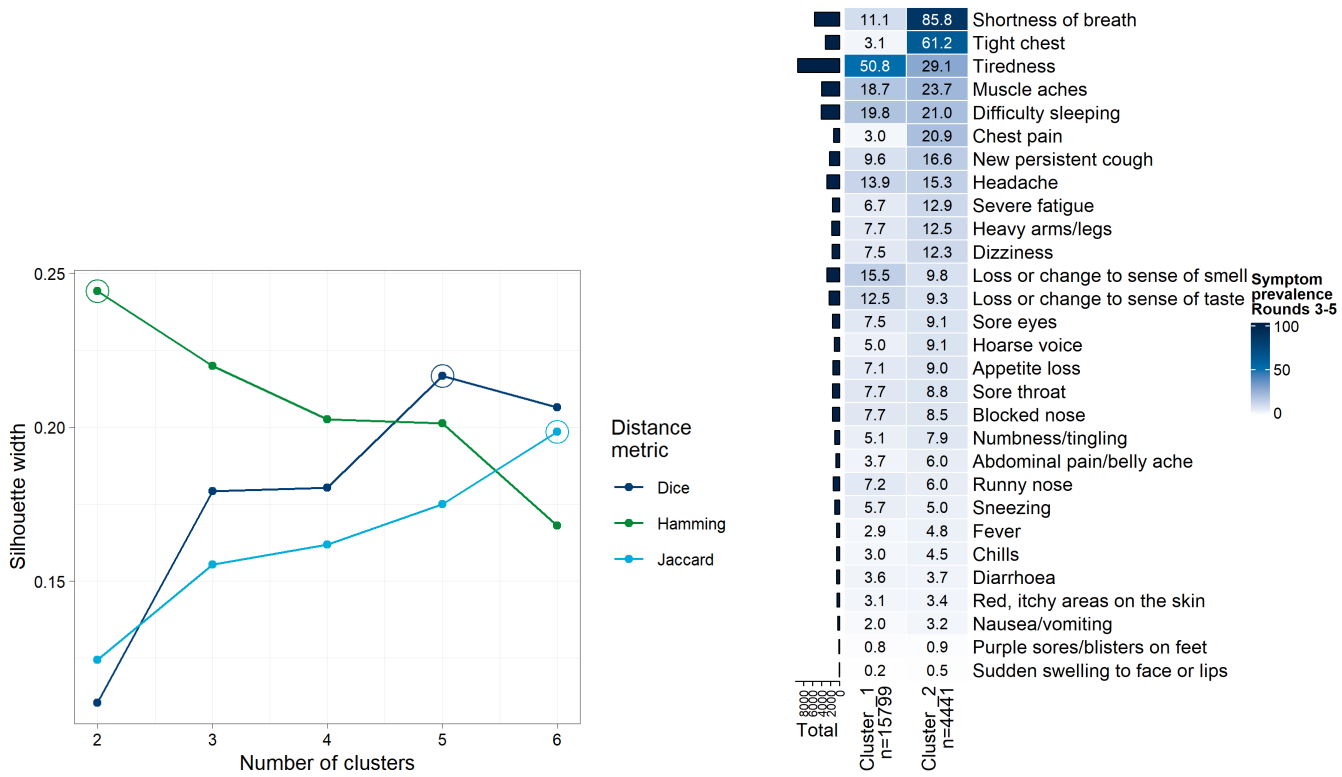
Cluster analysis

Cluster analysis was performed on 20,240 participants in rounds 3–5 who were still symptomatic 12 weeks after symptom onset, based on symptom profiles at 12 weeks.

Before performing the final clustering analysis, the number of clusters (k) needed to be chosen. The optimal distance metric was also determined at this stage: a range of different values of k were tested using three different distance metrics appropriate to binary data (Dice, Hamming and Jaccard). Using the silhouette score to determine the optimal specifications, 2 clusters was found to be optimal, using the Hamming distance 4.6. Bootstrapped stability analysis recovered the same clusters in 100% of bootstrap runs.

Cluster L1 (n=15,799) was characterised primarily by a high prevalence of persistent tiredness. Cluster L2 (n=4,441) was characterised by a high prevalence of respiratory symptoms including shortness of breath and tight chest (figure 4.6). The cluster medoids — the representative observations which form the nucleus of each cluster — were a participant with tiredness (only) at 12 weeks (Cluster L1) and a participant with tight chest and shortness of breath at 12 weeks (Cluster L2). A greater proportion of people in the respiratory cluster (Cluster L2) reported severe symptoms at the time of their COVID-19 illness (43.5% [42.0,44.9]) than in Cluster L1 (27.4% [26.7,28.1]). Nearly three times as many participants in Cluster L2 reported having been hospitalised with COVID-19 (2.9% [2.5–3.5]) as in Cluster L1 (1.1% [0.9,1.3]).

In replication in the round 6 data, the algorithm again identified a subset of respondents (Cluster R6L2, n=1,582) who reported high prevalence of shortness of breath, with chest pain and tight chest. In contrast to the main analysis, this cluster also had high prevalence of tiredness. The second cluster (Cluster R6L1, n=1,263) had high prevalence of loss of sense of smell and taste (Appendix C, figure C.1).



(a) Silhouette plot

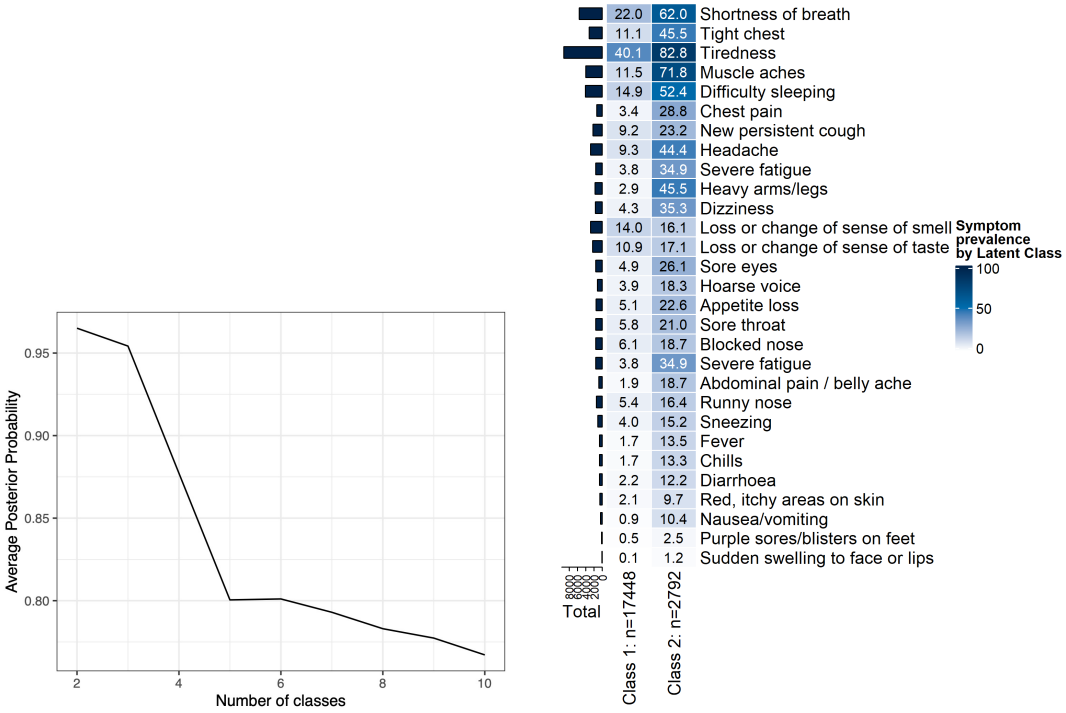
(b) Symptom prevalence by cluster

Figure 4.6: Results of clustering analysis using the CLARA PAM (Partitioning around medoids) algorithm. Using the silhouette score to identify the optimal number of clusters and distance metric, two clusters with Hamming distance was found to be optimal (panel 4.6a. Cluster L1 ("tiredness cluster") had high prevalence of tiredness. Cluster L2 ("respiratory cluster") was a smaller subset of 4,441 participants who had high prevalence of shortness of breath and tight chest as well as chest pain. Reproduced with modification from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270]

Latent Class Analysis

As with the clustering analysis, Latent Class Analysis (LCA) requires the choice of an optimal number of classes (k). This was done by testing a range of values of k between 2 and 10 and calculating the mean posterior probability (the probability of an observation being assigned to the correct class by the algorithm) across all observations. In this analysis a k of 2 was optimal, with the average posterior probability exceeding 0.95 (with 0.9 being a generally accepted cut-off for robust classes) (Figure 4.7).

As with the cluster analysis, LCA identified one smaller class characterised by respiratory symptoms and higher overall symptom prevalence, and one larger class characterised predominantly by tiredness (Figure 4.7).



(a) Mean posterior probability across a range of number of classes (k)

(b) Symptom prevalence by cluster

Figure 4.7: Results of Latent Class Analysis (LCA) on symptom profile at 12 weeks in rounds 3–5, showing (a) average posterior probability across a range of models with different numbers of classes, and (b) symptom prevalence across classes in the two-class model. The two-class model is shown to be optimal. Similar to the main clustering analysis, LCA identifies one larger group in which tiredness is the predominant symptom and one smaller group with multiple symptoms including high prevalence of respiratory symptoms. Reproduced with modification from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270]

Estimating background symptom prevalence

All COVID-19 cases and symptoms in this analysis are self-reported, introducing the possibility of false positives and overestimation of COVID-19 cases and COVID-19-related symptom prevalence. To investigate the background prevalence of persistent symptoms during the study period, REACT-1 data from the same study period was used. The REACT-1 study asked participants what symptoms they had experienced in the past 7 days, and how long before that each symptom had started, up to a maximum of 11+ days. Taking the proportion of PCR-negative people who said they were experiencing one or more symptoms that had started 11 or more days previously gives an upper bound estimate of the prevalence of non-COVID-related persistent symptoms in the community during this period.

The data from rounds 2 to 14 of REACT-1, which were collected between June 2020 and September 2021, showed that the average weighted prevalence of any of the 26 symptoms lasting 11 or more days was 3.06% (95% CI 3.04-3.09) among 1,879,842 PCR-negative adults (figure 4.8).

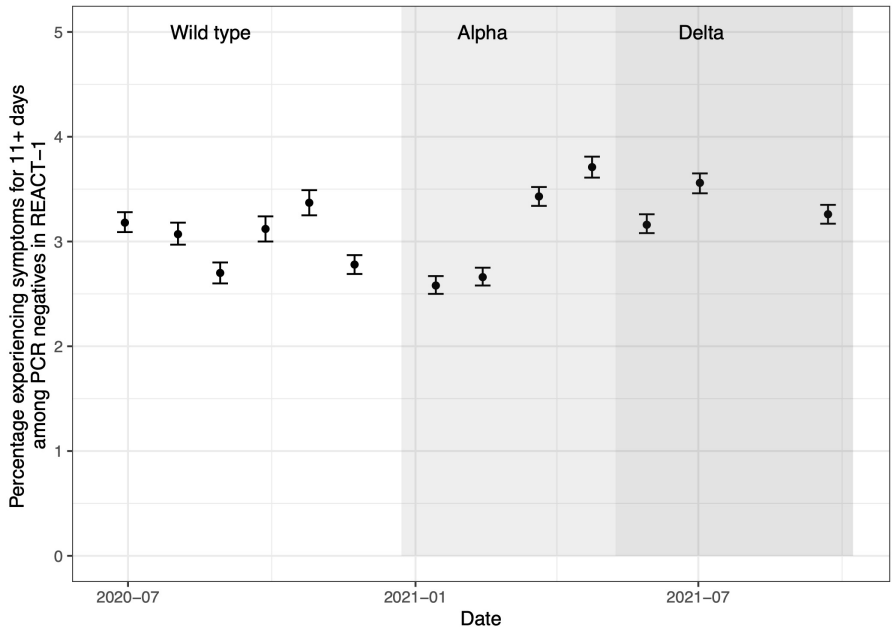


Figure 4.8: Prevalence of symptoms persisting for 11 or more days in 1,879,842 PCR-negative respondents to REACT-1. Data collected between July 2020 and September 2021 (rounds 2–14). Average weighted prevalence of 11+ days persistent symptoms was 3.06% (3.04,3.09) across rounds 2–14 of REACT-1. Error bars indicate 95% binomial confidence intervals of the prevalence. Reproduced from Whitaker *et al*, *Persistent COVID-19 symptoms in a community study of 606,434 people in England*, Nature Communications 2021 [270]

4.4 Discussion

In this chapter, representative samples of the population during autumn and winter 2020 and spring 2021 revealed high prevalence of persistent symptoms lasting 12 weeks or more in the English population. The prevalence estimates pointed to 5.8% of the adult population experiencing, or having experienced, one or more persistent symptoms post-COVID-19 during rounds 3–5 (this corresponds to more than 2 million adults in England), and 2.2% experiencing or having experienced three or more persistent symptoms (just under a million adults). In round 6, these figures were 3.1% and 1.6% respectively. This compares with 3.06% respondents reporting symptoms lasting 11 or more days among people testing negative for SARS-CoV-2.

The best source for external validation of these estimates is the Office for National Statistics data, which used repeated surveys and tests of a representative sample of households in the UK to estimate COVID-19 prevalence and symptom prevalence throughout the pandemic. The earliest period for which data on persistent symptoms is available is April 2021 [274], and they estimated that 1.69% (95% CI: 1.64,1.74) of the population was living with self-reported long COVID of any duration longer than 4 weeks at that time. Their estimates are lower than the estimates derived here for the corresponding time period (round 6) and the differences may arise from several factors: they survey children as well as adults, and prevalences were lower in children (0.18% [0.13,0.23] in 2–11 year olds, for instance, compared to 2.49% [2.36,2.62]) in 35–49 year olds); their data include the whole of the UK, not just England, and prevalences were lower in Scotland and Northern Ireland. There are also factors that should cause their estimates to be inflated relative to the estimates in this chapter: they did not restrict participants to reporting specific symptoms from a list, merely asked whether they were experiencing any persistent symptoms related to suspected COVID-19. They also used a threshold of 4 weeks rather than 12.

Ultimately the variation may be attributable to survey question framing. The REACT-2 surveys used in this chapter did not introduce the topic of long COVID: they asked what symptoms had been part of the respondent's suspected COVID-19, and how long they had lasted. By contrast, the ONS survey phrased the question "Would you describe yourself as having 'long COVID', that is, you are still experiencing symptoms more than 4 weeks after you first had COVID-19, that are not explained by something else?"[274]. It is possible that a person may

still be experiencing mild tiredness some weeks after COVID-19, for instance, but not classify themselves as having long COVID.

As with the population prevalence estimates, the estimates presented here of the likelihood of experiencing persistent symptoms conditional on a symptomatic infection are higher than in some studies, although other estimates have varied widely. At the lowest end, Sudre *et al* found that only 2.3% of people with COVID-19 still reported symptoms at 12 weeks [162]. Other studies estimated that 13.7% of people were symptomatic at 84 days [274], 14.8% at 90 days [167], 35% at 2 months [275], 34.7% at 7 months [276], 46% at 6 months [168], and as high as 51–52% at 6 months [195, 277]. A 2022 meta-analysis of 194 papers and 735,006 participants suggested that that 45% of COVID-19 survivors, including hospitalised and non-hospitalised individuals, were experiencing a range of unresolved symptoms at 4 months, and commented on the variability of prevalence estimates in the studies [278].

The results presented here indicate that 37.7% of individuals who contracted COVID-19 exhibited one or more symptoms at the 12-week mark during the autumn/winter of 2020-2021, and 21.6% in spring 2021. These estimates may be partially influenced by the extensive list of symptoms surveyed, many of which are common and not exclusive to COVID-19. However, the analysis of over 1.8 million PCR-negative respondents from the REACT-1 study revealed a background prevalence of persistent symptoms lasting 11 or more days to be approximately 3%. This serves as an upper bound for the prevalence of persistent symptoms not related to COVID-19. Therefore, the estimate of the prevalence of persistent symptoms associated with COVID-19 is roughly ten times greater than the background prevalence. This concurs with a study conducted by the Office for National Statistics, which investigated 26,922 UK residents between April and August 2021 and found the point prevalence of any of 12 symptoms to be 3.4% in non-COVID-19 positive individuals, with 0.5% reporting symptoms lasting 12 weeks or more [279].

The decrease in persistent symptoms between rounds 3–5 and the sixth round was primarily due to a substantial decline in persistent tiredness, falling from 16.8% to 8.0%. There are various explanations for this trend. The majority (60%) of historical infections reported during the sixth round occurred prior to July 2020, which may indicate that the decline in reported prevalence of persistent symptoms could be attributed to recall bias: ie individuals who experienced persistent symptoms in 2020 and then recovered were more likely to recall this when surveyed in late

2020 than they would be in sprint 2021. Although seasonal effects may influence symptom prevalence, the background symptom prevalence remained relatively constant throughout the study period. Studies have identified associations between lockdown measures and increased levels of tiredness[280] and stress[281, 282], and while the third to fifth rounds were conducted primarily during lockdown restrictions in the UK, the sixth round was conducted during the transition from the second step of reopening (when schools, retail, and outdoor hospitality were open) to 'step three' (when indoor venues were allowed to reopen and the 'rule of six' [3] was implemented). Finally, the survey structure was altered in the sixth round and the list of symptoms surveyed was modified. In the analysis of the round 6, prevalence estimates were based on the 27 out of 29 symptoms that were common to the earlier rounds, which may have slightly underestimated the symptom reporting prevalence in comparison to the earlier rounds.

The phenotypical characteristics identified as risk factors for persistent symptoms – increasing age, female sex, high BMI and having been hospitalised with COVID-19 – have previously been identified as risk factors in other research [162, 283, 284]. Other studies have also pointed to the same socio-economic factors identified here (measures of deprivation and household income) as risk factors for persistent symptoms[196, 197].

The finding of a smooth linear association between age and risk of persistent symptoms, conditional on symptomatic infection, has been replicated in other studies, but also contrasts with some other published research. A 2022 study by Subramanian *et al* [197], which used a similar study design focusing on a large (N=486,149) cohort of non-hospitalised adults and defining their outcome as one or more persistent symptom at 12 weeks, identified the same risk factors, including a protective effect of South Asian ethnicity in unadjusted models (although the effect went away after mutual adjustment). However, the risk of persistent symptoms across age groups followed a different trajectory, generally decreasing from age 30 to age 70, albeit with lower risk in the 18-19 age group and higher risk in the oldest age group (70+). After mutual adjustment, there was a clear declining trend in risk of persistent symptoms as age increased. By contrast, a meta-analysis by Thompson et al [196], which aggregated Electronic Health Care Records (ECHR) from ten UK-based studies, found a strikingly similar pattern to that reported here, both in risk of long COVID in the population (which followed an inverted U shape), and of risk of long COVID conditional on initial infection (which followed a smooth

[3]The 'rule of six' prohibited any social gathering of more than six people, except for in a set of limited circumstances

increasing linear trend) – see figure 1 in their paper.

Two distinct and stable symptom clusters were identified at 12 weeks post-symptom-onset in rounds 3-5. The results of additional analyses performed using different clustering methods revealed consistent patterns of symptoms in each cluster. This finding suggests that long COVID may have subgroups with specific symptom profiles. One of these subgroups, Cluster L2, was characterised by a high prevalence of respiratory symptoms including shortness of breath and chest tightness/pain. The replication data from round 6 also showed a high prevalence of these and related symptoms in Cluster L2R6. Other studies have used unsupervised approaches to identify subtypes of long COVID, though the time points at which the subtypes were characterised differed, as did symptom lists and clustering and partitioning approaches. For example, Sudre *et al* identified two symptom clusters at 28 days post-symptom-onset, which were different from the clusters identified in this chapter. Huang*et al* [186] characterised five symptom clusters at 61 days post-symptom-onset, two of which showed a high prevalence of respiratory symptoms similar to those seen in Cluster L2 (and Cluster L2R6) in this chapter. Subramanian *et al* [197] used Latent Class Analysis to partition their data on the basis of symptom profile at 12 weeks and found three clusters: one large cluster containing 80% of the participants and characterised by a range of symptoms including pain, fatigue and rash, one smaller cluster (14.2% of participants) characterised by neurological and psychological symptoms (depression, anxiety, insomnia and brain fog), and one cluster (5.8% of participants) characterised by cough, shortness of breath and phlegm. While the symptoms surveyed differ from those surveyed in REACT-2, and include a range of psychological symptoms that were not covered in REACT surveys in relation to COVID-19, the broad partitioning of a large group with low overall prevalence of symptoms and a smaller group with high prevalence of respiratory symptoms is consistent with the analysis presented here. Reese *et al* [187] found six clusters based on semantic quantification of electronic health records (EHRs) in patients with persistent COVID-19 symptoms, which could broadly be characterised as two multi-system clusters with high prevalence of symptoms across multiple systems (comprising 30% of study participants), and four single-system clusters characterised by symptoms from a specific biological system (neurological, pulmonary, neuropsychiatric and pain/fatigue). While the methods are very different to those presented here, the clustering results have some commonalities: the presence or absence of respiratory symptoms is a common differentiating characteristic of the clusters found, as is a higher proportion of people who had severe acute COVID (including having been hospitalised)

in the clusters characterised by respiratory symptoms and multi-system symptoms.

4.5 Conclusion

The results presented in this chapter show that there was a high burden of persistent symptoms in the population of England during winter 2020/2021, which had declined somewhat by spring 2021. The risk factors identified support findings from other research, and show that, as with the burden of acute disease and risk of hospitalisation and death, the burden of persistent COVID-19 symptoms disproportionately affected the economically disadvantaged.

The sub-types identified by clustering also support the emerging picture of long COVID/post-COVID syndrome as a constellation of overlapping syndromes affecting different biological systems. Further research is needed to fully characterise these sub-types and elucidate the biological mechanisms underpinning them.

Chapter 5

Conclusions

The research questions in this thesis were driven by real-time public health demands during the COVID-19 pandemic. Chapter two arose from a personal need to keep on top of COVID-19 literature and an impression, reinforced by some early published research, that the supply chain of research output had been disrupted by COVID-19 and that this merited investigation and reflection. The findings present a mixed picture: the response in terms of research output has been remarkable, although there is some suggestion that this has come at the expense of research into other areas (some analysis of author-specific output could test this conjecture more convincingly). Broadly, there was a high degree of alignment between what was driving the public conversations and what was driving research interest, although the direction of effect, which cannot be detected with the data available, will determine whether the effect is a result of the public being more interested in scientifically credible papers, or of academic effort being more strongly directed towards what is generating public attention than is normal. There is some evidence that the large spike of preprints that appeared in the first half of 2020 was of lower average quality than non-COVID-19 preprints, and there is more convincing evidence that preprints were generating huge public attention in a way that they were not intended to. As we enter the 'enquiry' phase of the COVID-19 era, this algorithmic interrogation of the research ecosystem provides some insight into the function of the research economy under the stress of a pandemic which may be useful in planning for future health crises.

Chapters three and four were developments of analyses conducted as part of the REACT-1 and REACT-2 programmes. Chapter three originated from the need to identify potential COVID-

19 cases in the community, which turned into the research paper published in 2021 [151]. This paper identified three symptoms (appetite loss, muscle aches and muscle aches) in addition to the 'classic' four (loss or change of sense of smell, loss or change of sense of taste, cough and fever) that were jointly positively predictive of SARS-CoV-2 positivity with the original Wild-type strain of the virus, and demonstrated that the same symptoms were similarly predictive of infection with the Alpha variant. This analysis formed the basis of the variant-stratified analysis presented in chapter three, using a similar variable selection approaches to identify changing patterns of symptomatology with different SARS-CoV-2 variants. These findings, and those from many other research groups, fed into the eventual expansion of the NHS 'trigger' list of symptoms [155] during the Omicron wave in 2022. The extent to which these changing symptom profiles reflect changes in the SARS-CoV-2 virus versus changes in the host immune response is still an open question and one that merits further investigation. Similarly, the direct comparison of the symptomatology and severity of the BA.2 vs BA.1 Omicron variants and the finding of increased symptom burden associated with the BA.2 variant remains salient, even though these variants have been superseded almost everywhere by XBB.1.5, BQ.1, CH.1.1, XBB and other Omicron sub-variants, as a counterweight to the perception that COVID-19 is on an inevitable progression towards lower virulence [285].

Chapter four represented an early (on the timescale of the COVID-19 pandemic) investigation of the epidemiology of post-COVID syndrome. The prevalence estimates were based on an immune-naive population and partly influenced by the effects of lockdown and the psychological disturbances associated with living through a pandemic, and therefore – fortunately – substantially higher than the current population estimates. However, the risk factors identified have been broadly replicated in other studies since the research was conducted, and the picture of post-COVID syndrome as an overlapping constellation of distinct conditions arising from different host systems is congruent with current thinking. Undoubtedly, significant work remains to be done to broaden our understanding of the condition. The cross-sectional analysis presented in chapter four has fed into the ongoing REACT-GE / REACT-LC cohort study [286], which aims to identify biological signatures of Post-COVID syndrome and of severe COVID-19 using untargeted OMICs analysis from a number of platforms. The goal is to identify biological pathways and mechanisms of persistent symptoms and severe disease, and eventually to develop treatments.

