



Role of spatial and sociodemographic variation on SARS-CoV-2 transmission and public health interventions in England

Sumali Bajaj

St. Edmund Hall

Department of Biology

University of Oxford

Supervisors: Professor Moritz Kraemer, Professor Ben Lambert,
Professor Christl Donnelly, Professor José Lourenço

A thesis presented for the degree of *Doctor of Philosophy*
Trinity 2024

Abstract

This thesis investigates the dynamics of SARS-CoV-2 transmission in England from October 2020 to March 2022, focusing on how mobility patterns, sociodemographic factors, and public health interventions influenced the spread of the virus. The primary motivation for this research is to provide insights that can inform more equitable and effective public health responses during pandemics, particularly in the context of emerging infectious diseases.

Using Bayesian hierarchical models, I examined spatial differences in virus transmission, and the impact of human mobility on the spread of the Alpha and Delta variants. I developed novel methodologies for understanding testing behaviors and biases across different sociodemographic groups. I also introduced Bayesian nonparametric methods, to assess the spatial and sociodemographic drivers of the dynamic clustering in transmission patterns throughout England.

Key findings include the important role of human mobility in the early spread of variants of concern. Analysis of the national COVID-19 testing program identified disparities in testing use among different sociodemographic groups and delays in reporting positive results among ethnic minorities. Additionally, mobility between areas and spatial proximity influenced synchronous transmission patterns throughout the study period.

In conclusion, this thesis contributes to the broader understanding of how infectious diseases like SARS-CoV-2 spread within a population and underscores the importance of equitable public health interventions. The methodologies developed here can be applied to future pandemics, helping to ensure that public health responses are both effective and inclusive. The findings emphasize the critical need for real-time data and interdisciplinary approaches in addressing public health challenges.

Acknowledgments

What an immense privilege it has been to pursue something I am passionate about, with the freedom to explore and engage in scientific thought. Surrounded by individuals whose intellectual rigor and creativity I deeply admire, these past four years have been nothing short of a dream come true for me. In this space, I would like to extend my sincere gratitude to a few of the many individuals whose direct or indirect influence has made the completion of this thesis possible.

To the NERC Doctoral Training Programme, the Clarendon Scholarship, and St. Edmund Hall – without your generous support, this journey would not have been possible. Your contributions have been instrumental in every step of this process.

To José, thank you for believing in an excited and somewhat confused student, and for helping me make my way to Oxford. Your support, especially in getting me involved with teaching early on, has meant so much to me. I hope I get a chance to learn more from you in the future. To Moritz, you have been an incredible supervisor, and the opportunities you provided have shaped my DPhil journey in more ways than I can count. Thank you for your trust, for encouraging my ideas, inspiring me to think creatively, and for always being excited and supportive, whether in scientific discussions or about my badminton matches. I look forward to many more *walk-and-talk science* chats in the future. To Christl, thank you for your steadfast support and for always making time, often at short notice, to provide invaluable insights. I am excited about the possibility of working more closely with you in the future. To Ben – A *counterfactual* where you were not my supervisor is one I’d never want to consider – it perfectly captures the *effect* you’ve had on this journey. From your early suggestions at Imperial to our latest discussions, I have learned something new from every conversation – whether it was about tacos, coffee, sandwiches, or the wonders of condensation. You’ve helped me build my statistical and computational skills, even when I dreaded learning something new. Thank you for your empathy, and for empowering me to embrace the value of teaching and knowledge-sharing as an impactful part of our work. You are not only a supervisor but a mentor whose kindness and wisdom have profoundly shaped

my academic path. I look forward to continuing to learn from you in the years ahead.

To my wonderful office mates – Rhys, Rosario, Prathyush, Bernardo, Yalda, Yan, Abhishek, and Simon – thank you for the scientific discussions, endless jokes, and warm conversations we’ve shared. You’ve made this journey so much fun, and I can’t wait to see the amazing work you all accomplish. Special thanks to Joseph, my DPhil companion, thank you for constantly challenging my methods, critiquing my color choices (and improving them), and providing much-needed emotional support at times. Your love for food (though you ruined Popeyes for me) and our debates about coffee have been a delightful part of this process. I also want to extend my thanks to the various students from the Department of Biology and other departments across the university with whom I’ve had the pleasure of discussing research and statistics in various settings; I’ve learned so much from each of you.

To my incredible collaborators and mentors – Sunetra, Lisa, Reshania, Proochista, and Seth, thank you for welcoming me so warmly into your worlds. You have provided opportunities that shaped this thesis and my future steps, and I have so much to learn from each of you. I also want to thank my other collaborators, Marina, Rajeev, Zulma, and Nico who have been kind and generous with their time to host me and engage in exciting work over the years. Additionally, thank you to Ioana, Katherine, and Everlyn for the enjoyable moments we’ve shared working together. A special thank you to Richard for his invaluable support with statistical and computational work. I’ve cherished our discussions and look forward to more opportunities to learn from you.

I have been fortunate to be guided and supported by exceptional staff members at the Department of Biology and NERC DTP – Andrea, Heather, and Victoria, who made this journey much easier with their kindness and encouragement. A special thank you to the brilliant librarians at Teddy Hall, who have always been there to help with a flick of a wand. Also, a big thank you to the staff at Teddy Hall, Grzegorz, Simon, and Jose, who magically solved any problems that came my way.

To the wonderful cafes on Cowley Road – MacKitten, Peloton, G&D, you have been my refuge during the final months of writing. Thank you for providing a warm and comforting environment to work in.

To Amma and Papa, thank you for giving me a life full of freedom, allowing me to follow whichever path I chose, and for a relaxed and laid-back interest in my academic journey. You’ve kept me grounded and in touch with the realities of life, while also being my unwavering safety

net, allowing me to explore the world and pursue my dreams with confidence. Papa, thank you for teaching me to find joy in everything I do – I wish you could see that I have done just this with my DPhil! And Amma, thank you for always reminding me that health is the foundation of everything, and Dharmender is the best actor. To Mummy and Daddy, my grandparents, you have been the other half of the safety net that has sustained me throughout my life. Daddy, you’ve shown me the value of independence and self-sufficiency in all aspects of life, and I will always treasure your diligent interest in my work and your ceaseless support. I miss you. Mummy, you did all of this and with such grace – your big smile, loud laugh, and warm hugs are a constant source of inspiration. You make life look effortless, and I aspire to follow your example every day. To Chachu and Chachi, my favorite couple, thank you for your endless and unconditional love and support. I have so much to learn from you both. To Sumati, Shreya, and Ashi, my sometimes mischievous, always loving, and often annoying sisters! Sumati and Shreya, thank you for the endless laughs and for making Oxford feel like home. Ashi, thank you for teaching me how to use Illustrator (and makeup!). I am in awe of the three of you and feel incredibly lucky to have you as my sisters. To Ma and Dita, thank you for demonstrating that pursuing one’s passion is always possible at any stage of life – I’ve learned that all it takes is making time for it. I hope I have absorbed a small part of your confidence and curiosity to continue stepping out of my comfort zone. To Prashant – we’ve somehow completed yet another degree “together” – can you believe it? Thank you for distracting me, engaging with my work, and excelling so quietly in your own. You made these four years at Oxford all the more enjoyable. When do we meet again? Thank you, Seema Chachi, for being my favorite teacher and making maths fun. Your patience and guidance during our school lessons helped me not only enjoy the subject but also gain confidence. I’ll always be grateful for the time and effort you put into teaching me.

Much love to my furry companions – Hero, Nancy, Tommy, Simba, Kep, Kiki, Annie, Cookie. Though you’ll never read this, I hope you know the joy and comfort you’ve brought into my life.

I have been incredibly lucky to be surrounded by my wonderful badminton family and friends from different parts of my life. Special thanks to Isaac, Jenny, Shefali, Cori, Jake, Kari, Samvida, Matthew, Ruth, Oliver, Tiggy, Qiqi, and Kevin for your friendship, laughter, and life chats over the past four years and more.

A heartfelt thank you to my family for always encouraging me and making me laugh – a lot. Pramod chachu, Chunchun didi, Ram bhaiya, Jaggs, Abhishek dada, Aishu didi, Krishan mausaji, Rita mausi, Vaibhav dada, Rita, Raghav dada, Devika bhabhi, Shivi, Shubhangi,

Abhishek, Jishnu, Suman – thank you making this journey feel like home and celebrating my achievements as your own. A special shoutout to Sujith bhaiya for being my stats buddy.

I also feel incredibly lucky to have friends who are family – Himanshu, Nikita, Shailaja, Anika, Danny, Amu, and Aditya, thank you for your constant love and support, and for showing little interest in my DPhil, which has kept me grounded! You all make my life wonderful.

Last but by no means least, to Raghav – my best friend and partner for life. Thank you for sharing this journey with me (perhaps even more enthusiastically than I did!). Your inquisitive nature, infectious energy, and ability to make every aspect of life fun are truly remarkable, and have played a big role in how I have (tried to) approach these four years. Thank you for making sure I attended the social events, and lifted the extra weights. We’ve experienced countless joys and a few losses, but every day has been brighter with you by my side. These four years have been filled with exploration and joy, like a walk in the rain, and I can’t wait to see where we go next.

I would like to dedicate this thesis to my Daddy and Papa. I hope you are happy and playing tennis and golf, wherever you are.

Pal pal dil ke paas tum rehte ho

Statement of contributions and publications

This thesis presents the research I have developed over the past three years. When I began as a student in the NERC doctoral training program in October 2020, my initial aim was to investigate the role of population-level immunity in the dynamics of vector-borne pathogen epidemics. However, by 2021, I became involved in several projects focused on understanding the spread of SARS-CoV-2 in England. Of these studies, two have been published in peer-reviewed journals (included in the appendix), one has been accepted for publication, and another is currently in preparation. In the following sections, I outline my specific contributions to these scientific papers, highlighting the collaborative nature of idea generation and method development, as well as the invaluable mentorship I received from my colleagues. However, this thesis will focus solely on the work I carried out.

Chapter 2

The two papers which form this chapter were part of a collaborative working group. Below, I describe my specific contributions to each paper, which I will present in this thesis.

My co-authors on the paper led the sections on the phylogeographic analysis, and I was responsible for quantifying the spatial heterogeneity in the spread of the Alpha variant across England, including developing the statistical methodology, analysis, interpretation, and writing.

Moritz U G Kraemer*, Verity Hill*, Christopher Ruis*, Simon Dellicour*, **Sumali Bajaj***, John T McCrone, Guy Baele, Kris V Parag, Anya Lindström Battle, Bernardo Gutierrez, Ben Jackson, Rachel Colquhoun, Áine O'Toole, Brennan Klein, Alessandro Vespignani, COVID-19 Genomics UK (COG-UK) Consortium, Erik Volz, Nuno R Faria, David M Aanensen, Nicholas J Loman, Louis du Plessis, Simon Cauchemez, Andrew Rambaut, Samuel V Scarpino, Oliver G Pybus. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* (2021).

*Joint first authors

<https://doi.org/10.1126/science.abj0113>

I led the statistical analysis used to understand the factors that contributed to the growth of the Delta variant. My co-authors on the paper were responsible for the phylogeographic analysis. I developed the statistical methodology in collaboration with BL. I was responsible for coding up the methods, interpretation of results, writing, and reviewing the manuscript.

John T McCrone*, Verity Hill*, **Sumali Bajaj***, Rosario Evans Pena*, Ben C Lambert, Rhys Inward, Samir Bhatt, Erik Volz, Christopher Ruis, Simon Dellicour, Guy Baele, Alexander E Zarebski, Adam Sadilek, Neo Wu, Aaron Schneider, Xiang Ji, Jayna Raghvani, Ben Jackson, Rachel Colquhoun, Áine O'Toole, Thomas P Peacock, Kate Twohig, Simon Thelwall, Gavin Dabrera, Richard Myers, The COVID-19 Genomics UK (COG-UK) Consortium, Nuno R Faria, Carmen Huber, Isaac I Bogoch, Kamran Khan, Louis du Plessis, Jeffrey C Barrett, David M Aanensen, Wendy S Barclay, Meera Chand, Thomas Connor, Nicholas J Loman, Marc A Suchard, Oliver G Pybus, Andrew Rambaut, Moritz U G Kraemer. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. *Nature* (2022). <https://doi.org/10.1038/s41586-022-05200-3>

Chapter 3 and 4

These two chapters are part of a single manuscript that has been accepted for publication. The division highlights the breadth and depth of the methods developed and the analyses conducted to understand biases in SARS-CoV-2 testing among various sociodemographic groups in England over 18 months. I led this work including conceiving the research plan and developing methods in collaboration with BL, SC, and RC. I performed all the analyses, interpretation, writing, and revisions with BL, RN.

Sumali Bajaj*, Siyu Chen*, Richard Creswell*, Reshania Naidoo*, Joseph L.-H Tsui*, Olu-mide Kolade, George Nicholson, Briec Lehmann, James A Hay, Moritz U G Kraemer, Ricardo Aguas, Christl A Donnelly, Tom Fowler, Susan Hopkins, Liberty Cantrell, Prabin Dahal, Lisa J White, Kasia Stepniewska, Merryn Voysey, Ben Lambert, EY-Oxford Health Analytics Consortium. COVID-19 testing and reporting behaviours in England across different sociodemographic groups: a population based study using testing data and data from community prevalence surveillance surveys *The Lancet Digital Health (In press)* (2024). medRxiv <https://doi.org/10.1101/2023.10.26.23297608>

Chapter 5

I was responsible for all aspects of the work presented in this chapter, including conception and design, analysis and interpretation (with RC and BL), and writing. This chapter is currently being prepared for submission to a journal.

Sumali Bajaj, Ioana Bouros, Katherine Shepherd, Moritz U G Kraemer, Richard Creswell, Ben Lambert. Evolution of the spatial scales of transmission of SARS-CoV-2 in England between 2020 and 2022. *In preparation* (2024).

Related publications

There are also related manuscripts that have been published, but their results are not directly tied to the content of this thesis.

1. **Sumali Bajaj**, Robin Thompson, Ben Lambert. A mechanistic approach to estimating R_t and infectious disease case counts in the presence of reporting delays. *Under review* (2024).
2. Zhiyuan Chen, Joseph L-H Tsui, Bernardo Gutierrez, Simon Busch Moreno, Louis du Plessis, Xiaowei Deng, Jun Cai, **Sumali Bajaj**, Marc A Suchard, Oliver G Pybus, Philippe Lemey, Moritz U G Kraemer, Hongjie Yu. COVID-19 pandemic interventions reshaped the global dispersal of seasonal influenza viruses. *Science (Accepted)* (2024). medRxiv-<https://doi.org/10.1101/2023.12.20.23300299>
3. Bernardo Gutierrez*, Joseph L-H Tsui*, Giulia Pullano*, Mattia Mazzoli*, Karthik Gangavarapu*, Rhys P D Inward, **Sumali Bajaj**, Rosario E Pena, Simon Busch-Moreno, Marc A Suchard, Oliver G Pybus, Alejandra Dunner, Rodrigo Puentes, Salvador Ayala, Jorge Fernandez, Rafael Araos, Leo Ferres, Vittoria Colizza, Moritz U G Kraemer. Routes of importation and spatial dynamics of SARS-CoV-2 variants during localised interventions in Chile. *PNAS Nexus (Accepted)* (2024). medRxiv-<https://doi.org/10.1101/2024.01.18.24301504>
4. Siyu Chen, Richard Creswell, Rachel Hounsell, Liberty Cantrell, **Sumali Bajaj**, Prabin Dahal, Lok Hei Tsui, Olumide Kolade, Ma'ayan Amswych, Reshania Naidoo, Tom Fowler, Susan Hopkins, Kasia Stepniewska, Merryn Voysey, Lisa White, Rima Shretta, Ben Lambert, EY-Oxford Health Analytics Consortium. COVID-19 mass testing in adult social

- care in England. *medRxiv* (2024). <https://doi.org/10.1101/2024.06.06.24308563>
5. Siyu Chen, Rachel Hounsell, Libert Cantrell, Joseph L-H Tsui, Reshania Naidoo, Prabin Dahal, Richard Creswell, **Sumali Bajaj**, Jennifer A. Flegg, Tom Fowler, Susan Hopkins, Ben Lambert, Merryn Voysey, Lisa J. White, EY-Oxford Health Analytics Consortium, Kasia Stepniewska, Rima Shretta. Assessing the impact of testing for COVID-19 using lateral flow devices in NHS acute trusts in England. *medRxiv* (2024). <https://doi.org/10.1101/2024.06.06.24308561>
 6. Joseph L-H Tsui*, John T McCrone*, Ben Lambert*, **Sumali Bajaj***, Rhys P. D. Inward, Paolo Bosetti, Rosario Evans Pena, Houriiyah Tegally, Verity Hill, Alexander E Zarebski, Thomas P Peacock, Luyang Liu, Neo Wu, Megan Davis, Isaac I Bogoch, Kamran Khan, Meaghan Kall, Nurin Iwani Binti Abdul Aziz, Rachel Colquhoun, Áine O’Toole, Ben Jackson, Abhishek Dasgupta, Eduan Wilkinson, Tulio de Oliveira, The COVID-19 Genomics UK (COG-UK) consortium, Thomas R Connor, Nicholas J Loman, Vittoria Colizza, Christophe Fraser, Erik Volz, Xiang Ji, Bernardo Gutierrez, Meera Chand, Simon Dellicour, Simon Cauchemez, Jayna Raghvani, Marc A Suchard, Philippe Lemey, Andrew Rambaut, Oliver G Pybus, Moritz U G Kraemer. Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science* (2023). <https://doi.org/10.1126/science.adg6605>
 7. Solveig A van der Vegt, Liangti Dai, Ioana Bouros, Hui Jia Farm, Richard Creswell, Oscar Dimdore-Miles, Idil Cazimoglu, **Sumali Bajaj**, Lyle Hopkins, David Seifert, Fergus Cooper, Chon Lok Lei, David Gavaghan, Ben Lambert. Learning transmission dynamics modelling of COVID-19 using comodels, *Mathematical Biosciences* (2022). <https://doi.org/10.1016/j.mbs.2022.108824>
 8. David J Haw, Giovanni Forchini, Patrick Doohan, Paula Christen, Matteo Pianella, Robert Johnson, **Sumali Bajaj**, Alexandra B. Hogan, Peter Winskill, Marisa Miraldo, Peter J White, Azra C Ghani, Neil M Ferguson, Peter C Smith, Katharina D Hauck. Optimizing social and economic activity while containing SARS-CoV-2 transmission using DAEDALUS. *Nature Computational Science* (2022). <https://doi.org/10.1038/s43588-022-00233-0>
 9. Verity Hill, Christopher Ruis, **Sumali Bajaj**, Oliver G Pybus, Moritz U G Kraemer. Progress and challenges in virus genomic epidemiology. *Trends in Parasitology* (2021). <https://doi.org/10.1016/j.pt.2021.08.007>.
 10. Carolin Vegvari, Federica Giardina, **Sumali Bajaj**, Veronica Malizia, Robert J Hardwick,

James E Truscott, Antonio Montresor, Sake J de Vlas, Luc E Coffeng, Roy M Anderson. Deworming women of reproductive age during adolescence and pregnancy: what is the impact on morbidity from soil-transmitted helminths infection?. *Parasites Vectors* (2021). <https://doi.org/10.1186/s13071-021-04620-w>

11. Veronica Malizia, Federica Giardina, Carolin Vegvari, **Sumali Bajaj**, Kevin McRae-McKee, Roy M Anderson, Sake J de Vlas, Luc E Coffeng. Modelling the impact of COVID-19-related control programme interruptions on progress towards the WHO 2030 target for soil-transmitted helminths. *Transactions of The Royal Society of Tropical Medicine and Hygiene* (2021). <https://doi.org/10.1093/trstmh/traa156>

Other publications

Other manuscripts not related to this work, which I was involved in:

1. Moritz U G Kraemer*, Joseph L-H Tsui*, Serina Y Chang*, Spyros Lytras, Samantha Vanderslott, **Sumali Bajaj**, Elizaveta Semenova, Mengyan Zhang, H Juliette T Unwin, Mark P Khurana, Jacob Liam Curran-Sebastian, Neil Scheidwasser, Oliver J. Watson, Cathal Mills, Abhishek Dasgupta, Luca Ferretti, Samuel V Scarpino, Etien Koua, Oliver Morgan, Houriiyah Tegally, Ulrich Paquet, Serge Belongie, Christophe Fraser, Neil M Ferguson, Edward C Holmes, Eric J Topol, David A Duchêne, Tanja Stadler, Patricia Kingori, Michael J Parker, Francesca Dominici, Nigel Shadbolt, Marc A Suchard, Oliver Ratmann, Seth Flaxman, Manuel G Rodriguez, Bernhard Schölkopf, Christl A Donnelly, Oliver G Pybus, Simon Cauchemez, Samir Bhatt. Artificial Intelligence for Infectious Disease Epidemiology. *Under review* (2024).
2. Joseph L-H Tsui*, Rosario E Pena*, Monica Moir*, Rhys P D Inward, Eduan Wilkinson, James Emmanuel San, Jenicca Poongavanan, **Sumali Bajaj**, Bernardo Gutierrez, Abhishek Dasgupta, Tulio de Oliveira, Moritz U G Kraemer, Houriiyah Tegally, Prathyush Sambaturu. Impacts of climate change-related human migration on infectious diseases. *Nature Climate Change* (2024). <https://doi.org/10.1038/s41558-024-02078-z>
3. Qiqi Yang, Ben Wang, Phillipe Lemey, Lu Dong, Tong Mu, R Alex Wiebe, Fengyi Guo, Nídia Sequeira Trovão, Sang Woo Park, Nicola Lewis, Joseph L-H Tsui, **Sumali Bajaj**, Yachang Cheng, LuoJun Yang, Yuki Haba, Bingying Li, Guogang Zhang, Oliver G

- Pybus, Huaiyu Tian, Bryan Grenfell. Synchrony of Bird Migration with Global Dispersal of Avian Influenza Reveals Exposed Bird Orders. *Nature Communications* (2024). <https://doi.org/10.1038/s41467-024-45462-1>
4. Bernardo Gutierrez, Darlan da Silva Candido, **Sumali Bajaj**, Abril Paulina Rodriguez Maldonado, Fabiola Garces Ayala, María de la Luz Torre Rodriguez, Adnan Araiza Rodriguez, Claudia Wong Arámbula, Ernesto Ramírez González, Irma López Martínez, José Alberto Díaz-Quinónez, Mauricio Vázquez Pichardo, Sarah C Hill, Julien Thézé, Nuno R Faria, Oliver G Pybus, Lorena Preciado-Llanes, Arturo Reyes-Sandoval, Moritz U G Kraemer, Marina Escalera-Zamudio. Convergent trends and spatiotemporal patterns of Aedes-borne arboviruses in Mexico and Central America. *Plos Neglected Tropical Diseases* (2023). <https://doi.org/10.1371/journal.pntd.0011169>
 5. Jeffrey W Eaton, Laura Dwyer-Lindgren, Steve Gutreuter, Megan O'Driscoll, Oliver Stevens, **Sumali Bajaj**, Rob Ashton, Alexandra Hill, Emma Russell, Rachel Esra, Nicolas Dolan, Yusuf O Anifowoshe, Mark Woodbridge, Ian Fellows, Robert Glaubius, Emily Haeuser, Taylor Okonek, John Stover, Matthew L Thomas, Jon Wakefield, Timothy M Wolock, Jonathan Berry, Tomasz Sabala, Nathan Heard, Stephen Delgado, Andreas Jahn, Thokozani Kalua, Tiwonge Chimpande, Andrew Auld, Evelyn Kim, Danielle Payne, Leigh F Johnson, Richard G FitzJohn, Ian Wanyeki, Mary I Mahy, Ray W Shiraishi. Naomi: a new modelling tool for estimating HIV epidemic indicators at the district level in sub-Saharan Africa. *Journal of the International AIDS Society* (2021). <https://doi.org/10.1002/jia2.25788>

As Sumali's supervisor, I confirm that the information presented here is representative of her contributions to the work listed in this thesis.



Moritz Kraemer, DPhil
 Professor of Epidemiology & Data Science
 Department of Biology
 University of Oxford

*No one sets out to ignore equity, but the way we frame issues of causality and response typically fails to give it due consideration.**

*Paul Farmer, Kim Yong Jim, Arthur Kleinman, and Matthew Basilio. Reimagining Global Health: An Introduction. 1st ed. University of California Press, 2013. <http://www.jstor.org/stable/10.1525/j.ctt46n4b2>.

Contents

1	Introduction	1
1.1	Inequity in the global variation of pathogens	1
1.2	SARS-CoV-2	2
1.2.1	Dispersal	2
1.2.2	Clinical manifestation	3
1.2.3	Interventions	3
1.3	Heterogeneity in risk of infection	5
1.3.1	Susceptibility	6
1.3.2	Exposure	6
1.4	Key epidemiological and statistical quantities	7
1.4.1	Growth Rate	7
1.4.2	Incidence	8
1.4.3	Prevalence	8
1.5	Methodological advancements in the thesis	8
1.6	Thesis outline	10
1.6.1	Chapter 2	10
1.6.2	Chapter 3	10
1.6.3	Chapter 4	11
1.6.4	Chapter 5	11
2	Impact of human behaviour on the invasion of SARS-CoV-2 variants of concern	13
2.1	Abstract	14
2.2	Introduction	15
2.3	Methods	17

2.3.1	Estimating the growth rate of Alpha	17
2.3.2	Modelling the relative growth of Delta	20
2.4	Results	27
2.4.1	Spatial heterogeneity in the expansion of Alpha	27
2.4.2	Factors contributing to the growth of Delta	30
2.5	Discussion	35
A	Appendix - Impact of human behaviour on the invasion of SARS-CoV-2 variants of concern	37
3	SARS-CoV-2 transmission dynamics between sociodemographic groups	41
3.1	Abstract	42
3.2	Introduction	43
3.3	Methods	44
3.3.1	Data pre-processing	44
3.3.2	Variant time periods	47
3.3.3	Causal debiasing approach for obtaining time-varying estimates of SARS-CoV-2 prevalence at fine scale resolution	48
3.3.4	Association between LTLA prevalence and PCR testing behaviour	50
3.3.5	Model comparison	50
3.3.6	Incidence and case detection ratio estimation	51
3.4	Results	55
3.4.1	Trends in testing uptake and reporting	55
3.4.2	Prevalence	58
3.4.3	Incidence and case detection ratio trends	59
3.4.4	Prevalence model checks	60
3.4.5	Incidence model checks	63
3.5	Discussion	66
B	Appendix - SARS-CoV-2 transmission dynamics between sociodemographic groups	69
4	Differences in SARS-CoV-2 testing behaviours by sociodemographic groups	79
4.1	Abstract	80

4.2	Introduction	82
4.3	Methods	83
4.3.1	Testing bias	83
4.3.2	Estimating false positivity, test sensitivity and specificity	84
4.3.3	Modelling the decline in PCR detection rate subsequent to reporting an LFD positive	86
4.4	Results	87
4.4.1	Confirmatory PCR, false positivity	87
4.4.2	Test sensitivity and specificity, PCR detectability	90
4.4.3	Trends in testing biases	92
4.4.4	Model fit	93
4.5	Discussion	94
C	Appendix - Differences in SARS-CoV-2 testing behaviours by sociodemo- graphic groups	97
5	Evolution of the spatial scales of transmission of SARS-CoV-2 in England between 2020 and 2022	99
5.1	Abstract	100
5.2	Introduction	101
5.3	Methods	102
5.3.1	Data pre-processing	102
5.3.2	Clustering	104
5.3.3	Model setup	109
5.3.4	Epidemiological analysis	111
5.4	Results	114
5.4.1	Clustering and trends	114
5.4.2	Factors driving synchrony in transmission	120
5.5	Discussion	121
D	Appendix - Evolution of the spatial scales of transmission of SARS-CoV-2 in England between 2020 and 2022	126

6 Discussion	138
6.1 Summary	138
6.2 Future directions	139
6.2.1 Statistical, mechanistic and causal models	139
6.2.2 Sociodemographic factors in epidemiological models	143
6.2.3 Modelling human behaviour	144
Appendix: Publication 1 for Chapter 2	174
Appendix: Publication 2 for Chapter 2	182
Appendix: Preprint for Chapters 3 and 4	208

List of Figures

1.1	Illustrative relationship between incidence and prevalence.	9
2.1	Simulation comparing known vs estimated relative growth rates for hypothetical locations.	26
2.2	Prior predictive simulations for the Delta frequency.	27
2.3	Weekly growth rates per LTLA for SGTF and non-SGTF cases.	29
2.4	Case growth rates of B.1.1.7 were correlated with human mobility and attack rates across the UK.	30
2.5	Variation in Delta growth rates across UTLAs in England.	31
2.6	Counterfactual scenarios for the impact of within-UTLA mobility.	33
2.7	Posterior predictive simulations for the Delta frequency.	34
A.1	Ratio of number of samples that have information about SGTF status compared to all reported cases.	38
A.2	Association between LTLA specific excess growth rates and mobility and imports beyond 20 December 2020.	38
A.3	Association between LTLA specific excess growth rates and mobility and imports for observed and scaled up cases.	39
A.4	Sensitivity results of only including data once Alpha crosses thresholds of 0.2 and 0.6 in an LTLA.	39
A.5	Observed (raw) vs posterior predicted growth rates.	40
3.1	Variant time periods.	48
3.2	The temporal profile of PCR detectability as a function of time since infection. .	54
3.3	Number of total (and positive) PCR tests performed and LFD tests reported in England.	55

3.4	Timelines, sociodemographic distribution, and testing data summary	57
3.5	Estimated prevalence for ethnic groups at the national level.	59
3.6	Estimated case detection ratios by sociodemographic groups.	61
3.7	Comparing LTLA level prevalence estimates from REACT-1, mass testing, and from the debiasing methodology.	62
3.8	Sensitivity analysis of ethnic group prevalences in LTLAs.	64
3.9	Comparing MCMC and MAP estimates of incidence per capita.	65
3.10	Weekly regional prevalence estimates by ethnic group.	66
B.1	Prevalence of SARS-CoV-2 in ethnic groups.	70
B.2	Prevalence of SARS-CoV-2 in individuals by (self-reported) age groups.	71
B.3	Prevalence of SARS-CoV-2 in England over time and by LTLA.	72
B.4	Relationship between deprivation, SARS-CoV-2 prevalence and PCR tests per- formed.	73
B.5	PCR and LFD positive test performed/reported per capita by ethnic groups. . .	74
B.6	Weekly number of PCR and LFD tests performed/reported per capita by age groups in England.	75
B.7	Weekly number of PCR and LFD tests performed/reported per capita by five ethnic groups in England.	76
B.8	Aggregated LTLA- and sociodemographic group-specific PCR tests performed and LFD tests reported per person.	77
B.9	Model fit for the ONS national prevalence model.	78
B.10	Model fits for the REACT-1 based prevalence model at LTLA level.	78
4.1	Confirmatory PCR tests over time.	89
4.2	Confirmatory PCR tests done overall and followed by an LFD positive, and false positive LFD tests by sociodemographic groups.	90
4.3	Decline in PCR detectability of PCR tests by ethnic groups.	91
4.4	Estimated testing bias by deprivation levels and age groups.	94
4.5	Posterior predictive check for modelling decay rate of PCR detectability after reporting a positive LFD test, by different ethnic groups.	95
C.1	Testing bias with uncertainty intervals by IMD.	97

C.2	Testing bias with uncertainty intervals by age-groups.	98
C.3	Estimated mean PCR testing bias by deprivation level and administrative regions.	98
5.1	SARS-CoV-2 prevalence estimates, alongside mobility and socio-demographic data.	115
5.2	Number of LTLAs in each cluster, average distance between cluster-specific LTLAs, and entropy.	117
5.3	Spatial clustering of prevalence trends from October 2020 to June 2021.	118
5.4	Spatial clustering of prevalence trends from July 2021 to March 2022.	119
5.5	Regression coefficient estimates of covariates explaining the probability of co-clustering of two LTLAs.	121
D.1	Spatial clustering of prevalence trends in Jan 2021 for different values of Σ	127
D.2	Monthly number of trips per capita across LTLAs.	128
D.3	Distance between cluster-specific LTLAs.	129
D.4	Convergence of MCMC chains for clustering.	130
D.5	Effect of concentration parameter choice on clustering.	131
D.6	Posterior predictive checks of prevalence estimates.	132
D.7	Posterior predictive checks, comparing eCDFS.	133
D.8	Adjusted R^2 from linear regressions.	134

Chapter 1

Introduction

1.1 Inequity in the global variation of pathogens

The spread of an infectious pathogen is fundamentally shaped by interactions between susceptible individuals and those carrying the infection. Over time*, distinct patterns emerge—spatial patterns indicating areas with higher or lower burdens of infectious cases, as well as spatiotemporal patterns revealing regions where transmission accelerates or decelerates. Initially, spatial connectivity may influence these patterns, with increasing global connectivity over the last century facilitating the dispersal of the pathogen across distances ranging from households to continents [1]. However, beyond contact patterns and mobility, these transmission patterns can often be traced back to underlying inequities within the population, stemming from layers of social and economic disparities [2].

Spanning over decades, this pattern has become evident in the three leading contributors to the global infectious disease burden [3]. HIV, tuberculosis (TB), and malaria caused by the human immunodeficiency virus, *Mycobacterium tuberculosis* bacteria, and *Plasmodium* parasites, respectively demonstrated distinct yet interconnected patterns of origin, spread, and current burden. HIV originated in Central Africa in the early-mid 20th century and spread internationally during the late 20th century through global travel and trade, with approximately 65% cases now concentrated in sub-Saharan Africa[4, 5]. TB, which has its origins dating back thousands of years in Africa, spread globally through human migration, and today remains predominantly concentrated in South Asia and sub-Saharan Africa accounting for two-thirds of the global cases

*A relative term which could be days, weeks, months, or even years, depending on a pathogen's characteristics and interventions restricting its spread.

[6, 7]. Similarly, malaria parasite discovered in the 1800s[†], intensified its global spread during the colonial era and remains a major health issue in the WHO African region, which accounts for approximately 95% of the cases and deaths [9, 10]. Unsurprisingly, these diseases have seen a much lower burden in wealthier nations, where advanced healthcare systems and improved living conditions have drastically reduced their presence.

Soon after spatial connectivity initiates the spread of infectious pathogens, social and economic inequalities become critical in shaping the networks where these pathogens entrench and thrive. They have profoundly impacted the burden of HIV, TB, and malaria, often disproportionately affecting marginalized populations. Factors including poor living conditions, social marginalisation, stigma, lack of care-seeking behaviour, weak health systems, poor nutrition are some of the key pathways through which these pathogens have established themselves in the socioeconomically vulnerable populations [11].

These global (and local) health challenges set the stage for the emergence of another critical disease – COVID-19. Like HIV, TB, and malaria, COVID-19 quickly became a major global health concern, highlighting the ongoing vulnerabilities in public health systems and further exacerbating disparities between populations.

1.2 SARS-CoV-2

1.2.1 Dispersal

In December 2019, the novel coronavirus disease (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (2019-nCoV, and subsequently referred to as SARS-CoV-2 in the literature and this thesis) was first identified in Wuhan, China [12]. A range of public health measures including enhanced testing and isolation of cases were implemented, followed by strict restrictions on mobility on 23 January 2020 in China. During this first month of 2020, as the pathogen spread to 6 countries worldwide, the WHO declared the outbreak as a Public Health Emergency of International Concern [13], and by 11 March as the virus spread to 114 countries with more than 118,000 cases and 4,291 deaths countries, the WHO characterised COVID-19 as a pandemic [14, 15]. The UK saw its first rise in cases in March - April 2020 [16]. As the pandemic pro-

[†]The scientific discovery of malaria and its causative agent, the *Plasmodium* parasite (both deeply intertwined with the history of colonialism), is credited to Alphonse Laveran, who first identified the parasite in 1880, and Ronald Ross, who later demonstrated in 1897 that mosquitoes transmit the parasite [8]. However, descriptions of malaria-like symptoms can be found in ancient texts dating back thousands of years, including in ancient Chinese, Indian, and Egyptian writings [9].

gressed, the spread of new variants further complicated the situation, driving additional public health responses and highlighting the challenges in controlling the virus’s transmission. Variants are versions of the virus that have undergone genetic mutations that can affect its transmissibility, disease severity, or the effectiveness of vaccines, diagnostics, and treatments. A variant of concern (VOC) is specifically identified when it leads to detrimental changes in clinical disease severity, causes significant shifts in COVID-19 epidemiology that strain health systems or require major public health interventions, or when there is a notable decrease in vaccine effectiveness, particularly against severe disease [17]. In Chapter 2, I will describe the invasion process of two VOCs: Alpha and Delta in England. Alpha, which likely originated in the UK, spread rapidly before the population had access to vaccines [18]. In contrast, the Delta variant arrived in the UK from other countries at a time when the population was getting vaccinated, and as the country began emerging from tiered restrictions [19].

1.2.2 Clinical manifestation

The clinical outcomes of SARS-CoV-2 infection vary widely, ranging from asymptomatic cases to severe, life-threatening conditions. Most individuals experience mild to moderate symptoms, including fever, cough, and fatigue, but some develop severe respiratory illness, such as pneumonia or acute respiratory distress syndrome, which can require hospitalization and mechanical ventilation [20]. High-risk groups, including the elderly and those with underlying health conditions like cardiovascular disease, diabetes, and obesity, are more likely to experience severe outcomes and higher mortality rates [20]. As new variants of the virus emerged, there were changes in clinical presentation, transmissibility, severity, and vaccine effectiveness. Additionally, long-term effects, often referred to as “long COVID” are known to occur in at least 10% of the cases, and include persistent symptoms such as fatigue, cognitive impairment, and respiratory issues, which can significantly impact quality of life [21].

1.2.3 Interventions

1.2.3.1 Vaccines

The development and global distribution of COVID-19 vaccines marked a pivotal moment in public health, characterized by both unprecedented speed and stark inequities. Following the identification of SARS-CoV-2 as the causative agent of COVID-19 disease in January 2020,

vaccine development advanced rapidly, with Pfizer-BioNTech and Moderna announcing successful Phase III trial results by November 2020. High-income countries swiftly secured large quantities of these vaccines, with the first doses administered in December 2020 [22]. By mid-2021, these nations, representing just 16% of the global population, had acquired over 70% of available vaccine doses, enabling them to vaccinate more than 70% of their populations by the year’s end [23]. In contrast, low- and middle-income countries (LMICs) faced severe delays, with low-income countries, for instance, reaching only about 1% vaccine coverage by this time [24]. Despite efforts like the COVAX initiative, the vaccine rollout remained deeply unequal [25]. This disparity not only prolonged the pandemic but also likely contributed to the emergence of new variants, underscoring the urgent need for a more equitable approach to global health [26]. By 1 October 2022, 77% of individuals in high- and upper-middle-income countries had completed their initial COVID-19 vaccination course, compared to only 50% coverage in LMICs [23].

1.2.3.2 Non-pharmaceutical interventions

In addition to vaccines, throughout the pandemic between 2020 and 2022, countries implemented varied intensities of non-pharmaceutical interventions (NPI) (including physical distancing measures, mask-wearing, hand-hygiene, etc.) to control the spread of the pathogen [27]. And unsurprisingly, adherence to NPIs like physical distancing, and lockdowns during the COVID-19 pandemic was closely tied to socioeconomic status. Individuals in lower socioeconomic groups often faced greater challenges in adhering to these measures. Those classified as key workers (employed in healthcare, retail, delivery services, etc.), were generally the ones living in crowded housing conditions and with poor access to healthcare and were subject to worse health outcomes [28]. Research showed that those with higher incomes were more likely to comply with NPIs, as they had the means to work from home and maintain physical distancing [29, 30]. This disparity in adherence contributed to higher infection rates and worse health outcomes among disadvantaged populations, further widening the gap in health inequities during the pandemic.

1.2.3.3 Testing

In March 2020, the government initiated testing efforts [31], and by May 2020, NHS Test and Trace (NHSTT) was officially established to manage a large-scale national testing and tracing effort [32]. NHSTT was responsible for implementing mass testing and tracing systems to quickly identify individuals infected with SARS-CoV-2 and their close contacts, reducing the spread of

the virus. The NHSTT program had four main objectives:

- test – to increase the speed and availability of testing;
- trace – to identify recent close contacts of positive cases and notify those at risk who need to self-isolate;
- contain – to control local outbreaks through coordinated local responses;
- enable – to gather information about the virus to help the government ease infection control measures as scientific understanding evolved.

The program was initially rolled out to patients, healthcare workers, and essential key workers, before expanding to the broader community. Through a project by the UK Health Security Agency (UKHSA), I had access to data related to community-wide testing, and I will discuss how the use of this intervention varied among different sociodemographic groups in Chapters 3 and 4.

1.3 Heterogeneity in risk of infection

Mortality can be as a key indicator for measuring the impact of an outbreak. In England, even during the early phase of the pandemic (1 March 2020 to 31 May 2020), people who were living in the most deprived areas of England were more likely to die from COVID-19 [33]. Ethnic minorities, particularly Black and South Asian communities, also faced higher mortality rates [34]. Age was another critical factor, with older adults experiencing significantly higher mortality, exacerbated in those from disadvantaged backgrounds [35]. Differences in mortality highlight the cumulative impact of socioeconomic disparities on the entire chain of events, from the increased risk of infection to the ultimate health outcomes. Understanding and quantifying the heterogeneities in transmission has long been a fundamental concern in disease dynamics [36]. It has been discussed in light of implementing effective control strategies [37, 38] at multiple scales, from individuals to species to geographies. In broad terms, variability in two primary sources – susceptibility and exposure to the pathogen, can give rise to heterogeneity in disease dynamics. These can further be divided into individual or population-level variabilities, which often act together in complex ways. In general, mechanisms that lead to heterogeneity in the early stages of transmission, ultimately leading to differences in health outcomes.

Generally, mechanisms that create heterogeneity in the early stages of transmission, ultimately leading to differences in health outcomes, can be categorized into heterogeneity in susceptibility and exposure to infection.