On a broader level the REACT-1 and REACT-2 programmes, of which the third and fourth chapter of this thesis represent just two outputs, have demonstrated the capacity for, and effectiveness of, large-scale data-gathering responses to a pandemic, even under lockdown conditions, and for close to real-time processing and sharing of the outcomes of the data to inform public policy and other research efforts. The REACT-1 programme was successful in its original research goal of providing reliable, near-real-time estimates of SARS-CoV-2 prevalence over time, and also generated multiple secondary insights into risk factors for transmission and infection, effectiveness of NPIs, behavioural responses to lockdowns and activity restrictions, vaccine effectiveness, and symptomatology and persistent symptoms of COVID-19. In the later rounds, when viral sequencing was conducted, the study was able to identify early spread of new variants and estimate transmission advantages and vaccine effectiveness against emergent variants [287]. The REACT-2 study, in addition to providing reliable, representative estimates of the prevalence of SARS-CoV-2 antibodies in 2020 and 2021, was able to demonstrate waning of antibodies, from both infection and vaccination, to make some early estimates of the prevalence of persistent symptoms, and to investigate the feasibility of using self-administered lateral flow tests for home testing for antibodies.

The REACT programmes ran in parallel to the ONS Coronavirus (COVID-19) Infection Survey, which had the same overarching research goal (surveillance of population infection trends in the UK), but different study design [288]. While the REACT programmes used non-overlapping cross-sectional samples of the population, the ONS study recruited a representative sample of households in the UK and repeatedly sampled the same members from these households at approximately monthly intervals. The differences in study design meant that the studies had different strengths. To take one specific example, the precision with which the REACT programme could investigate reinfection was limited by having to rely on individual recollection of prior infection, while the ONS study was able to accurately identify and investigate reinfections [289]. Conversely, elements of the REACT study design – at-home test administration, courier and postal collection of samples, and fully independent samples within each round – enabled surveillance to be conducted at a larger scale and lower per-sample cost, with a greater effective power than with a non-independent longitudinal study population using nurse visits to administer and collect test samples. Taken together, the two surveillance programmes represented a mutually complementary public health effort that effectively informed the response to the COVID-19 pandemic in England and the UK.

The REACT programme ended in March 2022, and the ONS Infection Survey in March 2023, as government policy moved towards conventional methods of monitoring disease spread and impact [290]. The large majority of the population now has some natural or vaccine-induced immunity to COVID-19, and successive waves of new Omicron sub-variants throughout late 2022 and early 2023 have not caused large spikes of hospitalisations and deaths. Nonetheless, the absence of any population surveillance programme of any scale inevitably increases the population's vulnerability to a new variant with substantially increased immune evasion or virulence, as the likelihood of early detection, and hence mitigation, is lower. Hopefully, the rapid planning and implementation of these surveillance projects in 2020 will function as templates for future large-scale rapid research efforts in response to infectious disease outbreaks.

Bibliography

1. World Health Organization. *Infection prevention and control during health care when novel coronavirus (nCoV) infection is suspected: interim guidance, 10 January 2020* tech. rep. WHO/2019-nCoV/IPC/2020.1 (2020).
2. Shrivastava, S. R. & Shrivastava, P. S. 2019-nCoV Outbreak Declared as Public Health Emergency of International Concern: What Next? *Int. J. Prev. Med.* **11,** 65 (June 2020).
3. *WHO Coronavirus (COVID-19) dashboard* `https://covid19.who.int/`. Accessed: 2022-10-26.
4. Msemburi, W. *et al.* The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature* **613,** 130–137 (Jan. 2023).
5. Richards, F. *et al.* Economic Burden of COVID-19: A Systematic Review. *Clinicoecon. Outcomes Res.* **14,** 293–307 (Apr. 2022).
6. Teng, B. *et al.* Economic recovery forecasts under impacts of COVID-19. *Econ. Model.* **110,** 105821 (May 2022).
7. Kaye, A. D. *et al.* Economic impact of COVID-19 pandemic on healthcare facilities and systems: International perspectives. *Best Pract. Res. Clin. Anaesthesiol.* **35,** 293–306 (Oct. 2021).
8. Blomberg, B., Cox, R. J. & Langeland, N. Long COVID: A growing problem in need of intervention. *Cron. Med.* **3** (Mar. 2022).
9. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382,** 727–733 (Feb. 2020).
10. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395,** 565–574 (Feb. 2020).

11. Shu, T. *et al.* Plasma Proteomics Identify Biomarkers and Pathogenesis of COVID-19. *Immunity* **53,** 1108–1122.e5 (Nov. 2020).

12. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (Mar. 2020).

13. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579,** 265–269 (Mar. 2020).

14. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181,** 271–280.e8 (Apr. 2020).

15. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395,** 1054–1062 (Mar. 2020).

16. Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* **323,** 1061–1069 (Mar. 2020).

17. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395,** 497–506 (Feb. 2020).

18. Tong, J. Y., Wong, A., Zhu, D., Fastenberg, J. H. & Tham, T. The Prevalence of Olfactory and Gustatory Dysfunction in COVID-19 Patients: A Systematic Review and Meta-analysis. *Otolaryngol. Head Neck Surg.* **163,** 3–11 (July 2020).

19. *Rhinology international journal* `https://www.rhinologyjournal.com/Abstract.php?id=2469`. Accessed: 2022-10-26.

20. Roberts, M. Coronavirus symptoms: UK adds loss of smell and taste to list. *BBC* (May 2020).

21. CDC. *Symptoms of COVID-19* `https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html`. Accessed: 2023-5-22. Dec. 2022.

22. Mizumoto, K., Kagaya, K., Zarebski, A. & Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro Surveill.* **25** (Mar. 2020).

23. Oran, D. P. & Topol, E. J. Prevalence of Asymptomatic SARS-CoV-2 Infection : A Narrative Review. *Ann. Intern. Med.* **173,** 362–367 (Sept. 2020).

24. Bai, Y. *et al.* Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA* **323,** 1406–1407 (Apr. 2020).

25. Kimball, A. *et al.* Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility - King County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69,** 377–381 (Apr. 2020).

26. Van Doremalen, N. *et al.* Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *N. Engl. J. Med.* **382,** 1564–1567 (Apr. 2020).

27. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26,** 672–675 (May 2020).

28. Zou, L. *et al.* SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients. *N. Engl. J. Med.* **382,** 1177–1179 (Mar. 2020).

29. Hamner, L. *et al.* High SARS-CoV-2 Attack Rate Following Exposure at a Choir Practice - Skagit County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69,** 606–610 (May 2020).

30. Lewis, D. Mounting evidence suggests coronavirus is airborne - but health advice has not caught up. *Nature* **583,** 510–513 (July 2020).

31. Haider, N. *et al.* Lockdown measures in response to COVID-19 in nine sub-Saharan African countries. *BMJ Glob Health* **5** (Oct. 2020).

32. *Timeline of UK government coronavirus lockdowns* `https://www.instituteforgovernment.org.uk/charts/uk-government-coronavirus-lockdowns`. Accessed: 2021-10-22. Apr. 2021.

33. *No More Lockdowns: China's New COVID Landscape* `https://publichealth.jhu.edu/2023/no-more-lockdowns-chinas-new-covid-landscape`. Accessed: 2023-4-6. Jan. 2023.

34. Mathieu, E. *et al.* Coronavirus Pandemic (COVID-19). *Our World in Data* (Mar. 2020).

35. Wang, M.-Y. *et al.* SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Front. Cell. Infect. Microbiol.* **10,** 587269 (Nov. 2020).

36. Sungnak, W. *et al.* SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* **26,** 681–687 (May 2020).

37. Meyerowitz, E. A., Richterman, A., Gandhi, R. T. & Sax, P. E. Transmission of SARS-CoV-2: A Review of Viral, Host, and Environmental Factors. *Ann. Intern. Med.* **174,** 69–79 (Jan. 2021).

38. Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581,** 465–469 (May 2020).

39. Harapan, B. N. & Yoo, H. J. Neurological symptoms, manifestations, and complications associated with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease 19 (COVID-19). *J. Neurol.* **268,** 3059–3071 (Sept. 2021).

40. Desforges, M. *et al.* Human Coronaviruses and Other Respiratory Viruses: Underestimated Opportunistic Pathogens of the Central Nervous System? *Viruses* **12** (Dec. 2019).

41. Xia, J., Tong, J., Liu, M., Shen, Y. & Guo, D. Evaluation of coronavirus in tears and conjunctival secretions of patients with SARS-CoV-2 infection. *J. Med. Virol.* **92,** 589–594 (June 2020).

42. Magor-Elliott, S. R. M. *et al.* Altered lung physiology in two cohorts after COVID-19 infection as assessed by computed cardiopulmonography. *J. Appl. Physiol.* **133,** 1175–1191 (Nov. 2022).

43. Khawaja, S. A. *et al.* COVID-19 and its impact on the cardiovascular system. *Open Heart* **8** (Mar. 2021).

44. Koralnik, I. J. & Tyler, K. L. COVID-19: A Global Threat to the Nervous System. *Ann. Neurol.* **88,** 1–11 (July 2020).

45. Hasan, L. K. *et al.* Effects of COVID-19 on the Musculoskeletal System: Clinician's Guide. *Orthop. Res. Rev.* **13,** 141–150 (Sept. 2021).

46. Gavriatopoulou, M. *et al.* Organ-specific manifestations of COVID-19 infection. *Clin. Exp. Med.* **20,** 493–506 (Nov. 2020).

47. Kaye, A. D. *et al.* COVID-19 impact on the renal system: Pathophysiology and clinical outcomes. *Best Pract. Res. Clin. Anaesthesiol.* **35,** 449–459 (Oct. 2021).

48. Elahi, S. Hematopoietic responses to SARS-CoV-2 infection. *Cell. Mol. Life Sci.* **79,** 187 (Mar. 2022).

49. Ramos-Casals, M., Brito-Zerón, P. & Mariette, X. Systemic and organ-specific immune-related manifestations of COVID-19. *Nat. Rev. Rheumatol.* **17,** 315–332 (June 2021).

50. Ye, Q., Wang, B. & Mao, J. The pathogenesis and treatment of the 'Cytokine Storm' in COVID-19. *J. Infect.* **80,** 607–613 (June 2020).

51. Fei, Y., Tang, N., Liu, H. & Cao, W. Coagulation Dysfunction. *Arch. Pathol. Lab. Med.* **144,** 1223–1229 (Oct. 2020).

52. Connors, J. M. & Levy, J. H. Thromboinflammation and the hypercoagulability of COVID-19. *J. Thromb. Haemost.* **18,** 1559–1561 (July 2020).

53. Pottegård, A. *et al.* Arterial events, venous thromboembolism, thrombocytopenia, and bleeding after vaccination with Oxford-AstraZeneca ChAdOx1-S in Denmark and Norway: population based cohort study. *BMJ* **373,** n1114 (May 2021).

54. Loo, J., Spittle, D. A. & Newnham, M. COVID-19, immunothrombosis and venous thromboembolism: biological mechanisms. *Thorax* **76,** 412–420 (Apr. 2021).

55. Ahmad, F., Kannan, M. & Ansari, A. W. Role of SARS-CoV-2 -induced cytokines and growth factors in coagulopathy and thromboembolism. *Cytokine Growth Factor Rev.* **63,** 58–68 (Feb. 2022).

56. Petrilli, C. M. *et al.* Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. *BMJ* **369,** m1966 (May 2020).

57. Zheng, Z. *et al.* Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *J. Infect.* **81,** e16–e25 (Aug. 2020).

58. Riley, S. *et al.* REal-time Assessment of Community Transmission (REACT) of SARS-CoV-2 virus: Study protocol. *Wellcome Open Res* **5,** 200 (2020).

59. Riley, S. *et al.* Resurgence of SARS-CoV-2: Detection by community viral surveillance. *Science* **372,** 990–995 (May 2021).

60. COVID-19 Forecasting Team. Variation in the COVID-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. *Lancet* **399,** 1469–1488 (Apr. 2022).

61. O'Driscoll, M. *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590,** 140–145 (Feb. 2021).

62. SeyedAlinaghi, S. *et al.* Genetic susceptibility of COVID-19: a systematic review of current evidence. *Eur. J. Med. Res.* **26,** 46 (May 2021).

63. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.* **23,** 3–20 (Jan. 2022).

64. RECOVERY Collaborative Group *et al.* Dexamethasone in Hospitalized Patients with Covid-19. *N. Engl. J. Med.* **384,** 693–704 (Feb. 2021).

65. Levin, M. J. *et al.* Intramuscular AZD7442 (Tixagevimab-Cilgavimab) for Prevention of Covid-19. *N. Engl. J. Med.* **386,** 2188–2200 (June 2022).

66. Weinreich, D. M. *et al.* REGN-COV2, a Neutralizing Antibody Cocktail, in Outpatients with Covid-19. *N. Engl. J. Med.* **384,** 238–251 (Jan. 2021).

67. Singh, A. K., Singh, A., Singh, R. & Misra, A. Molnupiravir in COVID-19: A systematic review of literature. *Diabetes Metab. Syndr.* **15,** 102329 (Oct. 2021).

68. Mahase, E. Covid-19: Pfizer's paxlovid is 89% effective in patients at risk of serious illness, company reports. *BMJ* **375,** n2713 (Nov. 2021).

69. *NIH clinical trial of remdesivir to treat COVID-19 begins* `https://www.nih.gov/news-events/news-releases/nih-clinical-trial-remdesivir-treat-covid-19-begins`. Accessed: 2023-4-6. Feb. 2020.

70. Beigel, J. H. *et al.* Remdesivir for the Treatment of Covid-19 - Final Report. *N. Engl. J. Med.* **383,** 1813–1826 (Nov. 2020).

71. WHO Solidarity Trial Consortium *et al.* Repurposed Antiviral Drugs for Covid-19 - Interim WHO Solidarity Trial Results. *N. Engl. J. Med.* **384,** 497–511 (Feb. 2021).

72. Thomson, K. & Nachlis, H. Emergency Use Authorizations During the COVID-19 Pandemic: Lessons From Hydroxychloroquine for Vaccine Authorization and Approval. *JAMA* **324,** 1282–1283 (Oct. 2020).

73. Boulware, D. R. *et al.* A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19. *N. Engl. J. Med.* **383,** 517–525 (Aug. 2020).

74. Schellack, N. *et al.* Social Media and COVID-19-Perceptions and Public Deceptions of Ivermectin, Colchicine and Hydroxychloroquine: Lessons for Future Pandemics. *Antibiotics (Basel)* **11** (Mar. 2022).

75. *Archived: WHO Timeline - COVID-19* `https://www.who.int/news/item/27-04-2020-who-timeline---covid-19`. Accessed: 2023-4-18.

76. Lurie, N., Saville, M., Hatchett, R. & Halton, J. Developing Covid-19 Vaccines at Pandemic Speed. *N. Engl. J. Med.* **382,** 1969–1973 (May 2020).

77. Tatsis, N. & Ertl, H. C. J. Adenoviruses as vaccine vectors. *Mol. Ther.* **10,** 616–629 (Oct. 2004).

78. *AstraZeneca and Oxford University announce landmark agreement for COVID-19 vaccine* `https://www.astrazeneca.com/media-centre/press-releases/2020/astrazeneca-and-oxford-university-announce-landmark-agreement-for-covid-19-vaccine.html`. Accessed: 2023-4-18. Apr. 2020.

79. Folegatti, P. M. *et al.* Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* **396,** 467–478 (Aug. 2020).

80. Sadoff, J. *et al.* Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. *N. Engl. J. Med.* **384,** 2187–2201 (June 2021).

81. Tanne, J. H. Covid-19: US authorises Johnson and Johnson vaccine again, ending pause in rollout. *BMJ* **373,** n1079 (Apr. 2021).

82. Wang, M., Jiang, S. & Wang, Y. Recent advances in the production of recombinant subunit vaccines in Pichia pastoris. *Bioengineered* **7,** 155–165 (Apr. 2016).

83. *COVID-19 vaccine tracker and landscape* `https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines`. Accessed: 2023-4-19.

84. Yang, J. *et al.* A vaccine targeting the RBD of the S protein of SARS-CoV-2 induces protective immunity. *Nature* **586,** 572–577 (Oct. 2020).

85. *Novavax: Nuvaxovid – COVID19 Vaccine Tracker* `https://covid19.trackvaccines.org/vaccines/25/`. Accessed: 2023-4-19.

86. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines - a new era in vaccinology. *Nat. Rev. Drug Discov.* **17,** 261–279 (Apr. 2018).

87. Schlake, T., Thess, A., Fotin-Mleczek, M. & Kallen, K.-J. Developing mRNA-vaccine technologies. *RNA Biol.* **9,** 1319–1330 (Nov. 2012).

88. Chen, Z. *et al.* Global diversity of policy, coverage, and demand of COVID-19 vaccines: a descriptive study. *BMC Med.* **20,** 130 (Apr. 2022).

89. Polack, F. P. *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N. Engl. J. Med.* **383,** 2603–2615 (Dec. 2020).

90. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* **384,** 403–416 (Feb. 2021).

91. Ledford, H., Cyranoski, D. & Noorden, V. *The UK has approved a COVID vaccine — here's what scientists now want to know* http://dx.doi.org/10.1038/d41586-020-03441-8. Accessed: 2023-4-19. Dec. 2020.

92. The-rollout-of-the-COVID-19-vaccination-programme-in-England.pdf.

93. Majeed, A., Pollock, K., Hodes, S. & Papaluca, M. Implementation of covid-19 vaccination in the United Kingdom. *BMJ* **378,** e070344 (Sept. 2022).

94. *AstraZeneca's COVID-19 vaccine authorised for emergency supply in the UK* https://www.astrazeneca.com/media-centre/press-releases/2020/astrazenecas-covid-19-vaccine-authorised-in-uk.html. Accessed: 2023-4-19. Dec. 2020.

95. *Regulatory approval of Spikevax (formerly COVID-19 Vaccine Moderna)* https://www.gov.uk/government/publications/regulatory-approval-of-covid-19-vaccine-moderna. Accessed: 2023-4-19. Jan. 2021.

96. Border, P. *Changing the UK COVID-19 vaccine dosing schedule* https://post.parliament.uk/changing-the-uk-covid-19-vaccine-dosing-schedule/. Accessed: 2023-4-19.

97. *JCVI statement on COVID-19 vaccination of children and young people: 22 December 2021* https://www.gov.uk/government/publications/jcvi-update-on-advice-for-covid-19-vaccination-of-children-and-young-people/jcvi-statement-on-covid-19-vaccination-of-children-and-young-people-22-december-2021. Accessed: 2022-10-26.

98. *JCVI statement on vaccination of children aged 5 to 11 years old* https://www.gov.uk/government/publications/jcvi-update-on-advice-for-covid-19-vaccination-of-children-aged-5-to-11/jcvi-statement-on-vaccination-of-children-aged-5-to-11-years-old. Accessed: 2023-4-20.

99. Mahase, E. Covid-19: UK will roll out Moderna's omicron BA.1 vaccine as part of autumn booster programme. *BMJ* **378,** o2038 (Aug. 2022).

100. Surie, D. *et al.* Early Estimates of Bivalent mRNA Vaccine Effectiveness in Preventing COVID-19-Associated Hospitalization Among Immunocompetent Adults Aged ≥65 Years - IVY Network, 18 States, September 8-November 30, 2022. *MMWR Morb. Mortal. Wkly. Rep.* **71,** 1625–1630 (Dec. 2022).

101. Wang, Q. *et al.* Antibody Response to Omicron BA.4-BA.5 Bivalent Booster. *N. Engl. J. Med.* **388,** 567–569 (Feb. 2023).

102. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372** (Apr. 2021).

103. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599,** 114–119 (Nov. 2021).

104. Elliott, P. *et al.* Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science* **374,** eabl9551 (Dec. 2021).

105. *GISAID - hCov19 variants* `https://www.gisaid.org/hcov19-variants/`. Accessed: 2022-5-23.

106. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603,** 679–686 (Mar. 2022).

107. Shema, H. & Shema, H. *The birth of modern peer review* `https://blogs.scientificamerican.com/information-culture/the-birth-of-modern-peer-review/`. Accessed: 2023-4-28.

108. Likhachev, M. & Stentz, A. R* search (2008).

109. Björk, B.-C. & Solomon, D. The publishing delay in scholarly peer-reviewed journals. *J. Informetr.* **7,** 914–923 (Oct. 2013).

110. Horbach, S. P. J. M. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quant. Sci. Stud.* **1,** 1056–1067 (Aug. 2020).

111. Squazzoni, F. *et al.* Gender gap in journal submissions and peer review during the first wave of the COVID-19 pandemic. A study on 2329 Elsevier journals. *PLoS One* **16,** e0257919 (Oct. 2021).

112. Horbach, S. P. J. M. No time for that now! Qualitative changes in manuscript peer review during the Covid-19 pandemic. *Res. Eval.* **30,** 231–239 (Jan. 2021).

113. Clark, J. How covid-19 bolstered an already perverse publishing system. *BMJ* **380,** 689 (Mar. 2023).

114. Schonhaut, L. *et al.* Scientific publication speed and retractions of COVID-19 pandemic original articles. *Rev. Panam. Salud Publica* **46,** e25 (Apr. 2022).

115. Burke, L. *Pandemic brought new attention to preprints* `https://www.insidehighered.com/news/2021/01/20/pandemic-brought-new-attention-preprints`. Accessed: 2022-11-23.

116. Else, H. How a torrent of COVID science changed research publishing - in seven charts. *Nature* **588,** 553 (Dec. 2020).

117. Callaway, E. Will the pandemic permanently alter scientific publishing? *Nature* **582,** 167–168 (June 2020).

118. Fraser, N. *et al.* The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biol.* **19,** e3000959 (Apr. 2021).

119. Ravinetto, R. *et al.* Preprints in times of COVID19: the time is ripe for agreeing on terminology and good practices. *BMC Med. Ethics* **22,** 106 (July 2021).

120. Yeo-Teh, N. S. L. & Tang, B. L. An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19). *Account. Res.* **28,** 47–53 (Jan. 2021).

121. Soltani, P. & Patini, R. Retracted COVID-19 articles: a side-effect of the hot race to publication. *Scientometrics* **125,** 819–822 (Aug. 2020).

122. *Infodemic* `https://www.who.int/health-topics/infodemic`. Accessed: 2022-11-21.

123. Pulido, C. M., Villarejo-Carballido, B., Redondo-Sama, G. & Gómez, A. COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. *Int. Sociol.* **35,** 377–392 (July 2020).

124. *Altmetric* `https://www.altmetric.com/`. Accessed: 2022-11-21.

125. Shamsi, A., Lund, B. D. & SeyyedHosseini, S. Sharing of retracted COVID-19 articles: an altmetric study. *J. Med. Libr. Assoc.* **110,** 97–102 (Jan. 2022).

126. Rosenberg, H., Syed, S. & Rezaie, S. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *CJEM* **22,** 418–421 (July 2020).

127. Rathore, F. A. & Farooq, F. Information Overload and Infodemic in the COVID-19 Pandemic. *J. Pak. Med. Assoc.* **70(Suppl 3),** S162–S165 (May 2020).

128. Grant, M. C. *et al.* The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries. *PLoS One* **15,** e0234765 (June 2020).

129. Guan, W.-J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **382,** 1708–1720 (Apr. 2020).

130. Grasselli, G. *et al.* Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA* **323,** 1574–1581 (Apr. 2020).

131. Lechien, J. R. *et al.* Clinical and epidemiological characteristics of 1420 European patients with mild-to-moderate coronavirus disease 2019. *J. Intern. Med.* **288,** 335–344 (Sept. 2020).

132. *Coronavirus (COVID-19) Infection Survey: methods and further information* `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseas methodologies/covid19infectionsurveypilotmethodsandfurtherinformation`. Accessed: 2023-6-14.

133. *Protocol and information sheets* `https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/protocol-and-information-sheets`. Accessed: 2023-6-14.

134. Steel, K. & Davies, B. *Coronavirus (COVID-19) Infection Survey May 2021 - Office for National Statistics* `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsinth latest`. Accessed: 2023-6-14. May 2021.

135. Koutsakos, M., Wheatley, A. K., Laurie, K., Kent, S. J. & Rockman, S. Influenza lineage extinction during the COVID-19 pandemic? *Nat. Rev. Microbiol.* **19,** 741–742 (Dec. 2021).

136. Islam, M. A. *et al.* Prevalence and characteristics of fever in adult and paediatric patients with coronavirus disease 2019 (COVID-19): A systematic review and meta-analysis of 17515 patients. *PLoS One* **16,** e0249788 (Apr. 2021).

137. Yamato, M. & Kataoka, Y. Fatigue sensation following peripheral viral infection is triggered by neuroinflammation: who will answer these questions? *Neural Regeneration Res.* **10,** 203–204 (Feb. 2015).

138. Brann, D. H. *et al.* Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. *Sci Adv* **6** (July 2020).

139. Butowt, R. & von Bartheld, C. S. Anosmia in COVID-19: Underlying Mechanisms and Assessment of an Olfactory Route to Brain Infection. *Neuroscientist* **27,** 582–603 (Dec. 2021).

140. Agyeman, A. A., Chin, K. L., Landersdorfer, C. B., Liew, D. & Ofori-Asenso, R. Smell and Taste Dysfunction in Patients With COVID-19: A Systematic Review and Meta-analysis. *Mayo Clin. Proc.* **95,** 1621–1631 (Aug. 2020).

141. Favas, T. T. *et al.* Neurological manifestations of COVID-19: a systematic review and meta-analysis of proportions. *Neurol. Sci.* **41,** 3437–3470 (Dec. 2020).

142. Caronna, E. *et al.* Headache: A striking prodromal and persistent symptom, predictive of COVID-19 clinical evolution. *Cephalalgia* **40,** 1410–1421 (Nov. 2020).

143. Groff, A. *et al.* Gastrointestinal Manifestations of COVID-19: A Review of What We Know. *Ochsner J.* **21,** 177–180 (2021).

144. Suresh Kumar, V. C. *et al.* Novelty in the gut: a systematic review and meta-analysis of the gastrointestinal manifestations of COVID-19. *BMJ Open Gastroenterol* **7** (May 2020).

145. Iacobucci, G. Covid-19: Runny nose, headache, and fatigue are commonest symptoms of omicron, early data show. *BMJ* **375,** n3103 (Dec. 2021).

146. Antonelli, M. *et al.* Risk factors and disease profile of post-vaccination SARS-CoV-2 infection in UK users of the COVID Symptom Study app: a prospective, community-based, nested, case-control study. *Lancet Infect. Dis.* **22,** 43–55 (Jan. 2022).

147. Department of Health & Care, S. *Further expansion of access to coronavirus testing helps protect the most vulnerable* `https://www.gov.uk/government/news/further-expansion-of-access-to-coronavirus-testing-helps-protect-the-most-vulnerable`. Accessed: 2022-11-2. Apr. 2020.

148. *Allocation of COVID-19 swab tests in England* https://www.gov.uk/government/publications/allocation-of-covid-19-swab-tests-in-england/allocation-of-covid-19-swab-tests-in-england. Accessed: 2023-5-3.

149. Menni, C. *et al. Loss of smell and taste in combination with other symptoms is a strong predictor of COVID-19 infection* Apr. 2020.

150. Antonelli, M. *et al.* Optimal symptom combinations to aid COVID-19 case identification: Analysis from a community-based, prospective, observational cohort. *J. Infect.* **82,** 384–390 (Mar. 2021).

151. Elliott, J. *et al.* Predictive symptoms for COVID-19 in the community: REACT-1 study of over 1 million people. *PLoS Med.* **18,** e1003777 (Sept. 2021).

152. Zens, M., Brammertz, A., Herpich, J., Südkamp, N. & Hinterseer, M. App-Based Tracking of Self-Reported COVID-19 Symptoms: Analysis of Questionnaire Data. *J. Med. Internet Res.* **22,** e21956 (Sept. 2020).

153. Graham, M. S. *et al.* Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *Lancet Public Health* **6,** e335–e345 (May 2021).

154. Mahase, E. Covid-19: Sore throat, fatigue, and myalgia are more common with new UK variant. *BMJ* **372,** n288 (Jan. 2021).

155. *COVID-19 symptoms and what to do* https://www.nhs.uk/conditions/covid-19/covid-19-symptoms-and-what-to-do/. Accessed: 2023-5-19.

156. Huang, C. *et al.* 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet* **397,** 220–232 (Jan. 2021).

157. Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nat. Med.* **27,** 626–631 (Apr. 2021).

158. Venkatesan, P. NICE guideline on long COVID. *Lancet Respir Med* **9,** 129 (Feb. 2021).

159. Sisó-Almirall, A. *et al.* Long Covid-19: Proposed Primary Care Clinical Guidelines for Diagnosis and Disease Management. *Int. J. Environ. Res. Public Health* **18** (Apr. 2021).

160. Davis, H. E. *et al.* Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine* **38,** 101019 (Aug. 2021).

161. Ayoubkhani, D. & Bosworth, M. *Self-reported long COVID after infection with the Omicron variant in the UK - Office for National Statistics* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/selfreportedlongcovidafterinfectionwiththeomicronvariant/6may2022. Accessed: 2023-1-23. May 2022.

162. Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nat. Med.* **27,** 626–631 (Apr. 2021).

163. HHS. *What is Long COVID?* https://www.covid.gov/longcovid/definitions. Accessed: 2023-1-23.

164. Yong, S. J. Long COVID or post-COVID-19 syndrome: putative pathophysiology, risk factors, and treatments. *Infect. Dis.* **53,** 737–754 (Oct. 2021).

165. Kara Steel And. *Coronavirus (COVID-19) Infection Survey* Dec. 2020.

166. Ayoubkhani, D. *Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK - Office for National Statistics* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/1april2021. Accessed: 2022-10-26. Mar. 2021.

167. Cirulli, E. T. *et al. Long-term COVID-19 symptoms in a large unselected population* Oct. 2020.

168. Klein, H. *et al.* Onset, duration and unresolved symptoms, including smell and taste changes, in mild COVID-19 infection: a cohort study in Israeli patients. *Clin. Microbiol. Infect.* (Feb. 2021).

169. Florencio, L. L. & Fernández-de-Las-Peñas, C. Long COVID: systemic inflammation and obesity as therapeutic targets. *Lancet Respir Med* **10,** 726–727 (Aug. 2022).

170. Kovarik, J. J. *et al.* A multi-omics based anti-inflammatory immune signature characterizes long COVID-19 syndrome. *iScience* **26,** 105717 (Jan. 2023).

171. Bastard, P. *et al.* Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370** (Oct. 2020).

172. Liew, F., Efstathiou, C. & Openshaw, P. J. Long Covid: clues about causes. *Eur. Respir. J.* (Mar. 2023).

173. Wong, T. L. & Weitzer, D. J. Long COVID and Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)-A Systemic Review and Comparison of Clinical Presentation and Symptomatology. *Medicina* **57** (Apr. 2021).

174. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21,** 133–146 (Mar. 2023).

175. Fukuda, K. *et al.* The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group. *Ann. Intern. Med.* **121,** 953–959 (Dec. 1994).

176. O'Donnell, J. S. & Chappell, K. J. Chronic SARS-CoV-2, a Cause of Post-acute COVID-19 Sequelae (Long-COVID)? *Front. Microbiol.* **12,** 724654 (Aug. 2021).

177. Gaebler, C. *et al.* Evolution of antibody immunity to SARS-CoV-2. *Nature* **591,** 639–644 (Mar. 2021).

178. Ayoubkhani, D. *et al.* Trajectory of long covid symptoms after covid-19 vaccination: community based cohort study. *BMJ* **377,** e069676 (May 2022).

179. Yong, S. J. Persistent Brainstem Dysfunction in Long-COVID: A Hypothesis. *ACS Chem. Neurosci.* **12,** 573–580 (Feb. 2021).

180. Barnden, L. R. *et al.* Intra brainstem connectivity is impaired in chronic fatigue syndrome. *Neuroimage Clin* **24,** 102045 (Oct. 2019).

181. Barnden, L. R. *et al.* Hyperintense sensorimotor T1 spin echo MRI is associated with brainstem abnormality in chronic fatigue syndrome. *Neuroimage Clin* **20,** 102–109 (July 2018).

182. Butterworth, P., Schurer, S., Trinh, T.-A., Vera-Toscano, E. & Wooden, M. Effect of lockdown on mental health in Australia: evidence from a natural experiment analysing a longitudinal probability sample survey. *Lancet Public Health* **7,** e427–e436 (May 2022).

183. Labrague, L. J. & Ballad, C. A. Lockdown fatigue among college students during the COVID-19 pandemic: Predictive role of personal resilience, coping behaviors, and health. *Perspect. Psychiatr. Care* **57,** 1905–1912 (Oct. 2021).

184. Porter, C. *et al.* Impact of the COVID-19 pandemic on anxiety and depression symptoms of young people in the global south: evidence from a four-country cohort study. *BMJ Open* **11,** e049653 (Apr. 2021).

185. Sudre, C. H. *et al.* Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app. *Sci Adv* **7** (Mar. 2021).

186. Huang, Y. *et al.* COVID Symptoms, Symptom Clusters, and Predictors for Becoming a Long-Hauler Looking for Clarity in the Haze of the Pandemic. *Clinical Nursing Research* (Sept. 2022).

187. Reese, J. T. *et al.* Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. *EBioMedicine* **87,** 104413 (Jan. 2023).

188. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28,** 100–108 (1979).

189. Zhang, H. *et al.* Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat. Med.* **29,** 226–235 (Jan. 2023).

190. Kaushal, R. *et al.* Changing the research landscape: the New York City Clinical Data Research Network. *J. Am. Med. Inform. Assoc.* **21,** 587–590 (May 2014).

191. Shenkman, E. *et al.* OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute With a Community-Based Distributive Medical Education Model. *Acad. Med.* **93,** 451–455 (Mar. 2018).

192. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classification* **31,** 274–295 (Oct. 2014).

193. Arnold, D. T. *et al.* Patient outcomes after hospitalisation with COVID-19 and implications for follow-up: results from a prospective UK cohort. *Thorax* **76,** 399–401 (Apr. 2021).

194. Carfì, A., Bernabei, R., Landi, F. & Gemelli Against COVID-19 Post-Acute Care Study Group. Persistent Symptoms in Patients After Acute COVID-19. *JAMA* **324,** 603–605 (Aug. 2020).

195. Evans, R. A. *et al.* Physical, cognitive, and mental health impacts of COVID-19 after hospitalisation (PHOSP-COVID): a UK multicentre, prospective cohort study. *Lancet Respir Med* **9,** 1275–1287 (Nov. 2021).

196. Thompson, E. J. *et al.* Long COVID burden and risk factors in 10 UK longitudinal studies and electronic health records. *Nat. Commun.* **13,** 3528 (June 2022).

197. Subramanian, A. *et al.* Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat. Med.* **28,** 1706–1714 (Aug. 2022).

198. Byambasuren, O., Stehlik, P., Clark, J., Alcorn, K. & Glasziou, P. *Impact of COVID-19 vaccination on long COVID: a systematic review and meta-analysis* June 2022.

199. Tran, V.-T., Perrodeau, E., Saldanha, J., Pane, I. & Ravaud, P. *Efficacy of COVID-19 Vaccination on the Symptoms of Patients With Long COVID: A Target Trial Emulation Using Data From the ComPaRe e-Cohort in France* Sept. 2021.

200. Fernández-de-Las-Peñas, C. *et al.* Long-COVID Symptoms in Individuals Infected with Different SARS-CoV-2 Variants of Concern: A Systematic Review of the Literature. *Viruses* **14** (Nov. 2022).

201. *Real-time assessment of community transmission (REACT) study* `https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/`. Accessed: 2022-10-26.

202. REACT Study Investigators *et al. Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study* July 2020.

203. *Real-time Assessment of Community Transmission findings* `https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/real-time-assessment-of-community-transmission-findings/`. Accessed: 2023-4-23.

204. UK Government. *COVID-19 Data Dashboard* `https://coronavirus.data.gov.uk/details/healthcare`. Accessed: 2021-11-17.

205. Whitaker, M. *OverReact*

206. *Brand colours* `https://www.imperial.ac.uk/brand-style-guide/visual-identity/brand-colours/`. Accessed: 2023-5-10.

207. *sharp: Stability-enHanced Approaches using Resampling Procedures* `https://cran.r-project.org/web/packages/sharp/index.html`. Accessed: 2022-11-10.

208. *PLXNA4 Gene - GeneCards — PLXA4 Protein — PLXA4 Antibody* `https://www.genecards.org/cgi-bin/carddisp.pl?gene=PLXNA4`.

209. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29,** 1189–1232 (2001).

210. Ibrahim, A. A., Raheem, Muhammed, Rabiat & Ganiyu. Comparison of the CatBoost classifier with other machine learning methods. *Int. J. Adv. Comput. Sci. Appl.* **11** (2020).

211. Hancock, J. T. & Khoshgoftaar, T. M. CatBoost for big data: an interdisciplinary review. *J Big Data* **7,** 94 (Nov. 2020).

212. Chen, T. & Guestrin, C. *XGBoost* 2016.

213. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30** (2017).

214. Bentéjac, C., Csörgő, A. & Martınez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* **54,** 1937–1967 (Mar. 2021).

215. Lundberg, S. M. & Lee, S.-I. Consistent feature attribution for tree ensembles. arXiv: `1706.06060 [cs.AI]` (June 2017).

216. Whitaker, M. *stabilityshap*

217. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **36,** 1–13 (Sept. 2010).

218. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26,** 1340–1347 (May 2010).

219. Hook, D. W., Porter, S. J. & Herzog, C. Dimensions: Building context for search and evaluation. *Front. Res. Metr. Anal.* **3** (Aug. 2018).

220. Ooms, J. Google's Compact Language Detector 2 [R package cld2 version 1.2.4] (Oct. 2022).

221. Dimensions Resources. *Dimensions COVID-19 publications, datasets and clinical trials* Mar. 2020.

222. Roychowdhury, K., Bhanja, R. & Biswas, S. Mapping the research landscape of Covid-19 from social sciences perspective: a bibliometric analysis. *Scientometrics* **127,** 4547–4568 (July 2022).

223. Hook, D. W., Porter, S. J., Draux, H. & Herzog, C. T. Real-Time Bibliometrics: Dimensions as a Resource for Analyzing Aspects of COVID-19. *Front Res Metr Anal* **5,** 595299 (2020).

224. Mondal, S., De, N. & Pal, A. On neighborhood Zagreb index of product graphs. *J. Mol. Struct.* **1223,** 129210 (Jan. 2021).

225. Lucas, K. L. & Roberts, S. A. Undergraduates' Sense of Belonging, Community, and Support at a Title V Center. *J. Stud. Aff. Res. Pract.* **59,** 529–543 (Oct. 2022).

226. Seseña, S., Rodrıguez, A. M. & Palop, M. L. Indoor air quality analysis in naturally ventilated university training laboratories: a health risk assessment. *Air Qual. Atmos. Health* **15,** 1817–1837 (Oct. 2022).

227. Cabanac, G., Oikonomidi, T. & Boutron, I. Day-to-day discovery of preprint-publication links. *Scientometrics* **126,** 5285–5304 (Apr. 2021).

228. Asmussen, C. B. & Møller, C. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data* **6,** 1–18 (Oct. 2019).

229. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. arXiv: `1706.09516 [cs.LG]` (June 2017).

230. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. arXiv: `1705.07874 [cs.AI]` (May 2017).

231. Larsen, P. O. & von Ins, M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* **84,** 575–603 (Sept. 2010).

232. Hutchins, B. I., Yuan, X., Anderson, J. M. & Santangelo, G. M. Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLoS Biol.* **14,** e1002541 (Sept. 2016).

233. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395,** 507–513 (Feb. 2020).

234. Fang, Z. & Costas, R. Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics* **123,** 1077–1101 (May 2020).

235. Bundgaard, H. *et al.* Effectiveness of adding a mask recommendation to other public health measures to prevent SARS-CoV-2 infection in danish mask wearers : A randomized controlled trial. *Ann. Intern. Med.* **174,** 335–343 (Mar. 2021).

236. Rosenbaum, L. Lead Researcher Behind Controversial Danish Study Says You Should Still Wear A Mask. *Forbes Magazine* (Nov. 2020).

237. Ioannidis, J. P. A., Salholz-Hillel, M., Boyack, K. W. & Baas, J. The rapid, massive growth of COVID-19 authors in the scientific literature. *R Soc Open Sci* **8,** 210389 (Sept. 2021).

238. Myers, K. R. *et al.* Unequal effects of the COVID-19 pandemic on scientists. *Nat Hum Behav* **4,** 880–883 (Sept. 2020).

239. Gao, J., Yin, Y., Myers, K. R., Lakhani, K. R. & Wang, D. Potentially long-lasting effects of the pandemic on scientists. *Nat. Commun.* **12,** 6188 (Oct. 2021).

240. Bagdasarian, N., Cross, G. B. & Fisher, D. Rapid publications risk the integrity of science in the era of COVID-19. *BMC Med.* **18,** 192 (June 2020).

241. Eysenbach, G. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *J. Med. Internet Res.* **13,** e123 (Dec. 2011).

242. Sathianathen, N. J. *et al.* Social Media Coverage of Scientific Articles Immediately After Publication Predicts Subsequent Citations - #SoME_Impact Score: Observational Analysis. *J. Med. Internet Res.* **22,** e12288 (Apr. 2020).

243. Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C. & Zhou, H. Early indicators of scientific impact: Predicting citations with altmetrics. *J. Informetr.* **15,** 101128 (May 2021).

244. Bornmann, L. & Haunschild, R. Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PLoS One* **13,** e0197133 (May 2018).

245. Van Schalkwyk, M. C. I., Hird, T. R., Maani, N., Petticrew, M. & Gilmore, A. B. The perils of preprints. *BMJ* **370,** m3111 (Aug. 2020).

246. Xu, F., Ou, G., Ma, T. & Wang, X. The consistency of impact of preprints and their journal publications. *J. Informetr.* **15,** 101153 (May 2021).

247. Whitaker, M. *et al.* Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England. *Nat. Commun.* **13,** 6856 (Nov. 2022).

248. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22** (Mar. 2017).

249. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **72,** 417–473 (July 2010).

250. Bodinier, B., Filippi, S., Nost, T. H., Chiquet, J. & Chadeau-Hyam, M. Automated calibration for stability selection in penalised regression and graphical models: a multi-OMICs network application exploring the molecular response to tobacco smoking. arXiv: 2106.02521 [stat.ME] (June 2021).

251. Wolter, N. *et al. Early assessment of the clinical severity of the SARS-CoV-2 Omicron variant in South Africa* Dec. 2021.

252. Geddes, L. Will Covid-19 become less dangerous as it evolves? *The Guardian* (Jan. 2022).

253. Doucleff, M. Fact check: The theory that SARS-CoV-2 is becoming milder. *NPR* (Jan. 2022).

254. Gaythorpe, K. *et al. Report 8: Symptom progression of COVID-19* tech. rep. (2020).

255. Walker, A. S. *et al.* Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time. *Elife* **10,** e64683 (July 2021).

256. Mishra, B. *et al.* High proportion of low cycle threshold value as an early indicator of COVID-19 surge. *J. Med. Virol.* **94,** 240–245 (Jan. 2022).

257. Marc, A. *et al.* Quantifying the relationship between SARS-CoV-2 viral load and infectiousness. *Elife* **10** (Sept. 2021).

258. Kawasuji, H. *et al.* Transmissibility of COVID-19 depends on the viral load around onset in adult and symptomatic patients. *PLoS One* **15,** e0243597 (Dec. 2020).

259. Lauring, A. S. & Hodcroft, E. B. Genetic Variants of SARS-CoV-2-What Do They Mean? *JAMA* **325,** 529–531 (Feb. 2021).

260. Hui, K. P. Y. *et al.* SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature* **603,** 715–720 (Mar. 2022).

261. Bhattacharyya, R. P. & Hanage, W. P. Challenges in Inferring Intrinsic Severity of the SARS-CoV-2 Omicron Variant. *N. Engl. J. Med.* **386,** e14 (Feb. 2022).

262. Latest insights team. *Coronavirus (COVID-19) latest insights - Office for National Statistics* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19latestinsights/infections. Accessed: 2022-5-23. May 2022.

263. Butowt, R., Bilińska, K. & von Bartheld, C. *Why Does the Omicron Variant Largely Spare Olfactory Function? Implications for the Pathogenesis of Anosmia in Coronavirus Disease 2019* 2022.

264. Graham, M. S. *et al.* Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *Lancet Public Health* **6,** e335–e345 (May 2021).

265. Zazhytska, M. *et al.* Non-cell-autonomous disruption of nuclear architecture as a potential cause of COVID-19-induced anosmia. *Cell* **185,** 1052–1064.e12 (Mar. 2022).

266. Wolter, N., Jassat, W., DATCOV-Gen author group, von Gottberg, A. & Cohen, C. Clinical severity of omicron lineage BA.2 infection compared with BA.1 infection in South Africa. *Lancet* **400,** 93–96 (July 2022).

267. Sievers, C. *et al.* SARS-CoV-2 Omicron variants BA.1 and BA.2 both show similarly reduced disease severity of COVID-19 compared to Delta, Germany, 2021 to 2022. *Euro Surveill.* **27** (June 2022).

268. Fonager, J. *et al.* Molecular epidemiology of the SARS-CoV-2 variant Omicron BA.2 sub-lineage in Denmark, 29 November 2021 to 2 January 2022. *Euro Surveill.* **27** (Mar. 2022).

269. Lewnard, J. A. *et al.* Clinical outcomes associated with SARS-CoV-2 Omicron (B.1.1.529) variant and BA.1/BA.1.1 or BA.2 subvariant infection in Southern California. *Nat. Med.* **28,** 1933–1943 (Sept. 2022).

270. Whitaker, M. *et al.* Persistent COVID-19 symptoms in a community study of 606,434 people in England. *Nat. Commun.* **13,** 1957 (Apr. 2022).

271. Sharot, T. *Weighting survey results* `http://www.redresearch.com/wp/wp-content/uploads/2016/01/Weighting-Survey-Results.pdf`. Accessed: 2022-10-26. 1986.

272. Hennig, C. & Imports, M. Package 'fpc'. *Available at: Available at: https://cran. r-project. org/web/packages/fpc/index. html ENT* **91** (2015).

273. Sørensen, T. J. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons* (I kommission hos E. Munksgaard, København, 1948).

274. Ayoubkhani, D. *Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK - Office for National Statistics* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/1april2021. Accessed: 2023-2-10. Mar. 2021.

275. Yomogida, K. *et al.* Post-Acute Sequelae of SARS-CoV-2 Infection Among Adults Aged ≥18 Years - Long Beach, California, April 1-December 10, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **70,** 1274–1277 (Sept. 2021).

276. Augustin, M. *et al.* Post-COVID syndrome in non-hospitalised patients with COVID-19: a longitudinal prospective cohort study. *Lancet Reg Health Eur* **6,** 100122 (July 2021).

277. Del Rio, C., Collins, L. F. & Malani, P. Long-term Health Consequences of COVID-19. *JAMA* **324,** 1723–1724 (Nov. 2020).

278. 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 (Jan. 2023).

279. Ayoubkhani, D., Pawelek, P. & Gaughan, C. *Technical article: Updated estimates of the prevalence of post-acute symptoms among people with coronavirus (COVID-19) in the UK - Office for National Statistics* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/technicalarticleupdatedestimates26april2020to1august2021. Accessed: 2022-10-26. Sept. 2021.

280. Ali, A., Siddiqui, A. A., Arshad, M. S., Iqbal, F. & Arif, T. B. Effects of COVID-19 pandemic and lockdown on lifestyle and mental health of students: A retrospective study from Karachi, Pakistan. *Ann. Med. Psychol.* **180,** S29–S37 (June 2022).

281. Morgül, E., Kallitsoglou, A. & Essau, C. A. Psychological effects of the COVID-19 lockdown on children and families in the UK. *Rev. psicol. clın. con niños adolesc.* **7,** 42–48 (Sept. 2020).

282. Odriozola-González, P., Planchuelo-Gómez, Á., Irurtia, M. J. & de Luis-Garcıa, R. Psychological effects of the COVID-19 outbreak and lockdown among students and workers of a Spanish university. *Psychiatry Res.* **290,** 113108 (Aug. 2020).

283. Sykes, D. L. *et al.* Post-COVID-19 Symptom Burden: What is Long-COVID and How Should We Manage It? *Lung* **199,** 113–119 (Apr. 2021).

284. Michelen, M. *et al.* Characterising long COVID: a living systematic review. *BMJ Glob Health* **6** (Sept. 2021).

285. Looi, M.-K. & Mahase, E. Has covid-19 become milder? *BMJ* **379,** o2516 (Oct. 2022).

286. *REACT GE* https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/studies/react-ge/. Accessed: 2023-5-18.

287. Elliott, P. *et al.* Design and Implementation of a National SARS-CoV-2 Monitoring Program in England: REACT-1 Study. *Am. J. Public Health* **113,** 545–554 (May 2023).

288. Steel, K. & Davies, B. *Coronavirus (COVID-19) Infection Survey - office for national statistics* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsinthecommunityinengland/characteristicsofpeopletestingpositiveforcovid19inengland27january2021. Accessed: 2022-10-26. Jan. 2021.

289. Steel, K. & Willis, Z. *Coronavirus (COVID-19) Infection Survey, characteristics of people testing positive for COVID-19, UK - Office for National Statistics* https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveycharacteristicsofpeopletestingpositivefo2february2022. Accessed: 2022-10-26. Feb. 2022.