1.3.1 Susceptibility

Individual-level variation can arise from biological differences in susceptibility to diseases. Various twin studies and genome-wide studies have shown that host genetic markers make individuals less or more susceptible to infectious diseases [39]. A classic example of this is the protective and amplifying relationship between sickle cell disease and the risk of contracting malaria [40]. Another example is the individual-level heterogeneity in vaccine-induced immune responses [41, 42]. At the population level, immunity plays a central role in answering questions about the level of immunity required in a population to achieve disease elimination, control, or eradication and governs the optimal immunization age [43]. However, the relationship between the level of the immunised population and the burden of disease is not always straightforward. The classic example of the rubella vaccine has shown that sub-optimal levels of vaccine coverage could lead to worse outcomes [44, 45], by increasing the average age of infection. In Greece, consistently low vaccination coverage in infant boys and girls, over 20 years, created a pool of susceptible women of child-bearing age. This led to the largest outbreak of congenital rubella syndrome in 1993 since the introduction of measles, mumps, and rubella vaccines. These studies highlight how asynchronous levels of immunity can impact disease persistence, also noted by Earn et al. [46]. In Chapter 5, I will explore the drivers of the synchronous and asynchronous patterns in the transmission of SARS-CoV-2 across England.

1.3.2 Exposure

Lloyd-Smith et al. characterised the role of individual variation in infectiousness on disease emergence for eight directly transmitted diseases [47]. The study showed that a disease with a high level of heterogeneity at this scale (i.e. at the individual level) could result in fewer but more explosive outbreaks. At the population level, contact patterns have been known to differ between individuals from different age groups [48], and studying its impact on disease dynamics is a common feature of various infectious diseases modelling studies [49]. Opatowski et al. showed that variation in H1N1 dynamics in eight countries could be explained primarily by age structures and levels of immunity in the populations. Countries that had a younger

population experienced higher transmission rates, perhaps due to pre-existing cross-immunity in older people [50]. In addition to age structure, incorporating the characteristics of the physical space in which individuals interact is important. In particular, House and Keeling showed that household size and its age composition are both important in predicting the early growth of cases, especially for influenza-like respiratory illnesses [48, 51]. Spatial heterogeneity is another pathway through which infectious diseases invade and persist in a population. The importance of spatial heterogeneity of host mixing in disease invasion and persistence process has been explored from various angles [52, 53]. Spatial structure in addition to age has been used to explain the well-documented persistence of measles in England [52]. More recently, the spatial distribution and growth of new cases of COVID-19 and its variants were predicted using high-resolution mobility data [54, 18, 19, 55]. I will be presenting some of the results from these studies in Chapter 2. Implications of spatial heterogeneity are not limited to directly transmitted pathogens. For an outbreak of chikungunya (a mosquito-borne virus) in Bangladesh, Salje et al. showed that individual characteristics (females spending more time at home) and localised social connections (how close individuals live to cases) were associated with a higher risk of infection [56]. More generally, accounting for behavioural patterns of how individuals interact with their contact network is of key importance in quantifying disease spread as well as for targeted control strategies. Anderson et al. showed that accounting for heterogeneity in sexual patterns can lead to estimating lower epidemic sizes [57]. Using detailed social network data of students from an elementary school in Pennsylvania, Cauchemez et al. showed that school-aged children were most important in introducing and spreading influenza in households [58]. During the 2002-03 outbreak of SARS in Hong Kong and Singapore, Li et al estimated that more than 70% of the infections were due to “super-spreading” events [59]. This was also seen more recently during the SARS-CoV-2 epidemic [60] and was attributed both to the social contact structures and infectiousness of individuals. In addition, there is also heterogeneity in access to diagnosis, and treatment, and in adherence to interventions [61, 62, 63].

1.4 Key epidemiological and statistical quantities

1.4.1 Growth Rate

The growth rate of a pathogen measures how quickly the number of cases is increasing or decreasing over time. The instantaneous growth rate is defined as the rate of change of the log-

transformed cases [64]. This quantity can give us an indication of the *speed* at which an epidemic is growing. When the growth rate is positive it indicates that the disease is spreading, and a negative growth rate suggests a decline in cases. Understanding the growth rate is crucial for assessing the potential spread of the disease and the effectiveness of interventions. I will estimate two variations of this metric in Chapter 2: (i) excess growth rate and (ii) relative growth rate.

1.4.2 Incidence

Incidence measures the number of *new* infections that occur in a population over a defined period (daily in this thesis). When I refer to incidence in this thesis, I mean the incidence of infection rather than just the reported cases. Incidence encompasses the actual number of new infections occurring in a population, including those that go undetected or unreported due to factors such as asymptomatic cases, limited access to testing, or delays in diagnosis. This distinction is crucial for accurately understanding the dynamics of disease transmission and assessing the overall impact of an epidemic.

1.4.3 Prevalence

Prevalence represents the total number of infectious individuals detectable by Polymerase chain reaction tests (PCR) in the population at a specific point in time (the number of individuals with a positive PCR test within a week, in this thesis). Prevalence is directly related to how long a person is detectable by the test, as a longer duration of detectable infection increases the number of active infectious individuals in the population (see Figure 1.1 illustrating the relationship between incidence and prevalence).

1.5 Methodological advancements in the thesis

Traditional prevalence estimation typically relies on gold-standard random community surveys, which provide unbiased estimates of infection prevalence by randomly sampling individuals. While accurate, these surveys are often limited by cost and logistical challenges, resulting in coarse spatial and temporal resolution. Nicholson et al. [65] improved this by incorporating mass testing data, which is more frequent and granular but biased. They used a causal debiasing framework to combine mass testing data with community survey results, producing more accurate, unbiased local-level prevalence estimates. Building on this, I extended their method



Figure 1.1: **Illustrative relationship between incidence and prevalence.** The numbers indicate how many individuals are detected as positive for a hypothetical infection, while the color gradient represents their pathogenic load indicating their current potential contribution to the prevalence estimate and/or true prevalence.

by using ethnic, age, and location information from the mass testing dataset to estimate weekly prevalence for different sociodemographic groups at a fine spatial scale.

The methods from *EpiForecast* and the ONS COVID-19 Infection Survey estimated incidence by using changes in prevalence over time, assuming an average infectious period [66, 67, 68]. Both approaches relied on PCR positivity data from ONS and therefore provide incidence estimates only for broad subgroups, such as the nine administrative regions or age groups at the national level. In contrast, the method I present in this thesis improves upon these by using unbiased prevalence estimates stratified by sociodemographic factors like age, ethnicity, and deprivation. I use these to estimate the incidence of different sociodemographic groups within local authorities in England. This allows for more granular, weekly incidence estimates that capture infection dynamics at a local-level across diverse communities.

We developed new methods to estimate the sensitivity and specificity of LFD tests using data from repeated testing, specifically positive LFD results followed by confirmatory PCR tests within three days. Additionally, I also presented a model that examines how the likelihood of a positive confirmatory PCR result changes over time and varies across ethnic groups, providing insights into different testing behaviors of groups of individuals.

In this thesis, I introduce a novel Bayesian nonparametric clustering method with two key advantages. First, unlike traditional clustering methods, this does not require the number of

clusters to be predefined. Second, our model is spatially agnostic and does not introduce any smoothing. We can use such methods for post-hoc regression analysis with covariates, which offers a more flexible approach to identifying drivers of clustering patterns.

1.6 Thesis outline

In this thesis, I investigated the dynamics of SARS-CoV-2 transmission in England from October 2020 to March 2022. I analysed spatial differences in transmission across the country, focussing on how mobility patterns and sociodemographic factors influenced exposure to the virus. I also explored variations in the uptake of public health interventions, particularly testing, which is critical for epidemic control. Additionally, I examined the spatial and sociodemographic drivers of transmission synchrony during this period. By quantifying these differences, my goal is to provide insights that can inform more equitable public health responses. I briefly describe the work done in the various chapters below:

1.6.1 Chapter 2

As discussed earlier, the global spread of various infectious diseases, including SARS-CoV-2, is facilitated by international travelers. Once emerging infectious diseases are seeded across different geographies, infections tend to occur within the community (for example, nursing homes, prisons, schools, and households). These settings are highly connected through multiple hierarchical layers (individuals nested in communities, cities, regions, countries, etc.) thereby sustaining the epidemics nationally [69, 70]. Any pandemic could be thought of as many small epidemics occurring in structured populations at the national and sub-national levels, all connected by the movement and interactions of individuals [71]. It is therefore important to understand how the accumulation and movement of individuals (immunologically naïve as well as infectious) within and into different geographies can shape how an epidemic unfolds, persists, or re-emerges. To this end, I explored the influence of mobility patterns in driving the initial invasion process of the Alpha and Delta VOCs of SARS-CoV-2 in England.

1.6.2 Chapter 3

Surveillance for infectious diseases is crucial because it allows public health authorities to identify outbreaks early, monitor the spread of the pathogens, and respond effectively to contain them.

By tracking the incidence and prevalence over time, surveillance provides essential data to guide public health interventions, allocate resources, and evaluate the effectiveness of control measures. Focussing surveillance for vulnerable groups is especially important because these populations might be at higher risk of not only severe outcomes but also at *risk of risk*, meaning they face compounded vulnerabilities due to socioeconomic disparities and health inequities [72]. For SARS-CoV-2, in Chile, people aged 0 to 40 years had an infection fatality rate around 3.1 times higher in the municipalities with the lowest socioeconomic status compared with the highest [28]. Studies also found a higher risk of infection associated with belonging to an ethnic minority group [73]. In this chapter, I estimated weekly prevalence and daily incidence at fine spatial scales and described how they varied by sociodemographic groups.

1.6.3 Chapter 4

The pandemic showed that the ability to adhere to non-pharmaceutical interventions such as physical distancing was a privilege correlated with socio-economic status [28, 74, 75], leaving those with lower incomes more vulnerable to infection and with worse health outcomes [76]. In England, the national testing programme was a crucial public health intervention that played an important role in controlling the spread of COVID-19 [77]. Under Pillar 2 of the testing strategy, free tests were made available to the general public, allowing widespread access to COVID-19 testing outside of healthcare settings. In this chapter, I examined whether there were any biases in testing across different sociodemographic groups and how testing behaviours varied among these groups.

1.6.4 Chapter 5

The dynamics of pathogen spread involve a complex interplay between rapid transmission or seeding of infection across different geographical areas and a moderated pace that prevents the complete depletion of susceptible individuals. While this balance is well-documented in the context of within-host dynamics and competition between pathogen strains, it also plays a crucial role in the broader synchrony and asynchrony that underpin sustained epidemics [78, 79]. In this chapter, I explored which areas in England had similar transmission levels, and how that changed over time. Additionally, I quantified how various spatial and sociodemographic factors were associated with these patterns.

Data is power, and telling a compelling story through insights generated from data is essential for informed decision-making in public health. Florence Nightingale famously used data to revolutionize healthcare, demonstrating how well-organized information could drive change and save lives.

In this thesis, I will present crucial statistics that describe the transmission dynamics of SARS-CoV-2 in England, offering an understanding of how the virus spreads across different regions and sub-populations. Additionally, I will introduce novel methods designed for epidemiologists, public health practitioners, and policymakers to enhance surveillance efforts and improve the equitable use of public health interventions. These contributions aim to not only deepen our understanding of the pandemic but also to equip those on the front lines with the tools necessary to respond effectively and equitably to future public health challenges.

Chapter 2

Impact of human behaviour on the invasion of SARS-CoV-2 variants of concern

During my first year of training in the NERC Doctoral Training Program (DTP), I became involved in a significant collaborative project focused on understanding the spatial spread of the SARS-CoV-2 Alpha variant of concern (VOC) in England. While my co-authors led the phylogeographic analysis sections, my responsibility was to quantify the spatial heterogeneity of the Alpha variant's spread across England. This included developing the statistical methodology, performing the analysis, interpreting the results, and contributing to the writing.

By summer 2021, I had completed the milestone for my *transfer to department* and began working on the newly emerging Delta variant. Specifically, I led the statistical analysis to understand the factors contributing to the growth of the variant. I contributed alongside other authors (including Prof. Moritz Kraemer) to the conception and design of the study. Professors Erik Volz and Samir Bhatt developed the initial model in [80], and I extended the statistical methodology, coded the methods, interpreted the results, and contributed to the writing and reviewing of the manuscript, in collaboration with Prof. Ben Lambert. My co-authors were responsible for the phylogeographic analysis.

In this chapter, I will present my contributions to two papers from this collaborative working group, both of which are included in the appendix.

Kraemer, Moritz U. G.*; Hill, Verity*; Ruis, Christopher*; Dellicour, Simon*; **Bajaj, Sumali***; et al. "Spatiotemporal Invasion Dynamics of SARS-CoV-2 Lineage B.1.1.7 Emergence." Science (2021). <https://doi.org/10.1126/science.abj0113>

McCrone, J.T.*; Hill, Verity*; **Bajaj, Sumali*** et al. "Context-specific emergence and growth of the SARS-CoV-2 Delta variant." Nature (2022). <https://doi.org/10.1038/s41586-022-05200-3>

2.1 Abstract

The emergence of SARS-CoV-2 variants of concern (VOCs) posed significant challenges to global public health, with each variant exhibiting changes in transmissibility, disease severity, and immune evasion. This chapter examines the spatial dynamics and growth rates of the Alpha and Delta variants, emphasizing the importance of genomic surveillance and mobility data in understanding their spread.

The Alpha variant (B.1.1.7), which emerged in the UK in late 2020, demonstrated higher transmissibility and hospitalization risk [80]. The main motivation of this study was that early in epidemics due to multiple seeding events, relative growth rates between the different variants might be inflated. By using mobility data we were able to explain these patterns. Using the S-gene target failure (SGTF) genomic signature as a proxy for Alpha, we estimated its growth rate and compared it with the growth rate of non-SGTF cases. We then used mobility data to highlight the role of human movement in the variant’s expansion. Our statistical analysis confirmed the phylogeographic finding that Alpha likely originated in Kent and spread across the UK, driven by mobility patterns from Kent and Greater London.

The Delta variant (B.1.617.2), first identified in India, became dominant in the UK by mid-2021. Its emergence in the UK took place within a diverse landscape of pre-existing immunity, both from previous infections and vaccinations, as well as various non-pharmaceutical interventions (NPIs). Genomic and mobility data indicated that Delta’s growth was influenced by human mobility, vaccination rates, and importation intensity. Using a Bayesian hierarchical model, we assessed the relative growth of Delta, accounting for various covariates such as mobility and vaccination coverage.

It is crucial to understand the drivers of the geographical spread of emerging VOCs to inform public health strategies, and our findings highlight the critical role of human mobility and public health measures in controlling the spread of SARS-CoV-2 variants. Effective response strategies require timely genomic surveillance and consideration of spatial dynamics to mitigate the impact of emerging VOCs.

2.2 Introduction

SARS-CoV-2 variants of concern (VOCs) [81] have played a critical role in the trajectory of the COVID-19 pandemic, presenting new challenges for public health responses worldwide. These variants, have demonstrated changes in transmissibility, disease severity, and the ability to evade immunity from prior infection or vaccination.

When a new variant emerges and starts to spread, it is crucial to understand where it originated and how it is spreading across geographies. Spatial heterogeneity can significantly impact the local likelihood and intensity of transmission, especially if the new VOC has higher transmissibility and poses a greater risk of adverse outcomes. Understanding these patterns helps in tailoring public health measures to effectively combat the spread of the virus, both locally and nationally or internationally.

2.2.0.1 Alpha

By late 2020, SARS-CoV-2 B.1.1.7 as known as the Alpha VOC started spreading rapidly across the UK. And while international travel restrictions were implemented to control the spread of this variant across countries, genomic surveillance found the VOC to be spreading in many countries [82]. It had also been shown that this variant had a higher transmissibility [80] and risk of hospitalisation [83] compared to other lineages. Using genomic and mobility data, my co-authors reconstructed the spatial dynamics of the emergence of the variant in the UK [18]. They found that the Alpha variant likely emerged in Kent in late September 2020. Between the start and end of the second English lockdown (5 November 2020 to 20 December 2020 [84]), the variant had spread across the country, driven by mobility patterns originating from Kent and Greater London. This initial spread was followed by local transmission [18], and mobility increased after the lockdown was lifted. The subsequent imposition of Tier 4 restrictions on 20 December 2020, which included a stay-at-home order and limitations on mixing, led to a reduction in exported Alpha variant cases from London and decreased mobility [84].

Using Polymerase chain reaction (PCR) tests with S-gene target failure(SGTF) as a proxy for B.1.1.7 infection, I estimated the growth rates of the Alpha VOC at the upper-tier local authority (UTLA) level. I focused particularly on comparing these growth rates in Greater London regions to those in the rest of the country. Additionally, I calculated the excess growth rate, defined as the difference between the growth rate of SGTF cases and that of non-SGTF cases. I then

analyzed how importation and human mobility influenced these excess growth rates across the UK.

2.2.0.2 Delta

By early-to-mid 2021, the SARS-CoV-2 (B.1.617.2 Pango lineage) Delta variant was rapidly spreading through India [85] and by May 2021 it had become the dominant lineage in the UK [86] and categorised as a VOC by the World Health Organisation.

The importation of the Delta variant coincided with a relaxation of social distancing measures in England [87]. More specifically, on 12 April 2021, outdoor dining and non-essential retail reopened. Using genomic and mobility data, my co-authors found that the majority of introductions of the Delta variant into England had originated from India [19]. Although mandatory hotel quarantine for arrivals from India was implemented on 23 April 2021, the phylogeographic analysis showed that the dominant transmission chains in England were already established before these international restrictions we enacted [19]. By 17 May 2021, restrictions on indoor dining and international travel were eased and amid the loosening of NPIs and increased social interactions, the virus spread rapidly throughout the country.

Building on results from the phylogeographic analysis, I then used a combination of epidemiological, aggregate human mobility and genomic data to test how the relaxation of non-pharmaceutical interventions (NPIs), estimated virus importations, and vaccination rates contributed to the growth rates of Delta in England. Prior to the work presented in this chapter, models for estimating the increase in transmissibility of new variants were typically based on increases in relative frequency [80] but rarely took into account other potential confounding factors, such as variation in population behaviour, vaccination rates or numbers of independent virus introductions. I modified and extended the Bayesian hierarchical model developed by Volz et. al [80], to estimate the effects of the above mentioned variables on the weekly relative growth of the Delta VOC.

2.3 Methods

2.3.1 Estimating the growth rate of Alpha

2.3.1.1 Data

Cases

Daily new cases by specimen date were downloaded from <https://coronavirus.data.gov.uk/details/download>. Data was last accessed on 29 January 2021. We also used the line-list data provided and collected by Public Health England (now UK Health Security Agency), which included information about the SGTF test result status. The details of how my collaborators merged the two datasets are provided in the supplementary methods section of the paper [18].

Estimated importation intensity (EII)

We used data from the COVID-19 Infection Survey by the Office for National Statistics (ONS) [88] as of January 29, 2021, to estimate the percentage of individuals testing positive for SARS-CoV-2 across UK. To calculate the importation intensity (EII) of the B.1.1.7 variant from Greater London, we multiplied the modelled daily positivity rates in Greater London by the estimated weekly movements to other UK locations. We then separated the EII into components for B.1.1.7 and non-B.1.1.7 variants using the weekly proportion of B.1.1.7 genomes found in Greater London.

Human mobility

The mobility data representing the number of trips between pairs of Local authority districts (LADs) were constructed using data from the mobile service provider O2, and the Google COVID-19 aggregated mobility research dataset. Details of this is the supplementary methods of the paper [18].

2.3.1.2 Model

To estimate the daily epidemic growth rates in each lower/upper-tier local authority (LTLA/UTLA), we fit a mixed effects generalised linear model (GLM), accounting for variations across different areas by including both fixed and random effects. This was done using the

R package lme4 v.1.1-21 [89].

For each 6-day sliding window, we analysed the log-transformed number of new cases, $\log(Y_i + 1)$ (where the "+1" adjustment accounts for zero cases), for LTLAs/UTLAs i using the regression model:

$$\log(Y_i) = \text{intercept}_i + r_i * \text{day}, \quad (2.1)$$

where:

- $\log(Y_i)$ represented the log-transformed number of new cases for LTLA/UTLA i .
- intercept_i was the baseline log-transformed number of new cases for LTLA/UTLA i , capturing the starting level of new cases.
- r_i was the growth rate for LTLA/UTLA i , indicating how the log of the number of new cases changes per day.
- day represented the time variable (in days).

The model included fixed effects: the average growth rate r representing the typical rate at which the log of the number of new cases increased per day across all LTLAs/UTLAs; and random effects, which accounted for the variability among LTLAs/UTLAs. Each LTLA/UTLA could have its own intercept (intercept_i) and its own growth rate (r_i), reflecting differences in baseline levels and growth rates of new cases.

In summary, this model estimated how the number of new cases evolved over time, allowing for variability in initial levels and growth rates across different LTLAs.

Excess growth rate

To calculate the instantaneous growth rates for both SGTF and non-SGTF cases, we fit a mixed effects log-linear GLM of new daily case counts as described above - stratified by SGTF test results. We estimate the total SGTF and non-SGTF in each LTLA/UTLA i and day d by adjusting counts upwards in proportion to total cases reported daily in each LTLA/UTLA.

$$\text{Estimated SGTF cases}_i(t) = \frac{\text{Observed SGTF}_i(t)}{\text{Total observed cases}_i(t)} \times \text{Total reported cases}_i(t),$$

$$\text{Estimated non-SGTF cases}_i(t) = \frac{\text{Observed non-SGTF}_i(t)}{\text{Total observed cases}_i(t)} \times \text{Total reported cases}_i(t), \quad (2.2)$$

where, $\text{Total observed cases}_i(t) = \text{Observed SGTF}_i(t) + \text{Observed non-SGTF}_i(t)$.

See Appendix Figure A.1 to see the proportion of total reported cases that were samples and classified into either STGF or non-SGTF cases.

We then calculated the excess growth rate; egr by subtracting the estimated growth rate for non-SGTF cases from SGTF cases at each day between 15 November and 20 December 2020 (for sensitivity we extended that period to 5 January 2021- see Appendix Figure A.2). We then used estimates from the phylogeographic reconstruction on the number of estimated imports of B.1.1.7 from Greater London, $phlyo_imp$, and estimated importation intensity as described in Section 2.3.1.1, EII , as predictors of the excess growth rate. We log transformed the mobility data to compresses the wide range of these values, making the model more stable and less sensitive to extreme values. The regression is described below:

$$egr_{ti} = \beta_{0i} + \beta_1 \log(EII)_{ti} + \beta_2 phlyo_imp_{ti} + \epsilon_{ti}, \quad (2.3)$$

where,

$$\beta_{0i} = \beta_0 + u_{0i},$$

are the random intercepts for LTLA i , and

$$\epsilon_{ti} \sim \mathcal{N}(0, \sigma_e^2),$$

$$u_{0i} \sim \mathcal{N}(0, \sigma_u^2).$$

These estimated growth rates were used in subsequent models in the paper (see section *Human mobility and prior outbreaks as predictors of B.1.1.7 growth* in [18]) to estimate the effects of mobility, and previous attack rates.

2.3.2 Modelling the relative growth of Delta

2.3.2.1 Data

UK population data

Population size data were used to calculate the proportion of the population that was partially or fully vaccinated in a location. UTLA-level 2020-mid-year population size estimates were downloaded from <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland>.

Incidence data from England

COVID-19 case count data for each LTLA were downloaded via <https://coronavirus.data.gov.uk/details/download>.

COVID-19 case count and vaccination data for the UK

COVID-19 cases and the cumulative vaccination data were downloaded for each UTLA from 30 January 2020 to 28 July 2021 by specimen and dosage date, respectively, via <https://coronavirus.data.gov.uk/details/download>. These data included positive laboratory-based polymerase chain reaction (PCR) tests and positive lateral flow tests, but did not include tests where the lateral flow test was positive and PCR follow-up tests were negative (further details at <https://coronavirus.data.gov.uk/details/about-data>). The COVID-19 case count at the UK country level was calculated by aggregating case data on the UTLA level. Additionally, to match the genomic data, the COVID-19 case count and vaccination data for some UTLAs were aggregated under an area code made up of these multiple UTLAs (Table 2.1). All entries with the recently discontinued area code E10000002 were assigned the new area code E06000060.

Area and area code	Constituents
Greater London, E13000001, E13000002	E09000007, E09000011, E09000012, E09000013, E09000019, E09000020, E09000022, E09000023, E09000028, E09000030, E09000032, E09000033, E09000001, E09000002, E09000003, E09000004, E09000005, E09000006, E09000008, E09000009, E09000010, E09000014, E09000015, E09000016, E09000017, E09000018, E09000021, E09000027, E09000024, E09000025, E09000026, E09000029, E09000031
West Midlands, E11000005	E08000026, E08000029, E08000025, E08000028, E08000030, E08000027, E08000031
South Yorkshire, E11000003	E08000019, E08000018, E08000017, E08000016
Tyne and Wear, E11000007	E08000037, E08000021, E08000022, E08000023, E08000024
Merseyside, E11000002	E08000012, E08000014, E08000011, E08000013, E08000015
Greater Manchester, E11000001	E08000003, E08000007, E08000008, E08000004, E08000005, E08000002, E08000001, E08000010, E08000006, E08000009
West Yorkshire, E11000006	E08000035, E08000036, E08000034, E08000033, E08000032

Table 2.1: **Grouping of UTLA area codes under higher level area codes as used in the analyses.**

Mobility data

Relative self-mobility, which measures the intensity of movements within a specific location, is calculated as the percentage of weekly movement during the period from 22 March 2020 to 6 June 2021 relative to the highest recorded movement in that location during the same timeframe. Details of the Google COVID-19 aggregated mobility research dataset used to produce estimates of the intensity of mobility within a location (as defined above) is given in our paper [19].

New Delta lineage introductions

Weekly new lineage introductions into the UK by UTLA were obtained from the continuous phylogenetic analysis and by aggregating daily estimates. This is described in our paper [19].

2.3.2.2 Data pre-processing

We kept data starting from the 13th epidemiological week, which corresponded to the week commencing 28 March 2021. These dates will be referred to as baseline elsewhere in this chapter.

We excluded weeks after the first time 95% of samples were observed to be Delta in each UTLA because, after this point, we could no longer estimate the relative growth rates reliably since Delta was effectively fixed in the population. We kept only those UTLAs which had data on Delta for at least 9 weeks. In the final dataset, we had 683 observations (across 64 UTLAs with approximately 11 weeks of non-missing data on average for each UTLA (Table 2.4).

2.3.2.3 Model

In this section, I describe how we modelled the dynamics of Delta penetration within a UTLA. In particular, we model how the number of Delta samples per UTLA (i), varied over time (t) (measured in weeks).

The underlying transmission conditions driving the observed number of Delta samples in a given UTLA may be similar to other UTLAs within the same region. We modelled this variation hierarchically and indexed variables at the UTLA level by $i[j]$ to indicate that UTLA i was nested within (the overarching) one of the nine administrative regions j . We used a binomial sampling distribution to model the number of Delta samples $Z_t^{i[j]}$:

$$Z_t^{i[j]} \sim \text{binomial}(Y_t^{i[j]}, p_t^{i[j]}), \quad (2.4)$$

where $Y_t^{i[j]}$ was the total number of sequenced samples, and $0 \leq p_t^{i[j]} \leq 1$ was the corresponding proportion of Delta samples in UTLA i in week t . We then transform this probability so that it is on the unconstrained logit scale:

$$\phi_t^{i[j]} = \text{logit}(p_t^{i[j]}). \quad (2.5)$$

A key quantity of interest was the relative growth in the proportion of Delta on the logit (i.e. log odds) scale, which we estimate weekly and we denote by $\rho_t^{i[j]}$, where

$$\phi_t^{i[j]} = \phi_{t-1}^{i[j]} + \rho_{t-1}^{i[j]}. \quad (2.6)$$

Relative growth for each UTLA was then modelled spatially depending hierarchically on the region, j , containing it. Relative growth was also assumed to depend on UTLA-specific covariates by the following relation:

$$\rho_t^{i[j]} = \rho_t^j + \beta' x_t^{i[j]} + \delta_t^{i[j]}, \quad (2.7)$$

where ρ_t^j was a region-level growth trend, $x_t^{i[j]}$ was a vector of covariates, and $\delta_t^{i[j]}$ was a UTLA- and week-specific term representing the deviation from the region-level growth. To account for temporal auto-correlation in the relative growth rate, a given region's relative growth was assumed to follow a random walk centred around its relative growth in the previous week:

$$\rho_t^j \sim \text{normal}(\rho_{t-1}^j, \sigma_2). \quad (2.8)$$

We estimated the model in a Bayesian framework with priors described in Table 2.2, and performed prior predictive checks to see how reasonable our choices were.

The computations were done in R and Stan [90] using four parallel chains with 50,000 to 60,000 iterations (depending on the model), half of which were discarded as warm-up iterations; the chains were subsequently thinned by a factor of 10. In all cases, MCMC sampling was diagnosed as converged with $\hat{R} < 1.01$, and bulk and tail effective sample sizes > 400 for all parameters. For 6 of 15 models used for model comparison (which included neither the no covariate model nor the best fit model), there remained 2 out of 4,410 parameters which had $\hat{R} > 1.01$ or had a tail effective sample size below 400; in all cases, the bulk effective sample sizes exceeded 400. In these models, the last two weeks were held-out from each UTLA to perform out of sample predictions, resulting in a smaller dataset, which likely explains the difficulty in obtaining convergence with 50,000 iterations.

Parameter	Prior distribution
ρ_1^j	normal(0,1)
$\delta_t^{i[j]}$	normal(0, σ_1)
$\phi_1^{i[j]}$	normal(a^j , b^j)
a^j	half-normal(-3,1)
b^j	half-normal(1.5,1)
σ_1	half-normal(0,5)
σ_2	half-normal(0,5)
β	half-normal(0,5)

Table 2.2: **Prior distributions for model parameters.**

2.3.2.4 Model checks

Counterfactuals

Our model outputted two sets of key quantities; (1) the weekly relative growth rate of Delta over time $\rho_t^{i[j]}$, and (2) the estimated ‘effect’ of a variable on the growth of Delta (β). To determine the implications of the effect sizes, we computed the estimated proportion of Delta samples when the covariates took factual versus counterfactual values.

Model comparison

To assess the importance of covariates, we compared the predictive performance of models which included different sets of covariates. All covariates were standardized by subtracting the mean and dividing by the standard deviation. Since the cumulative proportions vaccinated (considering either the cumulative proportion vaccinated with a 1st or 2nd dose) increased monotonically throughout the time period of observation, we included the UTLA-level mean of these variables in our regressions: i.e., we used time-invariant regressors. We did so to avoid the risk of spurious association due to both Delta and proportions vaccinated growing coincidentally. Covariates were chosen as important predictors if including them in the model improved the model fit on a hold-out set comprising the last two weeks of data for each UTLA.

Simulation and model robustness

The primary objective of the simulation process was to determine if the parameters within the model could be reliably recovered from hypothetical data generated under known conditions, providing validation for the model’s structure and robustness. The simulation was designed to mimic the hierarchical structure of the actual data. The following steps were implemented:

- Fixing parameter values: The parameters σ_1 and σ_2 were assigned values of 0.6 and 0.2, respectively. The parameters β s corresponding to the two covariates described below were fixed at 0.1 and 2.
- Data Structure: We simulated data for:
 - Five regions, j , each containing six UTLAs, totaling 30 UTLAs i .
 - 15 weekly time points, t .

- UTLA-Level covariates: For each of the 30 UTLAs we created $x_t^{i[j]}$. We first created a covariate for week and a corresponding covariate with values ranging from 0.2 to 0.5 as weeks progressed. This covariate mimics the trends of proportion of individuals immune over time.
- Relative Growth Simulation:
 - For each region, the initial regional growth ρ_j^1 rate was drawn randomly from a normal(0, 1), and the subsequent weekly regional growths ρ_j^t were normally distributed around the previous week's growth, and with standard deviation σ_2 , according to equation 2.8.
 - We then added UTLA specific deviations $\delta_t^{i[j]}$ and covariate information using fixed values of β s and $x_t^{i[j]}$ according to equation 2.7. This generated weekly UTLA-specific relative growth $\rho_t^{i[j]}$.
 - Then, to generate probability values on the logit scale for each UTLA, $\phi_t^{i[j]}$, we used equation 2.6, where initial values $\phi_1^{i[j]}$ for the five regions were fixed at $-1, -3, -4, -3, -2$.
 - Next, we used the inverse logit transformation on these unconstrained $\phi_t^{i[j]}$ to convert them to the probability scale, $p_t^{i[j]}$ (equation 2.5).
 - Like in our real data, we excluded the weeks after which the Delta proportion, $p_t^{i[j]}$ exceeded 95%.
 - Then, by fixing the number of samples $Y_t^{i[j]}$ at 200 for each UTLA, we drew the number of Delta samples $Z_t^{i[j]}$ following the binomial relation from equation 2.4.
- Model Implementation and parameter recovery: The simulated dataset was analyzed using the same modeling framework as the real data, estimating Delta samples, transforming probabilities to the logit scale, and modeling relative growth rates hierarchically while accounting for temporal autocorrelation.
- Evaluation: We compared the estimated parameters from the simulation to the original fixed values. We were able to recover our parameter values, and the model converged with $\hat{R} < 1.01$, bulk and tail effective sample sizes > 400 after 20,000 iterations, discarding 10,000 warm-up iterations and thinning by a factor of 10 (Figure 2.1 and Table 2.3).

This simulation process provided a structured framework for testing parameter identifiability, ensuring that the results obtained from real-world data analysis were reliable.

Parameter (example covariate)	Known value	Posterior mean (95% Bayesian credible interval)
Covariate 1 (baseline immunity)	2.0	1.67 (-0.43, 3.62)
Covariate 2 (week)	0.1	0.18 (0.13, 0.23)
σ_1	0.6	0.49 (0.43, 0.55)
σ_2	0.2	0.17 (0.09, 0.27)

Table 2.3: **Simulation: known vs estimated parameters.**

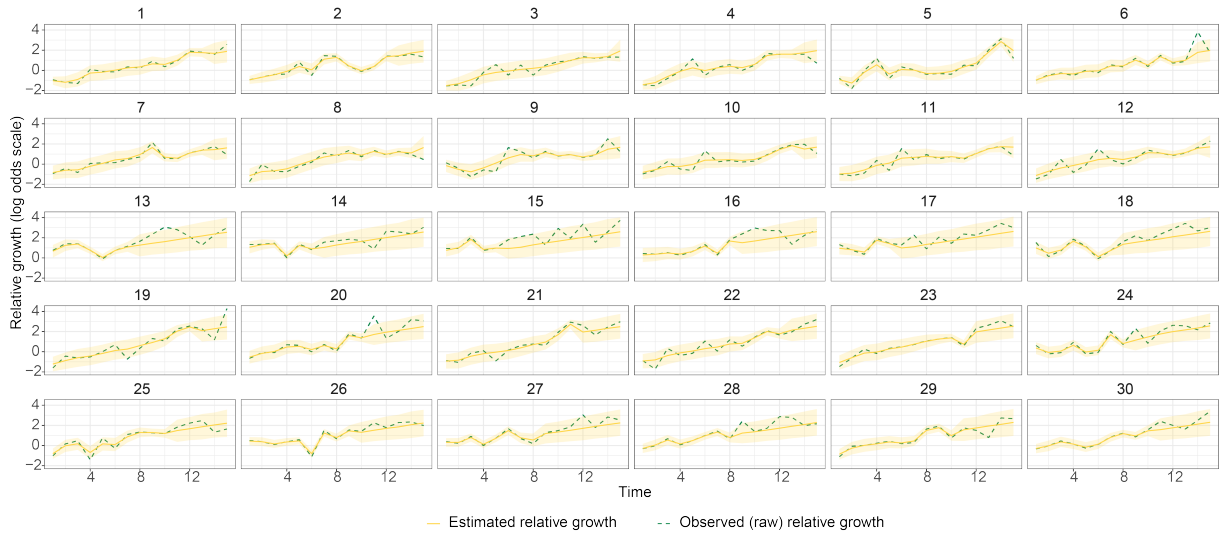


Figure 2.1: **Simulation comparing known vs estimated relative growth rates for hypothetical locations.** The solid yellow and dashed green line represent the estimated (posterior median) and observed relative growth rate of Delta simulated in thirty hypothetical areas. The yellow ribbons represent the 95% credible intervals.

Prior and posterior predictive checks

To validate the appropriateness of the priors used in our Bayesian model, we conducted prior predictive checks. These were performed by generating data from the model using only the prior distributions of the parameters, in the absence of any actual data. As described in Section 2.3.2.4, we chose priors (Table 2.2) so that a wide range of Delta proportions were possible yet were centred on low values in the absence of further information: our prior predictive distributions in Figure 2.2 illustrate these characteristics.

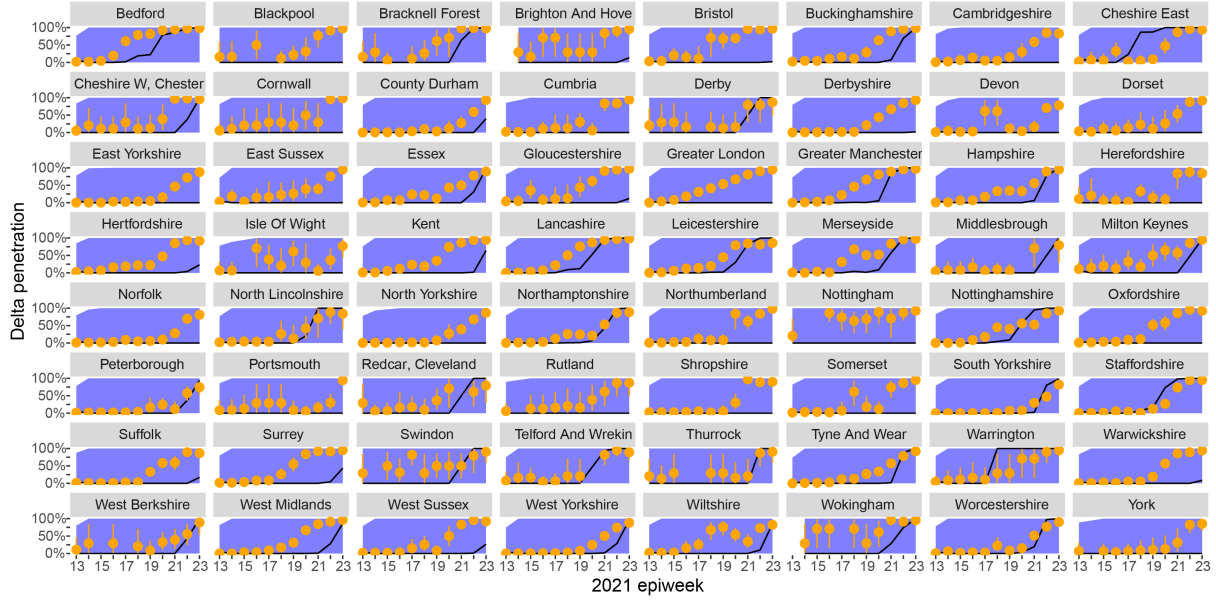


Figure 2.2: **Prior predictive simulations for the Delta frequency.** Orange point-ranges display the observed data (points indicate posterior medians and whiskers indicate 2.5%–97.5% posterior quantiles assuming a uniform prior for each independent data point). The blue shaded region represents the prior simulated 2.5%–97.5% quantiles; lines indicate median simulation values.

After the fitting the Bayesian model, we conducted posterior predictive checks. For this, we generated data from the model using the estimated posterior distributions (including the priors and observed data or likelihood) of the parameters.

2.3.2.5 Code availability

Code to reproduce the statistical analyses on Delta growth can be found at <https://github.com/sumalibajaj/Delta-Statistical-analysis-share>.

2.4 Results

2.4.1 Spatial heterogeneity in the expansion of Alpha

Using SGTF PCR-positive tests as an indicator for B.1.1.7 infection [86], we examined the daily growth rates of SARS-CoV-2 cases at the LTLA level for both SGTF and non-SGTF cases. Due to testing changes and the implementation of Tier 4 restrictions in the Southeast, we excluded case data from 25 to 31 January (See Appendix Figure A.2 to see the R^2 becoming noisy during this period). Following the lockdown in November 2020, case growth rates were highest in regions

of southeast England connected to Greater London and/or Kent. The acceleration in SGTF case growth rates in Greater London started in mid-November, preceding similar trends in other regions (Figure 2.4B). At the UTLA level, SGTF case growth rates were higher than those of non-SGTF cases (Figure 2.3A), supporting the increased transmissibility of B.1.1.7 [80, 91].

We extended these findings by quantifying the importation of B.1.1.7 cases from London and examining the relationship between importation trends and lineage-specific case growth rates. The phylogeographic analysis [18] showed that the increase in B.1.1.7 importation rates into lower-tier local authorities (LTLAs) closely matched the early growth rates of SGTF cases in those areas, such as Birmingham, Liverpool, and Manchester (Figure 2.4). We also calculated the regional difference between SGTF and non-SGTF case growth rates, and found an additive increase in SGTF growth rate of 0.0715, which corresponded to median multiplicative advantage of 1.576, assuming a generation time of 6.5 days. This was qualitatively similar to those reported previously [80, 91] with the caveat that generation times may differ between B.1.1.7 and other lineages [92, 93] (Figure 2.3B).

The positive correlation between the difference in case growth rates and the B.1.1.7 importation rate grew during the latter half of the November lockdown and remained very high ($R^2 > 0.75$) until mid-December, before declining. This trend persisted even when we accounted for uncertainties in the estimated number of infections across Greater London, and by restricting the LTLAs in the analysis once they reached different levels of SGTF sample proportions (Figure A.4). Our results remained robust across different data and methods used to estimate B.1.1.7 importation rates. The continued export of B.1.1.7 from Greater London and Kent partially explained the decline in growth advantage estimates for B.1.1.7 during the second half of December 2020, before the implementation of stricter control measures on December 20 (Tier 4).