290. Clarke, J., Beaney, T. & Majeed, A. UK scales back routine covid-19 surveillance. *BMJ* **376,** o562 (Mar. 2022).

Appendices

Appendix A

Appendix: The COVID-19 Research Landscape

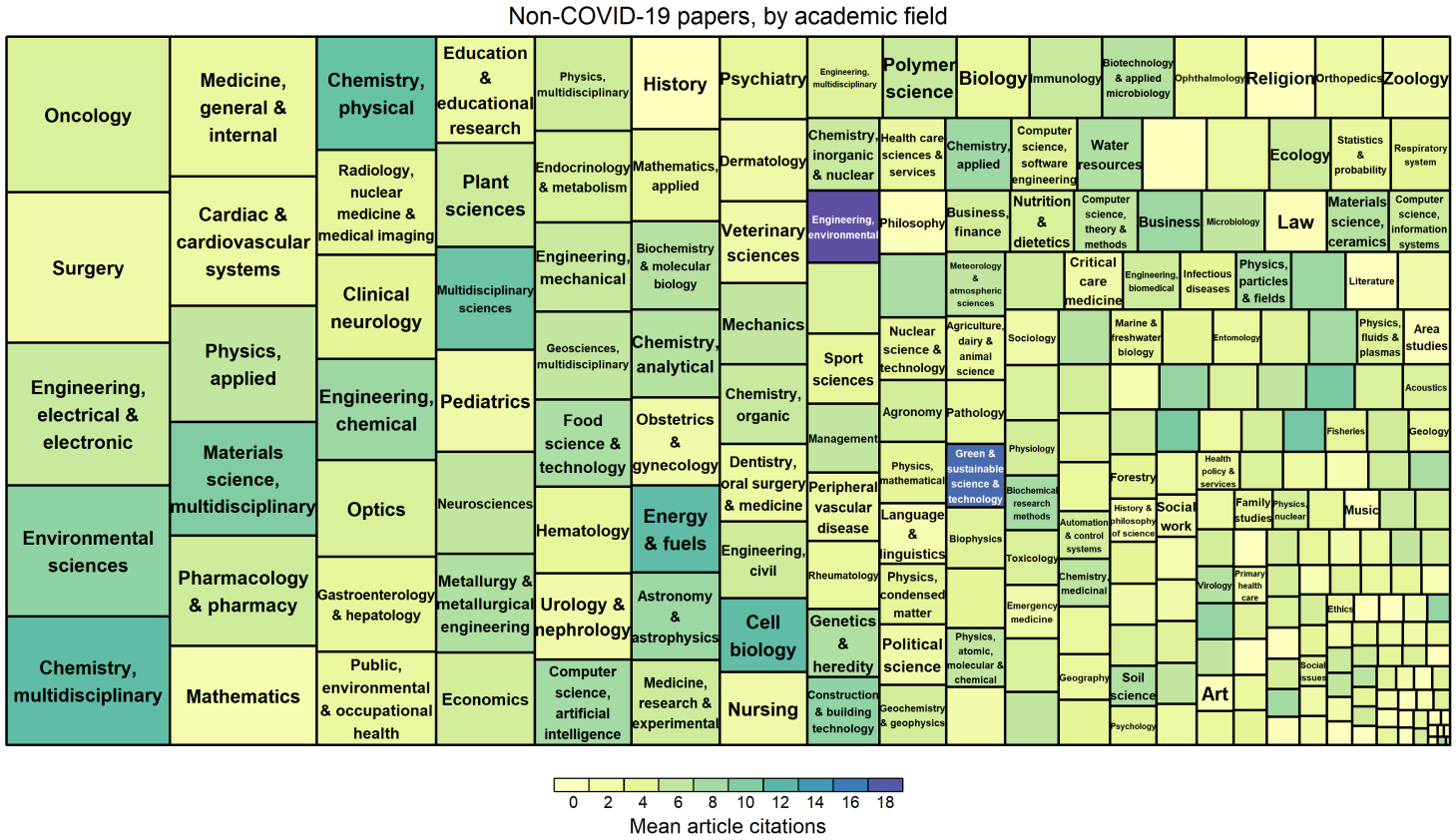


Figure A.1: Treemap plot showing journal fields among non-COVID-19 papers (2019–2022). The size of the box indicates the number of papers in this field; the colour scale indicate the average number of citations received by the articles in this field (as of March 2023).

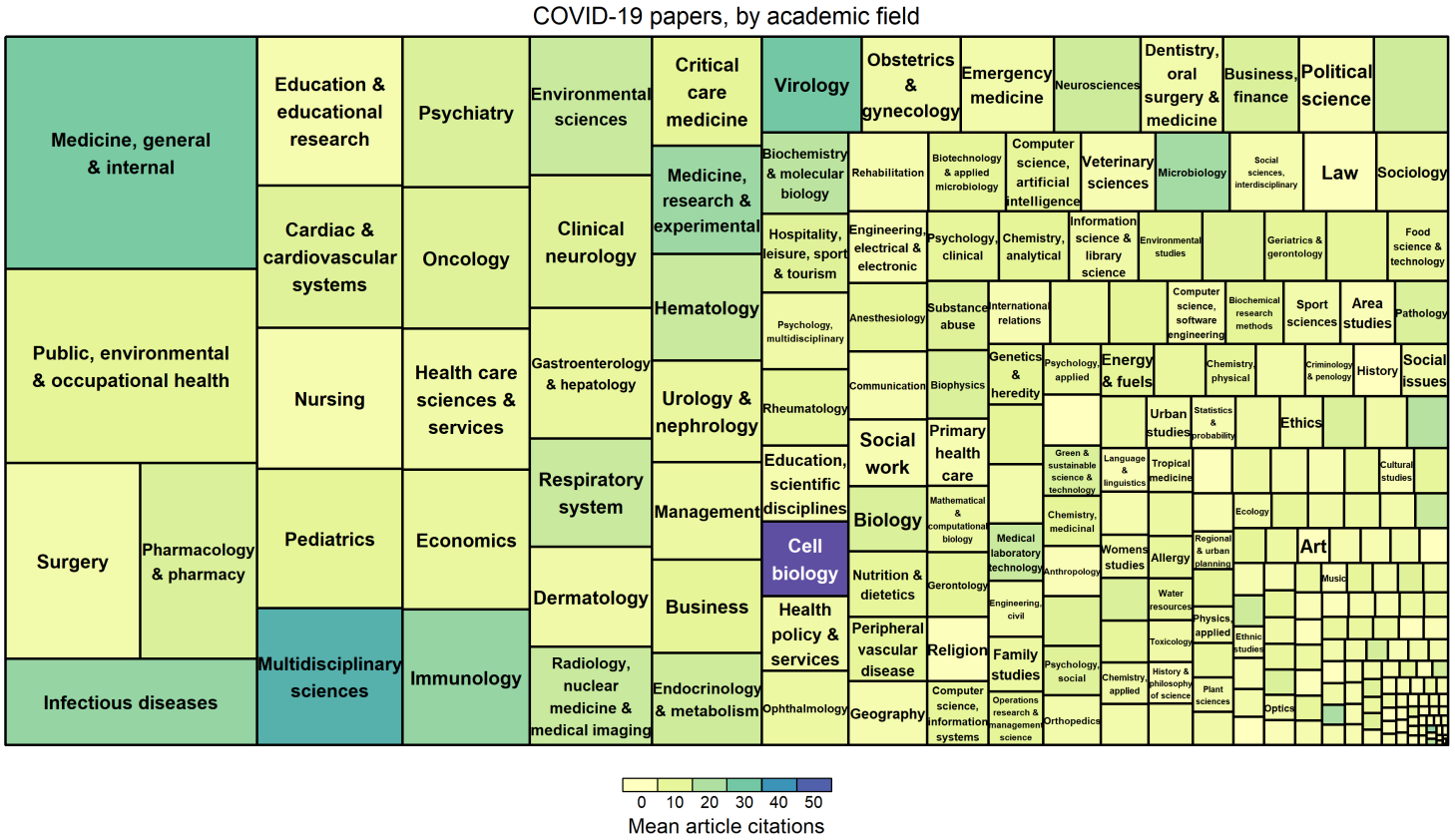


Figure A.2: Treemap plot showing journal fields among COVID-19 papers (2020–2022). The size of the box indicates the number of papers in this field; the colour scale indicate the average number of citations received by the articles in this field (as of March 2023).

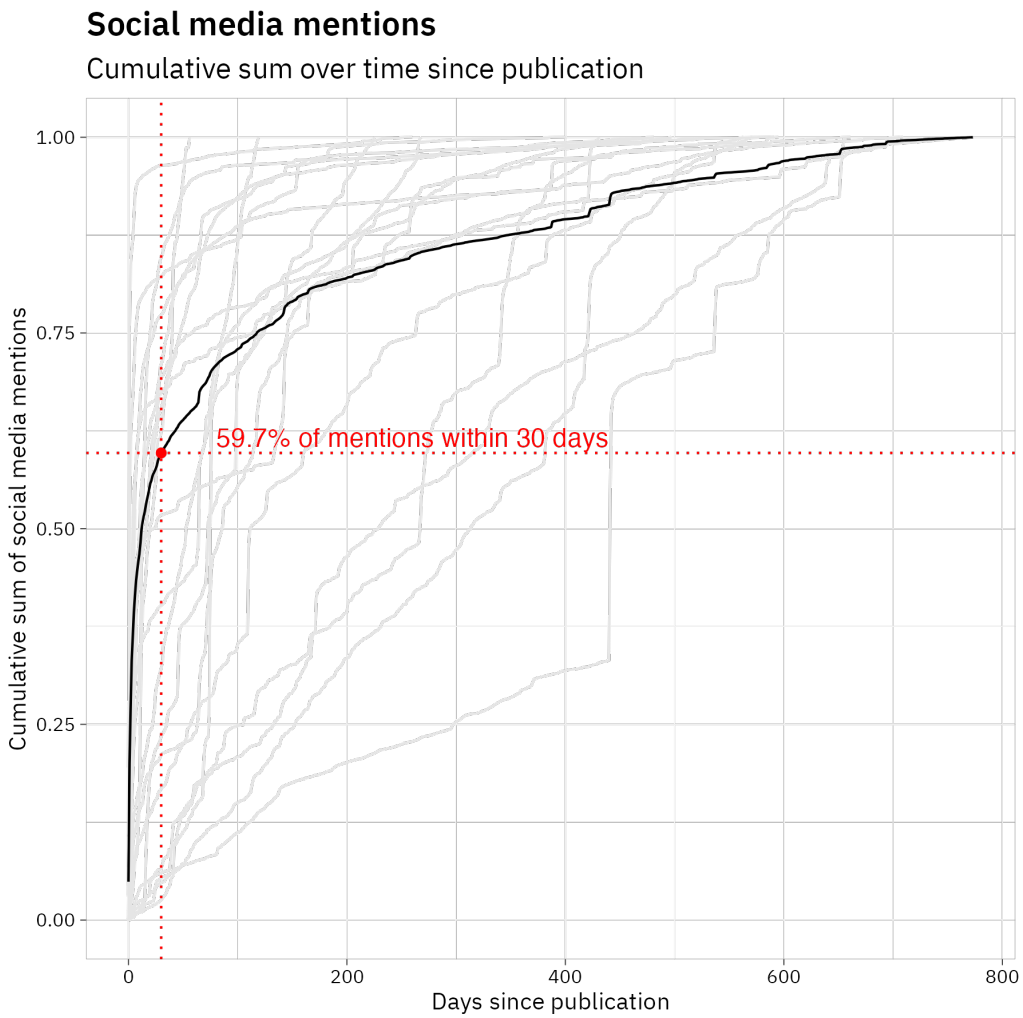


Figure A.3: Accumulation velocity of social media Altmetrics scores for the 20 COVID-19 papers with the highest Altmetrics scores. Average across all papers shown in black; individual paper accumulation curves shown in grey. Scores over time are normalised to a proportion of their 'final' score (as of March 2023). The papers accumulated an average of 59.7% of their 'final' scores within 30 days of publication.

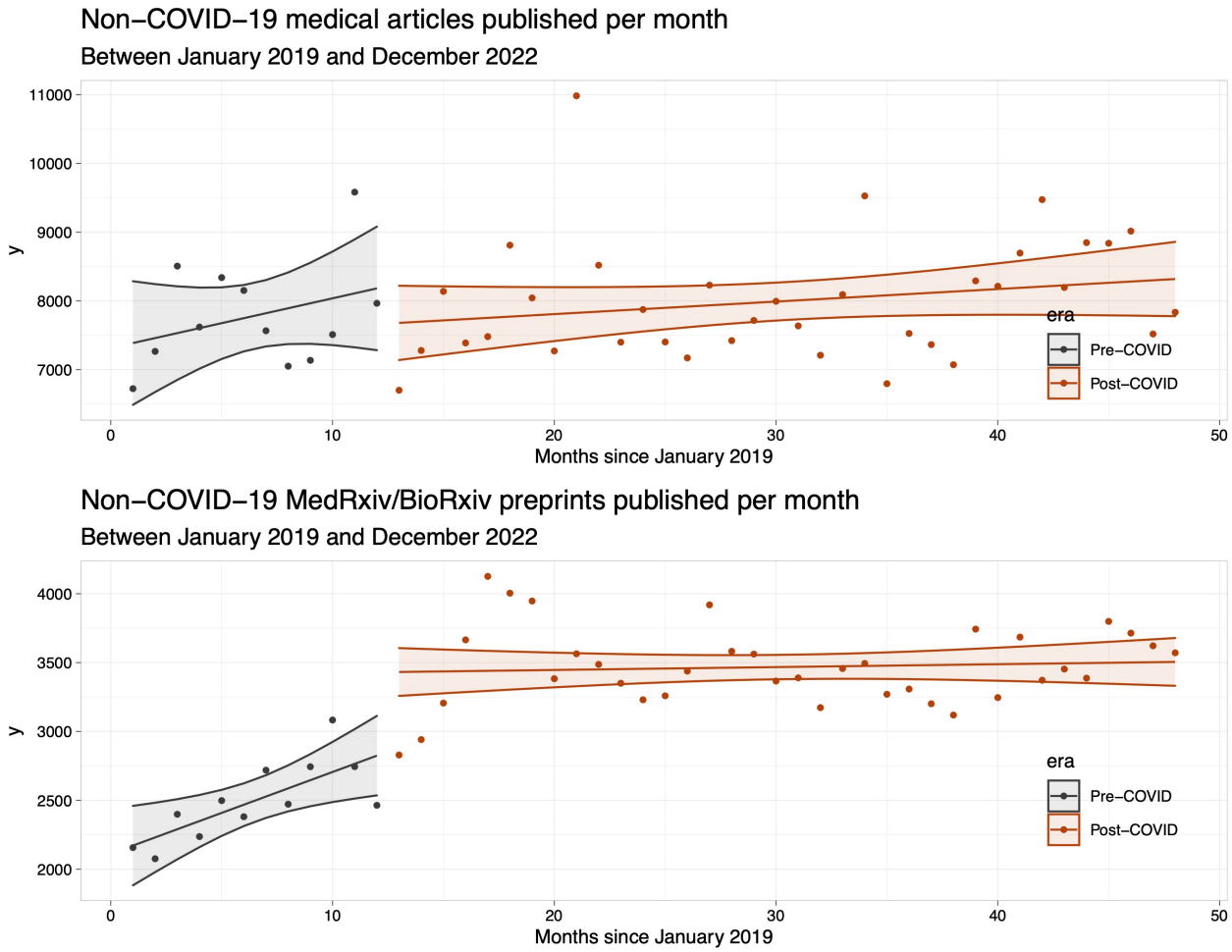


Figure A.4: Publication rates of non-COVID-19 medical papers in the pre-pandemic period (2019) and the pandemic period (2020-2022). Regression lines and 95% CIs are shown with articles/month regressed onto time (in months since January 2019) and a binary pre-pandemic/post pandemic variable, plus an interaction effect between these two independent variables. In journal articles, this interaction effect is not significant (p=0.46), suggesting a decline in growth rate during the pandemic period. In preprints, both the main effect for the pre/post-pandemic binary variable, and the interaction effect, are significant (p¡0.00001 and p=0.017 respectively), with a positive effect, suggesting a boost in preprint publishing and a subsequent lower growth rate.

Variable	**Data type**	**Data source**
Article title	Text	Dimensions
Article abstract	Text	Dimensions
Date of publication	Date	Dimensions
DOI	Text	Dimensions
Authors	Text	Dimensions
Journal title	Text	Dimensions
Issue number, volume, pages, etc	Text	Dimensions
PubMed ID	Text	Dimensions
Altmetrics ID	Text	Dimensions
Journal category (field)	Text	Dimensions
ISSN/ISBN	Text	Dimensions
Preprint / article status	Text	Dimensions
DOI of subsequent publication (for preprints)	Text	Dimensions
Journal Impact Factor (JIF)	Numeric	Thompson Reuters

Table A.1 continued from previous page

Variable	Data type	Data source
Journal Citation Ration (JCR)	Numeric	Thompson Reuters
Journal citations (count of all citations received by publishing journal)	Numeric	Thompson Reuters
Total citations	Numeric	CrossRef
Citations / day	Numeric	CrossRef
Relative citation ratio	Numeric	Dimensions
Field citation ratio	Numeric	Dimensions
Recent citations	Numeric	Dimensions
Total mentions online	Numeric	Altmetrics
Twitter mentions	Numeric	Altmetrics
Number of online accounts mentioning paper	Numeric	Altmetrics
Facebook mentions	Numeric	Altmetrics
Mainstream media mentions	Numeric	Altmetrics
Reddit user mentions	Numeric	Altmetrics
Wikipedia mentions	Numeric	Altmetrics
YouTube channel mentions	Numeric	Altmetrics
Blog mentions	Numeric	Altmetrics
Peer review site mentions	Numeric	Altmetrics
Policy document mentions	Numeric	Altmetrics
Cohort: science communicator mentions	Numeric	Altmetrics
Cohort: general public mentions	Numeric	Altmetrics
Cohort: scientist mentions	Numeric	Altmetrics
Cohort: doctors and medics mentions	Numeric	Altmetrics
Readers on Citeulike	Numeric	Altmetrics
Readers on Mendeley	Numeric	Altmetrics
Readers on Connotea	Numeric	Altmetrics
Readers on all reference managers	Numeric	Altmetrics
Rank: all journals	Numeric	Altmetrics
Rank: within journal	Numeric	Altmetrics
Rank: among similarly aged papers	Numeric	Altmetrics
Rank: among similarly aged papers within journal	Numeric	Altmetrics
Percentile: all journals	Numeric	Altmetrics
Percentile: within journal	Numeric	Altmetrics
Percentile: among similarly aged papers	Numeric	Altmetrics
Percentile: among similarly aged papers within journal	Numeric	Altmetrics
Altmetrics score	Numeric	Altmetrics

Table A.1: Data points available for each article, with sources

Topic	Top 8 terms	Classification
1	can, help, better, provide, understanding, understand, improve, lead	subject - general
2	use, digital, technology, based, technologies, users, using, used	method
3	decision, making, recommendations, guidelines, decision making, based, decisions, guidance	method
4	patients, treatment, therapy, treated, received, receiving, patient, diagnosed	method
5	evidence, review, literature, current, based, provide, available, potential	method
6	qualitative, interviews, experiences, study, analysis, themes, conducted, using	subject - general
7	post, covid, recovery, post covid, function, body, rehabilitation, mass	subject
8	case, patient, report, year, old, year old, presented, developed	method
9	number, increase, decrease, per, reduction, compared, average, significant	method
10	community, challenges, response, pandemic, crisis, communities, lessons, future	subject - general
11	many, new, one, world, way, place, take, people	subject - general
12	food, economic, income, pandemic, socio, vulnerable, household, loss	subject
13	control, prevention, epidemic, measures, covid, outbreak, disease, spread	subject
14	italy, european, italian, different, spain, germany, countries, main	subject - general
15	face, measures, preventive, mask, hand, masks, use, adherence	subject
16	role, important, play, covid, key, crucial, importance, well	subject - general
17	people, covid, residents, home, nursing, living, hiv, among	subject
18	cohort, outcomes, mortality, outcome, death, cohort study, primary, associated	method
19	economic, financial, market, business, crisis, tourism, sector, economy	subject
20	management, strategies, strategy, effective, based, approach, waste, mitigation	subject
21	data, collected, using, available, collection, used, information, analysis	method
22	cases, confirmed, number, case, covid, reported, confirmed cases, daily	subject
23	women, covid, pregnant, pregnant women, pregnancy, maternal, mothers, violence	subject
24	system, systems, healthcare, health, resources, capacity, pandemic, health system	subject - general
25	deaths, death, covid, million, per, number, population, excess	subject
26	b, variants, variant, omicron, delta, concern, ba, alpha	subject
27	global, pandemic, ongoing, global pandemic, caused, worldwide, global health, threat	subject - general
28	care, health care, patient, health, telemedicine, visits, telehealth, primary	subject
29	staff, protective, personal, equipment, safety, ppe, protective equipment, personal protective	subject
30	areas, urban, mobility, travel, rural, area, local, city	subject - general
31	days, median, time, day, onset, iqr, duration, range	method
32	model, learning, models, proposed, accuracy, using, based, performance	method
33	group, groups, significant, difference, differences, control, study, statistically	method
34	covid, critical, conditions, existing, illness, pre, pre existing, condition	subject
35	medical, covid, medicine, records, electronic, physicians, center, doctors	subject
36	one, two, three, four, five, least, six, seven	method
37	immune, response, cell, cells, responses, inflammatory, cytokine, t cell	subject
38	countries, south, country, africa, covid, across, middle, world	subject - general
39	work, working, pandemic, employees, burnout, job, home, professional	subject
40	studies, covid, well, however, several, needed, previous, recent	method
41	long, term, covid, long term, short, short term, effects, well	subject
42	changes, increased, use, change, increase, pandemic, decreased, significantly	method
43	states, united, united states, us, black, u.s, across, white	subject - general
44	covid, infections, co, ii, infection, bacterial, h, v	subject
45	study, findings, research, design, approach, paper, implications, purpose	method
46	cells, expression, host, cell, viral, angiotensin, entry, lung	subject
47	virus, india, world, spread, corona, various, like, corona virus	subject - general
48	cov, sars, sars cov, infection, confirmed sars, coronavirus sars, causes covid, compatible	subject - general
49	months, time, weeks, follow, month, two, points, week	method
50	participants, reported, self, study, self reported, among, completed, report	method
51	analysis, network, based, patterns, identify, using, structure, temporal	method
52	research, scientific, article, researchers, published, science, et al, field	method
53	covid, disease, world, diseases, infectious, organization, health, world health	subject - general
54	anxiety, depression, symptoms, stress, scale, psychological, disorder, depressive	subject
55	study, research, used, data, indonesia, aims, pandemic, carried	method
56	activity, physical, physical activity, time, exercise, activities, weight, lifestyle	subject
57	life, quality, covid, sleep, vitamin, satisfaction, vitamin d, threatening	subject
58	research, study, support, funding, national, approval, authors, interest	method
59	antibody, antibodies, igg, anti, sars, cov, sars cov, response	subject
60	training, program, virtual, programs, person, remote, based, implementation	subject
61	time, based, cost, real, can, effective, approach, tool	method
62	regression, associated, analysis, logistic, logistic regression, variables, predictors, multivariate	method
63	covid, correlation, significant, r, relationship, study, found, correlated	method
64	coronavirus, disease, coronavirus disease, novel, novel coronavirus, disease covid, coronavirus covid, caused	subject - general

Table A.2 continued from previous page

Topic	Top 8 terms	Classification
65	clinical, trial, trials, day, randomized, treatment, controlled, days	method
66	positive, testing, test, negative, tested, tests, tested positive, screening	subject
67	different, large, types, scale, type, three, two, based	method
68	pandemic, period, pre, compared, periods, pre pandemic, covid, time	subject - general
69	even, however, still, due, though, although, many, might	method
70	effect, interventions, intervention, effects, non, temperature, pharmaceutical, reduce	subject
71	across, uk, within, range, wide, including, national, canada	subject - general
72	covid, patients, cardiac, injury, liver, heart, cardiovascular, acute	subject
73	diabetes, disease, patients, chronic, comorbidities, kidney, covid, hypertension	subject
74	policy, government, measures, policies, state, pandemic, response, governments	subject
75	suggest, findings, likely, whether, less, indicate, may, however	method
76	population, prevalence, general, among, covid, general population, based, study	subject
77	factors, risk, associated, covid, risk factors, factor, identify, factors associated	subject
78	viral, virus, rna, respiratory, load, transmission, tract, viral load	subject
79	information, media, social, social media, news, communication, content, sources	subject
80	transmission, contact, quarantine, contacts, close, tracing, community, contact tracing	subject
81	social, distancing, social distancing, isolation, home, measures, physical, pandemic	subject
82	human, virus, influenza, viruses, respiratory, humans, coronaviruses, mers	subject
83	workers, healthcare, health, nurses, care, professionals, hcws, healthcare workers	subject
84	current, pandemic, need, covid, dental, future, attention, situation	subject
85	article, pandemic, crisis, political, paper, legal, rights, society	subject
86	symptoms, fever, pain, covid, cough, common, fatigue, symptom	subject
87	clinical, characteristics, laboratory, features, covid, course, outcomes, clinical characteristics	subject
88	model, models, number, epidemic, time, parameters, dynamics, data	method
89	years, age, older, adults, aged, among, groups, elderly	subject - general
90	perceived, behavior, behaviors, intention, trust, perception, self, influence	subject
91	ct, lung, pneumonia, patients, chest, covid, imaging, findings	subject
92	detection, sensitivity, samples, diagnostic, rapid, tests, testing, specificity	subject
93	studies, review, systematic, search, included, articles, meta, analysis	method
94	lockdown, air, period, reduction, pollution, quality, lockdowns, due	subject
95	molecular, drug, compounds, potential, binding, docking, protease, drugs	subject
96	study, showed, present, aimed, revealed, present study, analysis, study aimed	method
97	cancer, patients, surgery, surgical, treatment, covid, procedures, cancer patients	subject
98	vaccine, vaccination, vaccines, dose, vaccinated, mrna, doses, hesitancy	subject
99	new, development, process, article, main, state, problem, new coronavirus	subject - general
100	survey, respondents, online, questionnaire, questions, reported, responses, online survey	subject - general
101	early, covid, phase, stage, initial, stages, late, phases	subject - general
102	knowledge, covid, towards, practice, practices, regarding, attitudes, attitude	subject - general
103	levels, level, high, higher, low, lower, significantly, associated	method
104	ci, ratio, odds, confidence, interval, associated, aor, adjusted	method
105	protein, spike, sars, cov, sars cov, binding, rbd, spike protein	subject
106	icu, care, patients, intensive, intensive care, ventilation, unit, mechanical	subject
107	march, april, period, january, june, february, december, july	subject - general
108	patients, respiratory, ill, critically, ards, failure, acute, distress	subject
109	neurological, manifestations, stroke, system, infection, associated, complications, covid	subject
110	china, outbreak, wuhan, december, spread, chinese, coronavirus, novel	subject - general
111	patients, blood, levels, dimer, d dimer, serum, count, crp	subject
112	infection, infected, infections, infected patients, patients infected, viral infection, reinfection, however	subject
113	treatment, drugs, drug, clinical, use, therapeutic, antiviral, therapy	subject
114	p, vs, lt, p lt, significantly, higher, compared, gt	method
115	covid, non, without, non covid, compared, similar, including, without covid	subject - general
116	severe, disease, severity, covid, severe covid, mild, moderate, patients	subject
117	supply, development, production, demand, products, chain, sustainable, energy	subject - general
118	rate, mortality, rates, incidence, covid, higher, mortality rate, high	subject
119	learning, online, education, students, teaching, teachers, online learning, face	subject
120	pandemic, impact, affected, covid, negative, impacts, significant, effects	subject - general
121	age, mean, years, score, gender, male, female, study	method
122	using, test, value, study, used, statistical, analysis, data	method
123	may, covid, potential, help, lead, increase, affect, contribute	subject - general
124	health, public, public health, emergency, health emergency, authorities, concern, response	subject
125	health, mental, mental health, pandemic, problems, physical, well, impact	subject
126	patients, hospital, hospitalized, admission, admitted, hospitalization, stay, discharge	subject - general
127	children, family, parents, pediatric, adolescents, families, child, caregivers	subject - general

Table A.2 continued from previous page

Topic	Top 8 terms	Classification
128	sars, cov, sars cov, genome, viral, mutations, sequencing, genetic	subject
129	emergency, covid, department, hospital, ed, new, march, emergency department	subject - general
130	individuals, covid, subjects, compared, controls, healthy, asymptomatic, symptomatic	method
131	stress, psychological, support, distress, coping, emotional, resilience, well	subject
132	students, school, university, academic, schools, college, student, pandemic	subject
133	effects, covid, events, adverse, safety, side, event, reported	subject
134	cross, study, sectional, cross sectional, among, sectional study, conducted, questionnaire	method
135	risk, high, covid, high risk, exposure, increased, risks, increased risk	subject
136	services, service, access, pandemic, health, delivery, support, needs	subject
137	coronavirus, respiratory, acute, syndrome, severe, acute respiratory, respiratory syndrome, severe acute	subject - general
138	pcr, rt, rt pcr, positive, reaction, chain, polymerase, polymerase chain	subject
139	covid, related, specific, pandemic related, including, health related, covid related, associated	subject - general
140	first, second, wave, third, two, first wave, covid, waves	subject - general

Table A.2: 140 topics identified by topic modelling of abstracts from 486,030 COVID-19 papers published in 2020-2022. Topics are classified into three categories: subject - general (for generic topics relating to the pandemic and healthcare), subject (for topics relating to specific aspects of COVID-19), and method (for topics that relate to the methodology of the paper). The top 8 terms associated with the topic (as measured by Gamma value) are shown for each topic.

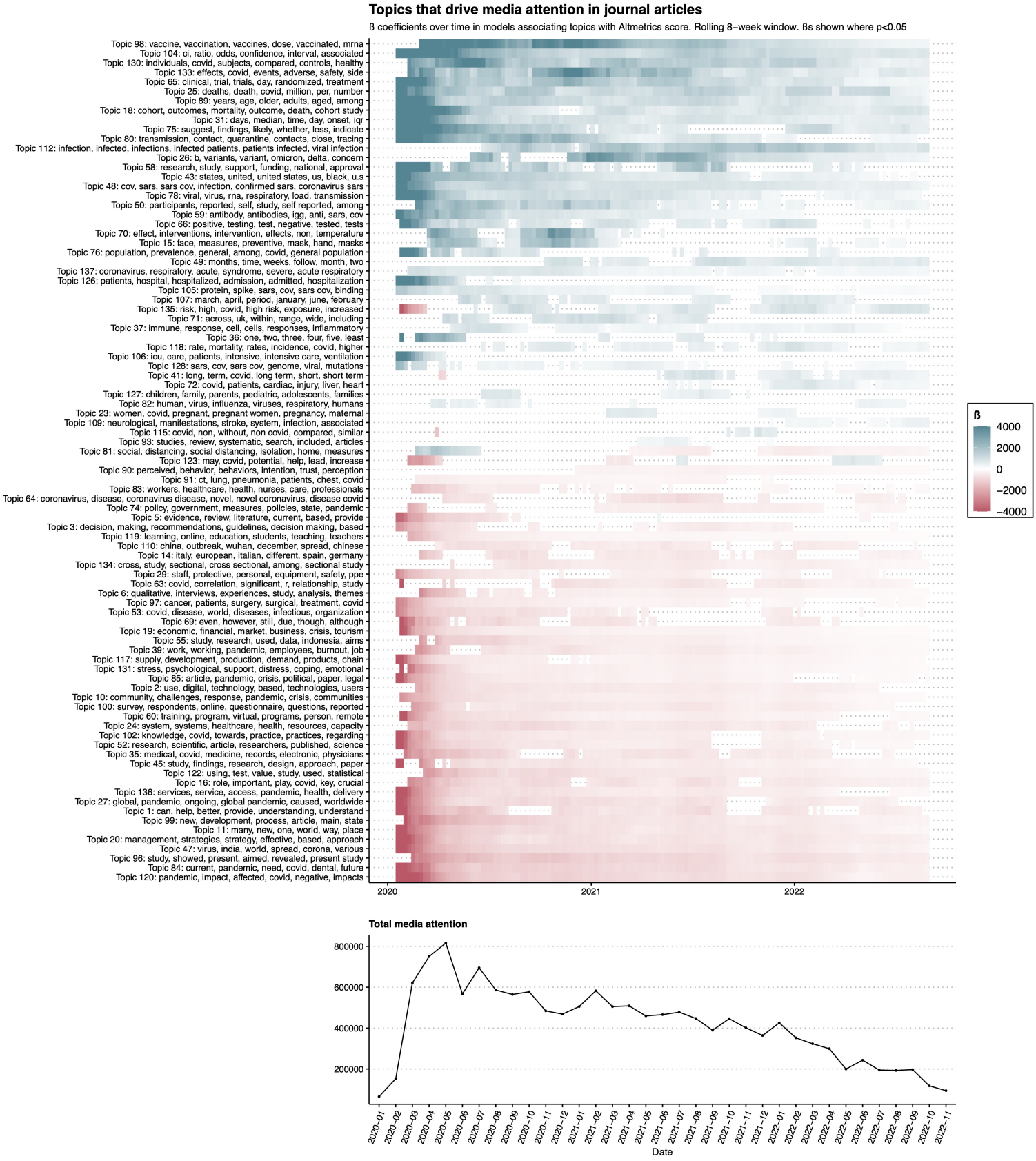


Figure A.5: Heatmap showing results of univariable regression of Altmetrics scores onto abstract topics on a rolling 8-week time window throughout the pandemic period 2020–2022, for journal articles only (excluding preprints). Colour represents beta values: blue = positive betas, red = negative betas and colour darkness represents beta value magnitude (darker hue = larger beta). Topics are ordered by average beta value across the three years, and only topics with at least one period of significant association with Altmetrics scores are shown.

Appendix B

Appendix: Symptom profiling

Variable	Category	Rounds 2-7 (Wild type)	Rounds 8-10 (Alpha)	Rounds 13-15 (Delta)	Round 17 (BA.1 Omicron)	Rounds 19 (BA.2 Omicron)	Pooled
All participants	All participants	713,202 (100%, [100-100])	357,960 (100%, [100-100])	228,934 (100%, [100-100])	83,114 (100%, [100-100])	86,162 (100%, [100-100])	1,469,372
Sex	Female	402,121 (56.4%, [56.3-56.5])	201,126 (56.2%, [56-56.3])	129,652 (56.6%, [56.4-56.8])	47,205 (56.8%, [56.5-57.1])	49,636 (57.6%, [57.3-57.9])	829,740
	Male	311,081 (43.6%, [43.5-43.7])	156,834 (43.8%, [43.7-44])	99,282 (43.4%, [43.2-43.6])	35,909 (43.2%, [42.9-43.5])	36,526 (42.4%, [42.1-42.7])	639,632
Age group	5-17	0 (0%, [0-0])	0 (0%, [0-0])	0 (0%, [0-0])	0 (0%, [0-0])	0 (0%, [0-0])	0
	18-24	32,298 (4.5%, [4.5-4.6])	15,805 (4.4%, [4.3-4.5])	7,268 (3.2%, [3.1-3.2])	2,346 (2.8%, [2.7-2.9])	6,984 (8.1%, [7.9-8.3])	64,701
	25-34	73,352 (10.3%, [10.2-10.4])	35,260 (9.9%, [9.8-9.9])	19,691 (8.6%, [8.5-8.7])	6,891 (8.3%, [8.1-8.5])	14,319 (16.6%, [16.4-16.9])	149,513
	35-44	103,317 (14.5%, [14.4-14.6])	51,269 (14.3%, [14.2-14.4])	30,996 (13.5%, [13.4-13.7])	11,253 (13.5%, [13.3-13.8])	13,864 (16.1%, [15.8-16.3])	210,699
	45-54	137,175 (19.2%, [19.1-19.3])	68,057 (19%, [18.9-19.1])	43,205 (18.9%, [18.7-19])	15,697 (18.9%, [18.6-19.2])	11,662 (13.5%, [13.3-13.8])	275,796
	55-64	157,180 (22%, [21.9-22.1])	80,550 (22.5%, [22.4-22.6])	53,663 (23.4%, [23.3-23.6])	19,949 (24%, [23.7-24.3])	16,042 (18.6%, [18.4-18.9])	327,384
	65-74	144,612 (20.3%, [20.2-20.4])	74,535 (20.8%, [20.7-21])	50,980 (22.3%, [22.1-22.4])	18,344 (22.1%, [21.8-22.4])	15,386 (17.9%, [17.6-18.1])	303,857
	75+	65,268 (9.2%, [9.1-9.2])	32,484 (9.1%, [9-9.2])	23,131 (10.1%, [10-10.2])	8,634 (10.4%, [10.2-10.6])	7,905 (9.2%, [9-9.4])	137,422
Ethnicity	Asian	25,317 (3.6%, [3.6-3.6])	11,665 (3.3%, [3.3-3.4])	10,202 (4.6%, [4.5-4.6])	3,618 (4.5%, [4.3-4.6])	5,113 (6.1%, [5.9-6.2])	55,915
	Black	6,736 (1%, [0.9-1])	3,497 (1%, [1-1])	3,441 (1.5%, [1.5-1.6])	1,152 (1.4%, [1.3-1.5])	1,375 (1.6%, [1.5-1.7])	16,201
	Mixed	8,206 (1.2%, [1.1-1.2])	4,141 (1.2%, [1.1-1.2])	2,886 (1.3%, [1.2-1.3])	998 (1.2%, [1.2-1.3])	1,363 (1.6%, [1.5-1.7])	17,594
	Other	5,066 (0.7%, [0.7-0.7])	2,382 (0.7%, [0.7-0.7])	2,082 (0.9%, [0.9-1])	673 (0.8%, [0.8-0.9])	837 (1%, [0.9-1.1])	11,040
	White	658,752 (93.6%, [93.5-93.6])	330,027 (93.8%, [93.8-93.9])	205,368 (91.7%, [91.6-91.8])	74,662 (92.1%, [91.9-92.2])	75,494 (89.7%, [89.5-89.9])	1,344,303
PCR positive at time of survey	No	710,231 (99.6%, [99.6-99.6])	355,685 (99.4%, [99.3-99.4])	227,441 (99.3%, [99.3-99.4])	80,023 (96.3%, [96.2-96.4])	80,567 (93.5%, [93.3-93.7])	1,453,947
	Yes	2,971 (0.4%, [0.4-0.4])	2,275 (0.6%, [0.6-0.7])	1,493 (0.7%, [0.6-0.7])	3,091 (3.7%, [3.6-3.8])	5,595 (6.5%, [6.3-6.7])	15,425
Prior COVID-19 infection	No previous infection	604,198 (84.7%, [84.6-84.8])	301,856 (84.3%, [84.2-84.4])	189,854 (82.9%, [82.8-83.1])	62,249 (74.9%, [74.6-75.2])	54,243 (63%, [62.6-63.3])	1,212,400
	Previous infection greater than 28 days	3,308 (0.5%, [0.4-0.5])	10,374 (2.9%, [2.8-3])	14,428 (6.3%, [6.2-6.4])	8,654 (10.4%, [10.2-10.6])	19,209 (22.3%, [22-22.6])	55,973
	Previous infection with unknown time	26 (0%, [0-0])	56 (0%, [0-0])	34 (0%, [0-0])	10 (0%, [0-0])	26 (0%, [0-0])	152
	Previous infection within 28 days	1,276 (0.2%, [0.2-0.2])	1,906 (0.5%, [0.5-0.6])	1,659 (0.7%, [0.7-0.8])	4,735 (5.7%, [5.5-5.9])	5,332 (6.2%, [6-6.4])	14,908
	Suspected previous infection	103,703 (14.5%, [14.5-14.6])	43,483 (12.1%, [12-12.3])	22,805 (10%, [9.8-10.1])	7,410 (8.9%, [8.7-9.1])	7,288 (8.5%, [8.3-8.6])	184,689
	Unkown	691 (0.1%, [0.1-0.1])	285 (0.1%, [0.1-0.1])	154 (0.1%, [0.1-0.1])	56 (0.1%, [0.1-0.1])	64 (0.1%, [0.1-0.1])	1,250
Vaccination status	NA	-	-	912 (1.2%, [1.1-1.2])	819 (1%, [0.9-1.1])	1,096 (1.3%, [1.2-1.3])	2,827
	Not vaccinated	-	-	3,650 (4.6%, [4.5-4.8])	2,545 (3.1%, [2.9-3.2])	3,506 (4.1%, [3.9-4.2])	9,701
	One dose	-	-	519 (0.7%, [0.6-0.7])	288 (0.3%, [0.3-0.4])	620 (0.7%, [0.7-0.8])	1,427
	Three dose	-	-	14,388 (18.2%, [17.9-18.5])	69,714 (83.9%, [83.6-84.1])	69,741 (80.9%, [80.7-81.2])	153,843
	Two dose	-	-	56,135 (71%, [70.7-71.3])	6,156 (7.4%, [7.2-7.6])	7,375 (8.6%, [8.4-8.7])	69,666
	Unknown dose	-	-	3,433 (4.3%, [4.2-4.5])	3,592 (4.3%, [4.2-4.5])	3,824 (4.4%, [4.3-4.6])	10,849
Symptoms	Symptomatic	95,973 (13.5%, [13.4-13.5])	49,228 (13.8%, [13.6-13.9])	36,687 (16%, [15.9-16.2])	17,617 (21.2%, [20.9-21.5])	20,726 (24.1%, [23.8-24.3])	220,231
Number of reported symptoms	Mean (SD)	0.52 (1.66)	0.5 (1.63)	0.61 (1.84)	0.81 (2.12)	1.1 (2.66)	0.6 (1.83)
Sought medical attention for symptoms	No	87,252 (89.5%, [89.3-89.7])	43,636 (88.9%, [88.6-89.2])	31,933 (87.2%, [86.8-87.5])	15,864 (90.1%, [89.7-90.6])	18,405 (88.9%, [88.4-89.3])	197,090
	Yes	10,203 (10.5%, [10.3-10.7])	5,457 (11.1%, [10.8-11.4])	4,701 (12.8%, [12.5-13.2])	1,734 (9.9%, [9.4-10.3])	2,303 (11.1%, [10.7-11.6])	24,398
Symptoms affecting day-to-day activities	A lot	0 (0%, [0-0])	0 (0%, [0-0])	2,291 (1%, [1-1])	1,427 (1.7%, [1.6-1.8])	2,243 (2.6%, [2.5-2.7])	5,961

Table B.1 continued from previous page

Variable	Category	Rounds 2-7 (Wild type)	Rounds 8-10 (Alpha)	Rounds 13-15 (Delta)	Round 17 (BA.1 Omicron)	Rounds 19 (BA.2 Omicron)	Pooled
	A little	0 (0%, [0-0])	0 (0%, [0-0])	7,849 (3.4%, [3.4-3.5])	4,817 (5.8%, [5.6-6])	6,207 (7.2%, [7-7.4])	18,873
	Not at all	0 (0%, [0-0])	0 (0%, [0-0])	4,938 (2.2%, [2.1-2.2])	3,264 (3.9%, [3.8-4.1])	3,442 (4%, [3.9-4.1])	11,644
	Don't know / PNA / Non-response	95,973 (13.5%, [13.4-13.5])	49,228 (13.8%, [13.6-13.9])	21,609 (9.4%, [9.3-9.6])	8,109 (9.8%, [9.6-10])	8,834 (10.3%, [10.1-10.5])	183,753
	No symptoms reported	617,229 (86.5%, [86.5-86.6])	308,732 (86.2%, [86.1-86.4])	192,247 (84%, [83.8-84.1])	65,497 (78.8%, [78.5-79.1])	65,436 (75.9%, [75.7-76.2])	1,249,141
Days since onset of first symptom	Mean (SD)	7.93 (3.28)	8.09 (3.28)	7.89 (3.32)	8.26 (3.26)	7.7 (3.28)	7.95 (3.29)
Symptoms reported	Loss or change of sense of smell	2,453 (0.3%, [0.3-0.4])	1,392 (0.4%, [0.4-0.4])	1,460 (0.6%, [0.6-0.7])	712 (0.9%, [0.8-0.9])	1,021 (1.2%, [1.1-1.3])	7,038
	Loss or change of sense of taste	2,876 (0.4%, [0.4-0.4])	1,551 (0.4%, [0.4-0.5])	1,420 (0.6%, [0.6-0.7])	741 (0.9%, [0.8-1])	1,074 (1.2%, [1.2-1.3])	7,662
	New persistent cough	4,710 (0.7%, [0.6-0.7])	2,120 (0.6%, [0.6-0.6])	3,599 (1.6%, [1.5-1.6])	2,045 (2.5%, [2.4-2.6])	3,544 (4.1%, [4-4.2])	16,018
	Fever	5,666 (0.8%, [0.8-0.8])	2,515 (0.7%, [0.7-0.7])	2,619 (1.1%, [1.1-1.2])	1,135 (1.4%, [1.3-1.4])	2,305 (2.7%, [2.6-2.8])	14,240
	Runny nose	24,262 (3.4%, [3.4-3.4])	12,008 (3.4%, [3.3-3.4])	11,105 (4.9%, [4.8-4.9])	6,022 (7.2%, [7.1-7.4])	8,238 (9.6%, [9.4-9.8])	61,635
	Sneezing	24,496 (3.4%, [3.4-3.5])	10,909 (3%, [3-3.1])	9,673 (4.2%, [4.1-4.3])	4,733 (5.7%, [5.5-5.9])	6,453 (7.5%, [7.3-7.7])	56,264
	Blocked nose	17,806 (2.5%, [2.5-2.5])	8,504 (2.4%, [2.3-2.4])	8,067 (3.5%, [3.4-3.6])	3,883 (4.7%, [4.5-4.8])	6,043 (7%, [6.8-7.2])	44,303
	Sore eyes	14,341 (2%, [2-2])	6,777 (1.9%, [1.8-1.9])	4,628 (2%, [2-2.1])	2,257 (2.7%, [2.6-2.8])	2,894 (3.4%, [3.2-3.5])	30,897
	Sore throat	27,295 (3.8%, [3.8-3.9])	10,259 (2.9%, [2.8-2.9])	10,410 (4.5%, [4.5-4.6])	4,642 (5.6%, [5.4-5.7])	6,836 (7.9%, [7.8-8.1])	59,442
	Hoarse voice	8,322 (1.2%, [1.1-1.2])	3,292 (0.9%, [0.9-1])	4,509 (2%, [1.9-2])	2,122 (2.6%, [2.4-2.7])	3,395 (3.9%, [3.8-4.1])	21,640
	Headache	39,332 (5.5%, [5.5-5.6])	20,835 (5.8%, [5.7-5.9])	13,727 (6%, [5.9-6.1])	6,929 (8.3%, [8.2-8.5])	8,676 (10.1%, [9.9-10.3])	89,499
	Dizziness	12,857 (1.8%, [1.8-1.8])	6,128 (1.7%, [1.7-1.8])	4,372 (1.9%, [1.9-2])	1,761 (2.1%, [2-2.2])	2,732 (3.2%, [3.1-3.3])	27,850
	Appetite loss	6,676 (0.9%, [0.9-1])	3,318 (0.9%, [0.9-1])	2,736 (1.2%, [1.2-1.2])	1,206 (1.5%, [1.4-1.5])	1,934 (2.2%, [2.1-2.3])	15,870
	Nausea/vomiting	8,185 (1.1%, [1.1-1.2])	3,884 (1.1%, [1.1-1.1])	2,825 (1.2%, [1.2-1.3])	1,149 (1.4%, [1.3-1.5])	1,663 (1.9%, [1.8-2])	17,706
	Diarrhoea	11,107 (1.6%, [1.5-1.6])	4,799 (1.3%, [1.3-1.4])	3,573 (1.6%, [1.5-1.6])	1,400 (1.7%, [1.6-1.8])	1,889 (2.2%, [2.1-2.3])	22,768
	Abdominal pain / belly ache	14,707 (2.1%, [2-2.1])	6,934 (1.9%, [1.9-2])	4,619 (2%, [2-2.1])	1,801 (2.2%, [2.1-2.3])	2,506 (2.9%, [2.8-3])	30,567
	Shortness of breath	12,907 (1.8%, [1.8-1.8])	6,634 (1.9%, [1.8-1.9])	4,587 (2%, [1.9-2.1])	2,407 (2.9%, [2.8-3])	3,200 (3.7%, [3.6-3.8])	29,735
	Tight chest	8,977 (1.3%, [1.2-1.3])	4,514 (1.3%, [1.2-1.3])	2,969 (1.3%, [1.3-1.3])	1,556 (1.9%, [1.8-2])	1,935 (2.2%, [2.1-2.3])	19,951
	Chest pain	4,158 (0.6%, [0.6-0.6])	2,279 (0.6%, [0.6-0.7])	1,529 (0.7%, [0.6-0.7])	742 (0.9%, [0.8-1])	1,042 (1.2%, [1.1-1.3])	9,750
	Chills	7,077 (1%, [1-1])	5,429 (1.5%, [1.5-1.6])	2,652 (1.2%, [1.1-1.2])	1,770 (2.1%, [2-2.2])	2,695 (3.1%, [3-3.2])	19,623
	Difficulty sleeping	26,551 (3.7%, [3.7-3.8])	14,043 (3.9%, [3.9-4])	8,956 (3.9%, [3.8-4])	4,481 (5.4%, [5.2-5.5])	4,959 (5.8%, [5.6-5.9])	58,990
	Tiredness	39,897 (5.6%, [5.5-5.6])	19,756 (5.5%, [5.4-5.6])	14,813 (6.5%, [6.4-6.6])	7,257 (8.7%, [8.5-8.9])	9,508 (11%, [10.8-11.2])	91,231
	Severe fatigue	4,677 (0.7%, [0.6-0.7])	2,715 (0.8%, [0.7-0.8])	2,063 (0.9%, [0.9-0.9])	979 (1.2%, [1.1-1.3])	1,536 (1.8%, [1.7-1.9])	11,970
	Numbness/tingling	9,748 (1.4%, [1.3-1.4])	4,725 (1.3%, [1.3-1.4])	3,214 (1.4%, [1.4-1.5])	1,283 (1.5%, [1.5-1.6])	1,800 (2.1%, [2-2.2])	20,770
	Heavy arms/legs	8,527 (1.2%, [1.2-1.2])	4,017 (1.1%, [1.1-1.2])	2,707 (1.2%, [1.1-1.2])	1,101 (1.3%, [1.2-1.4])	1,916 (2.2%, [2.1-2.3])	18,268
	Muscle aches	21,342 (3%, [3-3])	11,069 (3.1%, [3-3.1])	7,488 (3.3%, [3.2-3.3])	3,537 (4.3%, [4.1-4.4])	5,126 (5.9%, [5.8-6.1])	48,562
First reported symptoms	Loss or change of sense of smell	873 (0.1%, [0.1-0.1])	570 (0.2%, [0.1-0.2])	451 (0.2%, [0.2-0.2])	235 (0.3%, [0.2-0.3])	239 (0.3%, [0.2-0.3])	2,368
	Loss or change of sense of taste	982 (0.1%, [0.1-0.1])	603 (0.2%, [0.2-0.2])	415 (0.2%, [0.2-0.2])	216 (0.3%, [0.2-0.3])	251 (0.3%, [0.3-0.3])	2,467
	New persistent cough	1,852 (0.3%, [0.2-0.3])	947 (0.3%, [0.2-0.3])	1,314 (0.6%, [0.5-0.6])	799 (1%, [0.9-1])	1,255 (1.5%, [1.4-1.5])	6,167