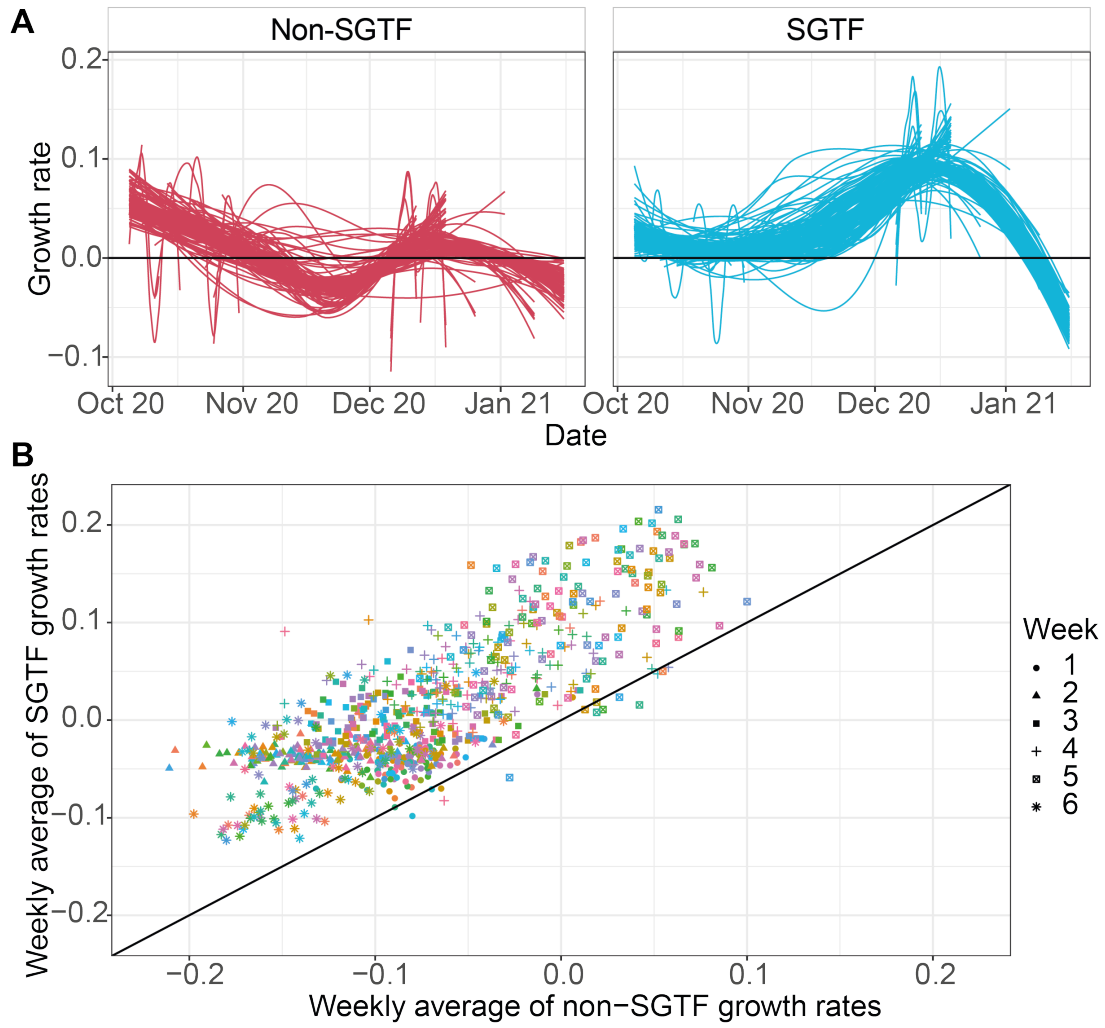


Figure 2.3: **Weekly growth rates per LTLA for SGTF and non-SGTF cases.** (A) Rolling 7-day average growth rates of SGTF and non-SGTF cases at the LTLA level. Each line represents an LTLA and shows a smoothed trend to illustrate the growth rates. (B) Symbols represent weeks between November 15 (week 1) and December 20 (week 6) and colors represent LTLAs.

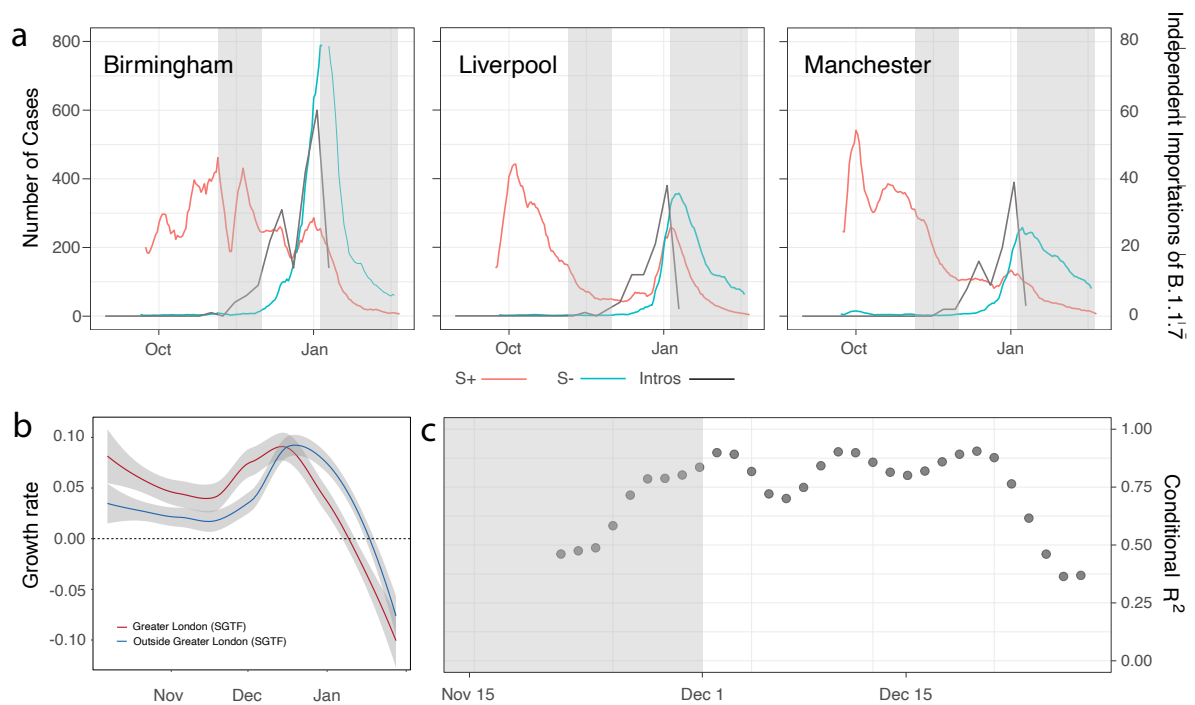


Figure 2.4: **Case growth rates of B.1.1.7 were correlated with human mobility and attack rates across the UK.** (A) Seven-day rolling average number of cases reported with SGTF (blue) and non-SGTF (red) for Birmingham, Liverpool, and Manchester. The black line represents the weekly number of independent introductions estimated from the phylogeographic analysis, and the gray shaded area indicates the timing of the second (5 November to 1 December) and start of the third (5 January) English lockdown. (B) Seven-day rolling average growth rates of SGTF cases in Greater London (red line) and outside of Greater London (dark blue line), and the grey areas represent the 95% confidence intervals. (C) Association between per-LTLA difference between SGTF and non-SGTF case growth rates (adjusted for variations in sampling intensity of SGTF cases over time) and number of B.1.1.7 importations into that region, as estimated from prevalence surveys and human mobility (gray dots) (See Appendix Figure A.3 for the R^2 when cases were not scaled up). The gray area shows the time of the second England lockdown. Conditional R^2 is the proportion of variance explained by both fixed and random effects in the model.

2.4.2 Factors contributing to the growth of Delta

Table 2.4 summarizes the UTLA-level data used in the analyses over roughly 10 weeks. On average, there were 139 estimated introductions of the Delta variant across UTLAs (range: 0-4777) during this period, according to the phylogeographic analysis. Average vaccine coverage of 1st and 2nd doses across UTLAs was 50.1% (range: 1.3%-65.3%), and 23.6% (range: 0.6%-34.8%), respectively.

Overall, growth rates of the Delta variant varied substantially across different locations and

weeks in England (Figure 2.5A, Figure 2.6). Our model indicated that the most important predictor of Delta’s growth (relative to Alpha) across UTLAs in England was the within-UTLA human mobility (Table A.1 shows the out-of-sample predictive performance of variables). This refers to the relative changes in weekly mobility within-UTLAs compared to the pre-pandemic period (Figure 2.5A, B). The critical role of within-UTLA mobility during the emergence of a new variant was unsurprising, as pre-emptive restrictions on movement and social interactions tend to slow the spread of new pathogens or variants (see counterfactual scenarios in Figure 2.6. The relaxation of NPIs therefore, increased both within- and among-region transmission (Figure 3C in [19]).

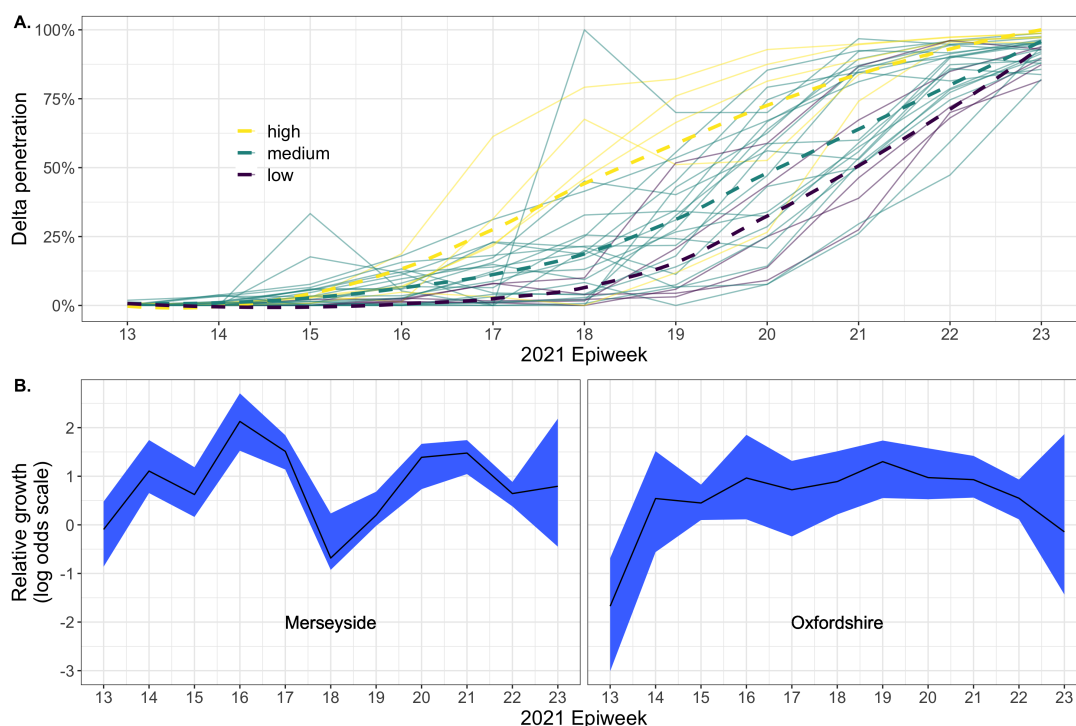


Figure 2.5: Variation in Delta growth rates across UTLAs in England. (A) The increase in Delta frequency compared with Alpha at the UTLA level. UTLAs are coloured according to the level of average within-UTLA mobility: high, the five UTLAs with the highest within-location mobility; low, the five UTLAs with the lowest within-location mobility; medium, the remaining UTLAs. Solid lines show data for given UTLAs; dashed lines show LOESS curves fit to the data for each mobility category. (B) Examples of weekly growth (the solid line corresponds to posterior medians) of UTLAs with high (left) and low (right) within-UTLA mobility. The shaded regions represent the corresponding 95% Bayesian credible intervals (2.5^{th} and 97.5^{th} quantiles of the posterior distribution). In (A), data are shown only for UTLAs with 500 or more sequenced samples.

To assess the impact of mobility as a covariate, we computed the estimated proportion of Delta samples under different scenarios of within-UTLA mobility, while holding all other covariates took

their observed values. Specifically, we considered counterfactual scenarios where within-UTLA mobility was set to its minimum value (0) and maximum value (1). The minimum value of 0 represents no movement within the location, while the maximum value of 1 corresponds to the highest recorded movement within the location during the period from 22 March 2020 to 6 June 2021 (as described in Methods section 2.3.2.1). The results of these counterfactual simulations are shown in Figure 2.6. In Cornwall, East Sussex, North Yorkshire, the Isle of Wight, and Rutland, observed mobility levels were notably lower than maximum potential mobility, likely because residents could afford to reduce their movements. The average Index of Multiple Deprivation score across these areas was approximately 16, indicating lower levels of deprivation compared to a score of 20 for other LTLAs, highlighting their relatively favorable conditions to reduce mobility.

Variable	Mean (minimum, maximum)
Number of weeks with data	10 (9, 11)
Number of introductions of Delta	139 (0, 4777)
Cumulative 1 st dose of vaccination (%)	50.1 (1.3, 65.3)
Cumulative 2 nd dose of vaccination (%)	23.6 (0.6, 34.8)
Within-UTLA mobility	0.8 (0.6, 0.9)

Table 2.4: **Summary of the data used in the excess growth analysis of Alpha VOC**
The mean (minimum, maximum) of the UTLA-level data over the course of the study.

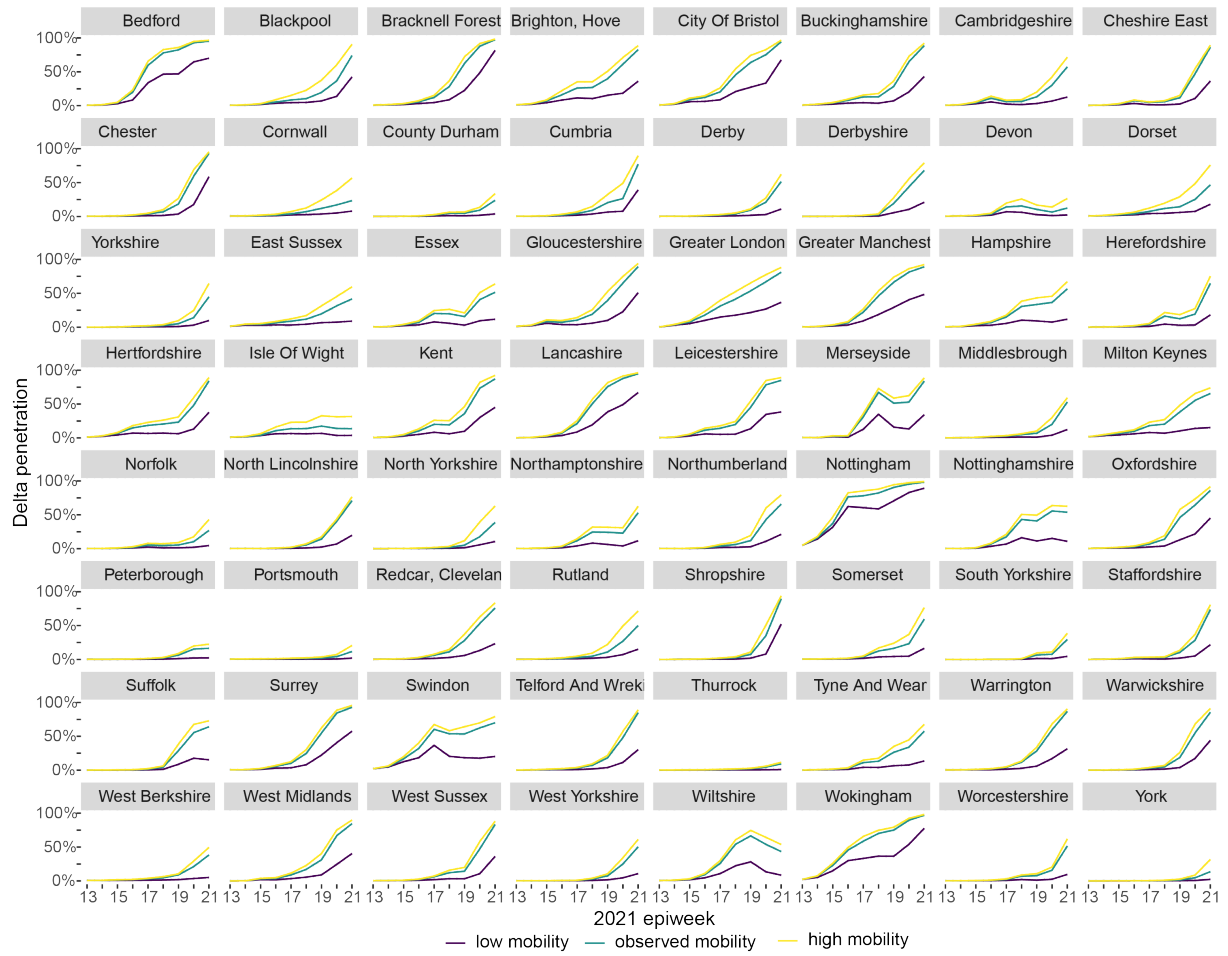


Figure 2.6: **Counterfactual scenarios for the impact of within-UTLA mobility.** (A) Estimated and observed proportions of Delta variant samples across UTLAs, for various counterfactual scenarios: minimum (purple) and maximum relative within-UTLA mobility (yellow), observed within-UTLA mobility (light blue). All lines represent median posterior estimates using the model with epiweek and within-UTLA mobility as covariates which was fit to data from 2021: epiweeks 13–21.

2.4.2.1 Model checks for Delta

Model comparison

Our best model included within-UTLA mobility and time since baseline, which outperformed the model where we included no covariates (Appendix Table A.1). A model including the cumulative proportion vaccinated with a second dose, time since baseline and within-UTLA mobility also outperformed the model with no covariates. However, the improvement in prediction accuracy was minimal, and this was the only model outperforming the no covariate model which included vaccinations, so we do not take this as strong evidence of the importance of vaccination in

slowing Delta growth. Our model's fit did not improve when incorporating weekly numbers of independent viral introductions estimated from genomic data or vaccination rates (Appendix Table A.1).

Prior and posterior predictive checks

The posterior predictive check (in Figure 2.7B) for delta frequency (proportion or penetration within a UTLA) demonstrated that majority of the observed values were within the 95% credible intervals of the model predictions, indicating that our model provided a good fit to the data.

As another measure of model appropriateness, we compared the raw estimates to the modelled growth rate (on the log odds scale) in Appendix Figure A.5. The goal was for the generated data to closely resemble the observed data, indicating that the model accurately captures the underlying processes. We observed substantial overlap between the observed growth rates and the posterior median of the modeled relative growth (falling within its 95% credible intervals).

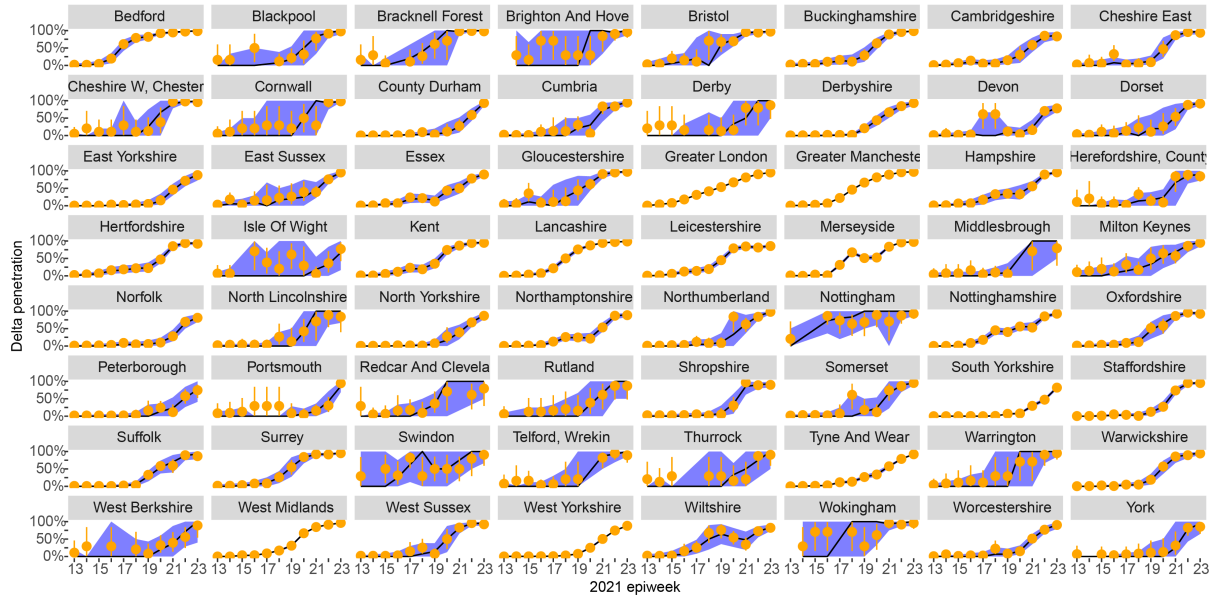


Figure 2.7: **Posterior predictive simulations for the Delta frequency.** The orange point-ranges display the observed data (points indicate posterior medians and whiskers indicate 2.5%–97.5% posterior quantiles assuming a uniform prior for each independent data point). The blue shaded region represents the posterior simulated 2.5%–97.5% quantiles and lines indicate median simulation values.

2.5 Discussion

Throughout the COVID-19 pandemic, genomic surveillance has played a crucial role in identifying and tracking the emergence of SARS-CoV-2 variants. Combining genomic data with statistical methods, as described in this paper, helps reveal the drivers of transmission and inform public health responses. This interdisciplinary approach is especially vital in time-sensitive situations, allowing for a more holistic understanding of the problem.

This chapter presents work done during the critical time of the rapid rise and spread of Alpha and Delta VOCs in the UK. It specifically focusses on using importation patterns from phylogeographic analysis done by my collaborators to be used as inputs into the statistical models which aim to describe drivers of growth rates. Additionally, it builds on previous work in the literature [80] by incorporating covariates to investigate the factors influencing the spatial variability in the growth of new VOCs.

For Alpha, we observed that its emergence in the UK was associated with high levels of its exportation from a source location within England. The early growth rates likely reflected the increased transmissibility of the variant, and connectivity of areas to the source location (Greater London and Kent).

In contrast to the Alpha variant, which spread from a single location in the south east of England due to its connectivity, we found that Greater London was played a lesser role in the expansion of Delta across the country. The phylogeographic analysis revealed that imports from India, and later other countries, found a fertile ground to expand as social restriction were easing. Soon after its introduction, local transmission patterns became more important in establishing this VOC.

Our analyses rely on the use of aggregated mobility data which have a number of limitations. This data is limited to smartphone users who have opted-in (the default is for these data not to be collected for a user) to sharing their location through the Google location history feature. And it may not be representative of the wider population.

We have conducted statistical analyses that revealed associations between a set of predictors and the outcome, specifically the growth of a VOC. While these insights are valuable in identifying relationships, it is crucial to recognize that these estimates of association do not imply causation. Causal inference is important because it allows us to understand the underlying mechanisms driving these relationships and assess the true impact of various drivers on VOC growth. An

essential extension of this work would involve developing causal models to derive insights into how different factors influence the growth of VOCs, thereby enabling more effective interventions and public health strategies. Given that mobility is a key predictor of transmission dynamics, another extension of this could be to develop intricate models that explore the dynamic feedback loop between individual behaviors in response to an epidemic and how these behaviors, in turn, influence the spread of the disease.

Our findings underscore that while genetic changes in emerging variants are crucial in determining if a variant will replace an existing lineage, mobility patterns governed by connectivity to high transmission areas, and changes in contact patterns due to the implementation of non-pharmaceutical interventions are critical in governing the spatial transmission dynamics of VOCs.

Chapter A

Appendix - Impact of human behaviour on the invasion of SARS-CoV-2 variants of concern

Covariates	log-likelihood
epiweek, within-UTLA mobility	-11,321
epiweek	-11,475
epiweek, within-UTLA mobility, cumulative proportion of 2 nd dose of vaccination	-11,487
no covariates	-11,546
epiweek, cumulative proportion of 2 nd dose of vaccination	-11,693
epiweek, cumulative proportion of 1 st dose of vaccination	-11,708
within-UTLA mobility	-11,865
cumulative proportion of 2 nd dose of vaccination	-12,090
cumulative proportion of 1 st dose of vaccination	-12,464
epiweek, introductions (2 week lag)	-35,702
introductions (2 week lag)	-36,210
introductions (1 week lag)	-122,778
epiweek, introductions (1 week lag)	-136,796
introductions	-150,302
epiweek, introductions	-166,290

Table A.1: **Out of sample prediction (leaving out the last two weeks of data).** Comparing models across a suite of covariate combinations.

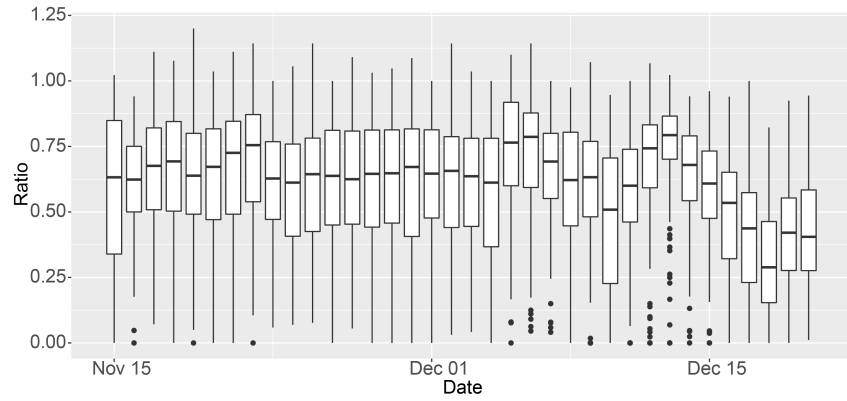


Figure A.1: **Ratio of number of samples that have information about SGTF status compared to all reported cases.** The box plots included ratios for each LTLA. Each box plot displays the median (middle horizontal line), interquartile range (top and bottom horizontal line), and potential outliers (dots).

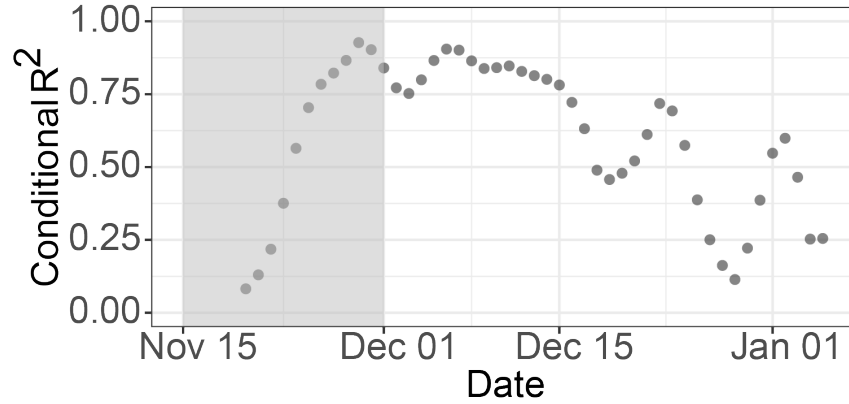


Figure A.2: **Association between LTLA specific excess growth rates and mobility and imports beyond 20 December 2020.** Association between per-LTLA difference between SGTF and non-SGTF case growth rates (adjusted for variations in sampling intensity of SGTF cases over time in grey) and number of B.1.1.7 importations into that region, as estimated from prevalence surveys and human mobility.

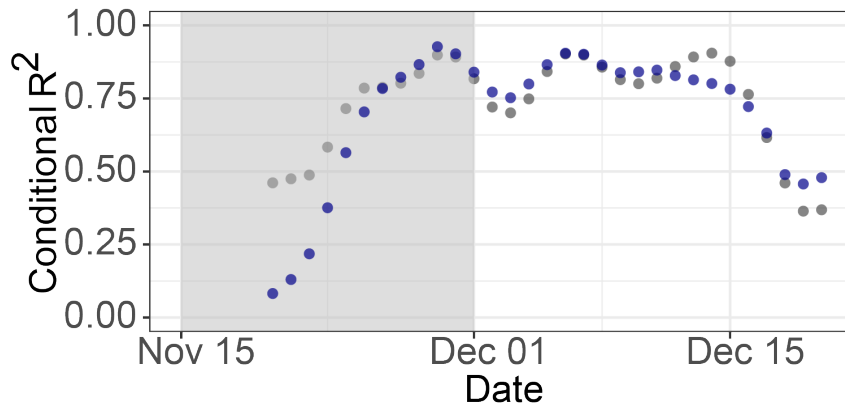


Figure A.3: **Association between LTLA specific excess growth rates and mobility and imports for observed and scaled up cases.** Association between per-LTLA difference between SGTF and non-SGTF case growth rates (adjusted for variations in sampling intensity of SGTF cases over time in grey and not adjusted in blue) and number of B.1.1.7 importations into that region, as estimated from prevalence surveys and human mobility.

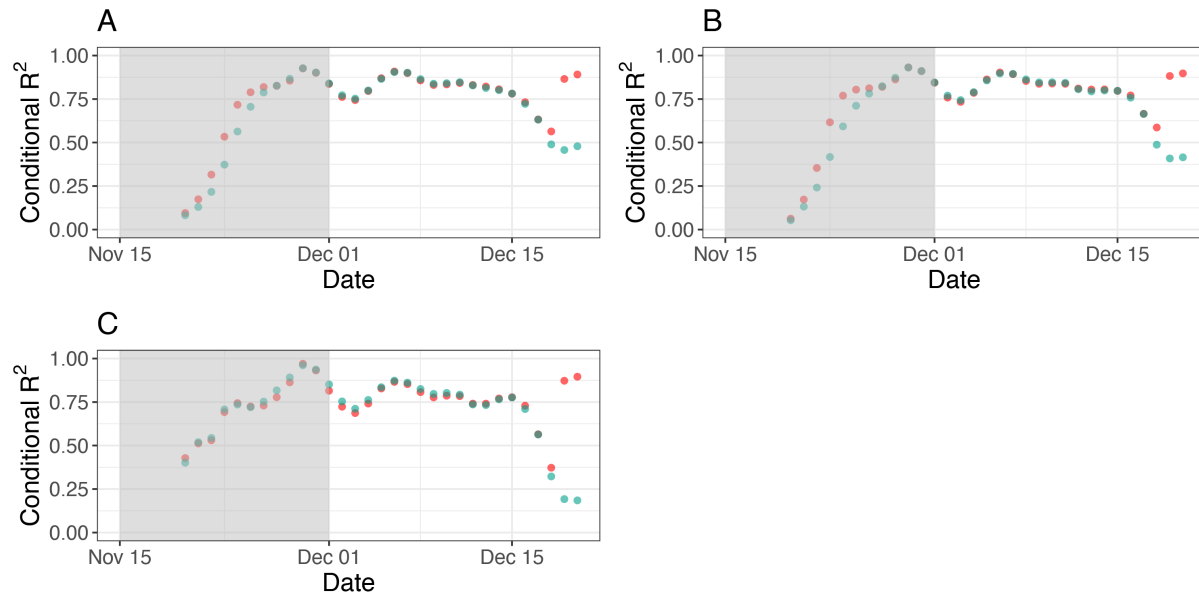


Figure A.4: **Sensitivity results of only including data once Alpha crosses thresholds of 0.2 and 0.6 in an LTLA.** (A) Association (conditional R^2) between LTLA specific difference between SGTF and non-SGTF case growth rates and number of importations as measured by estimated number of importations from prevalence surveys and human mobility (blue dots) and as measured by the genomic data (red dots). Grey area shows the time of the second English lockdown. (B) Same as (A) but only keeping subsequent rows for a given LTLA once 0.2 threshold of daily SGTF count/(SGTF count + non-SGTF count) is crossed. (C) Same as (A) but only keeping subsequent rows for a given LTLA once 0.6 threshold of daily SGTF count/(SGTF count + non-SGTF count) is crossed.

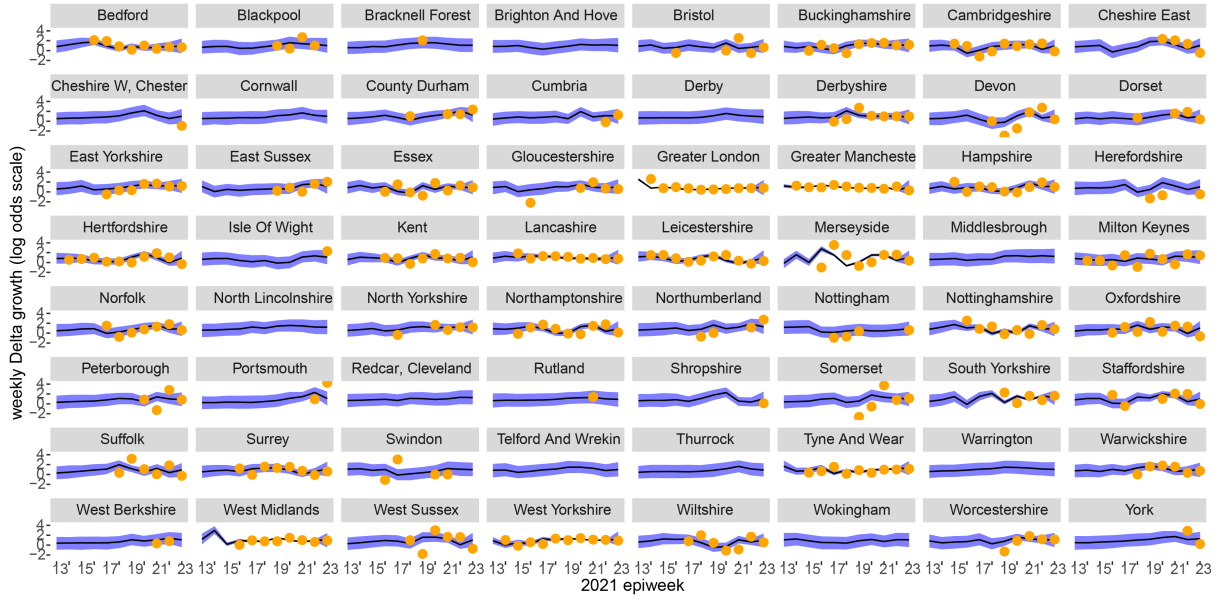


Figure A.5: **Observed (raw) vs posterior predicted growth rates.** Time-varying relative growth of Delta (on the log odds scale). The shaded regions represent the corresponding 2.5%–97.5% Bayesian credible intervals. The orange points indicate the raw growth rates calculated directed from the data (note, there are missing values in these data due to the presence of dates when no samples were taken or when the frequency of Delta remained at 0% or 100%).

Chapter 3

SARS-CoV-2 transmission dynamics between sociodemographic groups

In October 2022, I joined the EY-Oxford Health Analytics Consortium, appointed by the UK Health Security Agency to conduct a retrospective review of England’s COVID-19 testing programme. As part of the statistics team, I initially focused on evaluating three testing services: those for schools, healthcare workers, and adult social care. By January 2023, after gaining a clearer understanding of the extensive data, testing policies, and protocols, I shifted my focus to examining how test usage varied across different vulnerable groups in society. Chapters 3 and 4 detail the specific research questions and methods developed to address this issue. This work is in press with The Lancet Digital Health and is available as a preprint [94]. The manuscript is divided into two chapters presented here to highlight the range and depth of the methods and analyses used to understand biases in SARS-CoV-2 testing among various sociodemographic groups over the 18-month study period. I led this project, including conceiving the research plan and developing methods in collaboration with Prof. Ben Lambert (BL), Dr. Siyu Chen, and Dr. Richard Creswell. I performed all the analyses, interpretation, writing, and revisions, along with BL and Dr. Reshania Naidoo.

Bajaj, Sumali*; Chen, Siyu*; Creswell, Richard*; Naidoo, Reshania*, Tsui, Joseph L.-H*; et al. “Understanding COVID-19 testing behaviour in England through a sociodemographic lens.” medRxiv (2023). <https://doi.org/10.1101/2023.10.26.23297608>

In the previous chapter I quantified the spatial heterogeneity in the rise of Alpha and Delta variants of concern (VOCs). Building on these insights, in this chapter I quantify how transmission patterns varied not only spatially, but also across sociodemographic groups.

3.1 Abstract

Understanding patterns of transmission at fine scales is crucial for determining targeted and effective intervention strategies, particularly because transmission is thought to respond to local conditions.

We used mass testing data for lateral flow device (LFD; data for 290 million tests performed and reported) and polymerase chain reaction (PCR; data for 107 million tests performed and returned from the laboratory) tests made available for the general public, provided by date, self-reported age and ethnicity at lower tier local authority (LTLA) level. Using a mechanistic causal model to debias the PCR testing data, we obtained estimates of weekly SARS-CoV-2 prevalence by self-reported ethnic groups and age groups for LTLAs in England. This approach to debiasing the PCR (or LFD) testing data also estimated a testing bias parameter defined as the odds of testing in infected versus not infected individuals, which would be close to zero if the likelihood of test seeking (or seeking and reporting) was the same regardless of infection status. We also estimated the daily incidence allowing us to determine the fraction of cases captured by the testing programme.

From March 2021 onwards, individuals in the most deprived regions reported approximately half as many LFD tests per-capita than those in the least deprived areas (Median ratio [Inter quartile range, IQR]: 0.50 [0.44, 0.54]). During October 2020 – June 2021, PCR testing patterns were in the opposite direction (Median ratio [IQR]: 1.8 [1.7, 1.9]). Infection prevalences in Asian or Asian British communities were considerably higher than those of other ethnic groups during the Alpha and Omicron BA.1 waves. Our estimates indicate that the England COVID-19 mass testing program detected 26% - 40% of all cases (including asymptomatic cases) over the study period with no consistent differences by deprivation levels or ethnic groups.

Our work shows how mass testing data can be used in conjunction with surveillance surveys to identify gaps in the uptake of public health interventions at fine scale levels and by sociodemographic groups. It provides a framework for monitoring local interventions and yields valuable lessons for policymakers in ensuring an equitable response to future pandemics.

3.2 Introduction

The English national COVID-19 testing programme was rapidly implemented and significantly expanded, distributing a total of 2 billion lateral flow devices (LFDs) and registering 158 million polymerase chain reaction (PCR) tests, between October 2020 and March 2022 [77].

Although the testing programme was rolled out as a public health intervention with the primary aim of reducing onward transmission and a means to reopening society, it can also give us insights into the underlying transmission in the population. However, relying solely on testing data would yield biased estimates of the infection prevalence. This is because: (i) PCR tests were primarily used for symptomatic testing or as part of test-to-release protocols [95] and do not capture all the asymptomatic infections, and (ii) although LFD tests were recommended for asymptomatic testing, they likely suffer from reporting bias, unlike PCR tests.

On the other hand, community surveillance random sampling survey testing programmes such as REACT-1 [96] and ONS CIS [88] provided a gold-standard measure of infection prevalence; they had large sample sizes nationally and could produce precise estimates of prevalence at this level. However, at combinations of fine geographic or other scales (LTLA level and sociodemographic groups), the effective sample size in such surveys is much reduced, particularly because SARS-CoV-2 prevalence was generally low ($< 10\%$).

To integrate insights from both mass testing and community surveillance random sampling surveys, we implemented and extended the debiasing methodology of Nicholson et al. [65], leveraging the extensive mass testing dataset from the UK Health Security Agency (UKHSA). Our analysis focussed only on the Pillar 2 testing data, which includes PCR and LFD tests made available to the general public (not including tests made available to individuals with clinical need, and health and care workers through Pillar 1).

The debiasing method uses swab results from the REACT-1 study to correct for biases in the PCR mass testing data and so produces fine-scale PCR swab positivity rate estimates at a higher precision than possible through the community surveillance survey alone.

An uptick in cases may signal future increases in hospitalisations and deaths. However, case data are imperfect, being influenced by changes in both the true incidence of infection and changes in testing / test reporting behaviour. To address this, we developed a novel methodology to estimate the time-varying SARS-CoV-2 infection incidence for sociodemographic groups. Additionally, this allowed us to estimate the reach of the national testing programme.

This chapter provides a framework for estimating key epidemiological quantities: prevalence and incidence, at a fine spatial and sociodemographic scale, which can be used by local decision-makers to design and adapt equitable interventions. In this chapter and the next, when I refer to incidence and prevalence, I specifically mean *infection incidence* and *infection prevalence*.

3.3 Methods

3.3.1 Data pre-processing

3.3.1.1 LTLA codes

Due to LTLA boundary reassignments, some historical LTLA codes are now inactive, so we updated them and assigned regions (as given by the Office of National Statistics) where necessary according to Table 3.1 below.

Inactive LTLA codes	New LTLA codes
E07000004, E07000005, E07000006, E07000007	E06000060
E06000029	E06000058
E07000049	E06000059
E07000150, E07000152, E07000153, E07000156	E06000061
E07000151, E07000154, E07000155	E06000062
E07000205, E07000206	E07000244
E07000201, E07000204	E07000245
E07000190, E07000191	E07000246

Table 3.1: **LTLA code reassignment for inactive codes.**

LTLA code	Region
E07000004, E07000005, E07000006, E07000007	South-East
E07000009	East
E07000150, E07000151, E07000152, E07000153, E07000154, E07000155, E07000156	East Midlands

Table 3.2: **Regions assigned to LTLAs where this information was missing.**

The Isles of Scilly and City of London had no data on ethnicity (LTLA codes = E06000053 and E09000001 respectively) and were removed from the ethnicity analysis.

3.3.1.2 Deprivation

ONS published the Indices of Deprivation (IoD2019) based on 39 separate indicators, organised across seven distinct domains of deprivation (Income Deprivation (22.5%), Employment Deprivation (22.5%), Education, Skills and Training Deprivation (13.5%), Health Deprivation and Disability (13.5%), Crime (9.3%), Barriers to Housing and Services (9.3%), Living Environment Deprivation (9.3%)) which are combined and weighted to calculate the Index of Multiple Deprivation 2019 (which we abbreviate as ‘IMD’). This is an overall measure of multiple deprivation experienced by people living in an area and is calculated for every Lower-layer Super Output Area (LSOA) and was aggregated to other administrative levels by the ONS. To calculate the LTLA level measure by taking the population weighted average of all the LSOA scores in each LTLA. In this chapter, each LTLA was assigned to one of nine groups according to this aggregated ‘IMD- Rank of average score’ [97], with one and nine corresponding to most and least deprived, respectively. E06000060, E06000061, E06000062, do not have an IMD assigned to them in the latest dataset and were dropped from the analysis.