Table B.1 continued from previous page

Variable	Category	Rounds 2-7 (Wild type)	Rounds 8-10 (Alpha)	Rounds 13-15 (Delta)	Round 17 (BA.1 Omicron)	Rounds 19 (BA.2 Omicron)	Pooled
	Fever	1,984 (0.3%, [0.3-0.3])	827 (0.2%, [0.2-0.2])	716 (0.3%, [0.3-0.3])	339 (0.4%, [0.4-0.5])	598 (0.7%, [0.6-0.8])	4,464
	Runny nose	7,718 (1.1%, [1.1-1.1])	3,817 (1.1%, [1-1.1])	3,362 (1.5%, [1.4-1.5])	2,069 (2.5%, [2.4-2.6])	2,710 (3.1%, [3-3.3])	19,676
	Sneezing	7,783 (1.1%, [1.1-1.1])	3,486 (1%, [0.9-1])	2,982 (1.3%, [1.3-1.3])	1,532 (1.8%, [1.8-1.9])	1,856 (2.2%, [2.1-2.3])	17,639
	Blocked nose	5,327 (0.7%, [0.7-0.8])	2,746 (0.8%, [0.7-0.8])	2,285 (1%, [1-1])	1,240 (1.5%, [1.4-1.6])	1,741 (2%, [1.9-2.1])	13,339
	Sore eyes	4,304 (0.6%, [0.6-0.6])	2,011 (0.6%, [0.5-0.6])	1,287 (0.6%, [0.5-0.6])	642 (0.8%, [0.7-0.8])	804 (0.9%, [0.9-1])	9,048
	Sore throat	12,758 (1.8%, [1.8-1.8])	4,150 (1.2%, [1.1-1.2])	5,203 (2.3%, [2.2-2.3])	2,394 (2.9%, [2.8-3])	3,717 (4.3%, [4.2-4.5])	28,222
	Hoarse voice	1,983 (0.3%, [0.3-0.3])	787 (0.2%, [0.2-0.2])	1,053 (0.5%, [0.4-0.5])	596 (0.7%, [0.7-0.8])	817 (0.9%, [0.9-1])	5,236
	Headache	14,719 (2.1%, [2-2.1])	8,209 (2.3%, [2.2-2.3])	4,873 (2.1%, [2.1-2.2])	2,724 (3.3%, [3.2-3.4])	3,337 (3.9%, [3.7-4])	33,862
	Dizziness	3,648 (0.5%, [0.5-0.5])	1,751 (0.5%, [0.5-0.5])	1,214 (0.5%, [0.5-0.6])	461 (0.6%, [0.5-0.6])	726 (0.8%, [0.8-0.9])	7,800
	Appetite loss	1,542 (0.2%, [0.2-0.2])	812 (0.2%, [0.2-0.2])	627 (0.3%, [0.3-0.3])	299 (0.4%, [0.3-0.4])	384 (0.4%, [0.4-0.5])	3,664
	Nausea/vomiting	2,198 (0.3%, [0.3-0.3])	1,060 (0.3%, [0.3-0.3])	839 (0.4%, [0.3-0.4])	322 (0.4%, [0.3-0.4])	495 (0.6%, [0.5-0.6])	4,914
	Diarrhoea	3,029 (0.4%, [0.4-0.4])	1,322 (0.4%, [0.3-0.4])	1,015 (0.4%, [0.4-0.5])	377 (0.5%, [0.4-0.5])	509 (0.6%, [0.5-0.6])	6,252
	Abdominal pain / belly ache	4,650 (0.7%, [0.6-0.7])	2,293 (0.6%, [0.6-0.7])	1,471 (0.6%, [0.6-0.7])	548 (0.7%, [0.6-0.7])	716 (0.8%, [0.8-0.9])	9,678
	Shortness of breath	4,548 (0.6%, [0.6-0.7])	2,315 (0.6%, [0.6-0.7])	1,541 (0.7%, [0.6-0.7])	784 (0.9%, [0.9-1])	963 (1.1%, [1-1.2])	10,151
	Tight chest	2,552 (0.4%, [0.3-0.4])	1,306 (0.4%, [0.3-0.4])	786 (0.3%, [0.3-0.4])	423 (0.5%, [0.5-0.6])	452 (0.5%, [0.5-0.6])	5,519
	Chest pain	1,147 (0.2%, [0.2-0.2])	667 (0.2%, [0.2-0.2])	424 (0.2%, [0.2-0.2])	184 (0.2%, [0.2-0.3])	295 (0.3%, [0.3-0.4])	2,717
	Chills	1,905 (0.3%, [0.3-0.3])	1,799 (0.5%, [0.5-0.5])	738 (0.3%, [0.3-0.3])	532 (0.6%, [0.6-0.7])	829 (1%, [0.9-1])	5,803
	Difficulty sleeping	8,466 (1.2%, [1.2-1.2])	4,798 (1.3%, [1.3-1.4])	2,798 (1.2%, [1.2-1.3])	1,400 (1.7%, [1.6-1.8])	1,494 (1.7%, [1.6-1.8])	18,956
	Tiredness	15,210 (2.1%, [2.1-2.2])	7,679 (2.1%, [2.1-2.2])	5,741 (2.5%, [2.4-2.6])	2,745 (3.3%, [3.2-3.4])	3,728 (4.3%, [4.2-4.5])	35,103
	Severe fatigue	1,711 (0.2%, [0.2-0.3])	1,041 (0.3%, [0.3-0.3])	820 (0.4%, [0.3-0.4])	369 (0.4%, [0.4-0.5])	542 (0.6%, [0.6-0.7])	4,483
	Numbness/tingling	2,772 (0.4%, [0.4-0.4])	1,366 (0.4%, [0.4-0.4])	951 (0.4%, [0.4-0.4])	367 (0.4%, [0.4-0.5])	475 (0.6%, [0.5-0.6])	5,931
	Heavy arms/legs	2,204 (0.3%, [0.3-0.3])	1,111 (0.3%, [0.3-0.3])	736 (0.3%, [0.3-0.3])	286 (0.3%, [0.3-0.4])	472 (0.5%, [0.5-0.6])	4,809
	Muscle aches	6,554 (0.9%, [0.9-0.9])	3,739 (1%, [1-1.1])	2,379 (1%, [1-1.1])	1,166 (1.4%, [1.3-1.5])	1,636 (1.9%, [1.8-2])	15,474

Table B.1: Characteristics of study population. N (percent [95% CI]) or mean (SD). Reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

Variable	Category	Rounds 2-7 (Wild type)	Rounds 8-10 (Alpha)	Rounds 12-15 (Delta)	BA.1 (Omicron)	BA.2 (Omicron)	Pooled
All participants	All participants	2,971 (100%, [99.9-100])	2,275 (100%, [99.9-100])	1,493 (100%, [99.8-100])	2,836 (100%, [99.9-100])	3,560 (100%, [99.9-100])	13,135
Sex	Female	1,683 (56.6%, [54.9-58.4])	1,283 (56.4%, [54.3-58.4])	809 (54.2%, [51.7-56.7])	1,618 (57.1%, [55.2-58.9])	2,043 (57.4%, [55.8-59])	7,436
	Male	1,288 (43.4%, [41.6-45.1])	992 (43.6%, [41.6-45.7])	684 (45.8%, [43.3-48.3])	1,218 (42.9%, [41.1-44.8])	1,517 (42.6%, [41-44.2])	5,699
Age group	5-17	0 (0%, [0-0.1])	0 (0%, [0-0.1])	0 (0%, [0-0.2])	0 (0%, [0-0.1])	0 (0%, [0-0.1])	0
	18-24	239 (8%, [7.1-9.1])	178 (7.8%, [6.8-9])	60 (4%, [3.1-5.1])	153 (5.4%, [4.6-6.3])	235 (6.6%, [5.8-7.5])	865
	25-34	401 (13.5%, [12.3-14.8])	289 (12.7%, [11.4-14.1])	125 (8.4%, [7.1-9.9])	390 (13.8%, [12.5-15.1])	665 (18.7%, [17.4-20])	1,870
	35-44	475 (16%, [14.7-17.3])	401 (17.6%, [16.1-19.2])	298 (20%, [18-22.1])	484 (17.1%, [15.7-18.5])	688 (19.3%, [18.1-20.7])	2,346
	45-54	642 (21.6%, [20.2-23.1])	493 (21.7%, [20-23.4])	371 (24.8%, [22.7-27.1])	541 (19.1%, [17.7-20.6])	555 (15.6%, [14.4-16.8])	2,602
	55-64	618 (20.8%, [19.4-22.3])	510 (22.4%, [20.8-24.2])	307 (20.6%, [18.6-22.7])	618 (21.8%, [20.3-23.3])	646 (18.1%, [16.9-19.4])	2,699
	65-74	429 (14.4%, [13.2-15.7])	299 (13.1%, [11.8-14.6])	247 (16.5%, [14.7-18.5])	471 (16.6%, [15.3-18])	554 (15.6%, [14.4-16.8])	2,000
	75+	167 (5.6%, [4.8-6.5])	105 (4.6%, [3.8-5.6])	85 (5.7%, [4.6-7])	179 (6.3%, [5.5-7.3])	217 (6.1%, [5.4-6.9])	753
Ethnicity	Asian	158 (5.4%, [4.6-6.3])	137 (6.1%, [5.2-7.2])	84 (5.7%, [4.6-7])	137 (4.9%, [4.2-5.8])	173 (5%, [4.3-5.7])	689
	Black	40 (1.4%, [1-1.9])	42 (1.9%, [1.4-2.5])	31 (2.1%, [1.5-3])	50 (1.8%, [1.4-2.4])	45 (1.3%, [1-1.7])	208
	Mixed	26 (0.9%, [0.6-1.3])	30 (1.3%, [0.9-1.9])	14 (1%, [0.6-1.6])	50 (1.8%, [1.4-2.4])	73 (2.1%, [1.7-2.6])	193
	Other	23 (0.8%, [0.5-1.2])	20 (0.9%, [0.6-1.4])	19 (1.3%, [0.8-2])	26 (0.9%, [0.6-1.4])	37 (1.1%, [0.8-1.5])	125
	White	2,676 (91.5%, [90.5-92.5])	2,002 (89.7%, [88.4-90.9])	1,319 (89.9%, [88.3-91.3])	2,510 (90.5%, [89.4-91.6])	3,149 (90.6%, [89.5-91.5])	11,656
PCR positive at time of survey	Yes	2,971 (100%, [99.9-100])	2,275 (100%, [99.9-100])	1,493 (100%, [99.8-100])	2,836 (100%, [99.9-100])	3,560 (100%, [99.9-100])	13,135
Prior COVID-19 infection	No previous infection	1,840 (61.9%, [60.2-63.7])	1,011 (44.4%, [42.4-46.5])	563 (37.7%, [35.3-40.2])	816 (28.8%, [27.1-30.5])	877 (24.6%, [23.2-26.1])	5,107
	Previous infection greater than 28 days	35 (1.2%, [0.8-1.6])	86 (3.8%, [3.1-4.6])	51 (3.4%, [2.6-4.5])	148 (5.2%, [4.5-6.1])	198 (5.6%, [4.9-6.4])	518
	Previous infection with unknown time	3 (0.1%, [0-0.3])	4 (0.2%, [0.1-0.5])	0 (0%, [0-0.2])	2 (0.1%, [0-0.3])	1 (0%, [0-0.2])	10
	Previous infection within 28 days	586 (19.7%, [18.3-21.2])	766 (33.7%, [31.8-35.6])	722 (48.4%, [45.8-50.9])	1,666 (58.7%, [56.9-60.5])	2,188 (61.5%, [59.9-63])	5,928
	Suspected previous infection	500 (16.8%, [15.5-18.2])	403 (17.7%, [16.2-19.3])	152 (10.2%, [8.7-11.8])	200 (7.1%, [6.2-8.1])	291 (8.2%, [7.3-9.1])	1,546
	Unkown	7 (0.2%, [0.1-0.5])	5 (0.2%, [0.1-0.5])	5 (0.3%, [0.1-0.8])	4 (0.1%, [0.1-0.4])	5 (0.1%, [0.1-0.3])	26
Vaccination status	NA	-	-	11 (1.5%, [0.8-2.7])	36 (1.3%, [0.9-1.8])	33 (0.9%, [0.7-1.3])	80
	Not vaccinated	-	-	28 (3.8%, [2.6-5.4])	77 (2.7%, [2.2-3.4])	106 (3%, [2.5-3.6])	211
	One dose	-	-	11 (1.5%, [0.8-2.7])	15 (0.5%, [0.3-0.9])	23 (0.6%, [0.4-1])	49
	Three dose	-	-	73 (9.9%, [8-12.3])	2,276 (80.3%, [78.7-81.7])	2,925 (82.2%, [80.9-83.4])	5,274
	Two dose	-	-	579 (78.6%, [75.5-81.4])	285 (10%, [9-11.2])	265 (7.4%, [6.6-8.4])	1,129
	Unknown dose	-	-	35 (4.7%, [3.4-6.5])	147 (5.2%, [4.4-6.1])	208 (5.8%, [5.1-6.7])	390
Symptoms	Symptomatic	1,338 (45%, [43.3-46.8])	1,245 (54.7%, [52.7-56.8])	952 (63.8%, [61.3-66.2])	1,985 (70%, [68.3-71.7])	2,701 (75.9%, [74.4-77.2])	8,221
Number of reported symptoms	Mean (SD)	2.7 (4.1)	3.38 (4.5)	4.63 (5.21)	4.64 (4.72)	6.02 (5.26)	4.15 (4.83)
Sought medical attention for symptoms	No	1,187 (88.7%, [86.9-90.3])	1,091 (87.7%, [85.8-89.4])	850 (89.5%, [87.4-91.3])	1,866 (94.1%, [93-95.1])	2,522 (93.4%, [92.4-94.3])	7,516
	Yes	151 (11.3%, [9.7-13.1])	153 (12.3%, [10.6-14.2])	100 (10.5%, [8.7-12.6])	117 (5.9%, [4.9-7])	179 (6.6%, [5.7-7.6])	700
Symptoms affecting day-to-day activities	A lot	0 (0%, [0-0.1])	0 (0%, [0-0.1])	157 (10.5%, [9.1-12.2])	303 (10.7%, [9.6-11.9])	625 (17.6%, [16.3-18.8])	1,085
	A little	0 (0%, [0-0.1])	0 (0%, [0-0.1])	283 (19%, [17-21])	650 (22.9%, [21.4-24.5])	948 (26.6%, [25.2-28.1])	1,881
	Not at all	0 (0%, [0-0.1])	0 (0%, [0-0.1])	83 (5.6%, [4.5-6.8])	282 (9.9%, [8.9-11.1])	266 (7.5%, [6.7-8.4])	631

Table B.2 continued from previous page

Variable	Category	Rounds 2-7 (Wild type)	Rounds 8-10 (Alpha)	Rounds 12-15 (Delta)	BA.1 (Omicron)	BA.2 (Omicron)	Pooled
	Don't know / PNA / Non-reponse	1,338 (45%, [43.3-46.8])	1,245 (54.7%, [52.7-56.8])	429 (28.7%, [26.5-31.1])	750 (26.4%, [24.9-28.1])	862 (24.2%, [22.8-25.6])	4,624
	No symptoms reported	1,633 (55%, [53.2-56.7])	1,030 (45.3%, [43.2-47.3])	541 (36.2%, [33.8-38.7])	851 (30%, [28.3-31.7])	859 (24.1%, [22.8-25.6])	4,914
Days since onset of first symptom	Mean (SD)	7.08 (3.19)	7.58 (3.14)	6.85 (3.19)	6.62 (3.04)	6.06 (2.94)	7.06 (3.16)
Symptoms reported	Loss or change of sense of smell	399 (13.4%, [12.3-14.7])	323 (14.2%, [12.8-15.7])	367 (24.6%, [22.5-26.8])	214 (7.5%, [6.6-8.6])	349 (9.8%, [8.9-10.8])	1,652
	Loss or change of sense of taste	355 (11.9%, [10.8-13.2])	329 (14.5%, [13.1-16])	340 (22.8%, [20.7-25])	258 (9.1%, [8.1-10.2])	408 (11.5%, [10.5-12.5])	1,690
	New persistent cough	260 (8.8%, [7.8-9.8])	326 (14.3%, [12.9-15.8])	380 (25.5%, [23.3-27.7])	703 (24.8%, [23.2-26.4])	1,249 (35.1%, [33.5-36.7])	2,918
	Fever	292 (9.8%, [8.8-11])	262 (11.5%, [10.3-12.9])	305 (20.4%, [18.5-22.5])	498 (17.6%, [16.2-19])	922 (25.9%, [24.5-27.4])	2,279
	Runny nose	389 (13.1%, [11.9-14.4])	384 (16.9%, [15.4-18.5])	505 (33.8%, [31.5-36.3])	1,147 (40.4%, [38.7-42.3])	1,825 (51.3%, [49.6-52.9])	4,250
	Sneezing	411 (13.8%, [12.6-15.1])	404 (17.8%, [16.2-19.4])	436 (29.2%, [27-31.6])	1,018 (35.9%, [34.2-37.7])	1,572 (44.2%, [42.5-45.8])	3,841
	Blocked nose	387 (13%, [11.9-14.3])	401 (17.6%, [16.1-19.2])	402 (26.9%, [24.7-29.2])	805 (28.4%, [26.8-30.1])	1,383 (38.8%, [37.3-40.5])	3,378
	Sore eyes	240 (8.1%, [7.2-9.1])	215 (9.5%, [8.3-10.7])	199 (13.3%, [11.7-15.1])	358 (12.6%, [11.5-13.9])	568 (16%, [14.8-17.2])	1,580
	Sore throat	409 (13.8%, [12.6-15.1])	386 (17%, [15.5-18.6])	374 (25.1%, [22.9-27.3])	1,074 (37.9%, [36.1-39.7])	1,734 (48.7%, [47.1-50.4])	3,977
	Hoarse voice	194 (6.5%, [5.7-7.5])	184 (8.1%, [7-9.3])	221 (14.8%, [13.1-16.7])	631 (22.2%, [20.8-23.8])	1,043 (29.3%, [27.8-30.8])	2,273
	Headache	740 (24.9%, [23.4-26.5])	670 (29.5%, [27.6-31.4])	553 (37%, [34.6-39.5])	1,131 (39.9%, [38.1-41.7])	1,684 (47.3%, [45.7-48.9])	4,778
	Dizziness	237 (8%, [7.1-9])	223 (9.8%, [8.6-11.1])	187 (12.5%, [10.9-14.3])	309 (10.9%, [9.8-12.1])	536 (15.1%, [13.9-16.3])	1,492
	Appetite loss	308 (10.4%, [9.3-11.5])	319 (14%, [12.7-15.5])	238 (15.9%, [14.2-17.9])	293 (10.3%, [9.3-11.5])	562 (15.8%, [14.6-17])	1,720
	Nausea/vomiting	153 (5.1%, [4.4-6])	141 (6.2%, [5.3-7.3])	107 (7.2%, [6-8.6])	185 (6.5%, [5.7-7.5])	289 (8.1%, [7.3-9.1])	875
	Diarrhoea	180 (6.1%, [5.3-7])	171 (7.5%, [6.5-8.7])	122 (8.2%, [6.9-9.7])	209 (7.4%, [6.5-8.4])	280 (7.9%, [7-8.8])	962
	Abdominal pain / belly ache	180 (6.1%, [5.3-7])	165 (7.3%, [6.3-8.4])	103 (6.9%, [5.7-8.3])	192 (6.8%, [5.9-7.8])	328 (9.2%, [8.3-10.2])	968
	Shortness of breath	223 (7.5%, [6.6-8.5])	232 (10.2%, [9-11.5])	185 (12.4%, [10.8-14.2])	353 (12.4%, [11.3-13.7])	575 (16.2%, [15-17.4])	1,568
	Tight chest	204 (6.9%, [6-7.8])	203 (8.9%, [7.8-10.2])	139 (9.3%, [7.9-10.9])	271 (9.6%, [8.5-10.7])	423 (11.9%, [10.9-13])	1,240
	Chest pain	83 (2.8%, [2.3-3.4])	87 (3.8%, [3.1-4.7])	50 (3.3%, [2.5-4.4])	120 (4.2%, [3.6-5])	202 (5.7%, [5-6.5])	542
	Chills	309 (10.4%, [9.4-11.5])	330 (14.5%, [13.1-16])	252 (16.9%, [15.1-18.9])	496 (17.5%, [16.1-18.9])	880 (24.7%, [23.3-26.2])	2,267
	Difficulty sleeping	349 (11.7%, [10.6-13])	323 (14.2%, [12.8-15.7])	263 (17.6%, [15.8-19.6])	520 (18.3%, [17-19.8])	788 (22.1%, [20.8-23.5])	2,243
	Tiredness	727 (24.5%, [23-26])	700 (30.8%, [28.9-32.7])	534 (35.8%, [33.4-38.2])	1,074 (37.9%, [36.1-39.7])	1,658 (46.6%, [44.9-48.2])	4,693
	Severe fatigue	149 (5%, [4.3-5.9])	167 (7.3%, [6.3-8.5])	118 (7.9%, [6.6-9.4])	168 (5.9%, [5.1-6.9])	331 (9.3%, [8.4-10.3])	933
	Numbness/tingling	110 (3.7%, [3.1-4.4])	90 (4%, [3.2-4.8])	58 (3.9%, [3-5])	126 (4.4%, [3.7-5.3])	214 (6%, [5.3-6.8])	598
	Heavy arms/legs	196 (6.6%, [5.8-7.5])	193 (8.5%, [7.4-9.7])	100 (6.7%, [5.5-8.1])	229 (8.1%, [7.1-9.1])	419 (11.8%, [10.8-12.9])	1,137
	Muscle aches	549 (18.5%, [17.1-19.9])	463 (20.4%, [18.7-22.1])	372 (24.9%, [22.8-27.2])	767 (27%, [25.4-28.7])	1,198 (33.7%, [32.1-35.2])	3,349
First reported symptoms	Loss or change of sense of smell	109 (3.7%, [3.1-4.4])	82 (3.6%, [2.9-4.5])	73 (4.9%, [3.9-6.1])	36 (1.3%, [0.9-1.8])	46 (1.3%, [1-1.7])	346
	Loss or change of sense of taste	96 (3.2%, [2.7-3.9])	83 (3.6%, [3-4.5])	66 (4.4%, [3.5-5.6])	38 (1.3%, [1-1.8])	59 (1.7%, [1.3-2.1])	342
	New persistent cough	125 (4.2%, [3.5-5])	148 (6.5%, [5.6-7.6])	168 (11.3%, [9.7-13])	255 (9%, [8-10.1])	375 (10.5%, [9.6-11.6])	1,071
	Fever	112 (3.8%, [3.1-4.5])	104 (4.6%, [3.8-5.5])	83 (5.6%, [4.5-6.8])	133 (4.7%, [4-5.5])	232 (6.5%, [5.8-7.4])	664
	Runny nose	129 (4.3%, [3.7-5.1])	110 (4.8%, [4-5.8])	176 (11.8%, [10.2-13.5])	363 (12.8%, [11.6-14.1])	561 (15.8%, [14.6-17])	1,339
	Sneezing	125 (4.2%, [3.5-5])	122 (5.4%, [4.5-6.4])	146 (9.8%, [8.4-11.4])	304 (10.7%, [9.6-11.9])	442 (12.4%, [11.4-13.5])	1,139