3.3.1.3 Ethnicity

The ethnic population data were available only to the nearest thousand [98]. These estimates (published by the Office of National Statistics on 04 December 2019) were based on the Annual Population Survey (which is the Labour Force Survey plus various sample boosts), the mid-year population estimates and the 2011 Census, using the method described in [99]. A little over half (53%) of the LTLAs had zero as inputs for certain ethnic group populations, which meant that the true number could be a fraction between zero and one (in thousands). In this case, we assumed 0.5 (in thousands) or 500 individuals belonged to such a group in a given LTLA. We now describe the process of assigning ethnic population sizes by LTLAs to match the 2020 population estimates [100]. Let Y_{ij}^0 be the population (in thousands) of the i^{th} ethnic group in LTLA j , and $\sum_i Y_{ij}^0 = Y_j^0$ was the population of LTLA j from the ethnic population data from 2019 [98]. According to the assumption mentioned above, we updated Y_{ij}^0 as follows:

$$Y_{ij} = \begin{cases} Y_{ij}^0, & \text{if } Y_{ij}^0 > 0, \\ 0.5, & \text{otherwise,} \end{cases} \quad (3.1)$$

and $\sum_i Y_{ij} = Y_j$ was the updated population size of LTLA j . We then calculated the proportion

p_{ij} of ethnic group i in LTLA j as:

$$p_{ij} = \frac{Y_{ij}}{Y_j}. \quad (3.2)$$

These proportions were then multiplied by the mid-2020 population size estimates [100] for the LTLAs, Y_j^{new} , to obtain scaled up new ethnic group population sizes in LTLAs, Y_{ij}^{new} :

$$Y_{ij}^{new} = p_{ij} Y_j^{new}. \quad (3.3)$$

Reclassifying *Prefer not to say/Unknown* ethnicities

In the Pillar 2 mass PCR and LFD testing data, individuals self-reported their ethnicities and had the choice to select the option of “Prefer not to say/Unknown”. We reclassified these responses in the ethnicity column according to the proportion of other ethnicities for each day (or week) in a given LTLA.

For instance, let us consider reclassifying $X_{un,i,t}$, which represents the number of LFD-positive cases with unknown ethnicity reported in the i^{th} LTLA during week t . The number (and proportion) of LFD-positive cases in this LTLA with self-reported ethnicities for each of the eth ethnic groups – “Asian or Asian British”, “Black; African; Black British or Caribbean”, “White”, “Mixed or Multiple ethnic groups”, and “Other ethnic group” is defined as $X_{eth,i,t}$ (with proportion $p_{eth,i,t} = \frac{X_{eth,i,t}}{X_{i,t}}$), where $X_{i,t} = \sum_{eth} X_{eth,i,t}$ is the total number of LFD-positive cases with self-reported ethnicities in LTLA i during week t . Then $X_{un,i,t}$ was reclassified and the updated numbers for the known ethnic groups was given by:

$$X_{eth,i,t}^{new} = X_{eth,i,t} + (p_{eth,i,t} X_{un,i,t}). \quad (3.4)$$

For example, say there were 10 LFD positive cases ($X_{un,i,t}$) reported during a given week in an LTLA that were classified as “Prefer not to say/Unknown” ethnicity, and $X_{eth,i,t}$ was disaggregated as the following: say there were $X_{a,i,t} = 20$ LFD positives classified as “Asian or Asian British”; $X_{b,i,t} = 30$ as “Black; African; Black British or Caribbean”; $X_{w,i,t} = 30$ as “White”; $X_{m,i,t} = 10$ as “Mixed or Multiple ethnic groups”; and $X_{o,i,t} = 10$ as “Other ethnic group”. Then after the redistribution the number would be $X_{a,i,t}^{new} = 20 + (0.2 * 10) = 22$, $X_{b,i,t}^{new} = 30 + (0.3 * 10) = 33$, $X_{w,i,t}^{new} = 30 + (0.3 * 10) = 33$, $X_{m,i,t}^{new} = 10 + (0.1 * 10) = 11$, $X_{o,i,t}^{new} = 10 + (0.1 * 10) = 11$, respectively.

3.3.1.4 Age

Age group bounds were not always consistent between REACT-1, mass testing datasets, and population estimates, and we assumed equivalence between age groups as described in Table 3.3 below:

REACT-1	Testing and population size data
5-11	4-11
12-17	12-17
18-24	18-23
25-34	24-33
35-44	34-43
45-54	44-53
55-64	54-63
65-74	64-73
75+	74+

Table 3.3: **Age groups bins in REACT-1, mass testing, and population data.** Ages (in years) included in age-group bins in REACT-1 and in either the Pillar 2 testing data obtained from UKSHSA or the population size data.

3.3.2 Variant time periods

3.3.2.1 Data

To calculate the weekly frequencies of Variants of Concern (VOCs) in England, i.e., the weekly proportion of SARS-CoV-2 infections that were of a particular VOC, we used weekly lineage counts from genomes sequenced at the Wellcome Sanger Institute, downloaded from <https://covid19.sanger.ac.uk/downloads> (last accessed on 28 July 2023). Lineages were assigned according to the PANGOLIN nomenclature [101], and the number of each lineage is enumerated by week (according to sample dates) and LTLA.

3.3.2.2 Method

To obtain estimates of the weekly lineage frequencies at the national level, lineage counts were aggregated over all LTLAs in England. We considered only the major waves of VOCs that occurred during the study period, so weekly genome counts were further aggregated according to whether the lineage belongs to Alpha (all sub-lineages of B.1.1.7 and Q), Delta (all sub-lineages of B.1.617.2 and AY), and Omicron (all sub-lineages of B.1.1.529 and BA). Lineages that did not fall under these three VOCs were considered as “wild-types”. The weekly frequencies of these

three VOCs and the wild-type of SARS-CoV-2 were calculated at weekly intervals throughout the study period. To estimate the time periods during which these three VOCs were dominant, we used the first and last date when the frequencies of each VOC exceeded 90%. The time periods where each variant was dominant are as follows (Figure 3.4, 3.1):

- Alpha: 30 January 2021 to 24 April 2021
- Delta: 5 June 2021 to 4 December 2021
- Omicron: 1 January 2022 to 14 May 2022 (although our analysis period ended on 31 March 2022)

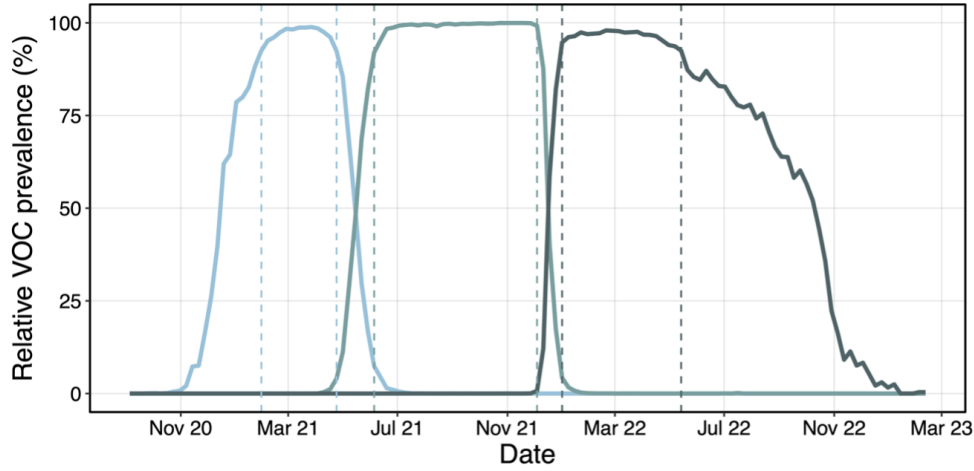


Figure 3.1: **Variant time periods.** Time periods during which the Alpha, Delta and Omicron (left-to-right in figure) variants of concern (VOC) were dominant ($\geq 90\%$ relative prevalence) in England. The solid curves are the VOC-specific weekly proportions using lineage counts from the Wellcome Sanger Institute [102], and the dashed coloured lines delineate the time periods when the relative proportion of each of the VOCs equalled or exceeded 90%.

3.3.3 Causal debiasing approach for obtaining time-varying estimates of SARS-CoV-2 prevalence at fine scale resolution

3.3.3.1 Method

This method compares REACT-1 and the Pillar 2 mass testing data at a “coarse” level (e.g., administrative regions, IMD regions, or age-group) and outputs a testing bias, which is given by:

$$\Omega = \log \left[\frac{\text{odds}(\text{tested} \mid \text{infected})}{\text{odds}(\text{tested} \mid \text{not infected})} \right]. \quad (3.5)$$

Ω at this coarse level was then used to correct Pillar 2 testing data at a lower level (e.g. LTLA or ethnicity within LTLA or age-group within LTLA) to produce estimates of SARS-CoV-2 prevalence at this scale. The coarse and fine scale levels for obtaining the different sets prevalence estimates are described in Table 3.4

In essence, the method estimates testing bias at a coarse scale where there are high numbers of both community surveillance survey tests and mass tests. These biases are then propagated to the nested finer spatial scales to estimate prevalence, making use of the sufficiently large volume but biased mass testing data at fine scales. Implicitly, this means that we assumed that biases at coarse scales were representative of those at finer scales, and violations in this assumption would result in biased estimates of infection prevalence.

The REACT-1 data consisted of numbers of tests conducted by week within each LTLA, along with the results of these tests. In LTLA-weeks with fewer than five observations, the data were censored if there were any positives to ensure the data were not identifiable.

For plotting purposes, we obtained national (or regional) ethnic-group prevalences by a weighted mean of our fine scale ethnic-group SARS-CoV-2 prevalence estimates and the respective population sizes of these groups in LTLAs (Figure 3.5A). For national-level plots of prevalences by age, we used a weighted mean of the fine scale SARS-CoV-2 prevalence estimates and the respective population size of each age-group within an LTLA (Figure 3.5B).

While it is not a key output of this chapter, we also estimated LTLA-specific prevalences when these fine levels were nested in their nine administrative regions of England (as done previously [65]). This was used for model comparison as described in the *Model comparison* section 3.3.5 below, to plot the spatial prevalence maps (Appendix Figure B.3), and to calculate national incidence and case detection ratios as well as to determine its association with IMD (see section 3.3.6 of *Incidence and case detection ratio estimation* and Figure 3.6A and B).

Sociodemographic group	Fine scale levels	Coarse levels
Deprivation (IMD)	LTLAs	National IMD levels
Ethnicity	Ethnic groups within LTLAs	Administrative regions of England
Age	Age groups within LTLAs	National age groups
NA	LTLAs	Administrative regions of England

Table 3.4: **Summary of fine scale and coarse level groupings for the different debiasing and incidence analyses.**

3.3.4 Association between LTLA prevalence and PCR testing behaviour

To test if PCR testing quantities were associated with prevalence, we performed weekly (t) linear regressions between LTLA (i) specific prevalences, $p_{t,i}$, and PCR tests performed per capita, $n_{pcr,t,i}$, as follows:

$$\mathbb{E}(n_{pcr,t,i}) = \beta_{0,t} + \beta_{1,t}p_{t,i}. \quad (3.6)$$

Then using the observed $p_{t,i}$ and estimated $\beta_{0,t}$ and $\beta_{1,t}$, we predicted the weekly LTLA-specific number of PCR tests performed per capita $\hat{n}_{pcr,t,i}$. We then used the weekly medians and interquartile ranges of these predicted values by IMD to plot Appendix Figure B.4C.

3.3.5 Model comparison

We compared three different groupings to determine if IMD levels were appropriate as coarse levels. This was done to compare the robustness of the prevalence estimates from our novel coarse level grouping to the published methodology and results [65].

For this comparison, we had access to weekly LTLA-specific data from the REACT-1 study, including the number of PCR tests performed and how many of those were positive. While we had access to this data, we did not use it to calculate LTLA-specific prevalences due to the small sample sizes in each LTLA. Instead, we used the data to compare the estimated prevalences generated by different grouping methods and to determine which method produced estimates most similar to the observed data.

We assessed the model's fit by summing the log-likelihoods for the weekly fine-scale LTLA prevalences. The likelihood was modeled using the Binomial distribution as follows:

$$X_{pos,i,t} \sim \text{binomial}(X_{total,i,t}, \text{prev}_{i,t}), \quad (3.7)$$

where $X_{total,i,t}$ is the observed count of PCR tests, $X_{pos,i,t}$ is the count of PCR positives from the REACT-1 survey in LTLA i during week t , and $\text{prev}_{i,t}$ is the posterior mean of the fine-scale prevalence estimates.

The model with the higher likelihood (or log-likelihood) is considered a better fit to the data, meaning its prevalence estimates for the LTLAs over the study period best explain the observed REACT-1 PCR positive data.

3.3.6 Incidence and case detection ratio estimation

3.3.6.1 Model

SARS-CoV-2 infection prevalence is dependent on recent histories of infection occurrence, which we modeled using the following equation (with both SARS-CoV-2 prevalence and infection occurrence measured on the same scale). We defined an individual to be infection-positive if their infection was detectable by PCR with a cycle threshold (Ct) value < 37 .

$$\theta_d = \sum_{\tau=1}^{\infty} w_{\tau} i_{d-\tau}, \quad (3.8)$$

where θ_d denotes prevalence on day d and i_d is the incidence on the same day; $0 \leq w_{\tau} \leq 1$ is the probability of being PCR positive given an infection which began τ days ago. If the set of probabilities denoting PCR positivity $\{w_{\tau}\}_{\tau=1}^{\infty}$ were precisely known, the above equation provides a linear system which can be inverted to determine an incidence time series $\{i_d\}_1^T$. But these probabilities are typically determined from regular testing data on individuals whose date of exposure is known or can be determined with a degree of certainty [103], and these probabilities likely changed throughout the course of the pandemic with the introduction of novel SARS-CoV-2 variants [104]. To account for these imperfections in our knowledge of $\{w_{\tau}\}_{\tau=1}^{\infty}$, we assumed a probabilistic relationship between incidence and prevalence of the form:

$$\theta_d \sim \text{normal}\left(\sum_{\tau=1}^{\infty} w_{\tau} i_{d-\tau}, \sigma_{inc}\right), \quad (3.9)$$

where $\text{normal}(\mu, \sigma)$ represents a normal distribution with mean μ and standard deviation σ ; we assumed that the θ_d were independent random variables, which would likely be violated in presence of substantial autocorrelation in the misspecification of $\{w_{\tau}\}_{\tau=1}^{\infty}$; σ_{inc} is estimated and characterises the degree of imperfections in our model.

We used a Bayesian framework to estimate infection incidence over time, $\{i_d\}_1^T$, and assumed a prior that the true infections were a constant to-be-estimated multiple of the reported cases given by reported PCR and LFD positive counts as follows:

$$i_d = \frac{\text{reported cases}_d}{\text{case detection ratio}_d}, \quad (3.10)$$

with case detection ratio $_d$ corresponding to the 15-day block that time d corresponds to. Within

the 15-day intervals, we assumed the case detection ratio to be piecewise constant.

Since the reported cases are the dates the cases appear in the testing data, the date of infections likely precedes these, and our assumption is that this lag is minimal relative to longer-term changes in case reporting behaviours.

Our approach to estimate the true case incidence as described above required daily estimates of prevalence and how the PCR detectability varied over the course of a typical infection throughout the study period. We also chose to use positive test results reported via UKHSA’s Pillar 2 PCR and LFD data to improve the precision of our incidence estimates (Equation 3.10). We elaborate on these three inputs below.

3.3.6.2 Daily SARS-CoV-2 prevalence estimates

For case detection ratios in Figure 3.6A and B we used LTLA-specific debiased prevalence estimates obtained using the methodology explained in the section 3.3.3 *Causal debiasing approach for obtaining time-varying estimates of SARS-CoV-2 prevalence at fine scale resolution*, where the LTLAs were nested within the nine administrative regions of England (Table 3.4). These were used to produce LTLA-specific incidences and then aggregated to produce national incidence estimates. We also used publicly available national prevalence estimates published by ONS [88] to obtain another set of national incidence estimates. For the ethnicity analysis, we used the mean prevalence estimates for each ethnic group within LTLAs of England.

The debiased SARS-CoV-2 prevalence was itself estimated with a margin of uncertainty, and we included this uncertainty in our analysis via a prior probability distribution which links the true prevalence with that estimated by our incidence model (<https://github.com/sumalibajaj/equitytesting>). The debiasing approach [65] estimated mean prevalence and with uncertainty bounds, and the prevalence estimates from ONS had upper and lower uncertainty bounds. We assumed that this uncertainty was captured by a normal distribution,

$$\theta_t \sim \text{normal}(\theta_{t,central}, \sigma_{t,central}), \quad (3.11)$$

where θ_t was the true prevalence for week t and $\theta_{t,central}$ was the mean prevalence estimate from the debiasing methodology or as published by ONS, and $\sigma_{\theta,t}$ was the uncertainty in these estimates. Standard errors were back-calculated from each of the lower ($\theta_{t,lower}$) and upper ($\theta_{t,upper}$) bounds for weekly prevalences of each fine scale level and national ONS estimates as

follows:

$$\begin{aligned}\sigma_{\theta,t,upper} &= \frac{\theta_{t,upper} - \theta_{t,central}}{1.96}, \\ \sigma_{\theta,t,lower} &= \frac{\theta_{t,central} - \theta_{t,lower}}{1.96}.\end{aligned}\tag{3.12}$$

We then took the average of these two standard errors and used it as the standard deviation of the specified normal priors for prevalence in our model:

$$\sigma_{\theta,t} = \frac{\sigma_{\theta,t,upper} + \sigma_{\theta,t,lower}}{2}.\tag{3.13}$$

Since the models we used were formulated at the daily level, we approximated daily prevalence (and its standard error) by linearly interpolating the weekly estimates. This likely understates the uncertainty in daily prevalence but this source of uncertainty was unlikely to have affected the qualitative trends we found across longer timescales.

3.3.6.3 The probability of detectable infection by PCR as a function of time after infection

For the pre-Omicron period (defined as the period before 31 Dec 2021 in our analysis, see section 3.3.2), we extracted the posterior median probability of PCR detectability for each day subsequent to infection up to a maximum of 30 days (after which the probability of testing positive via PCR was estimated as approximately zero) from Hellewell et al.[103] (Figure 3.2). We assumed that the PCR detectability time curve was the same for all time periods before Omicron; but since there was evidence that this changed for Omicron, we used updated estimates [104] starting from 1 Jan 2022 to end of our study period. For the Omicron period, we used the frequent testing data provided by Hay et al. [104] and estimated the proportion of individuals with detectable infection by SARS-CoV-2 Omicron VOC (PCR Ct value < 37) for 30 days since detection (Figure 3.2). Since these Omicron estimates represented the time since detection opposed to infection, we crudely shifted the Omicron estimates by one day forward and used Hellewell et al.'s [103] day one estimate of PCR detectability. Figure 3.2 provides a comparison between the two profiles of PCR detectability up until 30 days since infection. The method required at least 30 days of prevalence data to estimate incidence because we assumed that the probability of detecting infection via PCR after 30 days since infection was zero.

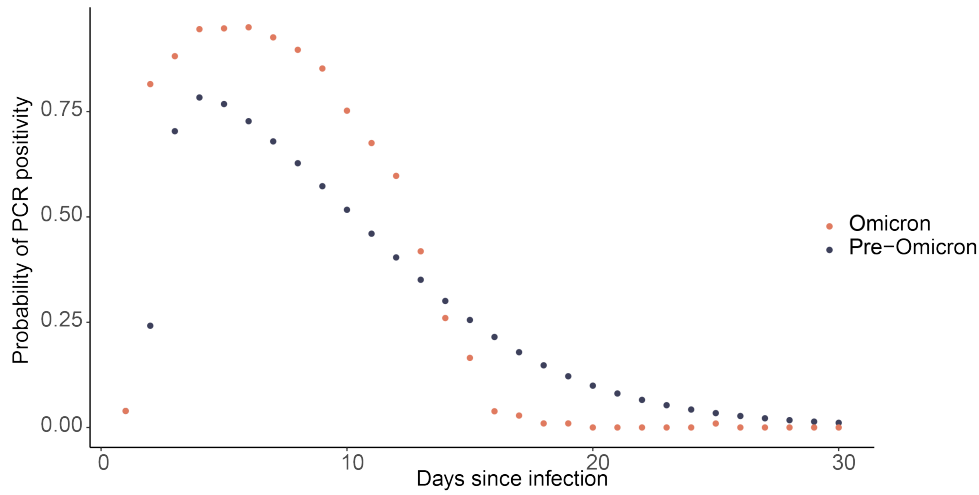


Figure 3.2: **The temporal profile of PCR detectability as a function of time since infection.** The dark blue and orange points represent probabilities before and after 15 December 2021, respectively. PCR positivity is defined as Ct value of less than 37.

3.3.6.4 Positive SARS-CoV-2 mass testing data

Since the Pillar 2 mass testing PCR and LFD positive data had weekly fluctuations, likely due to reporting biases on certain days of the week, we smoothed the daily reported cases data using a 7-day moving average centred on the day under consideration (Figure 3.3B).

REACT-1 and ONS incidence

The availability of two prevalence datasets (the REACT-1 surveys and the UK ONS CIS at the national level only) made it possible for us to infer the true incidence of infection in England and to compare this with the reported cases derived from the national mass testing programme. We used the LTLA-specific prevalences estimated using Pillar 2 PCR data and REACT-1 data as described above to estimate true incidence at the LTLA level and then aggregated it to get true national incidence estimates. Using the ONS prevalence estimates for England, we estimated another set of national incidences. We report time series of the average national case detection ratios derived from REACT-1 and ONS-CIS datasets defined as the fraction of PCR and LFD positive tests in the Pillar 2 dataset compared to the estimated true incidence, by LTLA-specific sociodemographic groups.

3.3.6.5 Model fitting

Because of the computational expense of full uncertainty quantification via Markov chain Monte Carlo (MCMC), we used optimisation to fit our model to data, resulting in a single set of parameters characterising the maximum a posteriori (MAP) estimates. The model was specified using the Stan language and fitted via their default optimisation algorithm [90] with the highest log-probability estimates amongst 5 separate optimisations constituting our parameter estimates.

3.4 Results

3.4.1 Trends in testing uptake and reporting

Over our study period, a total of 290 million LFD tests were reported with 2% of these results positive; and 107 million PCR tests were registered with 13% of tests positive (Figure 3.3, Appendix Table B.1). Distribution of PCR and LFD tests in the study data by ethnic and age-groups is given in Table 3.5.

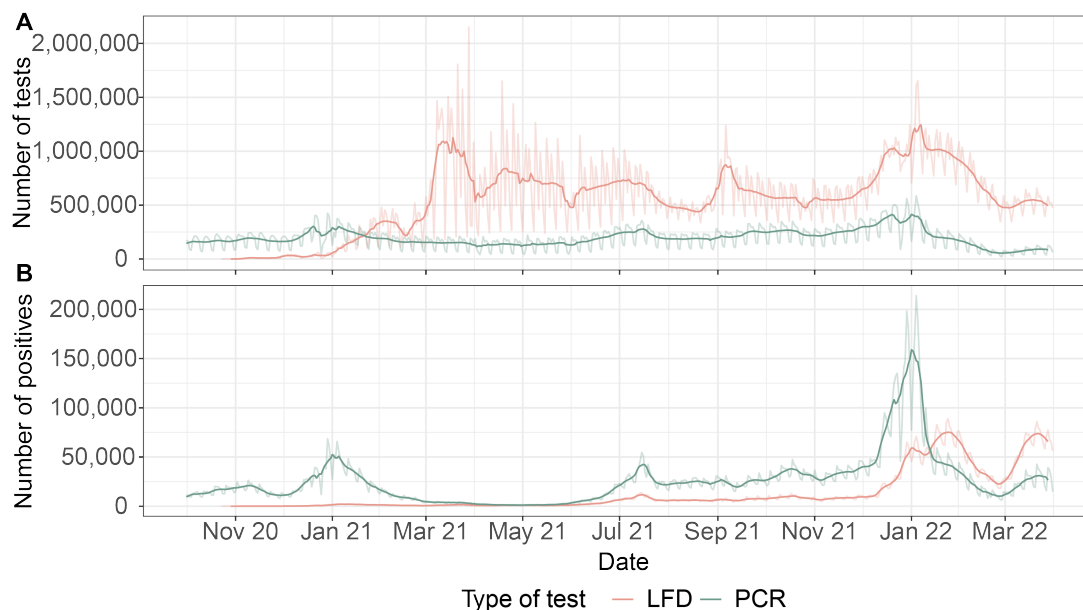


Figure 3.3: **Number of total (and positive) PCR tests performed and LFD tests reported in England.** The light-coloured curves are the raw numbers, and the darker curves display data that has been smoothed using a 7-day moving average (centred on the current day).

From October 2020 – June 2021, individuals in the most deprived areas (IMD=1) (Figure 3.4B) performed almost double (Median ratio [Inter quartile range, IQR]: 1.8 [1.7, 1.9]) the number of PCR tests per capita compared to those in the least deprived (IMD=9) (Figure

3.4E), and we speculated that this was due to the higher prevalence in deprived areas (Appendix Figure B.4) during this period. Linear regression models for the PCR tests per capita suggested that differences in prevalence explained some of the variation in testing across deprivation levels (Appendix Figure B.4) [105]. On the other hand, LFD test reporting displayed the opposite and a more pronounced trend, where those in the most deprived areas reported the fewest LFD tests per capita, which held even after the roll-out of free asymptomatic testing on 9 April 2021 (Figure 3.4A). From March 2021, the median number of LFD tests reported per capita in the most deprived areas was roughly half that in the least deprived (Median ratio [IQR]: 0.50 [0.44, 0.54]) despite more deprived areas having the highest prevalence (Appendix Figure B.4) [106]. Those in the Black ethnic groups had generally higher PCR testing coverages throughout the analysis period (Figure 3.4G) [107, 108].

Sociodemographic group	PCR tests (%)	LFD tests (%)
Asian or Asian British	9,255,814 (8.7)	17,101,937 (5.9)
Black; African; Black British or Carribean	4,971,930 (4.7)	8,612,138 (3.0)
Mixed or multiple ethnic groups	2,839,243 (2.7)	6,683,269 (2.3)
Other ethnic group	663,134 (0.6)	979,015 (0.3)
White	89,101,561 (83.4)	256,850,878 (88.5)
<i>Total</i>	106,831,682 (100)	290,227,237 (100)
0-5 years	7,035,817 (6.7)	9,982,533 (3.5)
12-17 years	5,027,175 (4.8)	45,235,975 (15.8)
18-24 years	8,406,948 (8.1)	19,057,785 (6.7)
25-34 years	18,207,635 (17.6)	33,996,218 (11.9)
35-44 years	17,817,989 (17.2)	41,322,993 (14.4)
45-54 years	17,660,539 (17.0)	52,744,974 (18.4)
55-64 years	15,697,028 (15.1)	52,018,760 (18.2)
65-74 years	5,872,716 (5.7)	23,081,627 (8.1)
75+ years	7,956,088 (7.7)	9,084,901 (3.2)
<i>Total</i>	103,681,935 (100)	286,525,766 (100)

Table 3.5: **Summary of testing data by sociodemographic groups.** Number (and percentage) of PCR tests performed, LFD tests performed and reported by ethnic and age groups between October 2020 and March 2022.

The roll-out of the universal testing offer sought to expand access to LFDs, but it corresponds with a separation in LFD testing patterns across ethnic groups (Figure 3.4H), and, subsequently, per capita LFD tests reported by White people were consistently higher than in other ethnic groups (from April 2021 per capita LFD tests were 34% (IQR: 22%, 44%) higher in White people versus Black people and 47% (IQR: 42%, 55%) higher in White people versus Asian people). Our

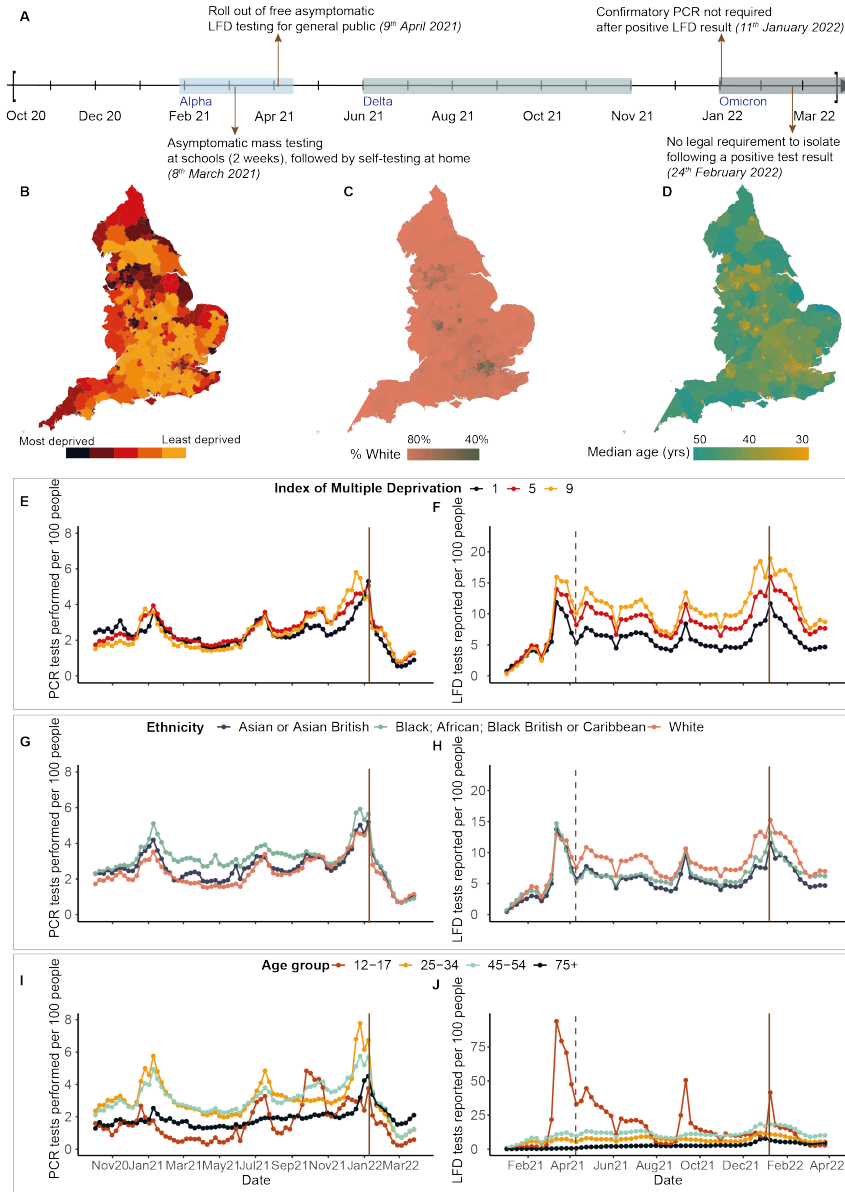


Figure 3.4: **Timelines, sociodemographic distribution, and testing data summary** (A) Timeline of the Alpha, Delta, and Omicron variants of concern and other important dates mentioned in our study (B) Spatial distribution of deprivation across England according to the rank of population weighted average of scores for the LSOAs in a given LTLA. (C) Proportion of White people across LTLAs. (D) Median age across LTLAs of England. (E, F) Weekly number of PCR, LFD tests per person by deprivation levels. All LTLAs were assigned deprivation levels from 1 to 9. (G, H) Weekly number of PCR, LFD tests per person by three ethnic groups (Appendix Figure B.7 for the other two ethnic groups). (I, J) Weekly number of PCR, LFD tests per person by four age groups (Appendix Figure B.6 for the other age groups). The solid vertical line corresponds to the date (11 January 2022) after which confirmatory PCR tests were no longer required following a positive LFD test, and the vertical dashed line corresponds to the date (9 April 2021) when free asymptomatic testing was made available to all individuals in England. See Appendix Figure B.8 for variation in tests per person by LTLAs and sociodemographic groups.

estimates of prevalence in White people were similar to or lower than for other ethnic groups (Figure 3.5A) and similarly so for LFD positivity (Appendix Figure B.5) suggesting that they performed and reported LFDs more frequently with asymptomatic or mild disease.

Working age populations had consistently the highest PCR testing coverages throughout the evaluation period (Figure 3.4I, Appendix Figure B.6) followed by that in the oldest group (those aged 75+). PCR testing coverages in secondary school-aged children (12-17) were more volatile, likely driven by school terms and the high variability in prevalence in these groups (Figure 3.4B). LFD test reporting trends were similar to those of PCR test seeking (Figure 3.4J), albeit with a large spike in LFD testing in school-aged children after the roll-out of asymptomatic testing in schools in March 2021 [109] (Figure 3.4A).

From 11 January 2022 onwards, confirmatory PCR tests were no longer required following a positive LFD test result due to the high rates of COVID-19 across the UK [110], which explains the sharp decline in PCR tests in (Figure 3.4E, G and I).

While the results described above are aggregated at the national level, there was variability between the sociodemographic groups at the LTLA level (appendix Figure B.8). This variation may be attributed to limited sample sizes, especially when we disaggregated the tests by self-reported ethnic groups in LTLAs.

3.4.2 Prevalence

Using the debiasing method described in Methods section 3.3.3, we generated two novel sets of prevalence estimates for each LTLA: one set, stratified by ethnic groups (Appendix Figure B.1); another, by age group (Appendix Figure B.2). We also used the method to estimate the LTLA-specific prevalence (Appendix Figure B.3) for deprivation levels.

During the Alpha wave, the SARS-CoV-2 prevalence in Asian people was considerably higher than for Black people which, in turn, was higher than for White people (Figure 3.4A) [111]. During the Omicron BA.1 and BA.2 variant waves, the prevalence in Asian groups remained highest and the lowest was observed in the Black ethnic group. This pattern was particularly striking in the North-West, West Midlands, and Yorkshire and the Humber regions (Appendix Figure 3.10) [18, 55, 19].

SARS-CoV-2 prevalence in secondary school-aged children surged much higher than in other age groups during the VOC waves (Figure 3.4B) [112, 113]. Prevalence in the oldest age group (75+) was generally lower than for others, although in early-mid 2022, it sharply increased and

coalesced with the others [55, 113].

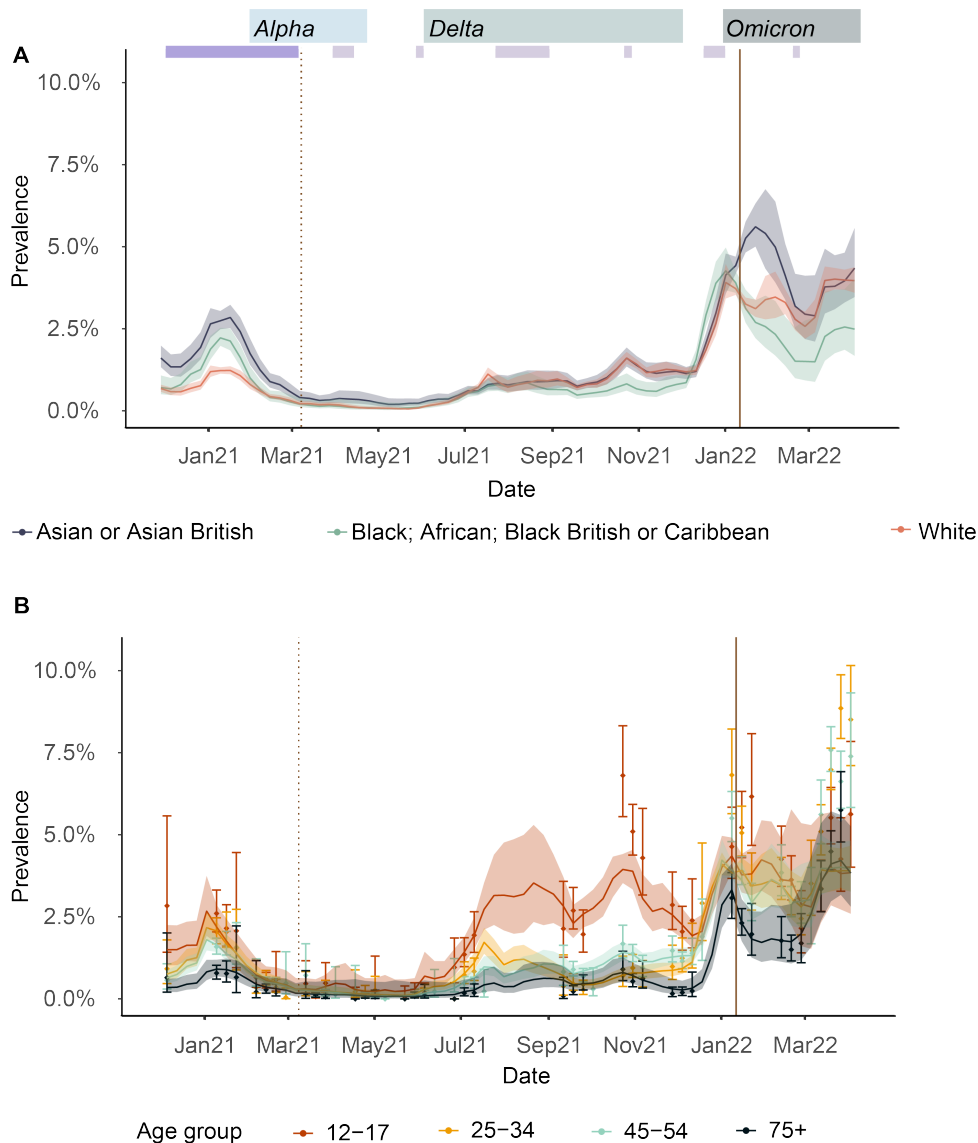


Figure 3.5: **Estimated prevalence for ethnic groups at the national level.** (A) Aggregated weekly national prevalence estimates for ethnic groups. This was the population-weighted average of posterior median estimates (solid lines) and 95% central credible intervals (shaded region) of ethnic group-specific debiased prevalences in LTLAs. (B) Aggregated weekly national prevalence estimates for age groups and 95% central credible intervals (solid and shaded areas respectively). National prevalence and 95% confidence intervals for age groups estimated and obtained directly from REACT-1 are shown as points and whiskers.

3.4.3 Incidence and case detection ratio trends

By comparing the estimated true underlying incidence (using REACT-1 and ONS estimates) to reported cases (defined as PCR and LFD positive tests), we found that the Pillar 2 national testing programme detected between 26%-40% of all infections across the entire evaluation pe-

riod, although this was likely an overestimate because of potential double counting of LFD and confirmatory PCR tests in the Pillar 2 data as well as multiple tests by the same person. This ratio varied substantially over time and increased until January 2022 after which PCR tests were no longer mandatory following a positive LFD test [110]. While the differences in the REACT-1 and ONS derived ratios (Figure 3.6A) are likely explained by the differences in their sampling strategies (and prevalence estimates), there were no notable persistent differences in case detection ratios according to the deprivation levels of LTLAs (Figure 3.6B) [107, 108].

3.4.4 Prevalence model checks

3.4.4.1 Model fit check

In Figure 3.7 we plot the prevalence estimates for a selection of LTLAs along with their corresponding REACT-1 and PCR estimates, highlighting how our debiasing method achieves a middle ground between the wide uncertainty intervals of the unbiased REACT-1 data and the precise but biased estimates from mass testing PCR data. The LTLA-specific data from REACT-1 were used solely for these plots (without being used in the debiasing method itself) to demonstrate how the debiasing method can estimate fine-scale prevalences by utilizing REACT-1 and mass testing data at the coarse scale and only the biased mass testing data at the fine scale.

3.4.4.2 Model comparison for deprivation analysis

We compared three different groupings to determine if IMD levels were appropriate as coarse levels. (This was done to compare the robustness of the prevalence estimates from our novel coarse level grouping to the published methodology and results [65].) The IMD-grouped model was a moderately better fit to the data compared to using administrative regions as the coarse level (Table 3.6). We also used a coarse grouping based on both IMD and administrative regions because both groupings in isolation gave similar model fits (Table 3.6) and we hypothesised that together they might be an even better fit to the data; in doing so we used only five deprivation levels (instead of nine like above) to ensure there were enough LTLAs in each category, resulting in 41 coarse groups overall. The models were fit at the coarse level (administrative region, IMD or administrative region and IMD) meaning that the predictions made at the LTLA level were largely independent of LTLA-specific characteristics (apart from the fact they belong to the larger group). However, as the size of coarse groups gets smaller, there is greater correspondence

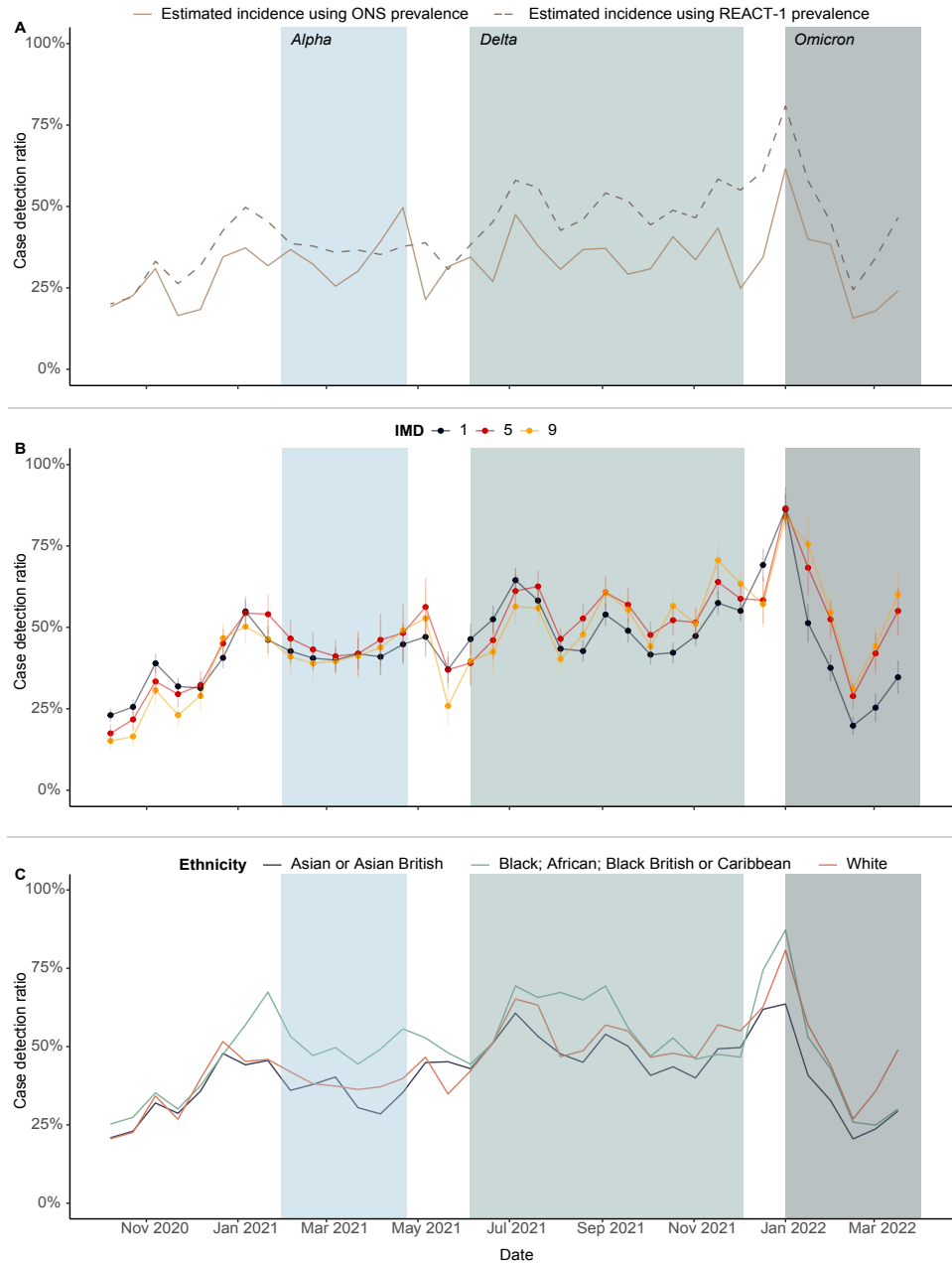


Figure 3.6: **Estimated case detection ratios by sociodemographic groups.** (A) Bi-monthly national estimated case detection ratio over time, using PCR and LFD positive and estimated incidence. The dark dashed and light brown curves represent ratios using REACT-1 and ONS prevalence data, respectively. (B) Case detection ratio over time by deprivation levels (IMD). The dots (and error bars) represent the coefficient estimates (and standard error) of a linear regression with LTLA specific case detection ratio as the dependent variable and IMD levels as predictors for each month. Here IMD level 1 (black) and 9 (yellow) refer to the most- and least-deprived areas, respectively. (C) Bi-monthly national case detection ratio tests by ethnic groups. Shaded areas from left to right, correspond to the time periods of Alpha, Delta, and Omicron VOCs respectively (defined in Methods section 3.3.2).

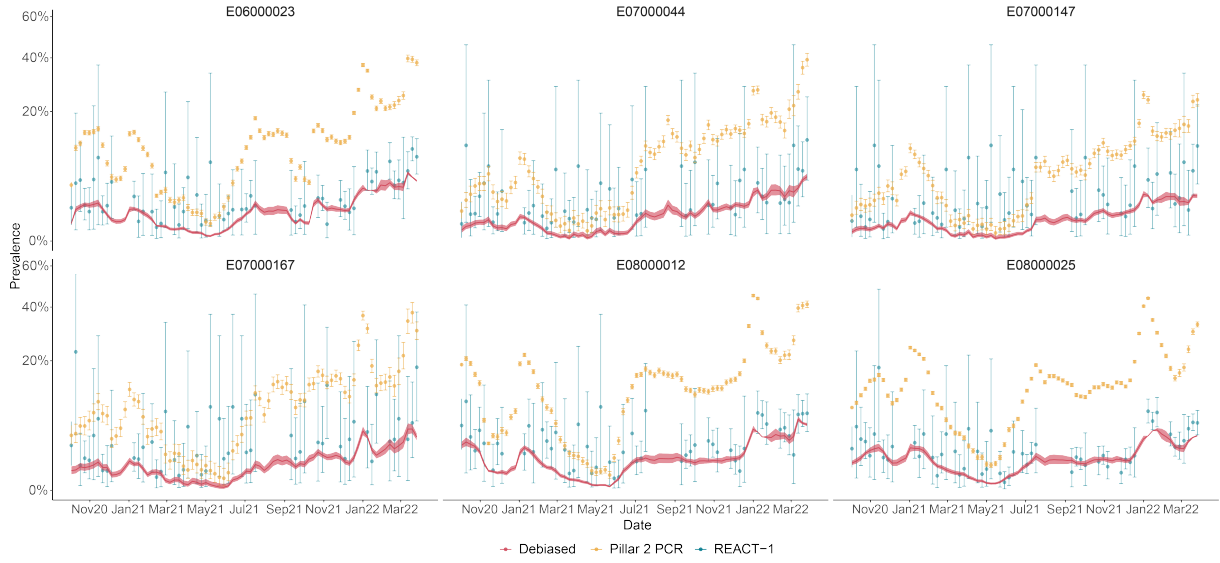


Figure 3.7: Comparing LTLA level prevalence estimates from REACT-1, mass testing, and from the debiasing methodology. Weekly prevalence estimates for the six LTLAs with highest (first row) and lowest (second row) average weekly prevalence. The colours correspond to different prevalences estimated using data from REACT-1 (blue), Pillar 2 mass testing PCR tests (yellow) and the debiasing method (red). The REACT-1 and Pillar 2 PCR prevalences are plotted with 95% central credible intervals, assuming uniform priors over the probability of testing positive. The shaded area around the debaised prevalence is the 95% central credible interval estimated by the model.

between the LTLA level characteristics and those of the coarse groups to which they belong. Because of this, our estimate of predictive accuracy between the model fit to 9 administrative regions and 9 IMD regions is likely to be more comparable than the model fit to 41 regions. This model had the highest predictive accuracy (Table 3.6) (see Methods section 3.3.5).

Coarse level	Log-likelihood
5 IMD regions and 9 regions	-106,594
9 IMD regions	-107,138
9 Regions	-107,330

Table 3.6: **Model comparison results across different coarse groupings.**

We were unable to do similar model comparisons for ethnicity and age-group prevalence analysis because we did not have access to fine scale data on the number of tests and positives by ethnic group or age group within LTLAs from REACT-1.

3.4.4.3 Sensitivity analysis for ethnic prevalence estimation

For LTLAs with zero (in thousands) population for a given ethnic group, we performed sensitivity analyses by considering 0.2 as well as 0.8 thousand as population sizes in the two LTLAs with highest and lower proportion of Whites each and saw similar prevalence estimates (Figure 3.8). Points were close to the 45-degree line, indicating comparable prevalence estimates when using 0.2 and 0.8 compared to 0.5, as population sizes for LTLAs with zero population in a given ethnic group.

3.4.5 Incidence model checks

3.4.5.1 MCMC vs MAP estimates

For computational efficiency, we used maximum a posteriori estimates of incidence instead of full uncertainty quantification using MCMC (Methods 3.3.6.5). Figure 3.9 compares the MAP and MCMC estimates for a subset of LTLAs with the highest and lowest proportion of individuals in the White ethnic groups, and shows that for these, the estimated uncertainty obtained using MCMC was minimal.

3.4.5.2 Model fit check

As discussed in the Methods section 3.3.6.4, we produced two sets of incidence estimates, one for each source of prevalence data: the prevalence estimates outputted from our debiasing method-

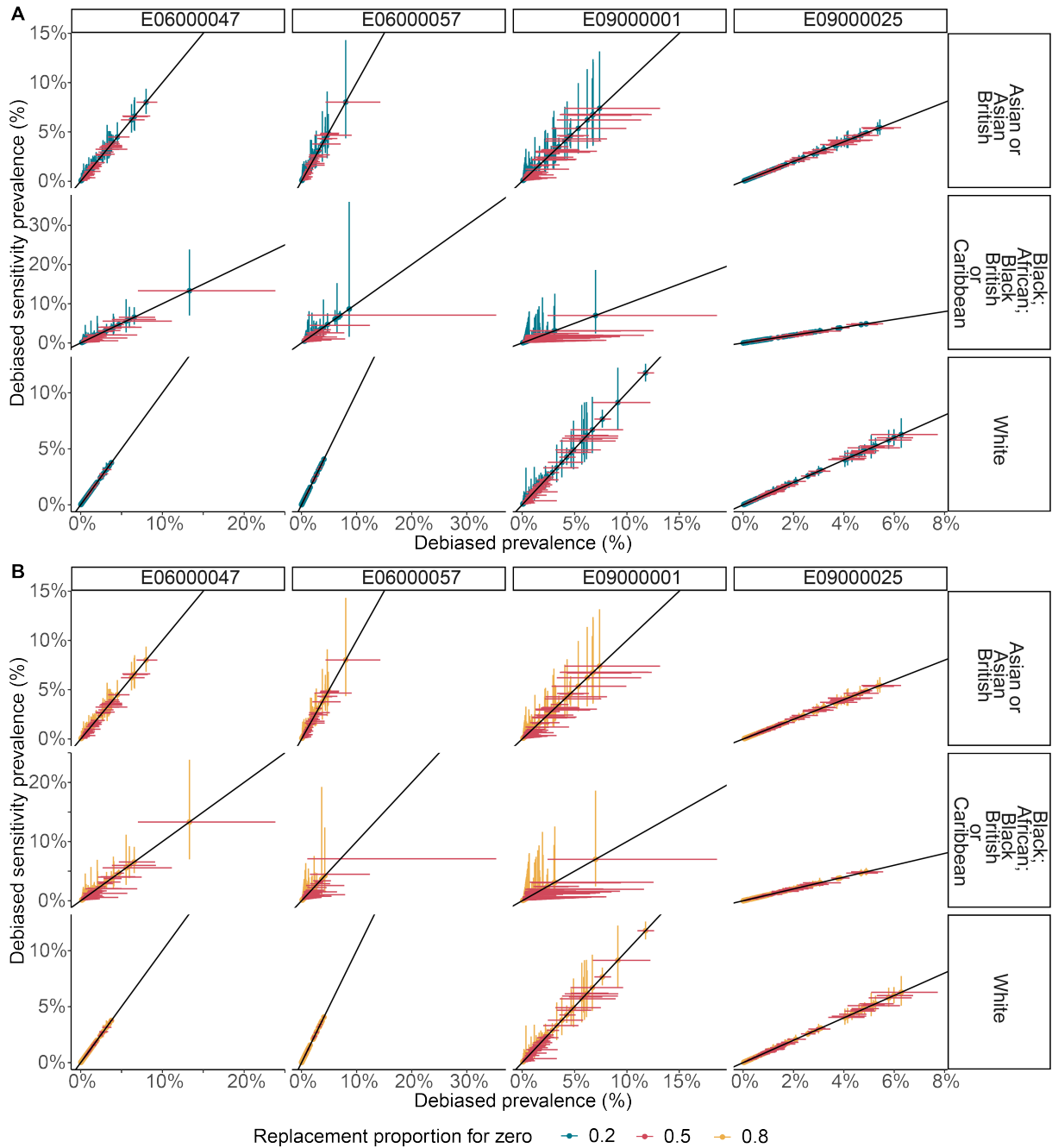


Figure 3.8: **Sensitivity analysis of ethnic group prevalences in LTLAs.** Comparison between ethnic group prevalences for four LTLAs (with first two LTLAs having the maximum proportion of individuals belonging to the White ethnic group; the last two having the minimum) when we replaced the zero proportion of ethnic groups in the data by 0.5 (red and is used in the main ethnic group analysis) by 0.2 (A: blue) and 0.8 (B: yellow). The horizontal and vertical bars are the 95% central credible intervals from the debiasing methodology. The solid 45-degree line is the line to compare how similar the estimates are.

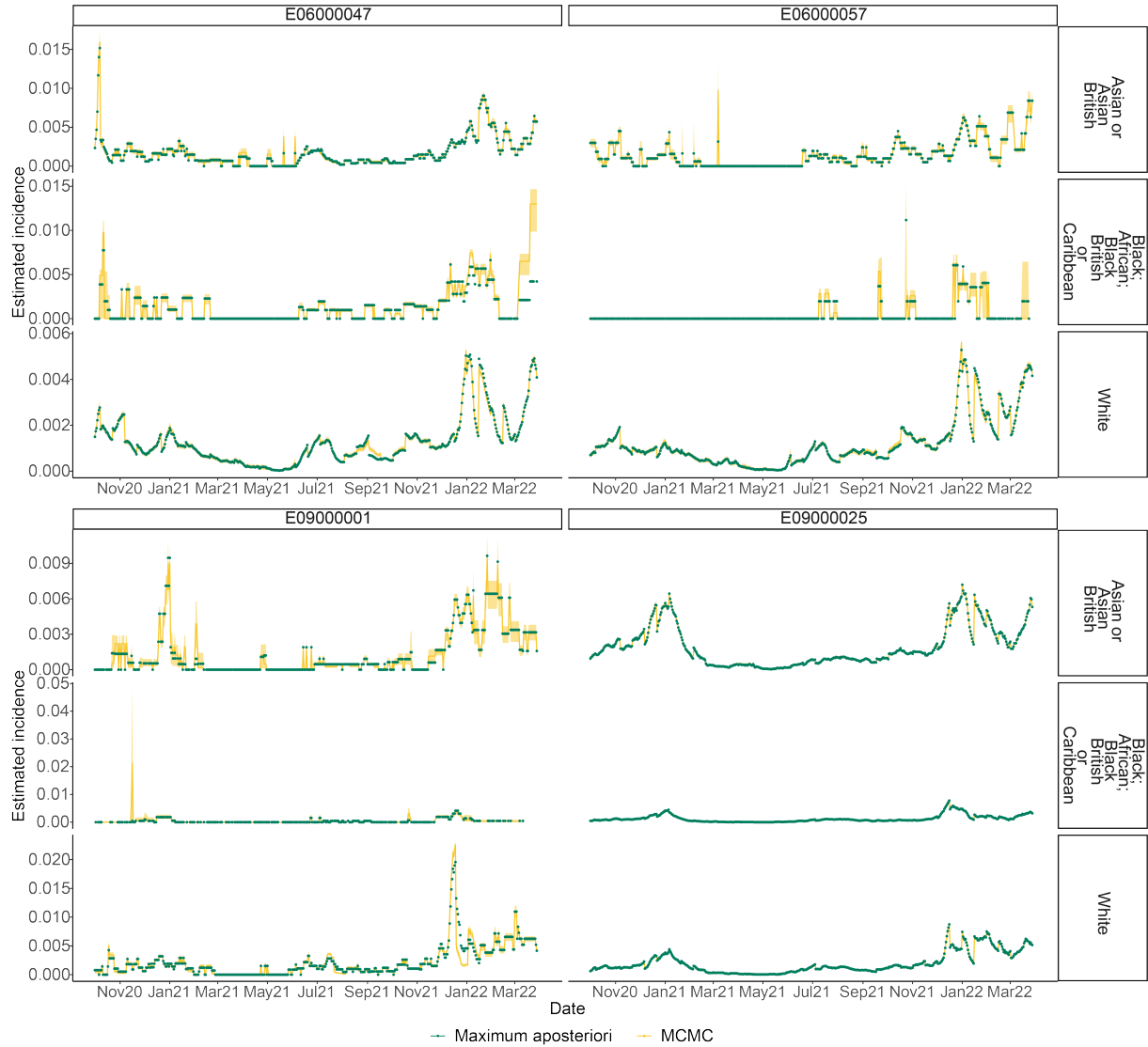


Figure 3.9: **Comparing MCMC and MAP estimates of incidence per capita.** Comparison between the estimated incidence using MCMC and MAP for two LTLAs each with (A) maximum and (B) proportion of individuals in the White ethnic group. The yellow solid curves are the posterior median estimates and the shaded yellow region is the 95% credible interval from the MCMC. The green dots are the parameter estimates with the highest log-probability across five optimisations using Stan’s default optimisation algorithm.

ology and another set from ONS. We were able to check the fit of our model to the data by comparing the prevalence estimates determined by our model with those prevalences used to fit it. For the ONS prevalence model, this was straightforward to visualise, and the fit to the data was reasonable (Figure B.9). For the LTLA-level prevalence model fits, we computed the root-mean-square error (RMSE) in predictions (using the central prevalence estimates from the debiased Pillar 2 prevalences). In Appendix Figure B.10, we overlay our model-estimated prevalences with the LTLA-level estimates used to fit our model for the four worst- and four best-fitting LTLAs. This showed that, generally, our modelled estimates were a reasonable fit to the data, although there was a general reduction in model performance during the Omicron period – either due to issues with the prevalence estimates used to fit the model or due to inappropriate estimates of the probability of detecting infection by PCR throughout the course of infection.

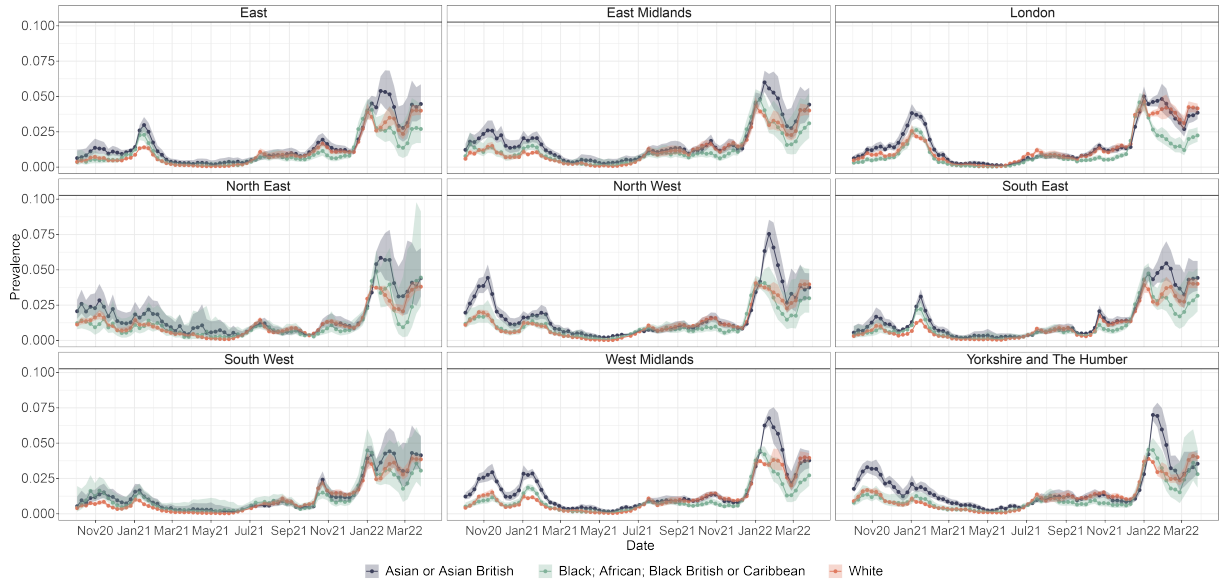


Figure 3.10: **Weekly regional prevalence estimates by ethnic group.** These were population-weighted average of posterior median estimates (solid curves) and 95% central credible intervals (shaded regions).

3.5 Discussion

Our work reveals the intricate and dynamic nature of COVID-19 testing patterns in England throughout our study period (October 2020 – March 2022), illustrating strident differences in test use and reporting between sociodemographic groups. These differences complicate and obfuscate the real-time assessment of transmission across non-traditionally reported groups, and

our prevalence estimates point towards an unequal burden of infection across them.

The PCR and LFD data from Pillar 2 which we explore offer complementary viewpoints of testing uptake and behaviour since PCR test results reflect test uptake whereas LFD test data were available only if individuals self-reported their results. However, only 15% of LFD tests distributed were reported [77], and data on subnational test distribution were unavailable, which prevented granular analysis of LFD uptake and reporting.

We found that prevalence levels explained variation in PCR tests per capita, but myriad other factors likely also influenced PCR testing, including individuals' age, their occupation, their risk perception and their vaccination status, and a more comprehensive approach would be necessary to probe further the observed differences in testing levels [105]. For LFD tests, we found that in the least deprived areas there were about twice as many tests reported compared to the most deprived areas – qualitatively similar to the results of the Liverpool COVID-19 study [106]. This may be partly explained by the disproportionately low digital access and resources to report LFDs, as well as lack of adequate support to isolate if testing positive. Black people had higher PCR testing coverage over the entire duration of the analysis. Since our prevalence estimates do not indicate a consistently higher prevalence in Black ethnic groups (Figure 3.5), it is possible that this reflects a greater risk of symptomatic disease, access to tests through workplace, and could also be partly explained by the ramping up of targeted community testing [107, 108]. The CORSAIR study, conducted between March 2020 and January 2021, also found that financial hardship, IMD, lower socioeconomic status and having a dependent child in the household were associated with poorer recognition of symptoms and a lower likelihood of booking a test [114].

Prevalence was highest in ethnic minority groups (Asian and Black) during the Alpha wave, as was found previously [111]. Additionally, during the Omicron wave, the prevalence in White ethnic groups was higher than in Black ethnic groups. It is unclear what drove these complex prevalence patterns which remained generally consistent across regions of England, but natural and vaccine-induced immunity along with differences in mixing behaviour are possible explanations [18, 55, 19]. The debiasing method implicitly assumed that all the LTLAs and their ethnic groups within a specific administrative region had the same PCR testing bias and behaviour since we did not have access to coarser ethnic group REACT-1 data. We also found that the prevalence of infection was highest in school-aged children during the waves of VOCs, as has been reported previously [112, 113], likely driven by the higher contact rates of this age group. Prevalence among older age groups was typically lower, but notably increased during the Omi-

cron period. This may have been partly due to the perceived lower risk of severe disease from the Omicron variant, increased social mixing, and the generally high levels of vaccine coverage in this group(27,30).

During the pandemic, a number of countries implemented mass testing as a control measure [115] and established population-based prevalence surveys [116]. The world-leading REACT-1 and ONS-CIS community prevalence surveys rolled-out in England and the UK, respectively, used representative samples of individuals to produce estimates of SARS-CoV-2 infection prevalence, while adjusting for selection bias due to non-response [117, 88]. Ours and previous work has shown how these high-quality, low-volume surveys can be combined with mass testing data to determine fine spatial scale estimates of disease prevalence for specific sociodemographic groups. These data could then be used to determine granular estimates of disease transmission, such as the time-varying reproduction number. Such nowcasting approaches could be deployed in real-time in future pandemics to improve disease surveillance, particularly to probe transmission in minority risk groups and spatial areas for which community surveillance random sampling surveys in isolation are insufficiently powered to estimate accurately. We also developed new methodologies to estimate incidence, which revealed the extent of the national testing program’s reach.

As part of the evaluation of the testing programme, we also had access to repeated testing data from individuals, which can be useful in understanding behavioural aspects of testing. In the next chapter, I will delve into the development and application of various metrics designed to identify and quantify biases in testing behaviors. I will shed light on the factors that influence how individuals use testing as a public health intervention and how these methods can reveal differences in transmission and barriers to uptake.

Chapter B

Appendix - SARS-CoV-2 transmission dynamics between sociodemographic groups

	Number of tests	Number of positives (%)
PCR tests performed with LTLA	106,986,026	14,021,052 (13.1)
PCR tests performed with ethnicity	106,831,682	13,933,106 (13.0)
PCR tests performed with age	103,681,935	13,554,671 (13.1)
LFD tests performed and reported with LTLA	290,462,387	7,007,391 (2.4)
LFD tests performed and reported with ethnicity	290,227,237	6,991,160 (2.4)
LFD tests performed and reported with age	286,525,766	6,441,966 (2.2)

Table B.1: **Summary of Pillar 2 mass testing data.** These include tests where there were data with information on LTLAs or ethnicity or age.

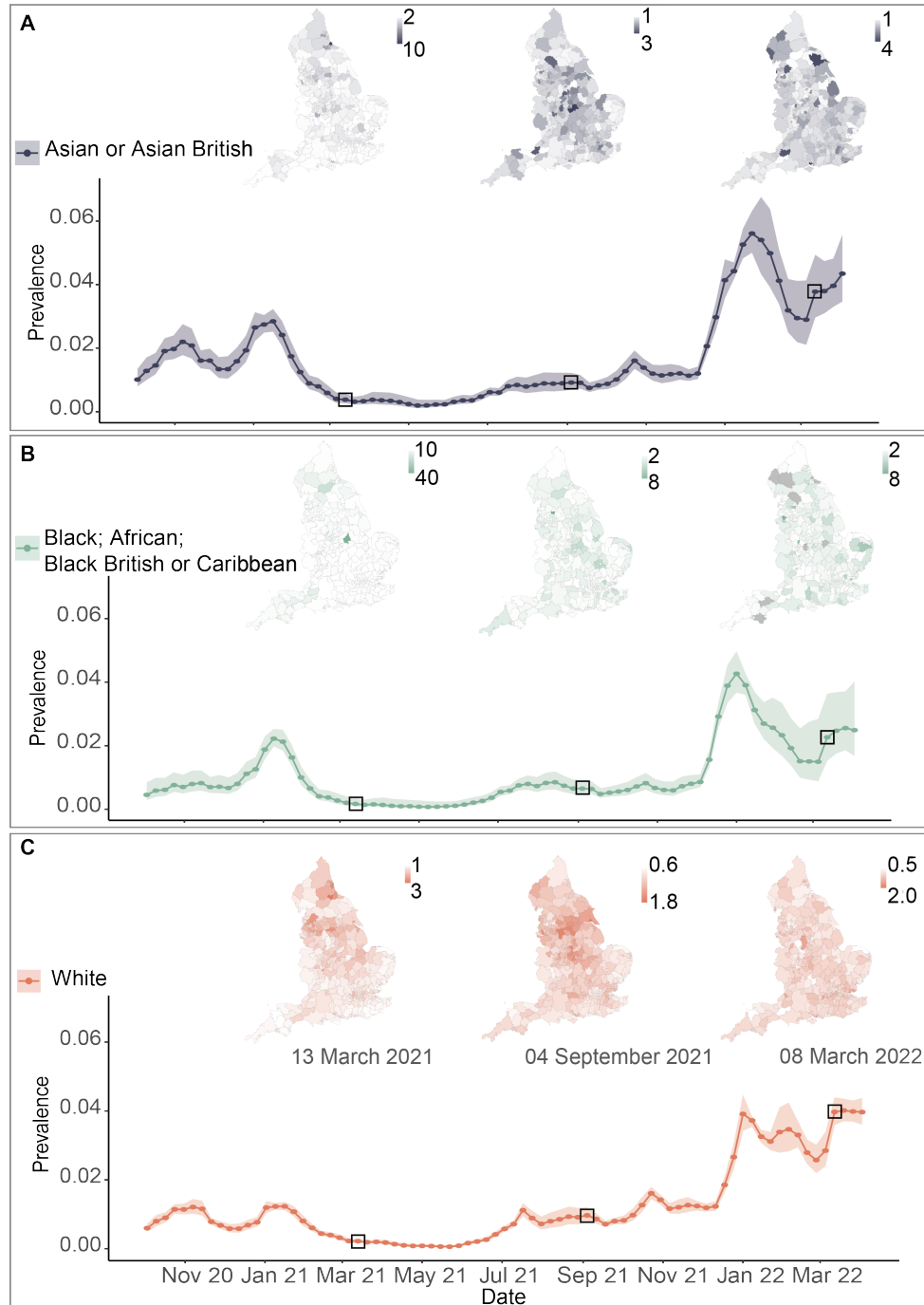


Figure B.1: **Prevalence of SARS-CoV-2 in ethnic groups.** (A) Asian ethnic group (B) Black ethnic group, and (C) White ethnic group over time calculated using a weighted mean of our fine scale prevalences and 95% central credible intervals of ethnic groups within LTLAs and their respective population sizes. The maps show the ethnic group specific prevalences between LTLAs during the mid-periods of Alpha, Delta, and Omicron VOCs as defined in section “Variant time periods” 3.3.2. The map shading is given by the ratio between the prevalence of an ethnic group in an LTLA and the national ethnic group specific prevalence during that week, indicating deviations from the average country-wide prevalence.

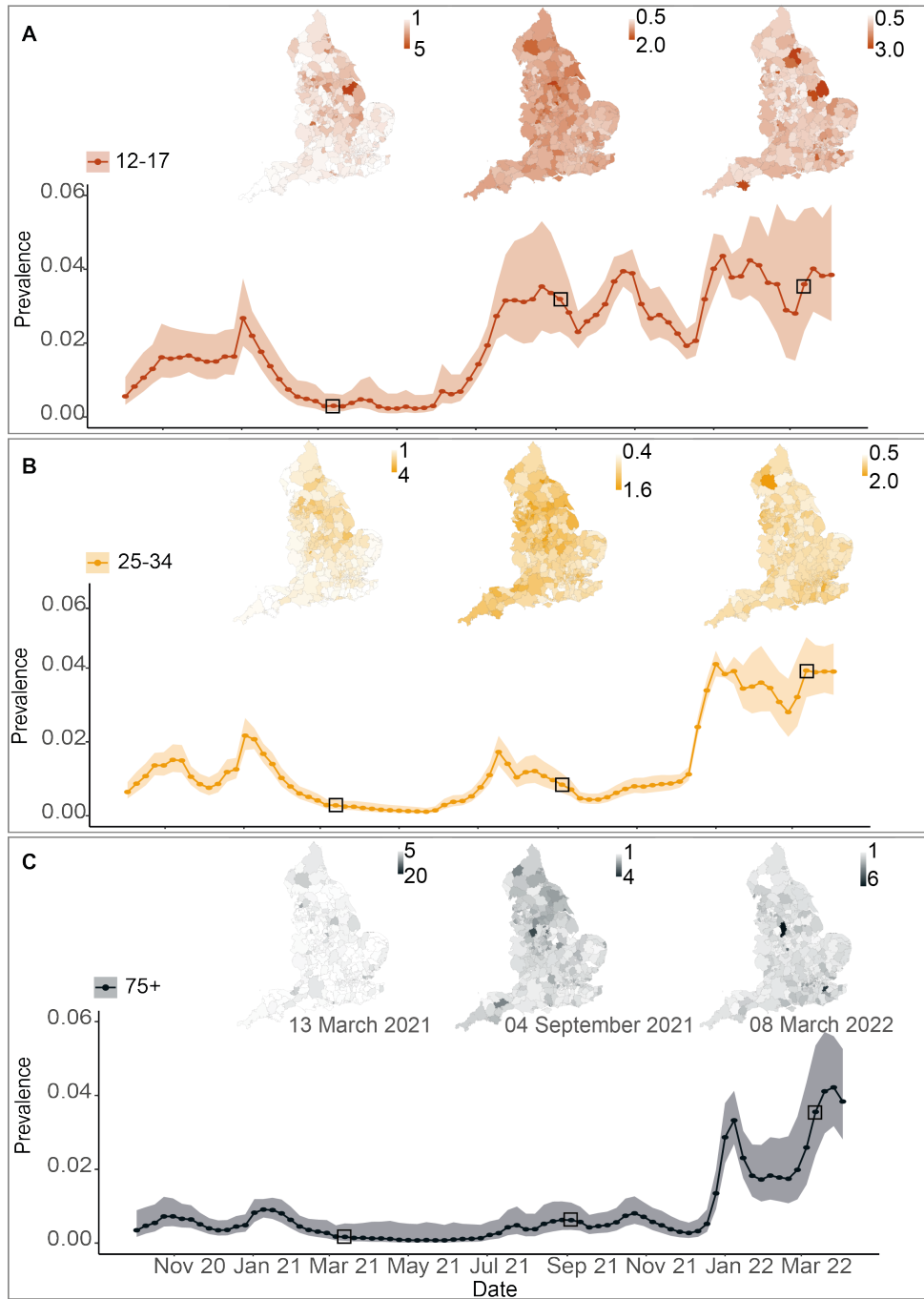


Figure B.2: **Prevalence of SARS-CoV-2 in individuals by (self-reported) age groups.** (A) 12-17 years (B) 25-34 years, and (C) 75+ years over time calculated using a weighted mean of our fine scale prevalences and 95% central credible intervals of age groups within LTLAs and their respective population sizes. The maps show the age group specific prevalences between LTLAs during the mid-periods of Alpha, Delta, and Omicron VOCs as defined in Supplementary methods section “Variant time periods”. The map shading is given by the ratio between the prevalence of an age-group in an LTLA and the national age-group specific prevalence during that week, indicating deviations from the average country-wide prevalence.

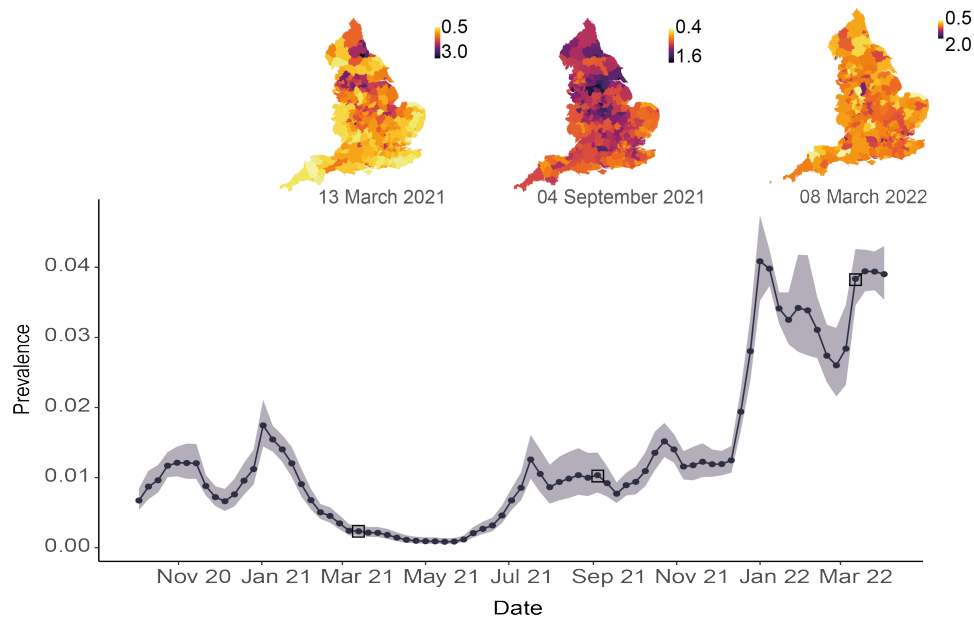


Figure B.3: **Prevalence of SARS-CoV-2 in England over time and by LTLA.** Using a weighted mean of our fine scale LTLA prevalences and 95% central credible intervals and their respective population sizes. The maps show the LTLA-specific prevalences during the mid-points of Alpha, Delta, and Omicron VOCs as defined in Supplementary methods section “Variant time periods”. The legend for the maps indicates a scale for the ratio between the prevalence of an LTLA and the national prevalence during that week, indicating which areas were higher or lower than the average country-wide prevalence.

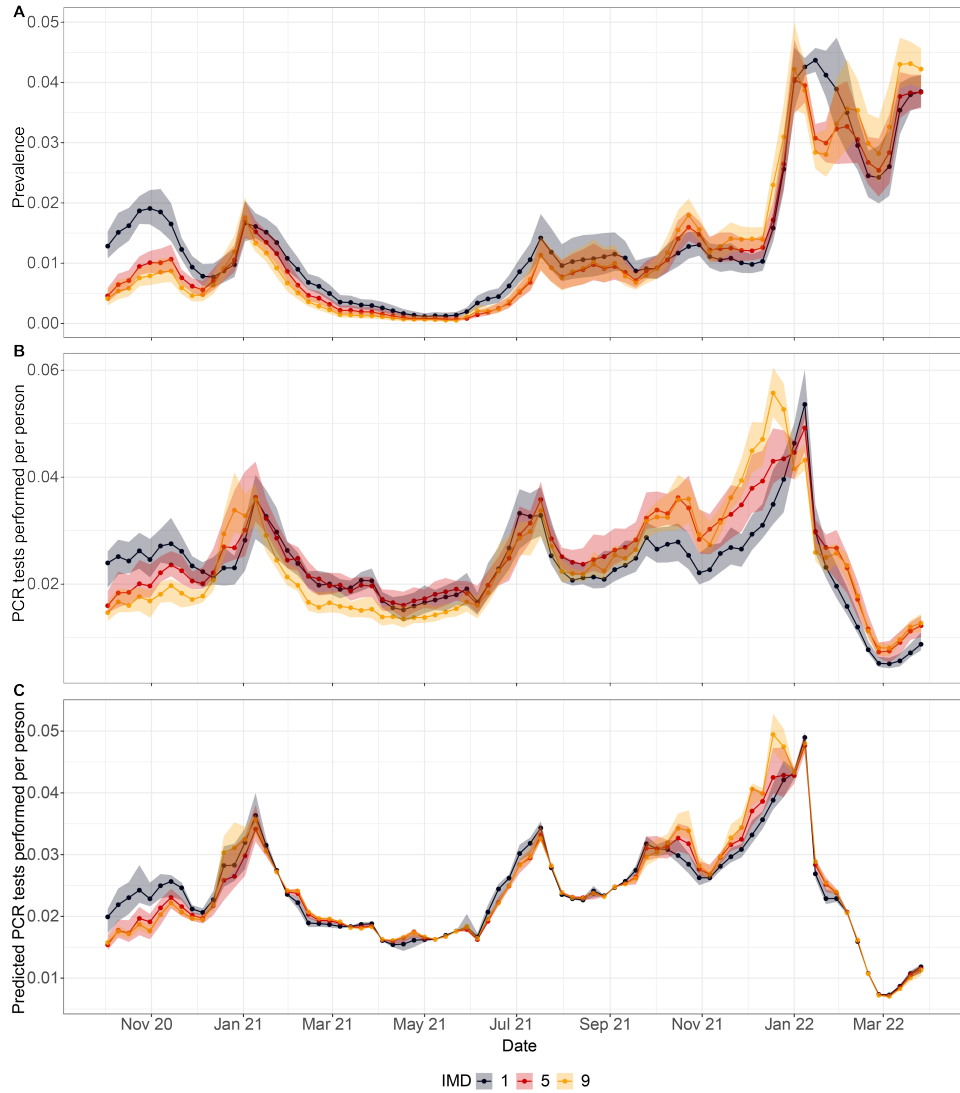


Figure B.4: Relationship between deprivation, SARS-CoV-2 prevalence and PCR tests performed. (A) Aggregated weekly national prevalence estimates for IMD levels (1=most deprived, 9 = least deprived). This was calculated by weighing the LTLA-specific debiased prevalences and 95% central credible intervals by their population sizes into IMD levels. The solid curves are the posterior medians and the shaded areas are the 95% central credible intervals. (B) Observed weekly median volume of PCR tests per capita per LTLA by three levels of IMD and shaded regions indicate the interquartile range across the LTLAs. (C) Predicted PCR volumes from weekly regressions between LTLA specific prevalences and PCR volume. See Methods section 3.3.4 “Association between LTLA prevalence and PCR tests” within the section 3.3.3 “Causal debiasing approach for obtaining time-varying estimates of SARS-CoV-2 prevalence at fine scale resolution” for details about the analysis. The solid curves show the median volume of predicted PCR tests and shaded areas indicate the interquartile range across all LTLAs according to their IMD level.

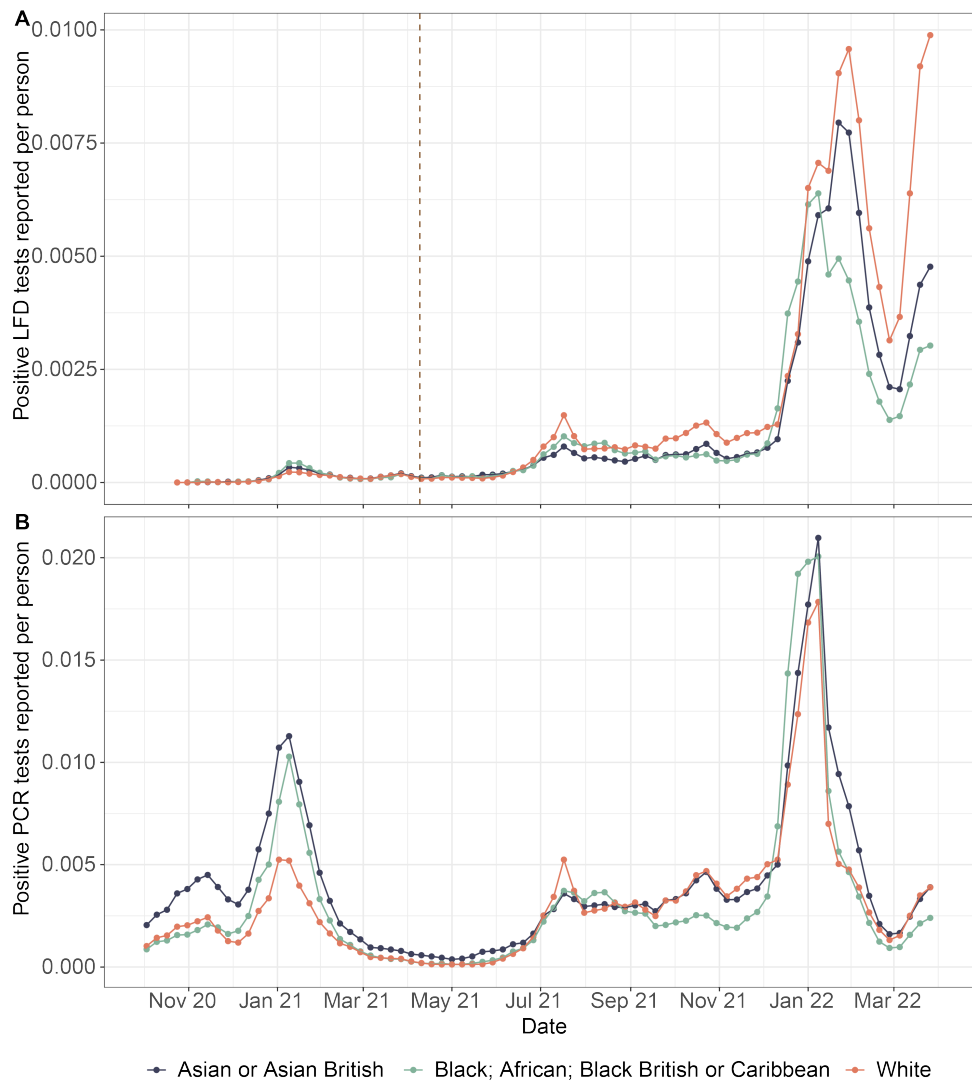


Figure B.5: **PCR and LFD positive test performed/reported per capita by ethnic groups.** The vertical dashed line corresponds to the date (9 April 2021) when free asymptomatic testing was made available to all of England.