Table B.2 continued from previous page

Variable	Category	Rounds 2-7 (Wild type)	Rounds 8-10 (Alpha)	Rounds 12-15 (Delta)	BA.1 (Omicron)	BA.2 (Omicron)	Pooled
	Blocked nose	120 (4%, [3.4-4.8])	138 (6.1%, [5.2-7.1])	98 (6.6%, [5.4-7.9])	203 (7.2%, [6.3-8.2])	304 (8.5%, [7.7-9.5])	863
	Sore eyes	77 (2.6%, [2.1-3.2])	65 (2.9%, [2.2-3.6])	42 (2.8%, [2.1-3.8])	91 (3.2%, [2.6-3.9])	147 (4.1%, [3.5-4.8])	422
	Sore throat	206 (6.9%, [6.1-7.9])	183 (8%, [7-9.2])	186 (12.5%, [10.9-14.2])	621 (21.9%, [20.4-23.5])	1,084 (30.4%, [29-32])	2,280
	Hoarse voice	50 (1.7%, [1.3-2.2])	38 (1.7%, [1.2-2.3])	59 (4%, [3.1-5.1])	159 (5.6%, [4.8-6.5])	212 (6%, [5.2-6.8])	518
	Headache	384 (12.9%, [11.8-14.2])	340 (14.9%, [13.5-16.5])	250 (16.7%, [14.9-18.7])	508 (17.9%, [16.5-19.4])	716 (20.1%, [18.8-21.5])	2,198
	Dizziness	76 (2.6%, [2-3.2])	65 (2.9%, [2.2-3.6])	41 (2.7%, [2-3.7])	64 (2.3%, [1.8-2.9])	116 (3.3%, [2.7-3.9])	362
	Appetite loss	69 (2.3%, [1.8-2.9])	78 (3.4%, [2.8-4.3])	48 (3.2%, [2.4-4.2])	49 (1.7%, [1.3-2.3])	85 (2.4%, [1.9-2.9])	329
	Nausea/vomiting	38 (1.3%, [0.9-1.8])	30 (1.3%, [0.9-1.9])	28 (1.9%, [1.3-2.7])	44 (1.6%, [1.2-2.1])	81 (2.3%, [1.8-2.8])	221
	Diarrhoea	47 (1.6%, [1.2-2.1])	46 (2%, [1.5-2.7])	30 (2%, [1.4-2.9])	46 (1.6%, [1.2-2.2])	69 (1.9%, [1.5-2.4])	238
	Abdominal pain / belly ache	58 (2%, [1.5-2.5])	44 (1.9%, [1.4-2.6])	18 (1.2%, [0.8-1.9])	43 (1.5%, [1.1-2])	70 (2%, [1.6-2.5])	233
	Shortness of breath	69 (2.3%, [1.8-2.9])	67 (2.9%, [2.3-3.7])	32 (2.1%, [1.5-3])	77 (2.7%, [2.2-3.4])	109 (3.1%, [2.5-3.7])	354
	Tight chest	45 (1.5%, [1.1-2])	48 (2.1%, [1.6-2.8])	31 (2.1%, [1.5-2.9])	51 (1.8%, [1.4-2.4])	77 (2.2%, [1.7-2.7])	252
	Chest pain	19 (0.6%, [0.4-1])	11 (0.5%, [0.3-0.9])	10 (0.7%, [0.4-1.2])	20 (0.7%, [0.5-1.1])	36 (1%, [0.7-1.4])	96
	Chills	103 (3.5%, [2.9-4.2])	131 (5.8%, [4.9-6.8])	60 (4%, [3.1-5.1])	139 (4.9%, [4.2-5.8])	271 (7.6%, [6.8-8.5])	704
	Difficulty sleeping	114 (3.8%, [3.2-4.6])	104 (4.6%, [3.8-5.5])	77 (5.2%, [4.1-6.4])	118 (4.2%, [3.5-5])	184 (5.2%, [4.5-5.9])	597
	Tiredness	314 (10.6%, [9.5-11.7])	264 (11.6%, [10.4-13])	202 (13.5%, [11.9-15.4])	385 (13.6%, [12.4-14.9])	602 (16.9%, [15.7-18.2])	1,767
	Severe fatigue	57 (1.9%, [1.5-2.5])	49 (2.2%, [1.6-2.8])	35 (2.3%, [1.7-3.2])	48 (1.7%, [1.3-2.2])	93 (2.6%, [2.1-3.2])	282
	Numbness/tingling	30 (1%, [0.7-1.4])	25 (1.1%, [0.7-1.6])	12 (0.8%, [0.5-1.4])	31 (1.1%, [0.8-1.5])	37 (1%, [0.8-1.4])	135
	Heavy arms/legs	58 (2%, [1.5-2.5])	53 (2.3%, [1.8-3])	21 (1.4%, [0.9-2.1])	44 (1.6%, [1.2-2.1])	86 (2.4%, [2-3])	262
	Muscle aches	249 (8.4%, [7.4-9.4])	197 (8.7%, [7.6-9.9])	121 (8.1%, [6.8-9.6])	261 (9.2%, [8.2-10.3])	378 (10.6%, [9.6-11.7])	1,206

Table B.2: Characteristics of swab positives in study population. N (percent [95% CI]) or mean (SD). NOTE: Individuals with Wild-type, Alpha and Delta variants were inferred by the time of infection; those with BA.2 and BA.1 variants were identified using sequencing. For this reason, the sum total of PCR positives with Omicron is lower in this table than the total reported in Table B.1, with the difference being those who either were not sequenced from rounds 17–19, or those who were sequenced but did not have BA.2 or BA.1. Table reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

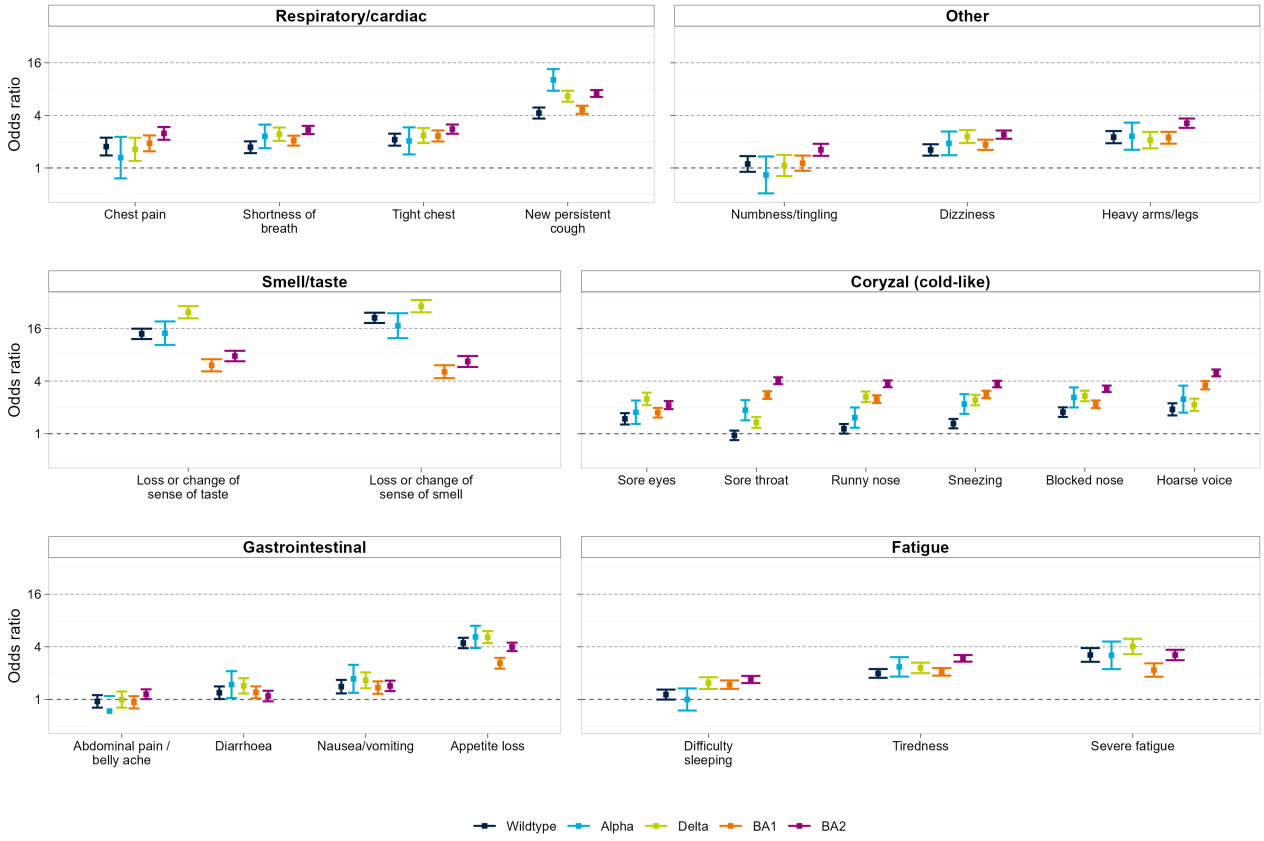


Figure B.1: Sensitivity analysis of the univariable symptom profiling among N=1,542,510 participants, additionally adjusted for time since symptom onset (as a categorical variable, 0–4, 5–10,11+ days). Error bars show 95% confidence intervals. Odds ratios are attenuated but the patterns are largely consistent with the main analysis. Reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

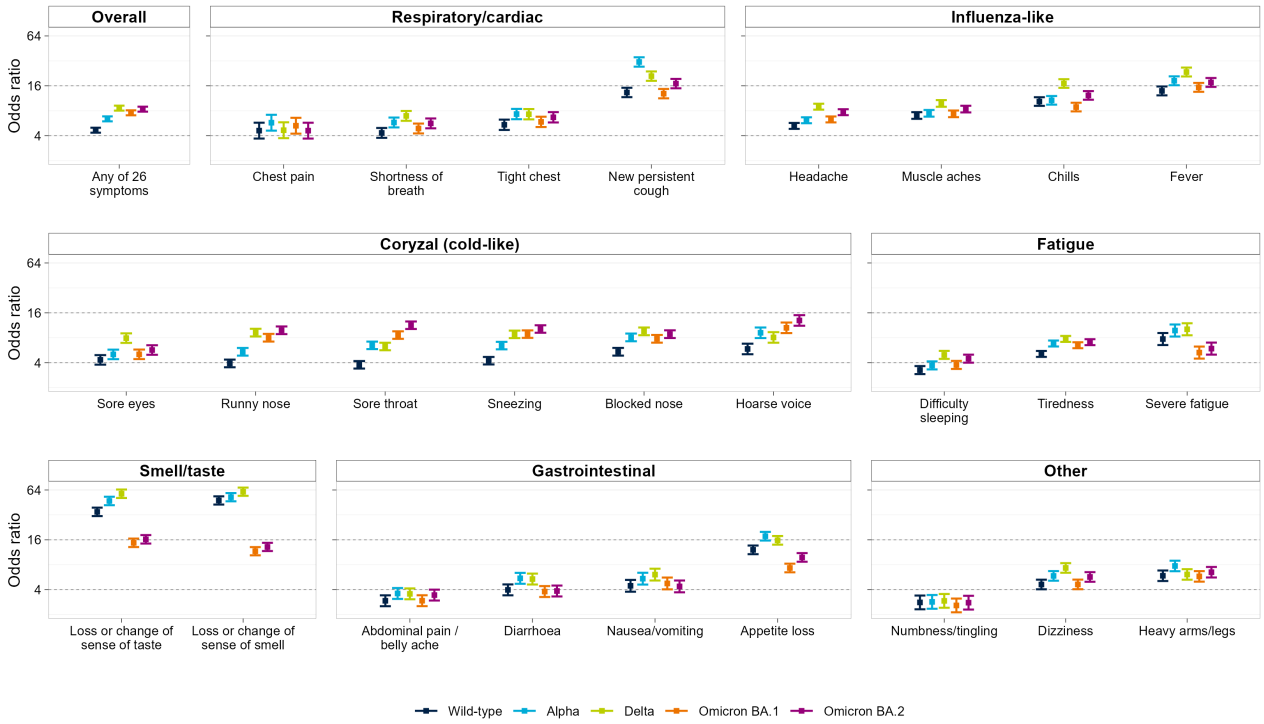


Figure B.2: Pooled logistic regression analysis of N=1,542,510 participants across five variant-phases. For each of the symptoms, a logistic model was fit with this specification: y[PCR pos/neg] covariates [age, sex, vaccination, calendar time] + variant [binary] + symptom [y/n] + variant:symptom[interaction] For the reference category variant (Wild-type) we report the odds ratio on the symptom. For subsequent variants, we report the product of the symptom odds ratio and the interaction effect with the relevant variant. Error bars show 95% confidence intervals. The patterns of changes in odds ratios between variants are consistent with the patterns shown in the stratified logistic regression analysis in Figure 2.: ORs are higher for BA.2 than BA.1 for nearly all symptoms. Fever and cough have the highest ORs for BA.2 and BA.1, while loss or change of smell or taste have the highest ORs in all previous variants. Reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

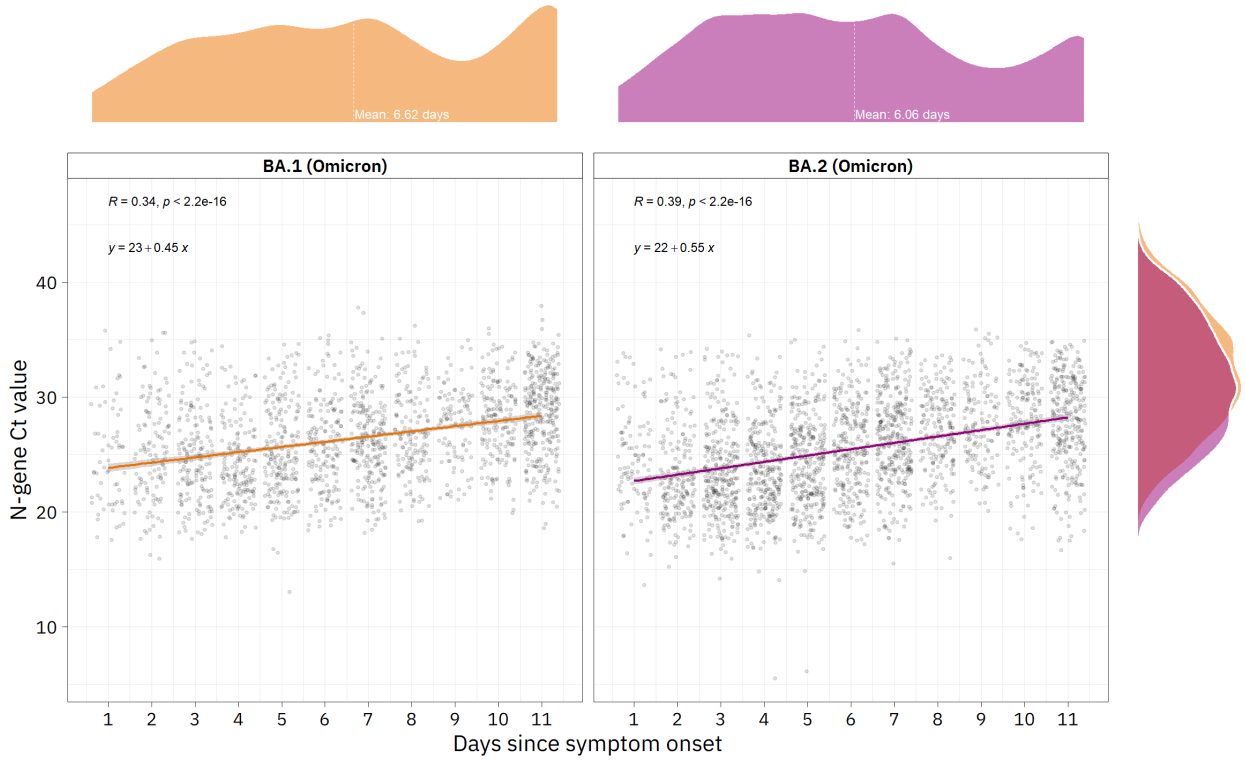


Figure B.3: Scatter plot showing (y-axis) N-gene Ct values among symptomatic swab-positive individuals and (x-axis) time (in days) since symptom onset, for BA.1 and BA.2. For each variant, a linear regression model is fitted with Ct value as the dependent variable and days since symptom onset (between 1 and 11 or more) as the predictor variable. Pearson's R value shown, with a p-value derived from a two-sided test with H0: r=0. For both variants there is a positive association between time since symptom onset and Ct value. Reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

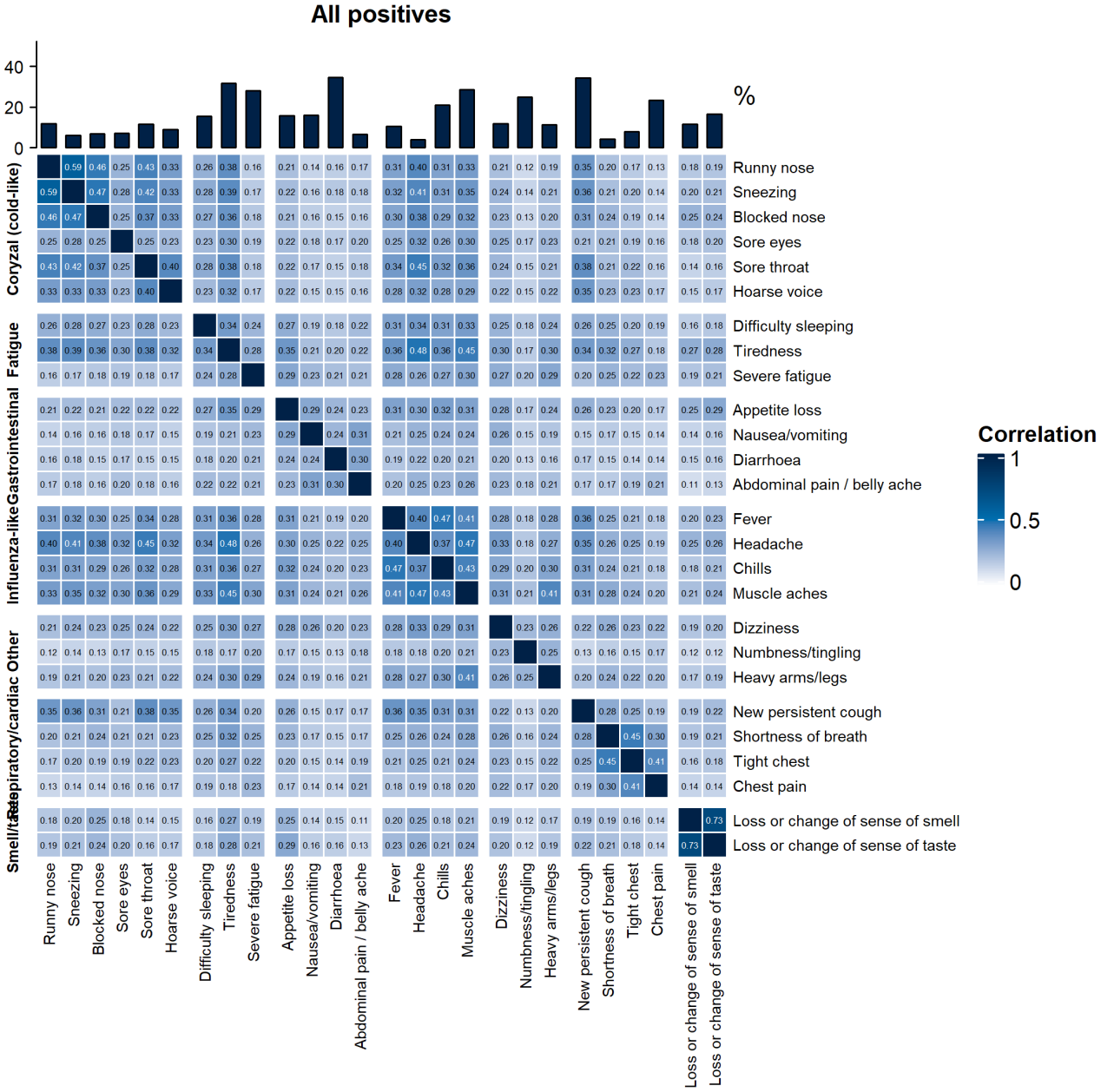


Figure B.4: Heatmaps showing co-occurrence of symptoms across variants among swab-positive individuals. Denominators for percentages are the number of swab positive individuals with each variant. Reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

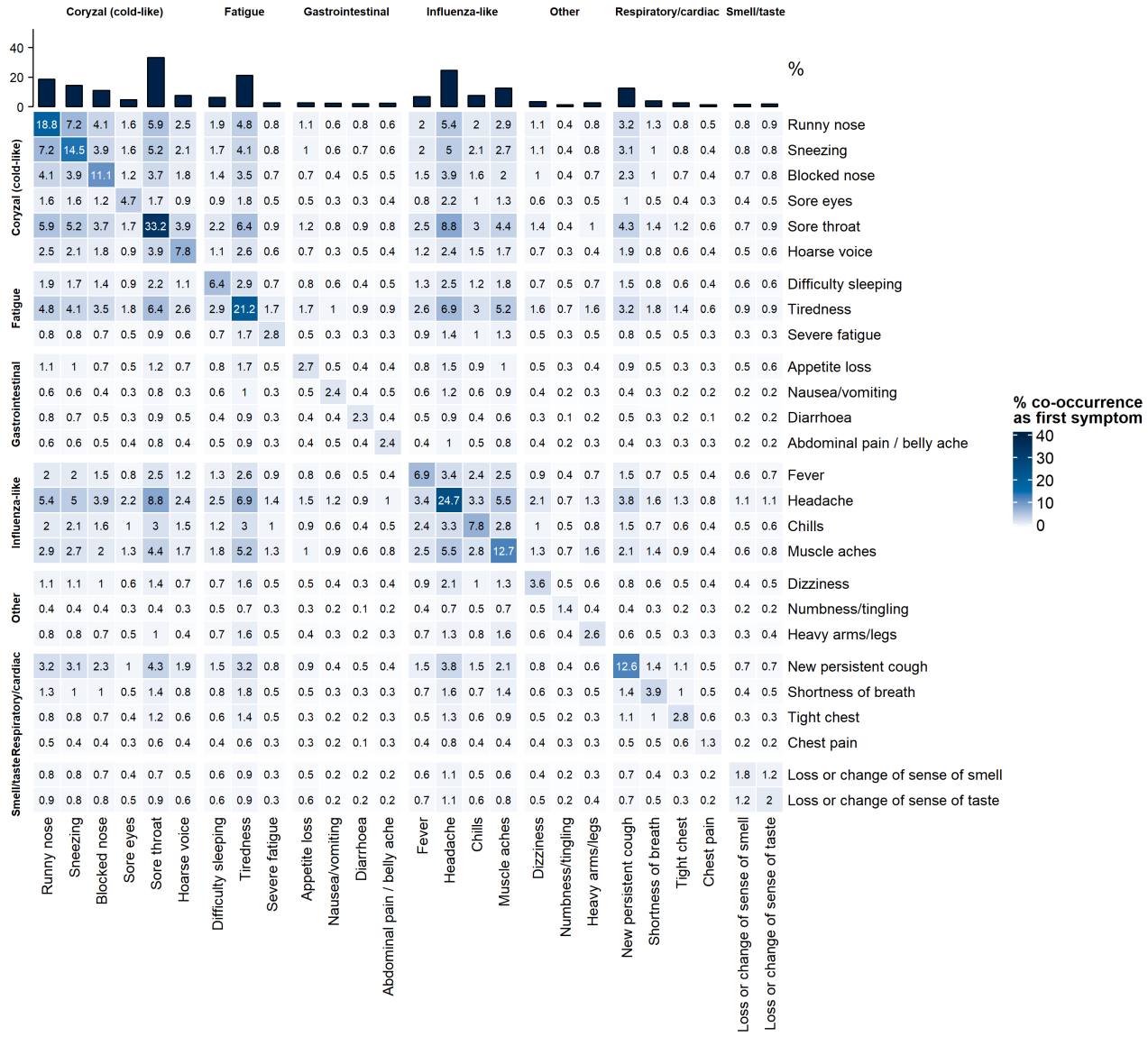


Figure B.5: Heatmap showing co-occurrence of first symptoms among swab positive individuals in rounds 17–19 (5 January to 31 March 2022), when Omicron was dominant. The denominator for percentages is the number of swab positive individuals who reported experiencing one or more of 26 symptoms in the week prior to testing in rounds 17–19 (n=7,176) of REACT-1. Sore throat, headache, and tiredness are the most common first symptoms and are the most commonly co-occurring pairs of first symptoms. Reproduced with modification from Whitaker *et al*, *Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England*, Nature Communications 2022 [247].