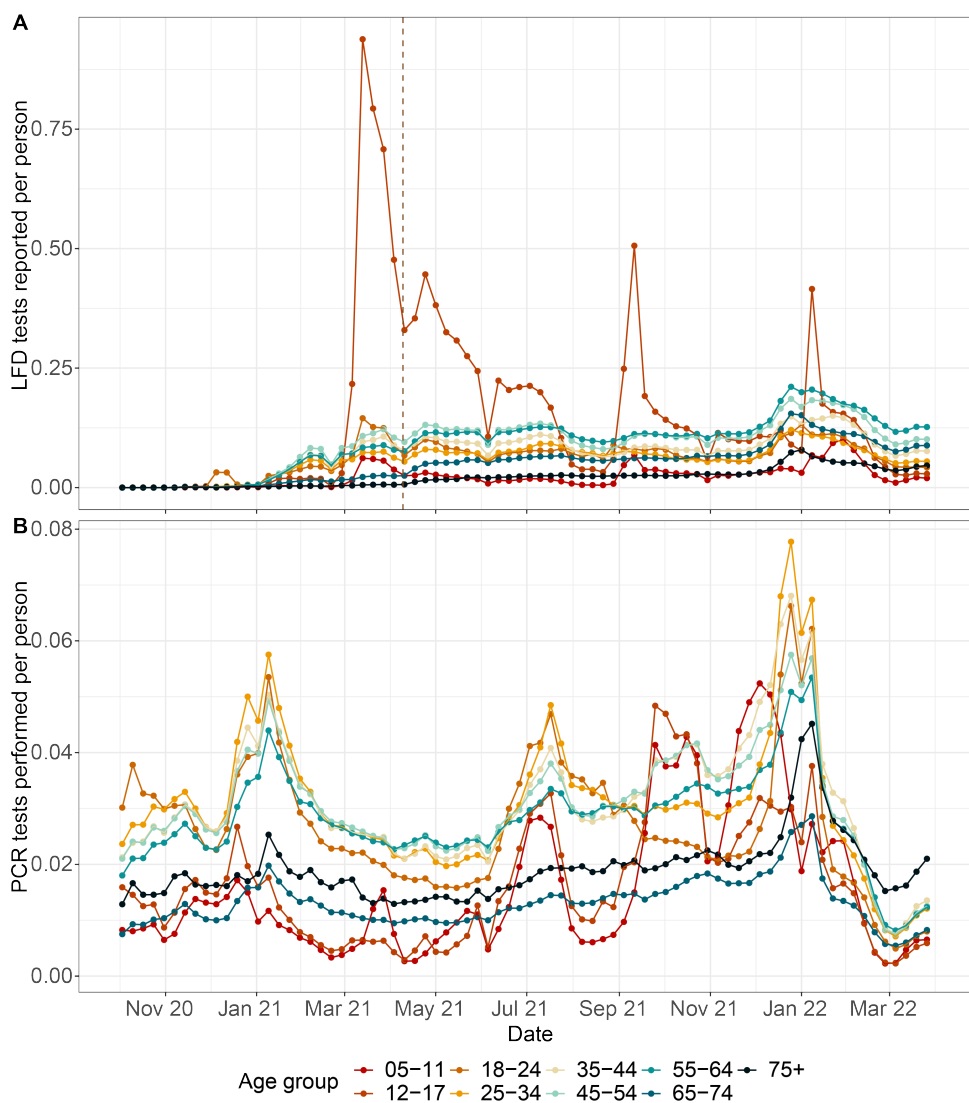


Figure B.6: **Weekly number of PCR and LFD tests performed/reported per capita by age groups in England.** The vertical dashed line corresponds to the date (9 April 2021) when free asymptomatic testing was made available to all of England.

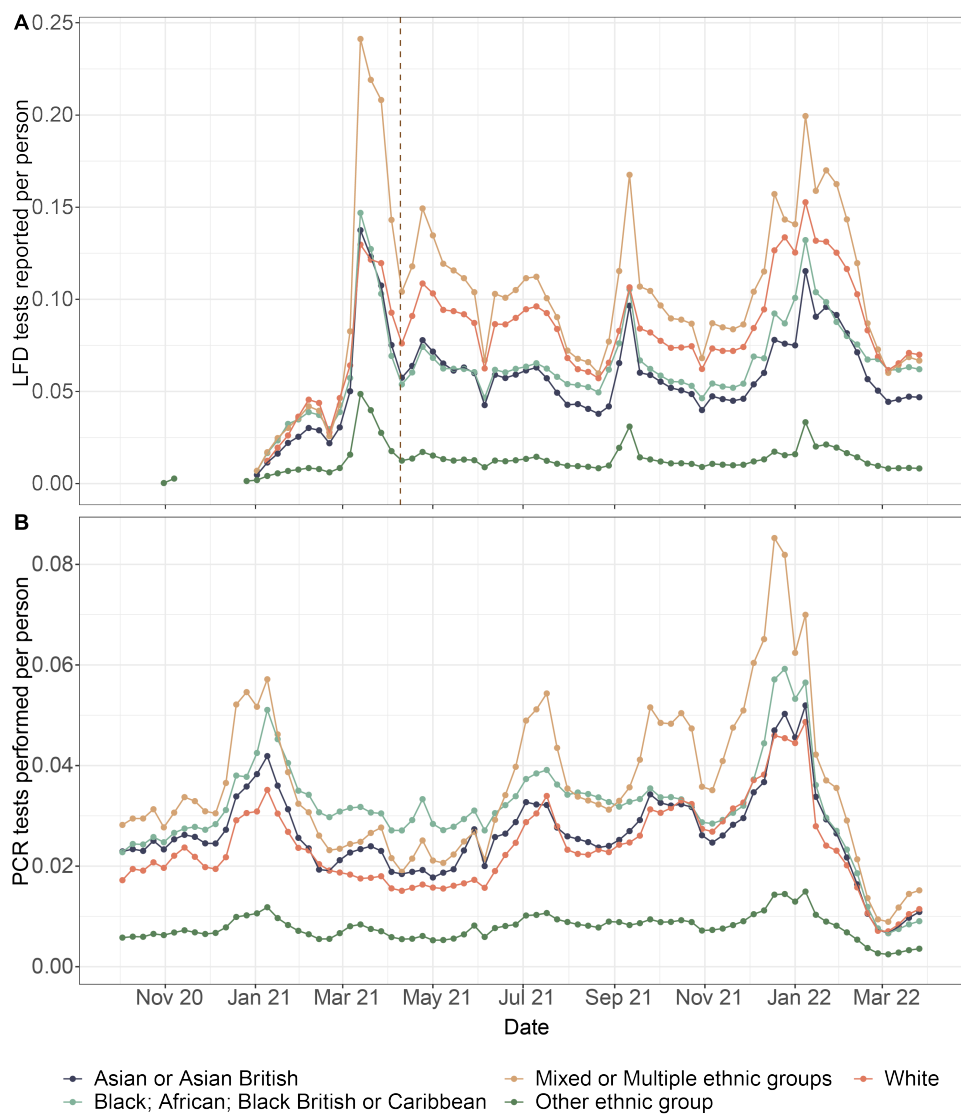


Figure B.7: **Weekly number of PCR and LFD tests performed/reported per capita by five ethnic groups in England.** The vertical dashed line corresponds to the date (9 April 2021) when free asymptomatic testing was made available to all of England.

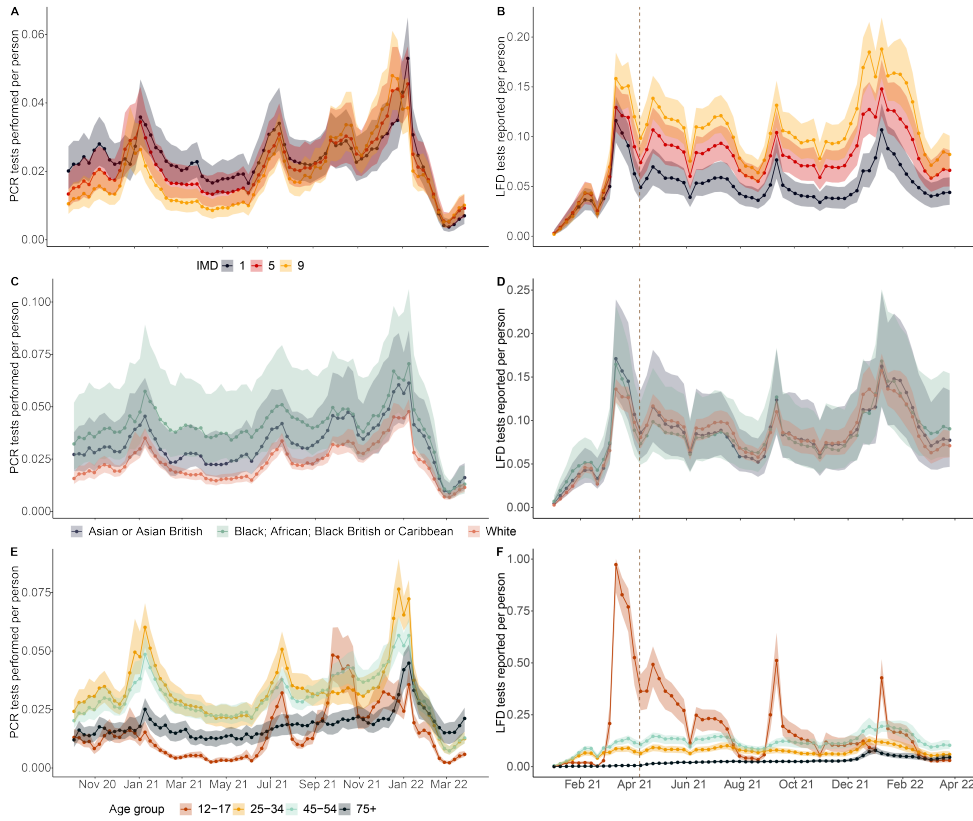


Figure B.8: **Aggregated LTLA- and sociodemographic group-specific PCR tests performed and LFD tests reported per person.** For each LTLA the number of PCR tests performed per person and LFD tests reported per person are calculated by (A, B) deprivation level, (C, D) ethnicity, and (E, F) age group. The solid lines and dots represent the median tests per person by each subgroup and shaded regions indicate the interquartile range across all LTLAs.

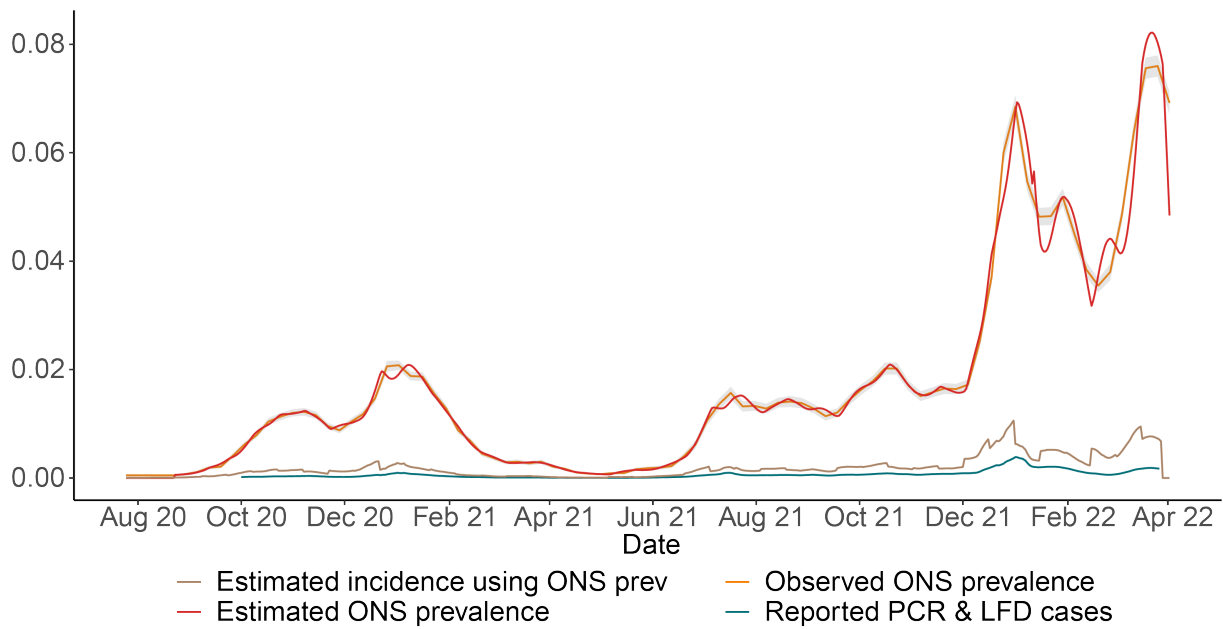


Figure B.9: **Model fit for the ONS national prevalence model.** Comparing country wide estimated (red curve) and observed (solid orange curve and shaded region for 95% central confidence intervals) prevalence from ONS data. The blue and brown curves are the Pillar 2 PCR and LFD positive tests and estimated incidence, respectively, both as a proportion of the population.

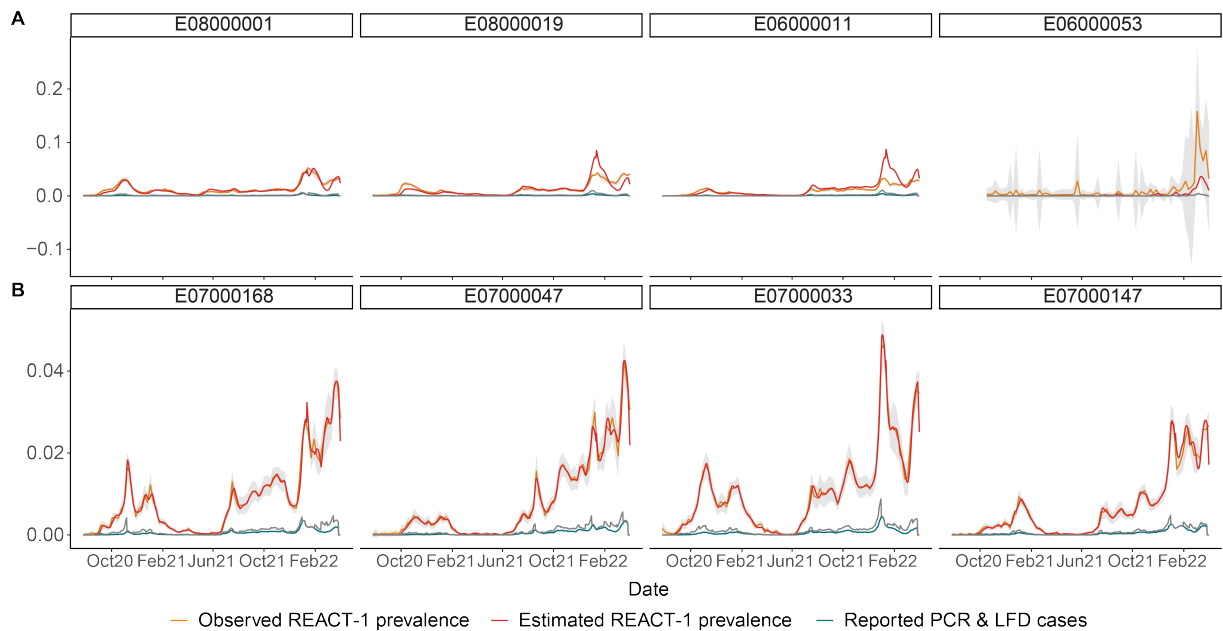


Figure B.10: **Model fits for the REACT-1 based prevalence model at LTLA level.** Comparing countrywide estimated (red curve) and observed (solid orange curve and shaded region for 95% central confidence intervals) prevalence for the (A) four worst and (B) best performing LTLAs, with the highest and lowest RMSEs respectively. The blue and brown curves are the Pillar 2 PCR and LFD positive tests and estimated incidence, respectively, both as a proportion of the population.

Chapter 4

Differences in SARS-CoV-2 testing behaviours by sociodemographic groups

Using the mass testing data from the collaboration with UKHSA, we estimated the weekly prevalence and daily incidence of SARS-CoV-2 for ethnic groups, ages and local authorities in England for 18 months as described in Chapter 3. Since PCR and LFD tests were freely available and reportable, our goal was to then determine if usage was consistent across different sociodemographic groups across the 18 months of our study, throughout this extensive and critical phase of the pandemic. This chapter details the metrics we developed from the extensive testing dataset, which gave us insight into the testing behaviours and biases. These measures can be utilized in real-time by local public health practitioners and policymakers to monitor and ensure the equitable use of public health interventions.

4.1 Abstract

Recognizing how access to and uptake of public health interventions such as testing in England led to disparities in health outcomes between sociodemographic groups is essential for designing and implementing equitable interventions in the future.

This chapter integrates and utilises data on repeated testing from the Pillar 2 mass testing dataset to identify and analyze disparities in public health intervention uptake among various sociodemographic groups.

Building on the debiasing analysis in the previous chapter, I present the biases in test uptake and reporting patterns among different age groups, ethnicities, and deprivation levels. Using confirmatory PCR data, we estimate false positivity rates, sensitivity, and specificity of the LFD tests. Additionally, we model how the likelihood of a positive confirmatory PCR test decreases with the time elapsed between an LFD positive result and the subsequent PCR test, highlighting differences among ethnic groups. While we would not expect considerable differences between ethnic groups, identifying any could reveal varying testing patterns.

These novel methods provide a framework to understand testing behaviors and biases, uncovering significant disparities. For instance, the analysis shows that PCR testing biases 3.3.3 were higher compared to LFD testing, meaning that people were more likely to perform LFD tests regardless of their infection status. During the invasion phases of the Delta and Omicron variants of concern, the PCR testing bias in the most deprived populations was roughly double (ratio: 2.2 and 2.7 respectively) of that in the least. These biases offer insight into how closely the naïve estimates of prevalence from PCR and LFD tests across different sociodemographic groups aligned with the actual infection rates, with lower testing bias suggesting that the naïve estimates were more reflective of the true situation. We also determined that ethnic minorities and older individuals were less likely to use confirmatory PCR tests through most of the study period and that there was possibly a longer delay in reporting a positive LFD test in the Black populations.

Differences in testing behaviours across sociodemographic groups may be reflective of the relatively higher costs of self-isolation to vulnerable populations, differences in test accessibility, digital literacy, and differing perception about the utility of tests and risks posed by infection.

The findings highlight the need for equitable public health strategies and demonstrate how real-time data can guide more targeted and effective interventions. These methodologies could

enhance disease surveillance and improve public health responses in future pandemics, especially for minority risk groups and regions where traditional surveillance is less effective. The results emphasize the importance of interdisciplinary collaboration in tackling public health challenges and ensuring fair access to health interventions.

4.2 Introduction

Throughout the COVID-19 pandemic, numerous studies identified sociodemographic factors including belonging to an ethnic minority group, older age, and social deprivation as being associated with an increased risk of SARS-CoV-2 infection and worse health outcomes in England [118, 119] and internationally [120]. It is crucial to understand how access to and/or uptake of public health interventions may contribute to these inequalities.

In England, testing for SARS-CoV-2 played a crucial role as a public health intervention during the COVID-19 pandemic, enabling the government to monitor disease transmission and forming the cornerstone of self-isolation policies. However, it remained unclear whether testing was delivered and taken up equitably across society [35, 121, 122, 123, 124]. The UK government initiated mass testing in response to the pandemic, with the initial testing efforts starting in March 2020 [31]. The testing programme, a key component of NHS Test and Trace (NHSTT), played an integral role through its various services.

From May 2020 onwards, PCR tests were available for suspected COVID-19 cases, while LFD tests for COVID-19 were rapidly developed and evaluated between December 2020 and April 2021 [125, 77]. Initially, LFD testing was rolled out to key populations, such as healthcare workers, and later to the wider population through a universal testing offer. This universal testing offer was a major component of the national programme, aimed at expanding asymptomatic LFD testing, and was delivered through various channels, with rollout to the general public beginning on 9 April 2021 [126].

As described in Chapter 3, we had unprecedented access to COVID-19 testing data for England as part of a UK Health Security Agency (UKHSA)-commissioned retrospective evaluation of England’s testing programme over the period from October 2020 to March 2022 [77].

In this chapter, I combine mass testing data with prevalence surveillance surveys to identify gaps in the uptake of public health interventions among sociodemographic groups at fine geographical scales. I present analyses that quantify biases in test uptake and reporting patterns across different sociodemographic groups, including age, self-reported ethnic group, and a measure of deprivation. Using confirmatory PCR data, I also present methods to estimate false positivity rates, sensitivity and specificity. We also modelled how the probability of a positive confirmatory PCR test declined with time elapsing between when the LFD positive was reported and the PCR test subsequently taken, by ethnic groups. This is important because, first, the

timing of positive LFD tests and the delay between confirmatory tests can impact the chances of a positive PCR result, potentially leading to false negatives. Second, although we wouldn't expect substantial differences between ethnic groups, identifying any could provide insight into varying testing patterns, such as earlier or later reporting of LFD results. And understanding the policies in place at the time may further explain these behavioural differences.

The framework and outputs from these novel methods could be used to inform public health policymaking, contributing towards more equitable responses to future outbreaks or pandemics.

4.3 Methods

4.3.1 Testing bias

The causal debiasing methods developed by Nicholson et al. [65] and our extension, as described in Methods section 3.3.3 of the previous chapter 3, compares the PCR (or LFD) swab positivity of the Pillar 2 mass testing data and PCR swab positivity from REACT-1 at a “coarse” level (e.g., administrative regions, IMD regions, or age-group) and outputs a testing bias, given by Equation 3.5 as follows:

$$\Omega = \log\left[\frac{\text{odds}(\text{tested} \mid \text{infected})}{\text{odds}(\text{tested} \mid \text{not infected})}\right].$$

The estimated testing biases Ω represent the log odds-ratio of testing (and of reporting for LFD tests) in infected versus not infected individuals. For example, if the probability of performing a PCR test was the same regardless of whether people were infected, then the biases would equal zero ($= \log(1)$). So, a large bias corresponds to a higher chance of an infected individual performing a PCR test compared with an uninfected individual, which would result in overestimated prevalence using the biased testing data. We also estimated the LFD test reporting bias using data on reported LFD tests at the coarse and fine scales. We term this as bias in context of using these data naively for surveillance purposes; however, we recognise that the testing programme was primarily a public health intervention rather than a surveillance programme and did not aim to use PCR or LFD testing data to determine SARS-CoV-2 prevalence. For PCR tests, this bias is the outcome of processes relating to test taking, whereas for LFD tests, it is due to those relating to both taking the test and reporting the test results.

We estimated the PCR and LFD testing biases for deprivation levels, ethnic groups and age groups, as described below using the coarse and fine levels described in Table 3.4:

4.3.1.1 Deprivation

We first estimated PCR and LFD testing biases (Figure 4.4A and C) for deprivation levels in England. We used the nine IMD levels in England as the coarse level. All LTLAs were assigned an IMD level (1 and 9 corresponding to most deprived and least deprived respectively) as described in the data pre-processing section. Testing data from these LTLAs were aggregated to form nine coarse IMD levels with corresponding REACT-1 and PCR Pillar 2 testing data on the number of tests performed and number of positives. The testing bias (Figure 4.4A) was estimated for each IMD level and propagated to the LTLAs nested within the coarse levels to produce fine-scale prevalence estimates. This assumes that the bias which is influenced by PCR test seeking behaviour, testing capacity and PCR sensitivity and specificity along with other confounding factors, is homogeneous within a given coarse level. Similarly, LFD test reporting bias was also estimated at the coarse IMD levels (Figure 4.4C).

4.3.1.2 Ethnicity

For ethnicity-related analysis, the coarse level comprised the nine administrative regions in England, and the five ethnic groups (as described in the previous chapter Section 3.3.1) in each LTLA made up the finer level of analysis. For example, the ethnic group of Whites in Oxford (LTLA code = E07000178) was nested in the South-East administrative region of England in this analysis.

4.3.1.3 Age

For the age-specific analysis, nine age-groups (as described in Table 3.3) at the national level were used as the coarse level; for the fine scale, we used the LTLA-specific age-groups. We estimated testing biases between age-groups in England (Figure 4.4B and D).

4.3.2 Estimating false positivity, test sensitivity and specificity

4.3.2.1 False positivity

We collated a subset of our Pillar 2 testing data comprising all reported LFD positive results which were followed up with a PCR test taken within three days of the reported positive. These data allowed us to estimate the false positive rate for LFD tests, assuming PCR tests were a gold standard. Using the total number of people reporting LFD positives, we could additionally

determine the probability that individuals reporting an LFD positive result subsequently took a PCR test, as was generally mandated throughout the period under observation, with two exceptions: from January to March 2021 and after January 2022 [110] when confirmatory PCR testing was temporarily suspended given the high COVID-19 prevalence meaning LFD tests likely had low false discovery rates. Using our estimates of prevalence and false positivity rates, we then estimated LFD test specificity and sensitivity across different age and ethnic groups throughout the pandemic, as described below.

4.3.2.2 Test sensitivity and specificity

I now detail our approach for estimating group-specific test sensitivities and specificities. First, we computed the probability of a positive confirmatory PCR test within 3 days of reporting a positive LFD test as follows:

For ease of notation, let δ indicate “reporting a positive LFD test” and γ indicate “seeking a PCR test within 3 days of reporting an LFD positive test”. Then using Bayes’ rule,

$$Pr(PCR+ | LFD+, \delta, \gamma) = \frac{Pr(LFD+ | PCR+, \delta, \gamma)Pr(PCR+ | \delta, \gamma)}{Pr(LFD+ | \delta, \gamma)}. \quad (4.1)$$

By making the simplifying assumptions that δ and γ do not influence the computed probabilities, we can then write the denominator as

$$[Pr(LFD+ | PCR+)Pr(PCR+)] + [Pr(LFD+ | PCR-)Pr(PCR-)].$$

Assuming that PCR tests are perfect, we denote:

$$Pr(PCR+ | LFD+) = Y,$$

$$Pr(PCR+) = \text{prevalence} = X,$$

$$Pr(LFD+ | PCR+) = \text{sensitivity} = a,$$

$$Pr(LFD+ | PCR-) = 1 - \text{specificity} = b.$$

We can then rewrite Equation 4.1 as:

$$Y = \frac{aX}{aX + b(1 - X)} = \frac{aX}{b + (a - b)X}. \quad (4.2)$$

To fit this model to our data, we let the number of positive PCR confirmatory tests during week t , given by $n_{PCR+,t}$ follow a binomial distribution:

$$n_{PCR+,t} \sim \text{binomial}(n_{LFD+,t}, Y_t), \quad (4.3)$$

where $n_{LFD+,t}$ is the number of positive LFD tests performed and reported during week t that were followed up with a confirmatory PCR test within 3 days, with Y_t modelled according to the relationship defined in 4.2, as follows:

$$Y_t = \frac{aX_t}{b + (a - b)X_t}. \quad (4.4)$$

We fit this model to national prevalence data at the group level. Where X_t is the SARS-CoV-2 prevalence for week t , with priors centred around national mean estimates μ_t and standard deviation σ_t . μ_t was obtained as a weighted mean of our fine scales LTLA prevalences and respective population sizes and σ_t were back-calculated from the corresponding weighted national 95% central credible intervals according to Equations 3.12, 3.13:

$$X_t \sim \text{normal}(\mu_t, \sigma_t). \quad (4.5)$$

The sensitivity and specificity parameters were estimated by fitting the model. These parameters were assigned the following priors:

$$\begin{aligned} a &\sim \text{uniform}(0, 1), \\ b &\sim \text{uniform}(0, 1). \end{aligned} \quad (4.6)$$

We let the sensitivity and specificity vary by ethnic group and age-group. Table 4.2 describes the estimates of sensitivity (a) and 1-specificity (b) in models with and without ethnicity or age-group.

4.3.3 Modelling the decline in PCR detection rate subsequent to reporting an LFD positive

We modelled how the probability of detecting SARS-CoV-2 infection via PCR after reporting a positive LFD result changed (by assumption, decayed) as a function of time and tested if this rate of decay was different between ethnic groups. Since viral loads are likely modal from the time of infection onwards, it is possible that an individual's probability of testing positive by

PCR could actually increase subsequent to reporting of the LFD positive. The overwhelming trends in our data, however, show that, on average, PCR detectability declined subsequent to the date of reporting of an LFD positive.

To calculate the decay rates, we aggregated two quantities by month m , ethnic group eth , and by number of days d (where $d = 0, 1, 2, 3$) between reporting a positive LFD test and subsequently seeking a PCR test:

- $n_{LFD+,d,m,eth}$ = number of reported LFD positives that were followed up with a PCR tests within d days for ethnic groups group eth during month m
- $n_{PCR+,d,m,eth}$ = corresponding PCR positive test count.

These were modelled as samples from a binomial distribution:

$$n_{PCR+,d,m,eth} \sim \text{binomial}(n_{LFD+,d,m,eth}, p_{d,m,eth}), \quad (4.7)$$

where $0 \leq p_{d,m,eth} \leq 1$ was the to-be-estimated probability of detection by a PCR test, d days since reporting a positive LFD test, during month m and for a given ethnic group eth . This probability was assumed to decay over time according to the following relationship:

$$p_{d,m,eth} = \text{Pr}(PCR+ | LFD+, \delta, \gamma, d, m, eth) = \Phi_{m,eth} e^{-\lambda_{m,eth} t}. \quad (4.8)$$

$\Phi_{m,eth}$ represents the differences in false positivity rates due to monthly changes in prevalences by ethnic groups.

We plot the monthly decay rates by ethnic groups $\lambda_{m,eth} t$ in Figure 4.3A.

4.4 Results

4.4.1 Confirmatory PCR, false positivity

Of the 7 million positive LFD tests reported, 2.4 million were confirmed by a PCR test with 68% of these on the same day (Table B.1 and Table Table 4.1, and Figure 4.2), and 89% returning a positive result. The proportion of PCR tests which were confirmatory varied substantially throughout the evaluation period and according to self-reported ethnicity and age group (Figure 4.2A and B). Generally, this proportion increased over time, and White people (Figure 4.2A)

and secondary school-aged children (Figure 4.2B) typically had considerably higher proportions of PCR tests which were confirmatory than other groups [127].

Between October 2020 – December 2021, 60% reported LFD positives were followed up with confirmatory PCR tests (Figure 4.1). From December 2020 – December 2021, White populations were 1.45 times (IQR: 1.34, 1.51) more likely than Black populations to take PCR tests following a reported LFD positive and 1.33 times (IQR: 1.29, 1.47) more likely than Asian people (Figure 4.2C). Adherence to confirmatory testing policy was generally similar across age groups apart from those aged 75+ during the period between the end of the Alpha wave and the start of the Delta wave (Figure 4.2D). (Figure 4.2E) shows that from March – May 2021, the false positive rate (proportion of negative confirmatory PCRs) for Black populations was roughly 1.3 times (IQR: 1.07, 1.61) that in White people and roughly 1.85 times (IQR: 1.66, 2.00) that in Asian populations. From May – June 2021, the false positive rate was substantially elevated in those aged 75+ (Figure 4.2F). At least transiently then, there were times when there were substantial differences across ethnic groups in the risk of isolating unnecessarily if following the governmental policy to self-isolate.

Confirmatory PCR tests	Number of LFD tests	Number of PCR positive results (%)
LFD positives followed by a PCR	2,328,414	2,069,391 (88.9)
Within same day	1,598,213	1,416,229 (88.6)
After one day	513,567	461,159 (89.8)
After two days	209,143	185,511 (88.7)
After three days	7,491	6,492 (86.7)

Table 4.1: **Summary of confirmatory PCR data.** Pillar 2 mass testing data for England between 1 October 2020 and 31 March 2022.

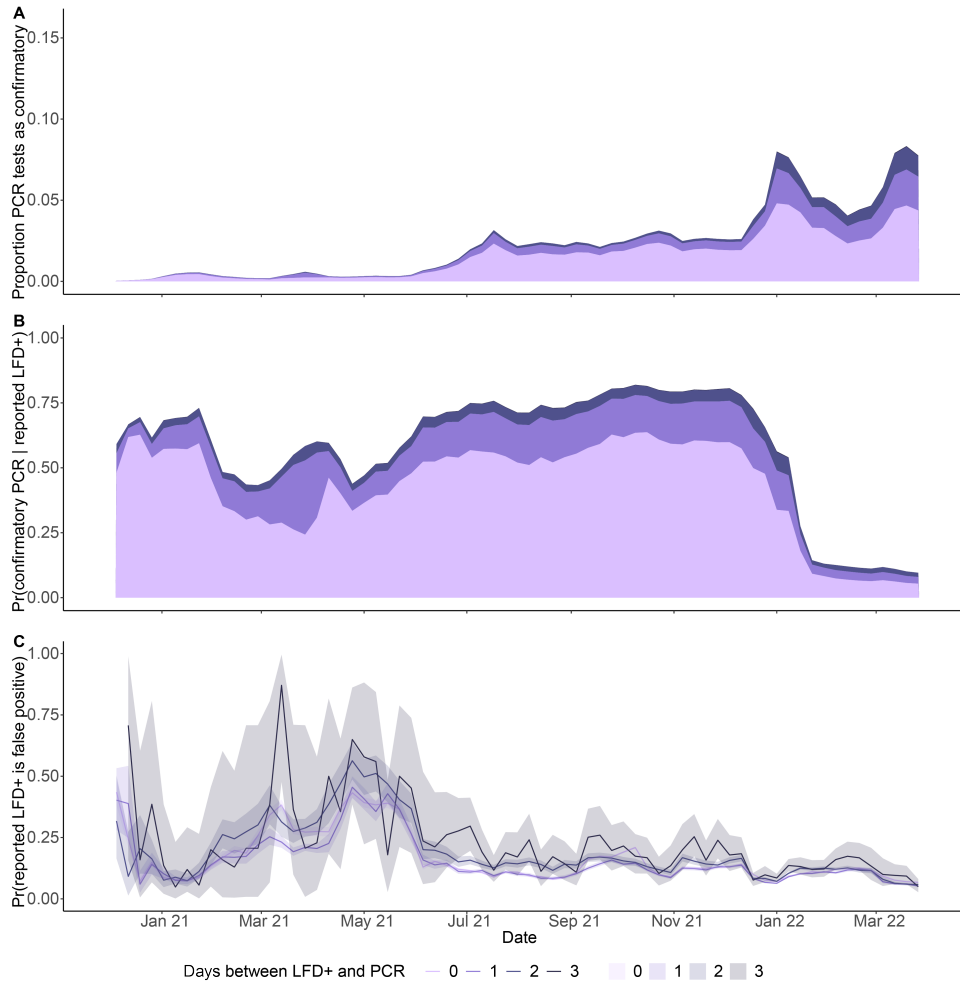


Figure 4.1: **Confirmatory PCR tests over time.** (A) Proportion of PCR tests performed as confirmatory tests within given number of days since reporting an LFD positive result. (B) Proportion of reported LFD positive tests followed by PCR confirmatory tests within given number of days. Both A and B are stacked charts, which show how the proportions and probabilities, respectively, increase as the window between the LFD test being reported and the PCR test occurring increases. (C) Posterior median probability of testing negative with a confirmatory PCR test after reporting a positive LFD across the study period and dependent on the time from when the LFD positive test was reported and the PCR test taken. The uncertainty intervals represent 95% central credible intervals, calculated assuming a uniform prior on the probability of a false positive.

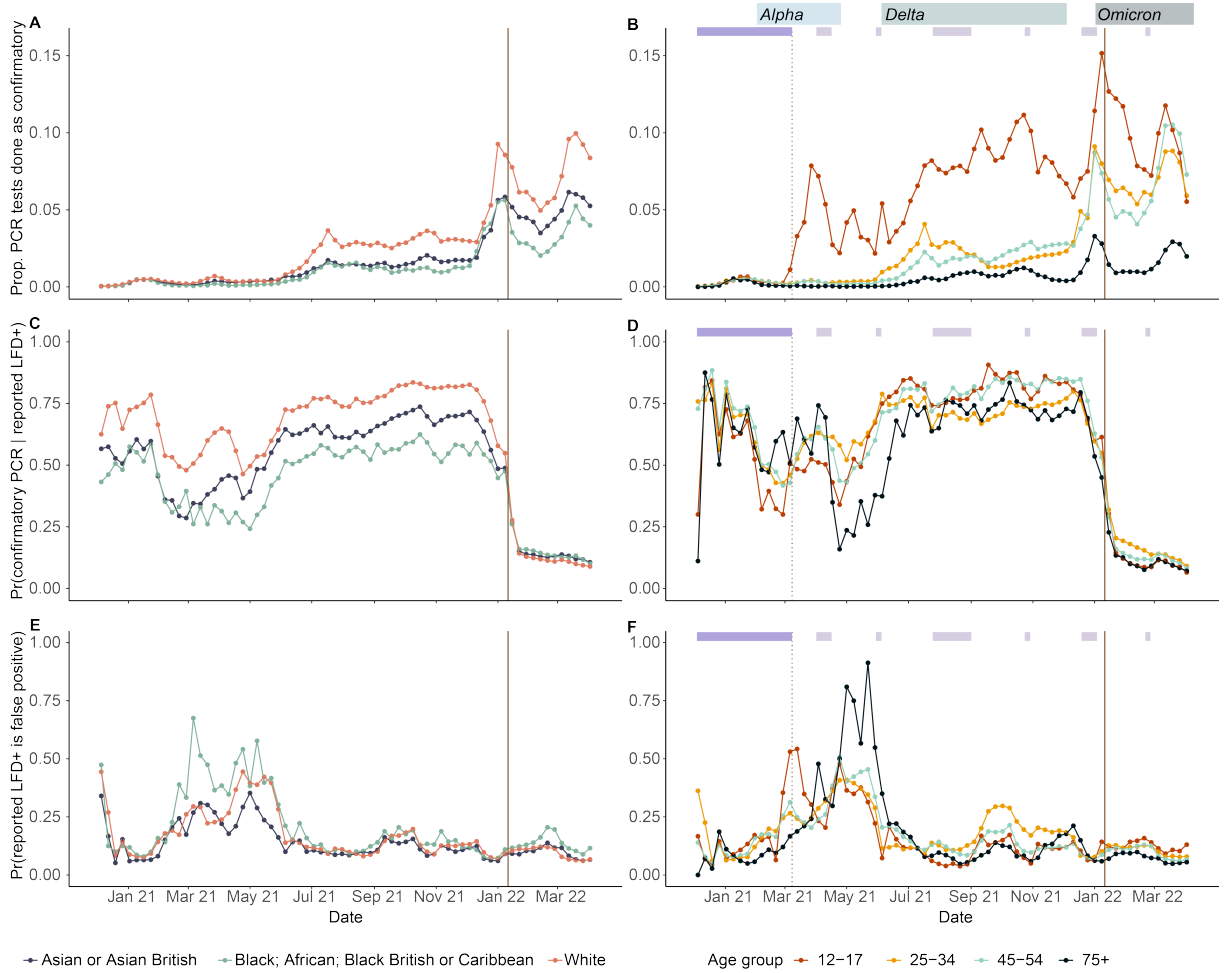


Figure 4.2: **Confirmatory PCR tests done overall and followed by an LFD positive, and false positive LFD tests by sociodemographic groups.** (A, B) Proportion of PCR tests performed as confirmatory tests by ethnic and age groups. (C, D) Proportion of reported LFD positive tests followed by PCR confirmatory tests within 72 hours among different ethnic and age groups. (E, F) The probability of testing negative with a confirmatory PCR test after reporting a positive LFD test, among different ethnic and age groups, conditional on testing positive with an LFD, reporting the positive LFD test and seeking a confirmatory PCR test. The solid vertical line corresponds to the date (11 January 2022) after which confirmatory PCR tests were no longer required following a positive LFD test, and the vertical dotted line corresponds to the date (8 March 2021) when onsite asymptomatic testing was made available in schools. Restrictions on face-to-face education and school holidays are indicated by the dark and light purple rectangles, respectively. The blue rectangles represent the time periods of Alpha, Delta, and Omicron VOCs as defined in Methods 3.3.2. The date axis starts in December 2020, when LFD tests started being reported.

4.4.2 Test sensitivity and specificity, PCR detectability

False positive rates are inversely related to disease prevalence, with the sensitivity and specificity of the test determining the exact relationship (Methods section 4.3.2.2). This allowed us to

estimate the sensitivity (range of central estimates: 66-72%) and specificity (99.5-99.9%) of LFDs by ethnicity and age group which we report in Table 4.2. While our posterior estimates had wide uncertainty intervals, the medians of these test characteristics were similar between age groups and differed across ethnic groups [128, 129].

Using the data on reported positive LFD tests followed by a PCR test within three days, we modelled how the probability of detecting SARS-CoV-2 infection by PCR after reporting a positive LFD result declined according to the time elapsing between reporting the LFD test and subsequent testing by PCR. We determined this decay rate for the different ethnic groups for monthly time windows throughout the study period, and found that this rate of decay was considerably higher in Black people during the Alpha wave and generally remained above the other groups throughout the study period (Figure 4.3A).

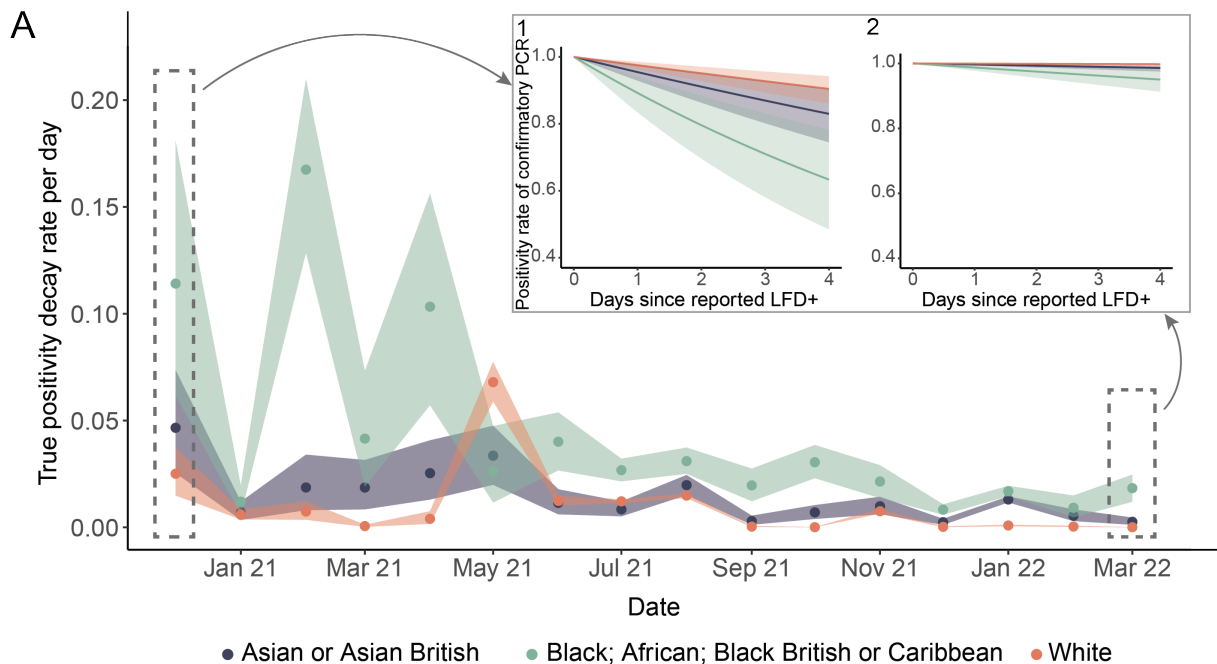


Figure 4.3: **Decline in PCR detectability of PCR tests by ethnic groups.** (A) Rate of decline in the probability of detecting a SARS-CoV-2 infection by PCR within 3 days from reporting a positive LFD test over the study period by ethnic groups. Solid dots (and shaded areas) in the main figures are the posterior median estimates (and 50% central credible intervals) of ethnic group specific decay rates. The inset plots translate what the rate estimates in A correspond to in terms of the rate of decline in probability of testing positive by PCR subsequent to reporting of an LFD positive (with 50% credible intervals) for the first (1: December 2020) and last month in our study (2: March 2022).

Model	Sensitivity	False positivity rate (1- Specificity)
No age	0.68 (0.19, 0.99)	0.0011 (0.0003, 0.0017)
<i>With age (in years)</i>		
5-11	0.67 (0.16, 0.99)	0.0017 (0.0004, 0.0025)
12-17	0.67 (0.16, 0.99)	0.0019 (0.0005, 0.0028)
18-24	0.67 (0.16, 0.99)	0.0013 (0.0003, 0.0019)
25-34	0.67 (0.16, 0.99)	0.0015 (0.0004, 0.0022)
35-44	0.67 (0.16, 0.99)	0.0012 (0.0003, 0.0018)
45-54	0.67 (0.16, 0.98)	0.0011 (0.0003, 0.0016)
55-64	0.67 (0.18, 0.99)	0.0008 (0.0002, 0.0012)
65-74	0.67 (0.15, 0.99)	0.0005 (0.0001, 0.0008)
75+	0.67 (0.20, 0.98)	0.0009 (0.0003, 0.0014)
No ethnicity	0.67 (0.20, 0.98)	0.0009 (0.0003, 0.0014)
<i>With ethnicity</i>		
Asian or Asian British	0.67 (0.18, 0.99)	0.0011 (0.0003, 0.0016)
Black; African; Black British or Caribbean	0.68 (0.18, 0.99)	0.0010 (0.0003, 0.0015)
Mixed or multiple ethnic groups	0.66 (0.16, 0.99)	0.0016 (0.0004, 0.0024)
Other ethnic groups	0.66 (0.16, 0.99)	0.0007 (0.0002, 0.0011)
White	0.72 (0.32, 0.99)	0.0009 (0.0004, 0.0013)

Table 4.2: **Estimated LFD test sensitivity and false positivity rate.** Posterior median (95% central credible interval) estimates of sensitivity and false positivity rate (= 1-specificity) in models without socio-demographic groups and with ethnic groups and age-groups. The possible values for the two quantities can range between 0 and 1. Since the number of data points in the age-group analysis and ethnicity analysis is not the same, as described in Table B.1, we produce two estimates for the model without sociodemographic variables (No age, no ethnicity).

4.4.3 Trends in testing biases

Our analysis indicated that infected individuals were much more likely to be tested by PCR than those uninfected. The testing biases for LFDs were considerably lower than for PCRs, meaning symptoms played a lesser role in shaping testing behaviours for LFDs. These results for both PCR and LFD tests are reflective of national testing policy throughout the study period. Throughout the analysis period, the relationship between testing biases (Figure 4.4C and D) and prevalence (Figure 3.5A and B) was not strong, but during the Delta wave and Omicron waves, there were visible increases in PCR testing biases (Figure 4.4C). In February 2022, the government removed the legal requirement for those in England to self-isolate following a positive test result and for close contacts who were fully vaccinated to conduct daily tests [130]. It is possible that this policy change prompted a shift towards increased usage of LFDs for symptomatic testing, which could explain the increasing bias in the use of LFDs in infected individuals (Figure 4.4C and D).

We found that deprivation and age were associated with testing biases on average, however the uncertainty intervals overlapped across deprivation levels (Appendix Figure C.1), although the age specific patterns were more distinct (Appendix Figure C.2). During the invasion phases of the Delta and Omicron VOCs, the most deprived populations were disproportionately more likely to test themselves via PCR only when infected versus when uninfected, and the same pattern held for LFD test-then-reporting. For example, in June 2021, at the start of the Delta wave (the second shaded box in (Figure 4.4A), the average odds of PCR testing in the most-deprived areas were $e^{3.21} \approx 25$ times higher in individuals with infection compared with individuals without infection; at the same time, the corresponding figure was only ≈ 11 in the least-deprived areas. From mid-February 2022, individuals in areas with different levels of deprivation had more similar PCR test-seeking behaviour. This was likely due in part to the lifting of the legal requirement to self-isolate [130].

Starting with the Alpha wave, the infection status for those aged 75+ played a lesser role in determining whether individuals tested themselves via PCR than for younger age groups (Figure 4.4B) [105]. We also found that the PCR testing bias was particularly high in school aged children (12-17 years) groups during March – June 2021 (Figure 4.4B) when LFD was extensively being used (corresponding to a low bias in LFD in (Figure 4.4D) [109]. LFD test reporting biases were more similar across age groups, although the oldest age group had a higher bias early on during the analysis period (Figure 4.4D).

4.4.4 Model fit

The model fit checks for the debiasing method are explained in the previous chapter’s results section 3.4.4. For deprivation analysis, we compared the various coarse groupings and found that the model with deprivation level and administrative regions had the highest predictive accuracy (Results section 3.4.4.2 and Table 3.6). In this grouping, the trends in PCR testing bias were more idiosyncratic, although there were regions and time periods where the bias was higher in more deprived areas. For example, during week of 1 June 2021 (Appendix Figure C.3) the estimated odds of testing were $e^{3.17} \approx 24$ and $e^{1.89} \approx 7$ times higher in individuals with infection compared to individuals without infection in most deprived ($\text{IMD} = 1$) and least deprived areas ($\text{IMD} = 5$) in North West England.

To check if our model for the decline in PCR detection subsequent to reporting an LFD positive test was a good fit to the data, we plotted the observed and predicted number of

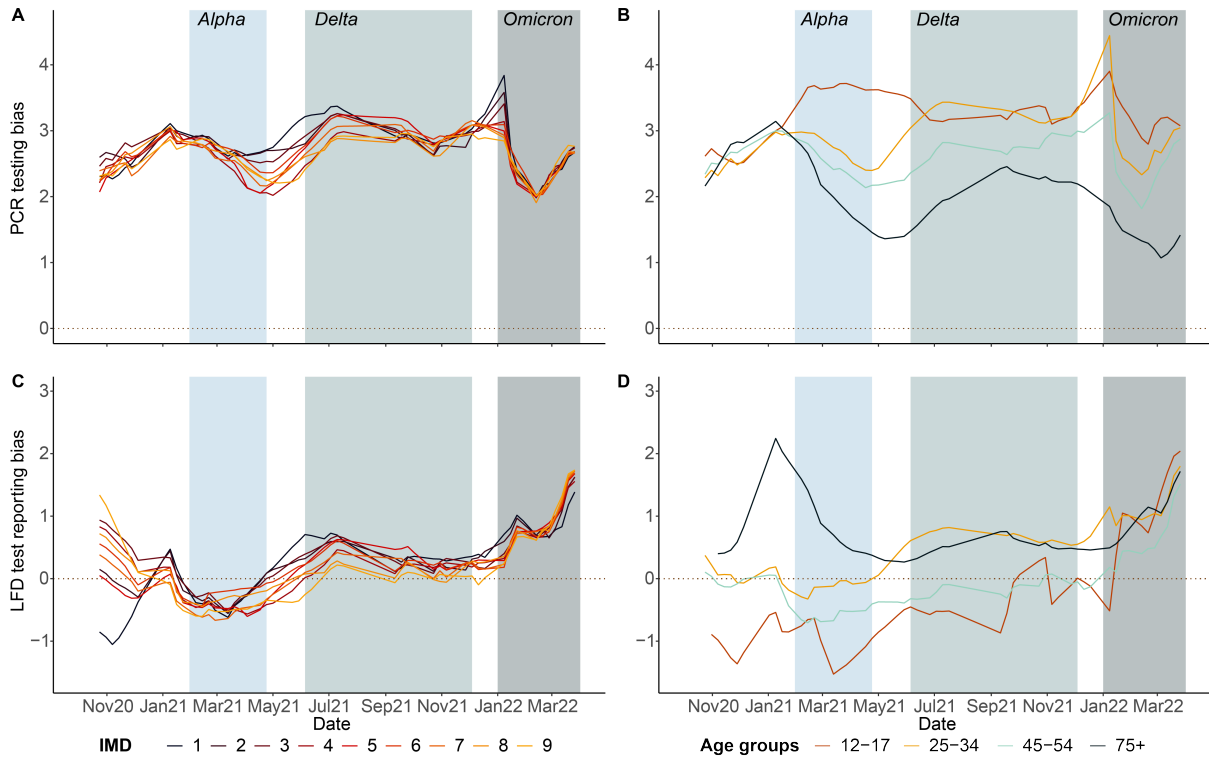


Figure 4.4: **Estimated testing bias by deprivation levels and age groups.** (A, B) Estimated mean PCR testing bias (log odds of taking a PCR test in infected versus uninfected subpopulations) by deprivation levels and age groups in England over time. (C, D) Estimated mean LFD test reporting bias in deprivation levels and age groups. 1 and 9 corresponding to most deprived and least deprived levels respectively. Shaded areas from left to right, correspond to the time periods of Alpha, Delta, and Omicron VOCs respectively (defined in Methods 3.3.2). Testing biases with uncertainty intervals and other age groups are provided in (Appendix Figure C.1, C.2).

$n_{PCR+,d,m}$ or “true positives”. For predicted estimates, we extracted the posterior median of $p_{d,m,eth}$ and multiplied it by $n_{LFD+,d,m,eth}$. Figure 4.5 shows that there was agreement between observed and predicted number of confirmatory PCRs, meaning that our model was appropriate.

4.5 Discussion

In this chapter, using data on reported positive LFD results which were followed by PCR tests, I presented and applied a method to estimate the sensitivity and specificity of LFD tests using mass testing data, which were in line with previous literature estimates [128, 129]. It is important to note that in addition to LFD tests being reported by individuals themselves, there was likely variation in how individuals performed tests, and there may be differences in laboratories and

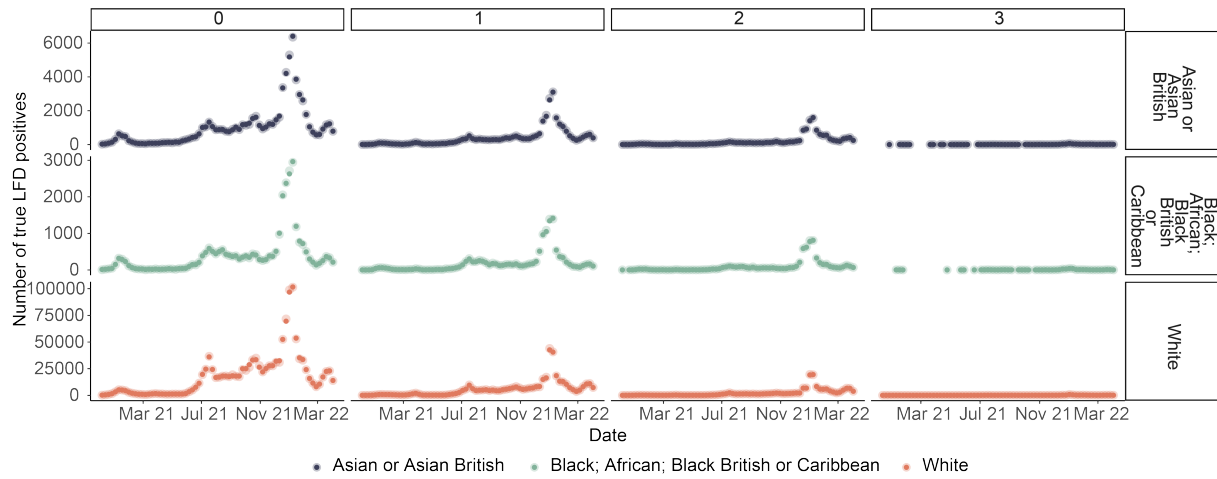


Figure 4.5: **Posterior predictive check for modelling decay rate of PCR detectability after reporting a positive LFD test, by different ethnic groups.** The lighter circles are the observed number of “true positives” i.e. positive PCR result within three days of reporting a positive LFD. Darker smaller circles are the estimates, for which we extracted the posterior median of probability of PCR positivity and multiplied it by the number of positive LFD reported.

test manufacturers. Our primary aim in this analysis was to illustrate how the sensitivity and specificity could be estimated using the data alone, without the use of pre-existing information, as might be the case during a future pandemic of a novel pathogen. However, using informative or weakly informative priors could provide us with more accurate and precise estimates.

Using the same confirmatory dataset, we modelled the decline in PCR detection rate after reporting a positive LFD test. The results hinted that individuals in the Black ethnic group may have tested and reported their positive LFD test later in their infection; however, these differences compared to other ethnic groups became smaller by the end of 2021. Lower values of this parameter later in the study period may indicate that infections were caught earlier by LFD tests or may reflect a longer period of PCR detectability for later VOCs. Other studies have found that the major barriers to testing for people in deprived areas were the financial and practical implications of having to self-isolate following a positive COVID-19 test [106], and those from ethnic minority backgrounds were disproportionately affected by this [131].

From the debiasing method described in detail in the previous chapter 3.3.3, we described the differences in testing biases by sociodemographic groups over time. In general, the PCR testing biases were higher compared to those of LFD, indicating that people were more likely to perform an LFD test regardless of infection status. This was in line with the general policy of symptomatic and asymptomatic use of PCR and LFD tests respectively (PCR tests were

also used as part of test-to-release policy and as confirmatory tests following a positive LFD result). During the invasion phases of Delta and Omicron, we found that the PCR testing bias was particularly marked in deprived areas. Many hypotheses could explain these behavioural differences, such as differential costs of self-isolation, differences in test accessibility and different views about the risks posed by infection, but our data cannot probe these. The PCR testing bias was consistently much lower in those aged 75+. This could be because older individuals were more at risk of severe disease, and they tested themselves more frequently by PCR due to caution [105]. It could also be because older people generally have more comorbidities, some of which may produce symptoms similar to COVID-19 causing them to seek PCR tests.

We considered three categories of sociodemographics in this chapter and the previous: self-reported ethnicities of individuals testing or reporting test results; their self-reported age; and a grouping of individuals according to the index of multiple deprivation of the LTLA where each test was taken (PCR) or taken and reported (LFD). These were chosen opportunistically according to criteria recorded in the testing data. These are suboptimal because they pool people according to somewhat arbitrary characteristics meaning they lack operational use; if it is known that self-reported Black individuals have a higher burden of infection, how should public health authorities actually respond? This group encompasses a huge diversity of different groups who live and interact in different communities. Existing categorisations also miss undocumented and presumably vulnerable individuals, such as undocumented migrants. Furthermore, there is compelling evidence suggesting a connection between belonging to an ethnic minority group and residing in economically disadvantaged areas, leading to an unequal burden of COVID-19 [132]. While our current study examines these factors in isolation, a potential future expansion could explore their combined impact [111]. Social scientists and local community leaders are better placed to understand these nuances, and interdisciplinary collaborations between these and those working in public health should be prioritised [133, 134].

The methods presented in this chapter collectively offer a toolkit for policymakers to monitor real-time equitable uptake of interventions and can help to design future surveillance studies.

Chapter C

Appendix - Differences in SARS-CoV-2 testing behaviours by sociodemographic groups

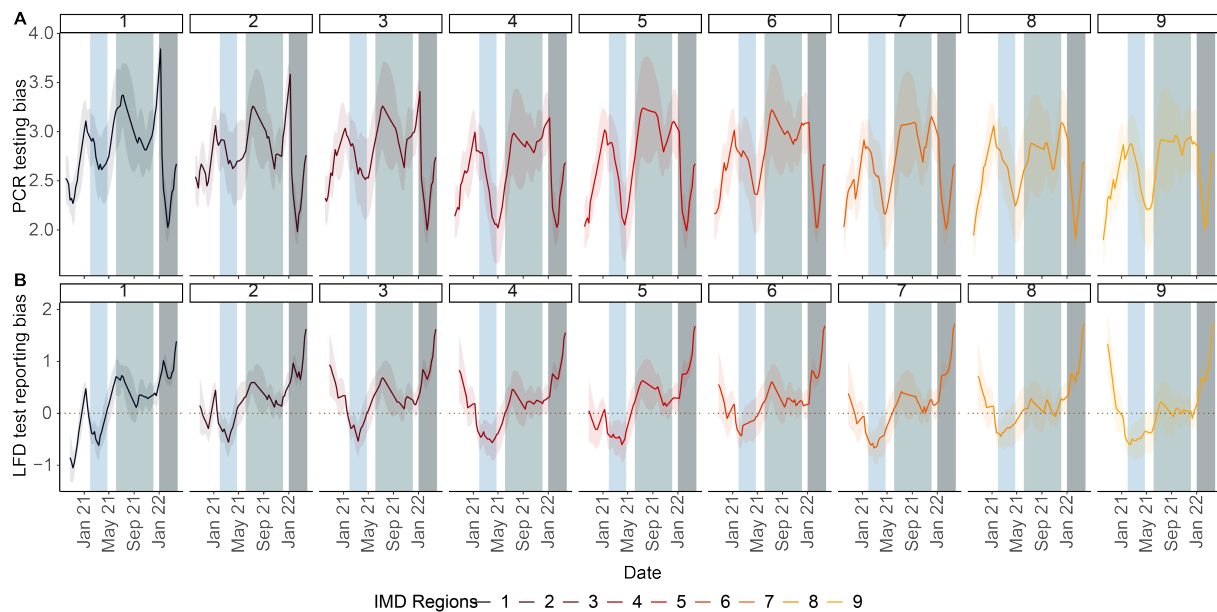


Figure C.1: **Testing bias with uncertainty intervals by IMD.** Estimated testing bias defined as the log odds of performing (or reporting) a PCR (or LFD) test in the infected versus uninfected subpopulations by deprivation levels. The solid curves are the posterior means, and the shaded areas are the 95% central credible intervals.

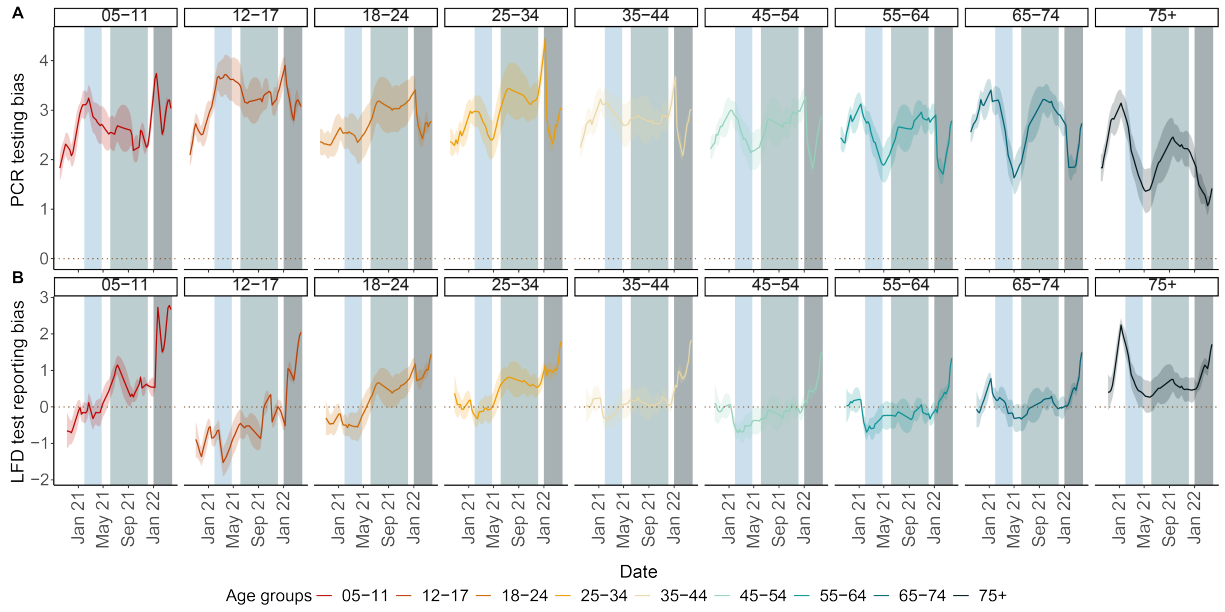


Figure C.2: **Testing bias with uncertainty intervals by age-groups.** Estimated testing bias defined as the log odds of performing (or reporting) a PCR (or LFD) test in the infected versus uninfected subpopulations by age groups. The solid curves are the posterior means, and the shaded areas are the 95% central credible intervals.

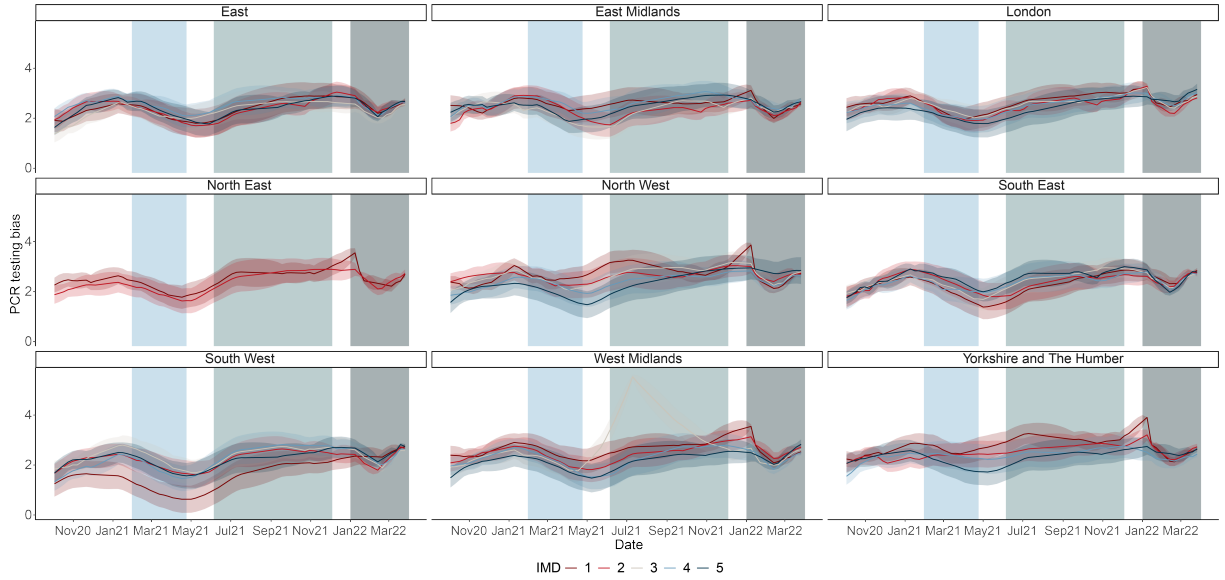


Figure C.3: **Estimated mean PCR testing bias by deprivation level and administrative regions.** The log-odds of performing a PCR test in the infected versus uninfected sub population by deprivation level (IMD 1 = most deprived, 5 = least deprived) and administrative regions of England over time and their 95% central credible intervals.

Chapter 5

Evolution of the spatial scales of transmission of SARS-CoV-2 in England between 2020 and 2022

After completing the previous project, I began studying Bayesian nonparametric methods as part of a reading group with researchers from the Departments of Computer Science and Statistics. As we read through the material, we started working on a tutorial paper to explain the mathematics and details of this rarely used methodology in epidemiology or infectious disease modeling. This inspired me to apply these methods to the fine spatial scale prevalence estimates from my earlier work. This chapter outlines the mathematics, code implementation, and inference process for using these flexible models to understand the drivers of spatial heterogeneity in SARS-CoV-2 transmission, applied to lower-tier local authority level prevalence estimates in England covering the months between October 2020 and March 2022.

5.1 Abstract

The COVID-19 pandemic revealed considerable variability in SARS-CoV-2 prevalence across England, influenced by demographic, socioeconomic, and mobility factors, alongside public health policies (Chapters 2, 3). In fact, the transmission of SARS-CoV-2 occurred at different levels of spatial scales - international, national, sub-national, throughout different phases of the pandemic in England, as discussed in 2. Understanding the distribution of SARS-CoV-2 prevalence is crucial for gaining insights into the underlying dynamics of virus transmission, which is essential for developing effective public health strategies and interventions. As the spatial scale of transmission can shift over time, particularly with the emergence of new variants, it becomes increasingly important to analyze these changes. Clustering techniques can provide valuable insights into these evolving patterns.

Traditional clustering methods, like K-Means and hierarchical clustering, have limitations in accurately capturing the dynamic nature of epidemiological data. This chapter introduces Bayesian nonparametric (BNP) methods, specifically Dirichlet process mixture models, to create flexible, data-driven clusters of SARS-CoV-2 prevalence without pre-specifying the number of clusters, which can scale with increasing data.

Our analysis of SARS-CoV-2 prevalence data from October 2020 to March 2022 highlights the effectiveness of BNP methods in identifying distinct transmission patterns and their drivers, such as mobility and spatial proximity. We found that periods of high prevalence were associated with more clusters, and factors like increased mobility between regions and spatial proximity significantly influenced synchronous transmission behaviors.

This approach offers nuanced insights into disease dynamics. The chapter also addresses the computational challenges of implementing BNP methods and provides code applicable to other epidemiological studies, demonstrating their broad utility in understanding and responding to infectious disease outbreaks. By enabling better estimation of epidemiological parameters at appropriate scales of aggregation, this approach will be particularly useful for tailoring optimal public health responses.

5.2 Introduction

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has profoundly impacted public health worldwide. In England, the trajectory of SARS-CoV-2 prevalence has shown considerable variability, with periods of rapid increase and decline influenced by a multitude of factors [94]. Demographic characteristics, socio-economic conditions, healthcare accessibility, mobility patterns, and public health policies have all likely played roles in shaping the geographic and temporal patterns of the spread of SARS-CoV-2 [119, 86].

Previous studies on understanding the role of synchrony in driving the spread of influenza [135] and measles [136] have highlighted the importance of spatial and temporal clustering in understanding disease spread and informing public health responses. Clustering analysis offers a valuable approach to examine the synchrony and asynchrony of SARS-CoV-2 transmission across different areas. By identifying clusters of high and low transmission over time, we can gain insights into the transmission dynamics and potential drivers of viral spread. In this chapter I was particularly interested in identifying groups of lower-tier local authorities (LTLAs) that had similar levels of transmission, defined as transmission clusters.

Traditional methods, such as K-Means [137, 138], and hierarchical clustering [139, 140], typically require the number of clusters to be specified in advance (or tuned in some approximate way, such as the elbow method), which can be a significant limitation when the true number of transmission clusters is unknown, uncertain, or varies over time. In contrast, Bayesian nonparametric (BNP) methods, such as Dirichlet Process Mixture Models (DPMMs) [141], allow for a flexible, data-driven determination of the number of clusters, adapting to the underlying distribution of the data. Additionally, using a spatially agnostic model, we can identify clusters based solely on the characteristics and similarities of the data, rather than based on their geographic proximity.

In this chapter, we first employ a DPMM to create clusters of SARS-CoV-2 prevalence using fine spatial scale data across England from October 2020 to February 2022, estimated in Chapter 3. After establishing these time-varying clusters, we investigate the factors driving synchronous transmission behaviours (or co-clustering) of LTLAs and how these factors evolved over time. A key challenge in implementing BNP methods lies in the inference and computation process, and here in addition to introducing and utilising this flexible class of models for public health questions, we also provide code which can be applied to other diseases and epidemiological

studies.

5.3 Methods

5.3.1 Data pre-processing

5.3.1.1 Prevalence estimates

I used the weekly, LTLA-specific SARS-CoV-2 prevalence between October 2020 and March 2022, estimated from the causal debiasing method introduced by [65], and was an output of chapter 3. In particular, I used the prevalence estimated in the scenario where the LTLAs were nested within the nine administrative regions of England (Table 3.4). In the prevalence data set, weeks run from Sunday to Saturday. We created a mapping between each date and which week it belongs to. For example, the dates 04/10/2020 (Sunday), 05/10/2020 (Monday) to 09/10/2020 (Friday), and 10/10/2020 (Saturday) belong to the same week. These prevalences were used in the clustering method described below.

We also use the national-level weekly estimates of prevalence reported by the Office of National Statistics (ONS) [88] to visualise the national prevalence trends in Figure 5.1.

5.3.1.2 Population density

Population density per square kilometer by LTLAs was accessed from ONS [100].

5.3.1.3 Mobility

The aggregated mobility data, provided by the O2 mobile phone carrier, includes weekday and weekend matrices detailing the volume of trips between all pairs of LTLAs as well as within individual LTLAs. Information on how to request access to this dataset can be obtained by contacting o2@businesso2.co.uk. The weekday data is provided for each week as the average daily number of trips over the five weekdays. For the weekend dataset, the data represents the average daily number of trips over two consecutive weekends.

By combining weekday and weekend trips, I calculated the number of trips within an LTLA i for a given week w (Sunday to Saturday), denoted as $M_{ii,w}$. Additionally, I calculated the weekly number of trips between pairs of LTLAs i and j . I now describe the process of calculating these weekly numbers of trips.

I assumed that the average number of weekday trips was constant across all five days, and the average number of weekend trips was consistent across all four days. Then, to ensure comparability between the prevalence and mobility datasets regarding the start and end days of a week, I mapped all dates in our mobility dataset to specific weeks, as described in the *Prevalence estimates* section above (section 5.3.1.1). For instance, the number of trips during the five weekdays starting on 5/10/2020 (Monday) was assigned to the week ending on 10/10/2020 (Saturday). We then added the estimated trips from 4/10/2020 (Sunday) and 10/10/2020 (Saturday) to determine the total number of trips for the week ending on 10/10/2020.

The number of trips between a pair of LTLAs say i and j , was calculated as the average number of weekly trips between LTLA i to j , and LTLA j to i :

$$M_{\bar{i}j,w} = \frac{M_{ij,w} + M_{ji,w}}{2}. \quad (5.1)$$

I then computed the number of weekly trips per capita (or mobility per capita) within an LTLA i and between pairs of LTLAs i and j as:

$$MPC_{ii,w} = \frac{M_{ii,w}}{P_{ii}}, \quad (5.2)$$

and

$$MPC_{\bar{i}j,w} = \frac{M_{\bar{i}j,w}}{P_{\bar{i}j}}, \quad (5.3)$$

where P_{ii} was the population of LTLA i , and $P_{\bar{i}j}$ was the mean of populations of LTLAs i and j .

5.3.1.4 Index of Multiple Deprivation (IMD)

ONS published the Indices of Deprivation (IoD2019) which are derived from 39 separate indicators, and are grouped into seven distinct domains of deprivation, then weighted and combined to create the IMD score, and rank. The IMD is an overall measure of multiple deprivation experienced by individuals in an area. They calculated this for each Lower-layer Super Output Area (LSOA) and then aggregated it to the LTLA level by taking the population-weighted average of all the LSOA scores within each LTLA. In this chapter, we use the aggregated "IMD-Rank of average score" for each LTLA in our analysis. Note: LTLAs with codes E06000060, E06000061, and E06000062 did not have an IMD assigned in the latest

dataset and were excluded from the analysis. The shapefile for the map in Figure 5.1C was downloaded from https://communitiesopendata-communities.hub.arcgis.com/datasets/29c47e93a802477eaca3d8b4d3a0c4b3_0/explore?location=52.723956%2C-2.489522%2C6.
58.

5.3.1.5 Age and population density

We used publicly available data on the population of LTLAs by one-year age groups published by ONS (for mid-2020 of this edition version), and calculated the proportion of the population older than 64 years for each LTLA [100].

5.3.1.6 Spatial proximity

The shapefile for LTLAs in England, which is publicly available by ONS, was used to calculate the distance (in km) between the centroids of all LTLAs and to identify neighbouring LTLAs. This is available at: <https://geoportal.statistics.gov.uk/datasets/ons::local-authority-districts-december-2022-boundaries-uk-bfc-2/explore>

5.3.2 Clustering

Overview

Here we describe a Bayesian nonparametric approach to clustering using Dirichlet processes mixture models. Let $p_{i,w}$ be the prevalence estimate of LTLA i (where $i \in 1, 2, \dots, N$ and $N =$ total number of LTLAs) during week w . We then create a multivariate vector consisting of all the prevalences in a given month m : $p_{i,m}$. We want to examine which LTLAs have similar multivariate prevalences or trends in a given month, i.e. which LTLAs belong to the same cluster. Unlike in the classical frequentist approach where nonparametric refers to not having a parametric form, Bayesian nonparametric models allow us to estimate a potentially infinite number of parameters: the finite number of parameters actually used to fit a particular dataset could potentially grow as the model sees more data. Now, instead of *a priori* specifying the number of clusters, this class of models lets the data inform how many clusters exist. Below, we describe the likelihood for our data, the Dirichlet process prior, and the inference algorithm which produces the clusters.

For notational ease, we will describe the model and process for a given month m , and refer

to LTLA i 's multivariate prevalence as p_i , without the time subscript. The components of the model are as follows.

5.3.2.1 Likelihood

We assume that our prevalences for LTLA i in a given month follow a multivariate normal distribution around their cluster mean θ_{z_i} and known (or fixed in our case) covariance matrix Σ :

$$p_i \sim \mathcal{N}(\theta_{z_i}, \Sigma), \quad (5.4)$$

where z_i is the cluster to which LTLA i 's prevalence belongs. It can take values $1, 2, \dots, K$ corresponding to the K distinct clusters. As discussed earlier, K is not fixed and is inferred by our clustering inference algorithm. We fix the covariance matrix of the likelihood to have diagonal elements equal to 1×10^{-5} , and off-diagonals equal to zero. The intuition is that as you increase this variance, you are likely to see fewer clusters. In Appendix Figure D.1 we show the choice of variance can impact the clustering.

Equation 5.4 is the likelihood for a given LTLA-specific prevalence. For the combined likelihood of a specific configuration of the clusters, we will require the overall marginal likelihood dependent on cluster assignments. In what follows, I derive the marginal likelihood from first principles.

By laws of conditional probability and marginalisation, we have:

$$P(A|B) = \int_{\gamma} P(A|B, \gamma) P(\gamma) d\gamma.$$

So, we can write our data generating process or (marginal) likelihood as:

$$\mathcal{L}(p_1, \dots, p_N | z_1, \dots, z_N) = \int_{\theta} P(p_1, \dots, p_N | \theta, z_1, \dots, z_N) P(\theta) d\theta \quad (5.5)$$

$$= \int_{\theta_1} \int_{\theta_2} \dots \int_{\theta_K} \underbrace{P(p_1 | \theta_{c_1}) P(p_2 | \theta_{c_2}) \dots P(p_N | \theta_{c_N})}_{\text{N data-point specific likelihoods}} \underbrace{P(\theta_1) P(\theta_2) \dots P(\theta_K)}_{\text{K cluster-specific parameter priors}} d\theta_1 \dots d\theta_K \quad (5.6)$$

$$= \int_{\theta_1} \int_{\theta_2} \dots \int_{\theta_K} \prod_{i=1}^N P(p_i | \theta_{z_i}) \cdot \prod_{j=1}^K P(\theta_j) d\theta_1 d\theta_2 \dots d\theta_K \quad (5.7)$$

$$= \prod_{j=1}^K \int_{\theta_j} P(\theta_j) \left(\prod_{\text{LTLA } i \text{ in cluster } j} P(p_i | \theta_j) \right) d\theta_j \quad (5.8)$$

$$= \prod_{j=1}^K \int_{\theta_j} \underbrace{\mathcal{N}(\theta_j | \mu_0, \Sigma_0)}_{\text{Base measure}} \left(\prod_{\text{prevalence } i \text{ in cluster } j} \underbrace{\mathcal{N}(p_i | \theta_j, \Sigma)}_{\text{likelihood from eq above}} \right) d\theta_j, \quad (5.9)$$

where θ is a vector of all cluster-specific means θ_{z_i} s, and $z_1, z_2, \dots, z_N$ is a vector of N clusters assignments, each of which can take values $1, 2, \dots, K$, describing which cluster each data point belongs to.

The base measure $\mathcal{N}(\mu_0, \Sigma_0)$ which acts as a prior for the cluster specific means is described below.

5.3.2.2 Prior

Concentration Parameter (α)

The concentration parameter α influences prior beliefs about the distribution of the cluster sizes. A higher α results in more, smaller clusters, while a lower α leads to fewer, larger clusters. In our case, we set α to 2. The rationale behind this choice is that, under the Dirichlet process prior, the expected number of clusters is given by equation 161 in [142] as $\sum_{i=1}^{\text{number of LTLAs}} \frac{\alpha}{i-1+\alpha}$. *A priori*, we make the assumption of expecting approximately 10 clusters, representing about one clusters per each of the nine administrative regions of England. This corresponds to an α value of approximately 2. We note that this choice merely specifies the location of the prior, and if the data suggests that significantly more or significantly fewer than 10 clusters are necessary, this will be reflected in our posterior results. We compared how the posterior number of clusters is affected by the choice of the concentration parameter.

The portion of the prior on that evaluates clustering is a partition function. The partition function tells us the probability of a particular arrangement (or partition) of LTLA specific prevalences into clusters. It is a way to represent all possible ways the data can be divided into clusters and the associated probabilities of these divisions. Let n_j indicate the number of LTLA prevalences assigned to cluster j . Then the partition function [143] can be written as:

$$p(n_1, \dots, n_K) = \frac{\alpha^{K-1} \prod_{j=1}^K (n_j - 1)!}{\prod_{i=1}^{N-1} (\alpha + i)}. \quad (5.10)$$

Base Measure

We already derived the implicit role of a base measure in the likelihood (section 5.3.2.1). This is used to describe the prior on our cluster-specific means. In our work, the base measure is a multivariate normal distribution. Specifically, if θ_{z_i} is the mean of the cluster z_i to which the i^{th} prevalence belongs, then $\theta_{z_i} \sim \mathcal{N}(\mu_0, \Sigma_0)$. Here, μ_0 is a vector of zeroes, and Σ_0 is the covariance matrix, defined as an identity matrix with diagonal elements equal to 1 and off-diagonal entries equal to zero.

5.3.2.3 Inference algorithm

Our model has a conjugate relationship between the multivariate normal likelihood and multivariate normal base measure from the prior on prevalence vectors. We thus marginalise out parameter values θ_{z_i} and employ the Gibbs sampler Neal’s Algorithm 3 [144] for inference. We want to use MCMC, specifically Gibbs sampling here to estimate the posterior cluster assignments z_i of each LTLA’s prevalences, every month. The algorithm presented here has been modified from [145].

To reiterate, each data point p_i is associated with an indicator variable z_i that specifies its cluster assignment. The Gibbs sampler updates the cluster assignment for each data point, conditional on all other cluster assignments. When updating z_i , we consider both, the posterior probability of assigning it to an existing cluster, say k , as well as the posterior probability of assigning it to a new cluster. In general, the Gibbs sampler can also update the cluster-specific parameter values (θ_{z_i}) conditional on the clustering configuration. But since we are not interested in estimating the cluster-specific means, we can ignore this, by marginalising out θ_{z_i} . So, the conditional posterior for z_i , conditional on the assignments of all other data points (z_{-i}) is given

by:

$$p(z_i = k | p_i, z_{-i}) \propto \begin{cases} \frac{\# \text{ LTLAs in cluster } k}{n - 1 + \alpha} \mathcal{N}(p_i | \frac{\bar{y}_k n_k \tau + \mu_0 \tau_0}{n_k \tau + \tau_0}, \frac{1}{n_k \tau + \tau_0} + \Sigma) & k \text{ has other LTLAs besides } i \\ \frac{\alpha}{n - 1 + \alpha} \mathcal{N}(p_i | \mu_0, \Sigma_0^2 + \Sigma^2) & k \text{ indicates a new cluster} \end{cases} \quad (5.11)$$

where $\bar{y}_k n_k$ is equivalent to the sum of all the prevalences belonging to cluster k , τ is the precision matrix of the multivariate normal likelihood (corresponding to the covariance matrix Σ), and similarly τ_0 is the precision matrix of the multivariate normal base distribution.

This has been implemented in R and can be accessed at: <https://github.com/ben18785/>

`dirichlet_process_tutorial/tree/main/inference_R.`

Algorithm 1: Gibbs Sampling for Dirichlet Process Clustering

Input: Data matrix *data*, concentration parameter α , mean vector μ_0 , covariance matrix Σ_0 , covariance matrix for likelihood Σ , initial cluster assignments *c_init*, maximum number of iterations *maxIters*

Output: Matrix of cluster assignments for each data point across all iterations

```

1 Initialize parameters and variables;
2 Compute number of data points data_n;
3 Compute  $\tau_0 = \text{inverse of } \Sigma_0$ ;
4 Compute  $\tau = \text{inverse of } \Sigma$ ; Set initial cluster assignments y_c;
5 Determine count of points in each cluster n_k;
6 Count total number of clusters c_n;
7 Create matrix (result) to store cluster assignments for each iteration ;
8 for iter ← 1 to maxIters do
9   for each data point i do
10     // Remove the data point from its current cluster
11     Identify the current cluster of data point i (c_i);
12     Decrement the count of points in the current cluster by one;
13     if the cluster becomes empty then
14       Replace the empty cluster with the last cluster;
15       Update the cluster count and assignments;
16     Temporarily set the cluster assignment of the current point to -1;
17     // Calculate log-probabilities for assigning the data point to each
18     cluster
19     Initialize a vector to store log-probabilities c_join_lp;
20     for each existing cluster do
21       Compute the sum of data points (prevalences) in the cluster;
22       Calculate the posterior mean and covariance for the cluster;
23       Compute the log-probability of the current data point joining the cluster;
24     Calculate the log-probability of the current data point starting a new cluster;
25     // Assign the data point to a cluster
26     Normalize the log-probabilities to obtain actual probabilities;
27     Sample a cluster for the current data point based on these probabilities;
28     if a new cluster is chosen then
29       Update the cluster count and assignments;
30     Assign the data point to the chosen cluster;
31     Update the count of points in the chosen cluster;
32   Store the current cluster assignments in the result matrix for the current iteration;
33 return result

```

If we were to multiply the likelihood and prior for a given cluster configuration, we can arrive at the unnormalised posterior which can be used in a Metropolis-Hastings (MH) implementation of an MCMC algorithm to infer cluster configurations. But in this work, we used a Gibbs sampler, because an MH algorithm would require consideration of proposal distributions, and we do not need that step in a Gibbs sampler, which only relies on conditional distributions for the parameters of interest. The conditional posterior for cluster assignments (not cluster-specific parameters in our case) can be derived by conditioning on all but one cluster assignment, but the derivation is based on results from Neal et al. in [144]. We plug in our base distribution and marginal likelihoods in the conditional posteriors. This is then used in the Gibbs sampler to infer cluster configurations.