Appendix C

Appendix: Persistent symptoms

Variable	Category	total	Number with 1 or more symptoms (weighted)	Number with 2 or more symptoms (weighted)	Number with 3 or more symptoms (weighted)	Prevalence 1 or more symptoms (weighted)	Prevalence 2 or more symptoms (weighted)	Prevalence 3 or more symptoms (weighted)
All participants	All participants	508707	26943	16368	10156	5.411 (5.350,5.474)	3.265 (3.216,3.314)	2.032 (1.993,2.071)
Sex	Male	223861	9942	5713	3369	4.602 (4.520,4.685)	2.612 (2.550,2.675)	1.541 (1.494,1.591)
	Female	284840	17001	10655	6787	6.186 (6.094,6.280)	3.890 (3.816,3.965)	2.500 (2.441,2.561)
Age group	18-24	28655	1463	832	488	5.379 (5.194,5.570)	3.000 (2.861,3.145)	1.749 (1.643,1.862)
	25-34	65159	3551	1912	1105	5.650 (5.499,5.805)	2.995 (2.884,3.110)	1.707 (1.623,1.794)
	35-44	85044	4873	2814	1633	5.770 (5.612,5.931)	3.336 (3.215,3.461)	1.953 (1.861,2.050)
	45-54	103478	6428	3965	2461	6.383 (6.224,6.545)	3.959 (3.833,4.089)	2.478 (2.378,2.582)
	55-64	105361	6035	3830	2444	5.922 (5.757,6.091)	3.826 (3.693,3.964)	2.477 (2.370,2.590)
	65-74	85297	3425	2202	1457	4.339 (4.185,4.499)	2.824 (2.699,2.955)	1.884 (1.782,1.992)
	74+	35713	1168	813	568	3.466 (3.314,3.623)	2.452 (2.325,2.586)	1.775 (1.666,1.890)
Ethnicity	Asian	17666	757	447	257	4.710 (4.496,4.934)	2.622 (2.462,2.792)	1.508 (1.387,1.639)
	Black	4091	259	126	75	6.174 (5.778,6.596)	3.140 (2.857,3.450)	1.773 (1.562,2.011)
	Mixed	6112	457	287	186	7.232 (6.588,7.933)	4.430 (3.926,4.995)	2.872 (2.469,3.337)
	Other	4434	283	167	110	6.381 (5.766,7.057)	3.843 (3.367,4.384)	2.556 (2.172,3.007)
	White	472750	24962	15200	9433	5.404 (5.338,5.471)	3.294 (3.242,3.347)	2.062 (2.020,2.104)
Adiposity	Underweight	6273	315	203	136	5.643 (5.150,6.181)	3.503 (3.115,3.937)	2.050 (1.757,2.391)
	Normal weight	183305	8271	4783	2807	4.674 (4.578,4.771)	2.679 (2.606,2.753)	1.563 (1.507,1.621)
	Overweight	161456	8356	4982	3058	5.368 (5.258,5.481)	3.211 (3.125,3.299)	1.995 (1.927,2.065)
	Obese	97597	6744	4360	2854	6.941 (6.782,7.105)	4.371 (4.243,4.503)	2.889 (2.785,2.998)
Healthcare or care home worker	Yes	27690	2133	1343	835	7.806 (7.493,8.130)	4.763 (4.516,5.022)	2.952 (2.757,3.159)
Index of multiple deprivation (IMD) quintile	1 - most deprived	50260	3288	2153	1435	6.452 (6.300,6.608)	4.112 (3.989,4.238)	2.700 (2.600,2.803)
	2	80894	4780	2984	1953	5.876 (5.737,6.020)	3.526 (3.417,3.639)	2.280 (2.192,2.372)
	3	109928	5944	3613	2250	5.214 (5.081,5.351)	3.133 (3.029,3.240)	1.908 (1.826,1.992)
	4	126237	6293	3747	2292	4.908 (4.776,5.042)	2.877 (2.776,2.982)	1.750 (1.671,1.833)
	5 - least deprived	141388	6638	3871	2226	4.604 (4.475,4.736)	2.679 (2.580,2.782)	1.521 (1.447,1.599)
Current smoker	Current cigarette smoker	50983	3404	2232	1513	6.772 (6.569,6.982)	4.329 (4.164,4.499)	2.923 (2.788,3.065)
	Not current cigarette smoker	453681	23301	13989	8551	5.233 (5.168,5.299)	3.123 (3.073,3.175)	1.916 (1.876,1.957)
	Prefer not to say	4043	238	147	92	5.880 (5.230,6.605)	3.883 (3.357,4.487)	2.240 (1.847,2.715)
Current vaper	Current vaper	28436	1905	1205	778	6.940 (6.663,7.229)	4.326 (4.105,4.558)	2.740 (2.564,2.928)
	Not current vaper	477729	24899	15074	9316	5.311 (5.247,5.375)	3.192 (3.142,3.242)	1.984 (1.945,2.024)
	Prefer not to say	2542	139	89	62	5.542 (4.765,6.437)	3.846 (3.204,4.611)	2.218 (1.740,2.822)
Severity of COVID-19 symptoms	No symptoms	416589	0	0	0	0.000 (0.000,0.001)	0.000 (0.000,0.001)	0.000 (0.000,0.001)
	Mild symptoms	27572	5147	2554	1268	17.565 (17.145,17.994)	8.653 (8.345,8.972)	4.281 (4.061,4.513)
	Moderate symptoms	45357	13610	8113	4904	29.344 (28.934,29.756)	17.303 (16.964,17.647)	10.563 (10.289,10.844)
	Severe symptoms	19187	8186	5701	3984	41.742 (41.057,42.431)	29.164 (28.535,29.801)	20.355 (19.800,20.922)
Treatment sought/ received for COVID-19	No medical attention sought	58330	14222	7862	4499	23.302 (22.970,23.637)	12.754 (12.494,13.020)	7.339 (7.136,7.547)
	Sought medical attention from pharmacist / by phone (NHS 111/GP)	21388	8196	5511	3712	36.996 (36.379,37.618)	24.574 (24.026,25.131)	16.735 (16.261,17.219)

Table C.1 continued from previous page

Variable	Category	total	Number with 1 or more symptoms (weighted)	Number with 2 or more symptoms (weighted)	Number with 3 or more symptoms (weighted)	Prevalence 1 or more symptoms (weighted)	Prevalence 2 or more symptoms (weighted)	Prevalence 3 or more symptoms (weighted)
	Visited GP/walk-in centre/A&E	6792	2556	1690	1089	36.612 (35.472,37.767)	24.708 (23.696,25.750)	15.640 (14.795,16.525)
	Hospital admission	739	386	291	214	50.417 (47.040,53.791)	37.902 (34.682,41.233)	27.294 (24.389,30.407)
Gross household income	0-14,999	32819	2205	1570	1132	6.755 (6.506,7.013)	4.802 (4.591,5.023)	3.513 (3.332,3.704)
	15,000-49,999	148039	8280	5110	3201	5.768 (5.650,5.888)	3.541 (3.448,3.637)	2.189 (2.115,2.265)
	50,000-149,999	128957	6520	3664	2041	5.180 (5.057,5.306)	2.827 (2.735,2.922)	1.570 (1.502,1.642)
	>150,000	16279	801	388	197	5.076 (4.729,5.447)	2.312 (2.078,2.570)	1.170 (1.007,1.359)
Organ transplant recipient	Organ transplant recipient	596	34	27	18	5.219 (3.701,7.312)	4.202 (2.862,6.129)	3.361 (2.186,5.135)
Diabetes (type I or II)	Diabetes (type I or II)	27358	1534	1054	780	5.650 (5.388,5.925)	3.819 (3.602,4.048)	2.949 (2.758,3.152)
Heart disease or heart problems	Heart disease or heart problems	21430	1459	1038	766	6.831 (6.501,7.176)	4.909 (4.628,5.206)	3.644 (3.401,3.902)
Hypertension	Hypertension	65792	3838	2576	1767	5.908 (5.726,6.095)	4.018 (3.867,4.174)	2.819 (2.692,2.951)
Stroke	Stroke	3352	200	141	103	5.862 (5.131,6.690)	3.975 (3.376,4.674)	3.030 (2.512,3.652)
Kidney disease	Kidney disease	3867	306	223	158	7.923 (7.123,8.806)	5.753 (5.070,6.522)	3.936 (3.374,4.588)
Liver disease	Liver disease	2107	169	125	100	8.230 (7.151,9.457)	5.670 (4.776,6.719)	4.527 (3.732,5.481)
Anemia	Anemia	7575	779	560	417	10.744 (10.107,11.415)	7.459 (6.924,8.033)	5.546 (5.082,6.049)
Asthma	Asthma	51337	4213	2947	2033	8.315 (8.082,8.555)	5.737 (5.540,5.939)	3.960 (3.796,4.131)
Other lung condition	Other lung condition	12500	1082	806	605	8.791 (8.318,9.289)	6.506 (6.096,6.942)	5.061 (4.698,5.450)
Cancer	Cancer	8042	391	248	184	5.024 (4.561,5.531)	3.242 (2.872,3.659)	2.550 (2.223,2.924)
Neurological condition	Neurological condition	5415	363	285	222	6.883 (6.238,7.588)	5.236 (4.673,5.862)	3.978 (3.489,4.532)
Immunocompromised*	Immunocompromised*	19302	1751	1333	1002	9.417 (9.005,9.846)	7.193 (6.829,7.574)	5.357 (5.042,5.691)
Depression	Depression	38125	3825	2757	2046	10.070 (9.783,10.365)	7.234 (6.988,7.488)	5.352 (5.139,5.574)
Anxiety	Anxiety	57140	5368	3727	2652	9.730 (9.497,9.967)	6.713 (6.517,6.914)	4.779 (4.613,4.951)
Psychiatric disorder (other)	Psychiatric disorder (other)	3385	359	272	216	10.422 (9.505,11.416)	7.778 (6.982,8.657)	5.923 (5.227,6.704)
LFIA test result		443429	20080	12129	7600	4.649 (4.586,4.712)	2.798 (2.749,2.847)	1.762 (1.723,1.801)
		32647	5805	3566	2110	16.518 (16.145,16.897)	9.979 (9.680,10.287)	5.914 (5.680,6.157)

Table C.1: Numbers and weighted prevalence of current persistent symptoms in the main study population (rounds 3–5). In the final three columns, prevalence is weighted to be representative of the adult population of England, by age, sex, ethnicity, lower-tier local authority population, and index of multiple deprivation (IMD).

Variable	Category	total	Number with 1 or more symptoms (weighted)	Number with 2 or more symptoms (weighted)	Number with 3 or more symptoms (weighted)	Prevalence 1 or more symptoms (weighted)	Prevalence 2 or more symptoms (weighted)	Prevalence 3 or more symptoms (weighted)
All participants	All participants	508707	28973	17966	11287	5.795 (5.732,5.860)	3.559 (3.508,3.610)	2.230 (2.190,2.271)
Sex	Male	223861	10697	6284	3746	4.939 (4.855,5.025)	2.863 (2.798,2.929)	1.692 (1.642,1.743)
	Female	284840	18276	11682	7541	6.615 (6.520,6.711)	4.225 (4.149,4.303)	2.746 (2.683,2.809)
Age group	18-24	28655	1535	886	514	5.650 (5.461,5.846)	3.192 (3.048,3.342)	1.836 (1.727,1.951)
	25-34	65159	3795	2110	1221	6.045 (5.890,6.205)	3.336 (3.219,3.457)	1.879 (1.791,1.971)
	35-44	85044	5276	3106	1824	6.249 (6.085,6.417)	3.655 (3.528,3.786)	2.151 (2.054,2.253)
	45-54	103478	6927	4338	2740	6.856 (6.692,7.024)	4.313 (4.181,4.448)	2.732 (2.627,2.841)
	55-64	105361	6536	4235	2731	6.388 (6.217,6.564)	4.192 (4.052,4.336)	2.750 (2.637,2.868)
	65-74	85297	3661	2420	1635	4.607 (4.448,4.772)	3.066 (2.936,3.202)	2.093 (1.985,2.206)
	74+	35713	1243	871	622	3.679 (3.523,3.841)	2.597 (2.466,2.735)	1.923 (1.810,2.042)
Ethnicity	Asian	17666	823	487	279	5.130 (4.907,5.363)	2.827 (2.661,3.004)	1.600 (1.475,1.735)
	Black	4091	285	141	86	6.745 (6.332,7.184)	3.531 (3.230,3.858)	2.066 (1.838,2.322)
	Mixed	6112	490	312	202	7.704 (7.041,8.425)	4.727 (4.206,5.308)	3.047 (2.632,3.525)
	Other	4434	306	185	121	6.817 (6.181,7.512)	4.170 (3.673,4.730)	2.810 (2.406,3.280)
	White	472750	26826	16688	10497	5.777 (5.709,5.846)	3.592 (3.538,3.647)	2.266 (2.223,2.310)
Adiposity	Underweight	6273	336	220	147	5.929 (5.423,6.478)	3.697 (3.299,4.142)	2.141 (1.841,2.488)
	Normal weight	183305	9041	5359	3201	5.081 (4.981,5.182)	2.974 (2.897,3.053)	1.758 (1.698,1.819)
	Overweight	161456	8987	5479	3407	5.764 (5.650,5.880)	3.503 (3.413,3.595)	2.191 (2.119,2.264)
	Obese	97597	7177	4725	3136	7.357 (7.192,7.524)	4.718 (4.585,4.855)	3.162 (3.052,3.275)
Healthcare or care home worker	Yes	27690	2311	1483	925	8.546 (8.220,8.884)	5.305 (5.045,5.578)	3.216 (3.013,3.432)
Index of multiple deprivation (IMD) quintile	1 - most deprived	50260	3477	2299	1538	6.853 (6.697,7.013)	4.411 (4.284,4.541)	2.900 (2.796,3.006)
	2	80894	5116	3248	2129	6.283 (6.138,6.431)	3.837 (3.724,3.955)	2.478 (2.387,2.574)
	3	109928	6390	3961	2512	5.595 (5.457,5.736)	3.411 (3.303,3.523)	2.114 (2.029,2.203)
	4	126237	6792	4149	2563	5.280 (5.144,5.420)	3.177 (3.071,3.287)	1.931 (1.848,2.017)
	5 - least deprived	141388	7198	4309	2545	4.964 (4.830,5.101)	2.961 (2.857,3.069)	1.728 (1.649,1.811)
Current smoker	Current cigarette smoker	50983	3569	2372	1613	7.050 (6.842,7.263)	4.604 (4.435,4.780)	3.109 (2.970,3.255)
	Not current cigarette smoker	453681	25154	15442	9573	5.632 (5.565,5.700)	3.421 (3.368,3.475)	2.116 (2.075,2.159)
	Prefer not to say	4043	250	152	101	6.167 (5.501,6.907)	4.037 (3.501,4.652)	2.418 (2.008,2.908)
Current vaper	Current vaper	28436	1997	1276	834	7.237 (6.954,7.530)	4.590 (4.363,4.829)	2.933 (2.751,3.127)
	Not current vaper	477729	26834	16596	10386	5.702 (5.637,5.769)	3.489 (3.437,3.541)	2.183 (2.142,2.225)
	Prefer not to say	2542	142	94	67	5.613 (4.831,6.514)	3.950 (3.299,4.724)	2.426 (1.924,3.053)
Severity of COVID-19 symptoms	No symptoms	416589	0	0	0	0.000 (0.000,0.001)	0.000 (0.000,0.001)	0.000 (0.000,0.001)
	Mild symptoms	27572	5478	2757	1354	18.690 (18.259,19.129)	9.288 (8.969,9.617)	4.476 (4.251,4.712)
	Moderate symptoms	45357	14618	8885	5440	31.396 (30.979,31.817)	18.844 (18.494,19.200)	11.570 (11.285,11.862)
	Severe symptoms	19187	8877	6324	4493	44.976 (44.284,45.669)	32.060 (31.413,32.713)	22.749 (22.170,23.338)
Treatment sought/ received for COVID-19	No medical attention sought	58330	15203	8585	4929	24.874 (24.534,25.216)	13.882 (13.612,14.157)	7.955 (7.744,8.171)
	Sought medical attention from pharmacist / by phone (NHS 111/GP)	21388	8809	6033	4140	39.572 (38.946,40.201)	26.642 (26.078,27.213)	18.365 (17.873,18.867)

Table C.2 continued from previous page

Variable	Category	total	Number with 1 or more symptoms (weighted)	Number with 2 or more symptoms (weighted)	Number with 3 or more symptoms (weighted)	Prevalence 1 or more symptoms (weighted)	Prevalence 2 or more symptoms (weighted)	Prevalence 3 or more symptoms (weighted)
	Visited GP/walk-in centre/A&E	6792	2855	1930	1271	40.408 (39.245,41.582)	27.884 (26.828,28.964)	18.077 (17.179,19.012)
	Hospital admission	739	413	320	242	53.874 (50.491,57.221)	41.478 (38.190,44.843)	31.108 (28.068,34.321)
Gross household income	0-14,999	32819	2346	1685	1213	7.199 (6.943,7.465)	5.089 (4.872,5.316)	3.715 (3.529,3.911)
	15,000-49,999	148039	8819	5580	3536	6.109 (5.988,6.233)	3.844 (3.747,3.943)	2.393 (2.316,2.472)
	50,000-149,999	128957	7156	4109	2334	5.689 (5.560,5.821)	3.167 (3.070,3.268)	1.778 (1.705,1.854)
	>150,000	16279	895	466	248	5.598 (5.234,5.986)	2.730 (2.475,3.009)	1.448 (1.265,1.657)
Organ transplant recipient	Organ transplant recipient	596	36	27	20	5.714 (4.118,7.879)	4.202 (2.862,6.129)	3.529 (2.320,5.335)
Diabetes (type I or II)	Diabetes (type I or II)	27358	1617	1126	841	5.978 (5.708,6.260)	4.058 (3.835,4.294)	3.139 (2.942,3.348)
Heart disease or heart problems	Heart disease or heart problems	21430	1535	1098	819	7.152 (6.815,7.504)	5.128 (4.841,5.431)	3.862 (3.613,4.128)
Hypertension	Hypertension	65792	4048	2762	1915	6.221 (6.035,6.413)	4.271 (4.116,4.432)	3.025 (2.894,3.162)
Stroke	Stroke	3352	214	149	110	6.234 (5.480,7.084)	4.146 (3.534,4.859)	3.145 (2.616,3.776)
Kidney disease	Kidney disease	3867	326	239	166	8.401 (7.577,9.305)	6.106 (5.403,6.895)	4.113 (3.538,4.777)
Liver disease	Liver disease	2107	175	126	104	8.509 (7.411,9.752)	5.807 (4.902,6.867)	4.755 (3.940,5.729)
Anemia	Anemia	7575	807	588	439	11.114 (10.467,11.795)	7.832 (7.283,8.418)	5.777 (5.304,6.290)
Asthma	Asthma	51337	4461	3137	2196	8.779 (8.540,9.025)	6.085 (5.883,6.294)	4.225 (4.056,4.401)
Other lung condition	Other lung condition	12500	1122	842	643	9.067 (8.587,9.571)	6.735 (6.319,7.178)	5.321 (4.949,5.719)
Cancer	Cancer	8042	412	270	196	5.216 (4.744,5.732)	3.485 (3.101,3.916)	2.653 (2.319,3.033)
Neurological condition	Neurological condition	5415	380	296	229	7.124 (6.469,7.841)	5.384 (4.813,6.018)	4.070 (3.575,4.631)
Immunocompromised*	Immunocompromised*	19302	1832	1407	1059	9.752 (9.333,10.188)	7.506 (7.135,7.895)	5.627 (5.304,5.968)
Depression	Depression	38125	3966	2885	2151	10.415 (10.124,10.714)	7.579 (7.327,7.839)	5.590 (5.372,5.816)
Anxiety	Anxiety	57140	5602	3948	2818	10.105 (9.869,10.346)	7.083 (6.883,7.289)	5.035 (4.865,5.211)
Psychiatric disorder (other)	Psychiatric disorder (other)	3385	374	285	226	10.829 (9.896,11.838)	8.109 (7.296,9.003)	6.126 (5.419,6.919)
LFIA test result		443429	21481	13252	8454	4.946 (4.882,5.010)	3.031 (2.980,3.082)	1.931 (1.890,1.972)
		32647	6247	3957	2358	17.769 (17.385,18.159)	11.049 (10.736,11.371)	6.568 (6.321,6.823)

Table C.2: Numbers and weighted prevalence of persistent symptoms (ie respondents who either currently still have persistent symptoms more than 12 weeks after symptom onset, or who did have persistent symptoms for 12+ weeks but have since recovered) in the main study population (rounds 3–5). The final three columns show prevalence weighted to be representative of the adult population of England, by age, sex, ethnicity, lower-tier local authority population, and index of multiple deprivation (IMD).

Predictor	Level	Crude	Age, sex adjusted	Mutually adjusted
Sex	Male [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Sex	Female	1.46 [1.42-1.51]	1.51 [1.46-1.55]	1.38 [1.32-1.45]
Age group	55-64 [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Age group	18-24	0.58 [0.54-0.62]	0.56 [0.53-0.6]	0.64 [0.57-0.72]
Age group	25-34	0.6 [0.57-0.63]	0.58 [0.55-0.61]	0.6 [0.56-0.65]
Age group	35-44	0.65 [0.62-0.68]	0.64 [0.61-0.67]	0.64 [0.59-0.68]
Age group	45-54	0.86 [0.83-0.9]	0.86 [0.82-0.9]	0.85 [0.8-0.91]
Age group	65-74	1.16 [1.1-1.22]	1.16 [1.1-1.23]	1.16 [1.01-1.33]
Age group	74+	1.51 [1.38-1.64]	1.47 [1.35-1.61]	1.19 [0.71-1.98]
Ethnicity	White [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Ethnicity	Asian	0.71 [0.65-0.77]	0.8 [0.74-0.88]	0.84 [0.74-0.96]
Ethnicity	Black	0.99 [0.85-1.15]	1.06 [0.91-1.24]	1.02 [0.81-1.28]
Ethnicity	Mixed	1.05 [0.94-1.18]	1.19 [1.06-1.33]	1.16 [0.98-1.38]
Ethnicity	Other	0.99 [0.86-1.14]	1.05 [0.91-1.22]	1.05 [0.83-1.32]
Adiposity	Normal weight [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Adiposity	Obese	1.52 [1.46-1.58]	1.53 [1.47-1.59]	1.39 [1.32-1.48]
Adiposity	Overweight	1.14 [1.1-1.18]	1.16 [1.12-1.21]	1.15 [1.09-1.21]
Adiposity	Underweight	1.15 [1.01-1.33]	1.17 [1.02-1.35]	0.99 [0.79-1.25]
Healthcare or care home worker	No [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Healthcare or care home worker	Yes	1.48 [1.39-1.56]	1.33 [1.26-1.41]	1.26 [1.17-1.36]
Index of multiple deprivation (IMD) quintile	3 [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Index of multiple deprivation (IMD) quintile	1 - most deprived	1.19 [1.12-1.25]	1.24 [1.17-1.31]	1.09 [1-1.19]
Index of multiple deprivation (IMD) quintile	2	1.05 [1-1.1]	1.08 [1.03-1.13]	1.04 [0.97-1.12]
Index of multiple deprivation (IMD) quintile	4	0.94 [0.9-0.98]	0.92 [0.88-0.97]	0.98 [0.92-1.05]
Index of multiple deprivation (IMD) quintile	5 - least deprived	0.85 [0.82-0.89]	0.84 [0.8-0.87]	0.91 [0.85-0.97]
Current smoker	Not current cigarette smoker [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Current smoker	Current cigarette smoker	1.23 [1.17-1.29]	1.35 [1.28-1.41]	1.11 [1.03-1.2]
Current smoker	Prefer not to say	0.92 [0.79-1.08]	1.03 [0.88-1.21]	0.84 [0.63-1.14]
Current vaper	Not current vaper [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Current vaper	Current vaper	1.12 [1.06-1.19]	1.26 [1.18-1.34]	1.11 [1.01-1.21]
Current vaper	Prefer not to say	0.84 [0.69-1.03]	0.94 [0.77-1.16]	0.94 [0.59-1.51]
Severity of COVID-19 symptoms	Mild symptoms [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Severity of COVID-19 symptoms	Moderate symptoms	1.82 [1.75-1.89]	1.74 [1.68-1.81]	1.51 [1.43-1.6]
Severity of COVID-19 symptoms	Severe symptoms	3.25 [3.11-3.39]	3.02 [2.89-3.15]	2.26 [2.1-2.42]
Treatment sought/ received for COVID-19	No medical attention sought [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Treatment sought/ received for COVID-19	Hospital admission	3.98 [3.37-4.69]	3.46 [2.93-4.09]	3.45 [2.57-4.64]
Treatment sought/ received for COVID-19	Sought medical attention from pharmacist / by phone (NHS 111/GP)	2.07 [2-2.14]	2.05 [1.98-2.12]	2.02 [1.92-2.13]
Treatment sought/ received for COVID-19	Visited GP/walk-in centre/A&E	1.81 [1.72-1.91]	1.67 [1.58-1.76]	1.57 [1.44-1.71]
Gross household income	15,000-49,999 [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Gross household income	>150,000	0.59 [0.54-0.64]	0.64 [0.58-0.69]	0.8 [0.73-0.88]
Gross household income	0-14,999	1.51 [1.42-1.61]	1.39 [1.3-1.48]	1.36 [1.22-1.51]
Gross household income	50,000-149,999	0.7 [0.67-0.72]	0.76 [0.73-0.79]	0.84 [0.8-0.88]
LFIA test result	Negative	-	-	-
LFIA test result	Positive	1.28 [1.24-1.33]	1.28 [1.24-1.33]	1.41 [1.33-1.49]
Hospitalised with COVID	No [reference]	NA [NA-NA]	NA [NA-NA]	NA [NA-NA]
Hospitalised with COVID	Yes	3.08 [2.61-3.63]	2.72 [2.3-3.21]	3.45 [2.57-4.64]

Table C.3: Odds ratios for logistic regression models with persistent symptoms y/n as the binary outcome variable.

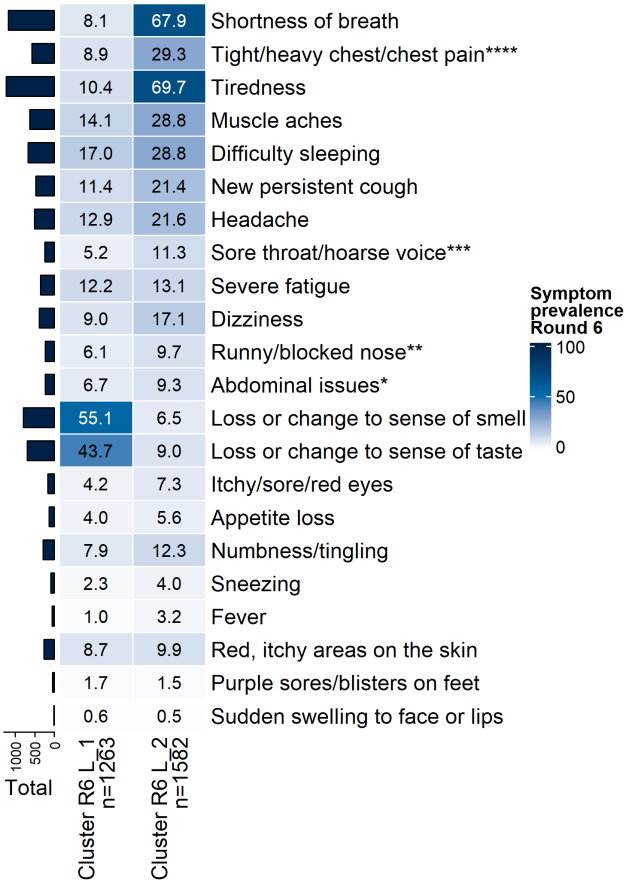


Figure C.1: Results of clustering analysis in the replication cohort (round 6). Cluster R6L1 had high prevalence of loss or change of sense of smell and taste; Cluster R6L2 had high prevalence of shortness of breath, tight chest and tiredness. As with the clustering in the main study population (rounds 3–5), participants with respiratory symptoms at 12 weeks were grouped together.