In summary, we are interested in learning the clusters z_i s to which each LTLA's monthly prevalence vector p_i belongs. Often in mixture models we are interested in knowing the parameters associated with the clusters, say the mean θ_{z_i} or shape of the clusters. But here in our model, we marginalise this out and only extract the indicator variable governing the cluster assignment of each LTLA, z_i .

5.3.3 Model setup

5.3.3.1 Initialisation

For a given month, we run the MCMC algorithm with four different chains, with each chain having a distinct initial setting consisting of a cluster assignment for LTLAs. The four initial configurations are as follows: the first chain assigns each LTLA to its own cluster (i.e. the number of initial clusters is equal to the number of LTLAs), the second chain assigns all LTLAs to a single cluster, and the third and fourth chains start with random assignments of LTLAs to a random number of clusters. Although it's possible to extend this random assignment to multiple chains, our study generates two such chains or initial cluster assignments.

5.3.3.2 Final cluster assignment

After running all four chains for a specified number of iterations (2000 in our case), we record the number of clusters generated by each chain in the final iteration and select a chain that has reached the modal number of clusters. For instance, if chain 1 results in 2 clusters, while chains

2, 3, and 4 result in 10 clusters each, our method will choose one of the latter three chains as the final chain whose cluster assignments are used for further analysis. We then use the last iteration to create the cluster maps and cluster trends, and all the iterations to calculate the probability of co-clustering for pairs of LTLAs (section 5.3.3.4, below).

MCMC chains can be slow to converge, so we made this pragmatic choice of selecting a final chain with cluster assignments which yield reproducible results.

5.3.3.3 Entropy

We calculated the entropy for each month to quantify the uncertainty and diversity within the clusters of LTLAs over time. This provided insight into how distinctly or uniformly LTLAs (or populations of LTLAs) are grouped each month. The entropy H for a given month was calculated as:

$$H = - \sum_{j=1}^K \text{prop}_j \log(\text{prop}_j), \quad (5.12)$$

where prop_j is the proportion of LTLAs (or population) in the j^{th} cluster and K is the total number of clusters. An entropy value of 0 indicates that all LTLAs (or population) are concentrated in a single, perfectly homogeneous cluster. The maximum possible entropy value $H_{\max}$ given by:

$$H_{\max} = \log(K), \quad (5.13)$$

occurs when LTLAs (or population) are evenly distributed across all clusters, reflecting maximum diversity in the cluster assignments.

5.3.3.4 Probability of cluster membership

Using the samples of cluster assignments from the selected chain, we calculate the co-clustering matrix for each month, which quantifies how frequently each pair of LTLAs cluster together. For this, we only use the second half of the samples, and discard the first half as part of the warm-up of the MCMC sampler.

After computing the raw counts in a matrix MS , we normalize each element by dividing by half the total number of iterations. This transformation produces a similarity matrix which can

be converted to a vector S (to be used in the regression) between all combinations of LTLAs, where S_{ij} represents the probability $\in [0, 1]$ that LTLAs i and j are assigned to the same cluster. Note that the matrix MS is a symmetric matrix, such that $MS_{ij} = MS_{ji}$. We therefore only include one probability value in S per pair of LTLAs i and j .

5.3.3.5 Posterior predictive checks

To evaluate model fit, we compare the empirical cumulative distribution function (eCDF) of observed and posterior predicted prevalence estimates for each cluster in a given month.

Ideally for this we would have the estimated cluster-specific means $\theta_{z_i S}$, (where $z_i = 1, 2, \dots, K$) which is the same as $\theta_k S$, (where $k = 1, 2, \dots, K$). But since these cluster parameters are marginalised out in our model (section 5.3.2.3), we do not have posterior estimates for them from our inference algorithm. Instead, we employ an approximate solution. We begin by calculating the posterior distribution of the mean prevalence of cluster k , (where $k \in \{1, 2, \dots, K\}$); θ_k , which is given by:

$$\theta_k \sim \mathcal{N} \left((n_k \Sigma^{-1} + \Sigma_0^{-1})^{-1} (\Sigma_0^{-1} \mu_0 + \Sigma^{-1} \sum_{z_j \in k} p_j), (n_k \Sigma^{-1} + \Sigma_0^{-1})^{-1} \right) \quad (5.14)$$

where n_k represents the number of LTLAs in cluster k , and $\sum_{z_j \in k} p_j$ is the sum of all observed prevalences in cluster k [146].

Next, we draw ten samples from the distribution described in equation 5.14. For each week within a given month, we compare the eCDF of these posterior samples with the observed prevalences. Ideally, these comparisons should align along the 45-degree line when plotted against each other.

5.3.4 Epidemiological analysis

Here, we describe the details of the statistical analysis used to understand the factors associated with the probability S_{ij} of two LTLAs i and j co-clustering, i.e. being in the same cluster, in a given month m . We only consider months with three or more clusters, otherwise the outcome variable lacks variability to explain. In our linear regression model, we examine the probability $S_{i,j}$ that LTLAs i and j belong to the same cluster, based on various predictors. Our model takes the form:

$$S_{ij} = X_{ij} * \beta + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, 1), \quad (5.15)$$

where X_{ij} is the matrix of the covariates described below, and β is the vector of parameters describing the relationship between the predictors and outcome, which we aim to estimate. We can expand equation 5.15 as:

$$\begin{aligned} S_{ij} = & \beta_0 + \beta_1 |MPC_{ii,w} - MPC_{jj,w}| + \beta_2 MPC_{\bar{ij},w} + \beta_3 \left| \frac{IMD_i}{10} - \frac{IMD_j}{10} \right| \\ & + \beta_4 |PD_i - PD_j| + \beta_5 \log |prop64_i - prop64_j| + \beta_6 D_{ij} + \beta_7 I(B_{ij}) + \epsilon \end{aligned} \quad (5.16)$$

Each coefficient in the model provides insight into how differences or similarities in these covariates between the two LTLAs influence the co-clustering probability. Here we describe the covariates and how to interpret their respective coefficients:

- **Absolute difference in monthly within-LTLA mobility per capita (monthly trips per capita)** ($|MPC_{ii,w} - MPC_{jj,w}|$): For one additional pairwise difference in monthly trips per capita within LTLA, the probability of co-clustering increases by the coefficient. i.e. a negative coefficient suggests that LTLAs with similar levels of local mobility have an increased probability of being in the same cluster.
- **Average mobility per capita between the LTLAs (monthly trips per capita)** ($MPC_{\bar{ij},w}$): For one additional (average number of) monthly trip per capita between a pair of LTLA, the probability of clustering increases by the coefficient. A positive coefficient implies that LTLAs with higher connectivity have an increased likelihood of being in the same cluster.
- **Absolute difference in IMD levels (divided by 10)** ($|\frac{IMD_i}{10} - \frac{IMD_j}{10}|$): If the difference in IMD scores between a pair of LTLAs is 10 more than the difference between another pair, then the probability of the former pair being co-clustered increases by the value of the coefficient. i.e a negative coefficient means that LTLAs with similar levels of deprivation (corresponding to a lower difference) have a higher probability of being co-clustered.
- **Absolute difference in population density (1000 people per square kilometer)** ($|PD_i - PD_j|$): The coefficient here indicates if an LTLA pair's difference in population

density is 1000 people per square km more than another pair, then the former's probability of co-clustering is higher than the latter by the value of the coefficient, i.e. a negative coefficient means that LTLAs with similar levels of population density have a higher probability of being co-clustered.

- **Logged absolute difference in the proportion of people over 64 years** ($\log |prop64_i - prop64_j|$): This coefficient shows how a 1% (or 10%) pairwise increase in the absolute difference in the proportion of the elderly population increases the probability of co-clustering by $\frac{\text{Coefficient}}{100}$ (or $\frac{\text{Coefficient}}{10}$). A negative coefficient means that LTLAs with similar levels of people older than 64 years of age have a higher probability of being co-clustered.
- **Distance between the centroids of the LTLAs (100 km) (D_{ij})**: The coefficient for this variable represents the impact of a 100 km increase in the distance between the centroids on the clustering probability. A negative coefficient indicates that as distance decreases i.e. if LTLAs are closer in geographical distance, they are more likely to be in the same cluster.
- **Shared boundary indicator ($I(B_{ij})$)**: The coefficient for this binary variable indicates the change in clustering probability when the two LTLAs share a boundary (1) versus when they do not (0). A positive coefficient suggests that sharing a boundary increases the likelihood of being in the same cluster.

In summary, each coefficient tells us how a one-unit change in the respective covariate (or a specified change, such as 100 km for distance between pairs LTLAs or ten points for IMD pairwise difference of LTLAs) affects the probability that two LTLAs are in the same cluster, holding all other variables constant. Positive coefficients indicate an increase in clustering probability with increasing values of the predictor, while negative coefficients indicate a decrease. This interpretation requires careful consideration, as some of our predictors represent differences between the characteristics of LTLAs. Therefore, an increase in the value of these predictors corresponds to more dissimilar LTLAs (see section 5.3.4).

5.4 Results

Figure 5.1A presents national-level prevalence from ONS, and LTLA specific prevalence trends using data from chapter 3, alongside aggregated mobility data 5.1B, which describes within- and between-LTLA movements from October 2020 to February 2022. A notable drop in movement was observed following the implementation of the second lockdown on 5 November 2020. Between the second lockdown and the third lockdown, which began on 6 January 2021, a tiered system of restrictions was in place across the country, during which we observed a continued decline in movements. Post-January 2021, a gradual increase in mobility was seen, with within-LTLA movements recovering more quickly. However, the introduction of “Plan B” measures on 8 December 2021 [55], which included mandatory face masks in most indoor public places, COVID passes for large events, and work-from-home guidance during the spread of the Omicron variant, corresponded with a significant drop in mobility. Additionally, Figures 5.1C, D, E highlight the geographic heterogeneity in deprivation levels, the proportion of individuals aged over 65 years, and population density across local authorities in England.

5.4.1 Clustering and trends

Using the LTLA specific weekly prevalence estimates, we created transmission clusters for each month using the DPMM described in the Methods section 5.3.2.

Our MCMC chains converged for all the months (see Appendix Figure D.4), except for the chains which started with all LTLAs assigned to a single cluster in certain months. However, due to the design of our inference process, this chain would not be selected for the final clusters unless all other chains also converge to a single cluster (as described in the Methods section 5.3.3.2, we select a chain having reached the modal number of clusters).

As detailed in the Methods section 5.3.2, the concentration parameter α reflects our prior belief about the expected number of clusters. Generally, a higher value of α favors configurations with a larger number of clusters. For our analyses, we used a default value of 2, which resulted in approximately 10 clusters, aligning with the nine administrative regions of England. To assess the robustness of our clustering results to changes in α , we conducted a sensitivity analysis using values of 1 and 5, which correspond to expected cluster numbers of approximately 6 and 21, respectively. The results of this analysis indicated that varying α had minimal impact on the final number of clusters (Figure D.5). This suggests that the choice of α does not significantly

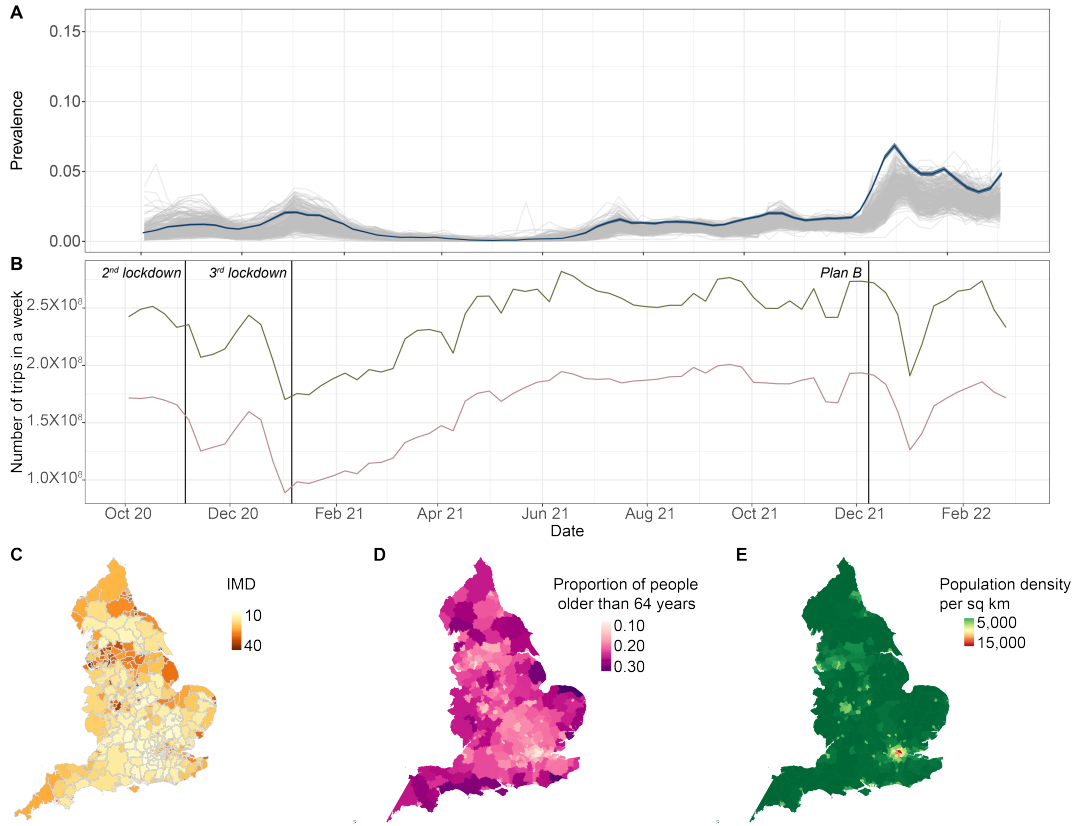


Figure 5.1: **SARS-CoV-2 prevalence estimates, alongside mobility and socio-demographic data.** (A) The grey lines show weekly LTLA-specific estimated prevalence from REACT-1 and mass testing data, while the blue line represents national prevalence from ONS data. (B) The green and pink lines indicate total weekly trips within and between LTLAs in England. See Appendix Figure D.2 for monthly trips per capita variation between LTLAs. Vertical bars indicate the timing of the second and third English lockdowns and the implementation of “Plan B.” Spatial distribution of (C) deprivation, (D) proportion of residents over 64 years, and (E) population density (per square kilometer) across LTLAs in England.

alter the clustering outcomes under the conditions tested.

In addition to the concentration parameter, the choice of likelihood function, particularly the noise parameter Σ , can also influence clustering. Specifically, Σ determines the sensitivity of the model to differences in the data, which refers to prevalences in a given month. In general, a smaller Σ value would result in the model assigning data points with smaller differences to separate clusters, thereby increasing the total number of clusters. In contrast, a larger Σ value would lead to fewer clusters, reflecting a model assumption that only substantially different prevalences should be grouped into distinct clusters. We evaluated the impact of varying Σ on the number of clusters formed, using data from January 2021 as an example (Appendix Figure

D.1). Our findings indicated that with the default parameter value of 1×10^{-5} , the data were divided into 5 clusters. Reducing Σ to 1×10^{-6} increased the number of clusters to 22, whereas increasing Σ to 1×10^{-4} resulted in only 2 clusters. These results demonstrate that while the choice of Σ may initially appear arbitrary, it aligns with our understanding of what constitutes a meaningful difference in prevalence.

We observed that the number of clusters varied between one and twelve (Figure 5.2A). Periods of low prevalence and lower variability between LTLAs, such as February 2021 to June 2021, were associated with fewer clusters. Conversely, periods with higher prevalence and greater variability between LTLAs, such as late 2020 and late 2021, were associated with a higher number of clusters, as seen in Figure 5.2A, 5.3, and 5.4. Our analysis showed that smaller clusters, defined as those with fewer LTLAs, generally consisted of LTLAs located closer to each other geographically (Figure 5.2B). This observation aligns with our regression analysis results, which identified spatial proximity as an important factor in explaining co-clustering. The entropy values, whether calculated using the population or the number of LTLAs in clusters, consistently indicate that clustering varies over time, with some months showing highly distinct and homogeneous clusters (low entropy), while others exhibit more dispersed and diverse clusters (higher entropy) (Figure 5.2C).

The clustering algorithm was able to identify distinct geographical patterns and trends within a given month. Our spatially agnostic method is able to pick out geographical trends and capture the spread of various variants of concern (VOC). For example, in December 2020, we observed a notable distinct trend in the Kent and Greater London regions during the invasion phase of the Alpha VOC, which were known to be the origin locations of this VOC [18]. Additionally, we detected the initial rise in infections of the Delta VOC from returning travelers in the northwest region in July 2021, followed by Greater London. This pattern was also evident in the number of independent introductions earlier in March 2021, aligned with mobility patterns [19]. In December 2021, we observed five distinct clusters with varying increasing prevalence trends. The cluster which exhibited the fastest growth, corresponded to the Greater London area, also noted by Elliott et al. [113].

To evaluate if our model was appropriate, we compared the observed prevalences with draws from the posterior distributions of cluster-specific means (section 5.3.3.5). The simulated prevalence values show a strong overlap with the observed prevalence data (see Appendix Figure D.6). We also compared the empirical cumulative distribution function (eCDF) of the observed and

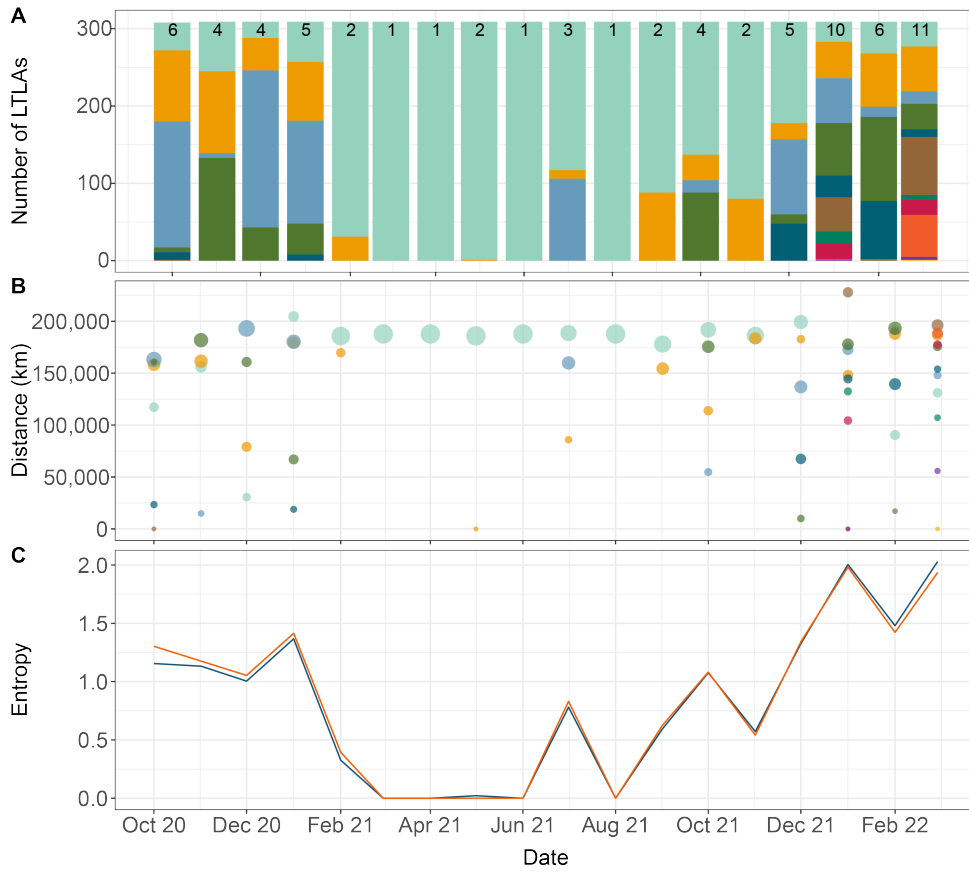


Figure 5.2: **Number of LTLAs in each cluster, average distance between cluster-specific LTLAs, and entropy.** (A) The bars show the distribution of LTLAs across distinct clusters for each month, with the numbers indicating the total number of clusters per month. (B) Average distance (in km) between LTLAs assigned to the same cluster. See Appendix Figure D.3 for the data points used to compute the average distance here. The size of the dots refers to the number of LTLAs assigned that particular LTLA. (C) Monthly entropy values for number of LTLAs (and population) in clusters in blue (orange). Entropy of zero corresponds to the configuration where all LTLAs are assigned to one cluster.

simulated prevalences (see Appendix Figure D.7). The eCDF is a function that represents the proportion of observations less than or equal to a particular value. It provides a non-parametric way to visualize the distribution of the data. The eCDF of the simulated data closely matches the eCDF of the observed data, i.e. it results in points lying near the 45-degree line. Both these checks indicate that our model effectively captured the underlying trends and variability in the observed data.

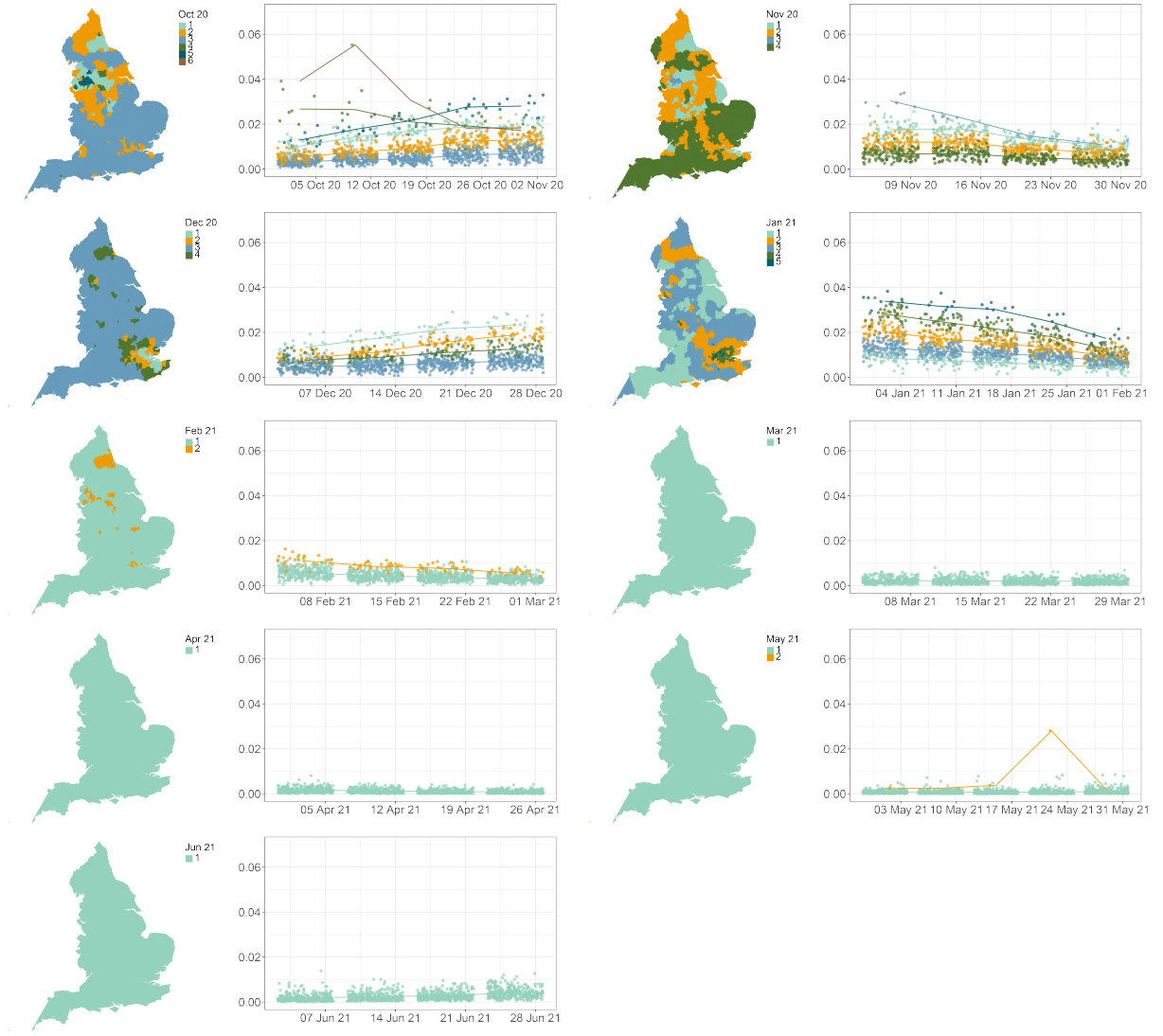


Figure 5.3: Spatial clustering of prevalence trends from October 2020 to June 2021. Each cluster is denoted by a distinct color. The points represent the observed weekly prevalences. The colors of the trends correspond to the cluster they belong to. The lines represent the mean of the observed prevalences for each cluster. The colors assigned to clusters have no meaning from one month to the next, i.e the cluster number assignments (and colors) can be interchanged in a given month, without affecting the interpretation.

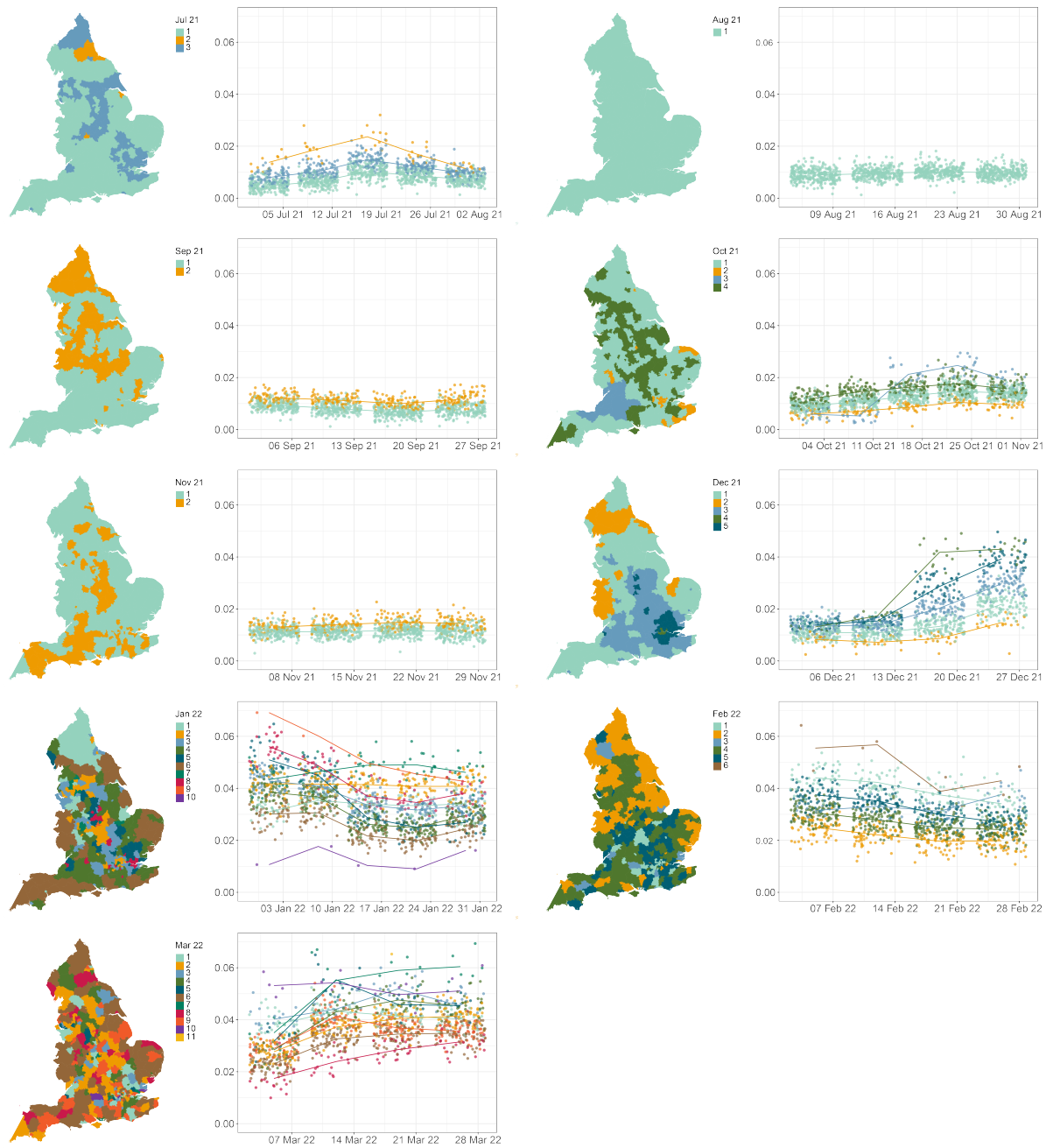


Figure 5.4: **Spatial clustering of prevalence trends from July 2021 to March 2022.** Each cluster is denoted by a distinct color. The points represent the observed weekly prevalences. The colors of the trends correspond to the cluster they belong to. The lines represent the mean of the observed prevalences for each cluster. The colors assigned to clusters have no meaning from one month to the next, i.e. the cluster number assignments (and colors) can be interchanged in a given month, without affecting the interpretation.

5.4.2 Factors driving synchrony in transmission

After creating time-varying clusters of prevalence trends, we wanted to assess the contribution of mobility, spatial, and demographic characteristics in determining transmission clusters. For this, we used a generalized linear model for the eight months between October 2020 and February 2022, during which we had at least three clusters. Increased movement between LTLAs, spatial proximity, and similar levels of population density were significant (i.e. 95% confidence intervals did not include the null parameter value of zero) predictors for the co-clustering of LTLAs. This is indicated by positive coefficients for monthly trips per capita between LTLAs and for the indicator variable for LTLAs being neighbors, negative coefficients for the distance between the centroids of the LTLAs and differences in population density between LTLAs (Figure 5.5).

Our analysis revealed evidence for the changing contributions of these factors over time. In particular, the coefficients of mobility between LTLAs which had the biggest effect size, initially increased over time (Figure 5.5), indicating that as time progressed, LTLAs with higher levels of connectivity were increasingly likely to be grouped together within the same transmission cluster. Although the effect size of the coefficients decreased in the later months of the study, they remained considerably larger compared to other factors. This may be explained by the implementation of “Plan B” restrictions and the acceleration of the nationwide booster vaccination program by December 2021 [55]. We also found that LTLAs with similar levels of deprivation were more likely to cluster together during most months, except in December 2020, and January 2021, corresponding to the invasion phase of Alpha VOC, and October 2021 and February 2022. The adjusted R-squared values ranged between 5% and 10%, indicating a relatively low level of explanatory power in the model. By January 2022, there was a notable drop in the variation in the probability of co-clustering explained by the covariates (Appendix Figure D.8). This was likely due to the high prevalence levels (Figure 5.4) and the corresponding increase in the number of clusters (Figure 5.2), making it challenging for the model to differentiate between signal and noise.

For model comparison, we fitted seven distinct models each month, with each model including an additional variable. We then performed ANOVA to evaluate the explanatory power of these models. The results indicated that the model with all seven variables explained the most variation in the outcome. This was supported by a high F-statistic and a significant p-value at 5% level of significance (see Appendix Tables D.1, D.2, D.3, D.4).

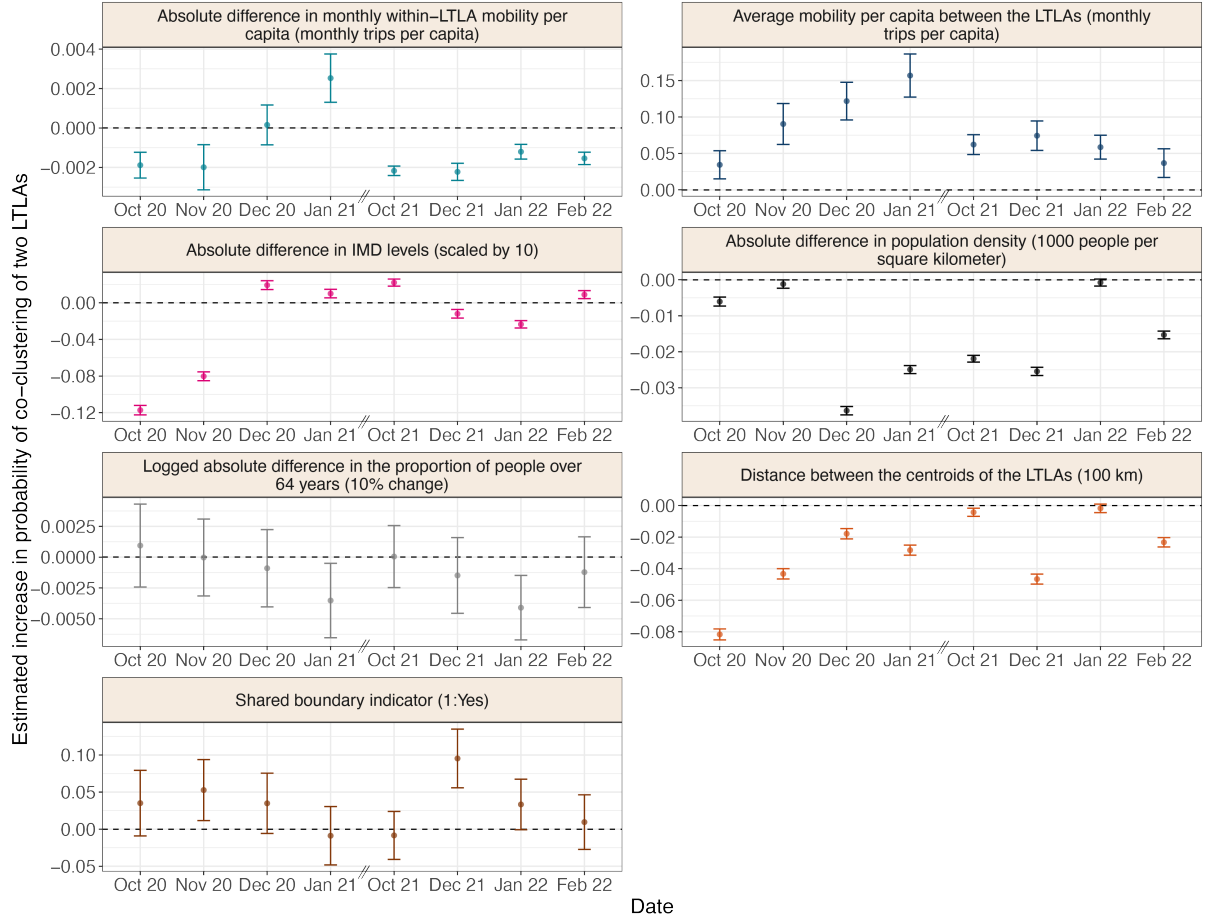


Figure 5.5: **Regression coefficient estimates of covariates explaining the probability of co-clustering of two LTLAs.** The dots represent the mean of the regression coefficient for a given covariate and the bars represent the 95% confidence intervals, for the months which estimated at least 3 clusters. The two slanted lines on the x-axis show a break in the months. Note - for logged absolute difference in the proportion of people over 64 years, we have divided the y-axis by 10, so that the coefficient corresponds an increase in the probability of co-clustering when the variable increases 10% instead of 1%.

5.5 Discussion

Identifying changes in clustering patterns of diseases and the drivers of such synchronous behaviours is crucial for understanding the transmission dynamics of pathogens across space and time. This is essential for implementing appropriate targeted public health responses, enhancing preparedness, and efficient resource allocation.

Our method leverages the flexibility of Bayesian nonparametrics, allowing the data to determine the number of clusters without pre-specifying it or incorporating geographical information about the areas. We demonstrate that our method effectively captures the spatial spread of

SARS-CoV-2 across LTLAs in England, even identifying areas that experienced a faster increase in various VOCs at different times, as highlighted by previous studies [18, 19, 55]. Additionally, we find that mobility, spatial proximity, and population density are important drivers of similar transmission dynamics between LTLAs.

These findings provide insights into how clustering patterns emerge and change during different phases of outbreaks, which may help inform future control strategies. Even when epidemiological data are subject to biases, routinely collected census and mobility data could still serve as reliable indicators for understanding transmission dynamics. Our results underscore that the scales of transmission are not static properties of geographical areas but are shaped by the dynamic responses of individuals to the spread of pathogens.

The clusters we identified did not align with the administrative boundaries of the nine regions in England. While the tiered system was implemented at the LTLA level, most publicly reported estimates of transmission levels, such as prevalence and incidence, were aggregated by administrative regions, which we have shown to be heterogeneous in terms of transmission dynamics [96, 88, 67, 147]. Studies, including our work in chapter 2, which used estimates at the LTLA level, often incorporated regions as a hierarchical structure under which LTLAs were assumed to be homogeneous [18, 19]. Our findings indicate that this assumption may obscure more localised variations in transmission, also noted by other studies which provided subnational estimates [147]. Future efforts should focus on refining the appropriate scale for hierarchical models and identifying the most suitable level for reporting epidemiological estimates to better capture the true transmission dynamics. This will be particularly valuable when case numbers are low, as hierarchical models typically borrow information from neighbouring areas.

BNP methods have found applicability in various subfields within epidemiology. Creswell et al. [148], introduced a Pitman-Yor process which is a generalisation of the DP to identify change points in the time-varying reproduction number (EpiCluster). The more commonly used BNP method, Gaussian processes have also been used to model the reporting-time delay for nowcasting of COVID-19 related mortality [149]. In spatial epidemiology, Gaussian processes have been used for disease risk prediction and mapping of malaria and dengue [150, 151], using spatial and environmental information. DPs have been used for creating antigenic clusters of different Influenza virus strains [152].

In this chapter, we used BNP for clustering and there are two broad advantages of the method we presented here. First, unlike traditional clustering methods like K-Means and Hierarchical

clustering, DPMMs have an advantage because the number of clusters does not have to be predefined. Second, during the SARS-CoV-2 pandemic, various studies also used Moran’s I to identify clustering patterns of transmission [153] and testing [154]. This is a measure of spatial autocorrelation, which gives an overall indication about the clustering pattern in a spatially explicit model but does not identify specific clusters. In comparison to this and Gaussian processes, our method relies on a spatially agnostic model, and we perform regression analysis with the covariates post-hoc after the clustering algorithm. This formulation can provide deeper insights into the underlying causes of disease clusters without making assumptions about the likelihood of areas being in the same transmission clusters, *a priori*.

The weekly LTLA-specific prevalence estimates used in the clustering is derived from the debiasing approach developed by Nicholson et al. [65] and implemented in Bajaj et al. [94]. The method assumes that biases (between true underlying prevalence estimate and naive prevalence estimate using Polymerase Chain Reaction testing data) at the regional level are reflective of those present within LTLAs nested within them, and violations in this assumption would result in biased estimates of infection prevalence. In regression analysis, we used IMD, which was an aggregated, time-invariant, and simplified measure of deprivation. Deprivation is a complex phenomenon with nuanced qualitative aspects that may not be fully captured by our quantitative measure. Covariates on population density and age structures were also assumed to be time-invariant.

Our clustering algorithm relies on setting the concentration parameter and defining the likelihood. We have observed that the number of clusters is more sensitive to the specification of the likelihood than to the concentration parameter. Intuitively, the standard deviation of the likelihood is related to what the user considers a significant difference in prevalence that justifies assigning data points to different clusters. Alternative approaches might include learning the standard deviation, although this may require an informative prior. While we employed a multivariate normal distribution for the likelihood (as the base measure of the Dirichlet prior) to allow efficient inference by analytically integrating out cluster-specific parameters, it is feasible to extend the method to non-conjugate prior-likelihood combinations for increased flexibility at the cost of more challenging inference. We used multivariable linear regression to quantify the association between various covariates and the probability of co-clustering of pairs of LTLAs. While Beta regression might seem appropriate, our outcome has values of 0 or 1, making it unsuitable. Even logistic regression is not suitable in our case, as the number of trials in the

binomial distribution would correspond to the number of iterations in the MCMC. Hence we continue with an interpretable multivariable linear regression. The variables used in our work could explain up to 10% of the variation in the probability of co-clustering.

Future research could focus on several key statistical advancements to enhance this work. One critical area is the development of formal convergence measures for the Markov Chain Monte Carlo (MCMC) algorithm, which could significantly improve the robustness and reliability of our statistical analyses. Additionally, incorporating temporal information directly into the clustering process rather than treating the model formulation as a multivariate distribution over a month could facilitate the identification of change points in the spatial evolution of the clustering process.

From an epidemiological perspective, a promising direction is to extend the current work by applying Hierarchical Dirichlet Processes. This approach would allow for the integration of data across multiple spatial scales, including sub-national, national, and international levels when answering questions spanning a range of geographies [155]. BNP methods could further our understanding of how transmission patterns shift with changing spatial scales throughout an outbreak. This would be particularly valuable in the early stages of an outbreak, where data are often limited, as it would allow us to borrow information from areas within the same transmission cluster, thereby improving the precision of estimates for critical parameters such as the reproduction number and growth rates. This approach could also be useful in identifying hotspots that might be facilitating the persistence of endemic transmission and in examining which areas might be acting as sources and sinks [156].

Future work could expand upon our statistical analysis of transmission synchrony by incorporating these dynamic spatial insights into complex agent-based models [157]. This integration would enable us to quantify additional factors influencing transmission similarity between areas, beyond the variables previously considered. Additionally, investigating how patterns of synchrony can be explained by typical data available versus the intrinsic randomness of outbreaks [158] would offer insights into the spatial dynamics of disease transmission, especially for pathogens that are associated with a shorter generation time and immunity duration. Such models would allow for a deeper exploration of the mechanistic drivers behind how a pathogen spreads in a given area, and how well can we predict its spread.

By bridging statistical analyses, epidemiological insights, and advanced modeling approaches, we can develop a more comprehensive understanding of disease dynamics and enhance our ability to inform public health strategies in real time. This interdisciplinary approach is vital for tackling

the complexities of infectious disease transmission and improving our preparedness for future outbreaks.

Overall, the flexible Bayesian nonparametric models presented in this work could serve as a valuable tool for health systems surveillance, offering insights into spatial patterns of transmission intensity and burden for various pathogens across different geographical contexts. However, further research and validation will be necessary to fully realize their applicability in real-world settings.

Chapter D

Appendix - Evolution of the spatial scales of transmission of SARS-CoV-2 in England between 2020 and 2022



Figure D.1: **Spatial clustering of prevalence trends in Jan 2021 for different values of Σ .** Each cluster is denoted by a distinct color. The points represent the observed weekly prevalences. The colors of the trends correspond to the cluster they belong to. The lines represent the cluster specific mean of the prevalences. The panels from top to bottom represent scenarios when diagonal elements of covariance matrix from the likelihood equal to 1×10^{-4} , 1×10^{-5} (default), 1×10^{-6} .

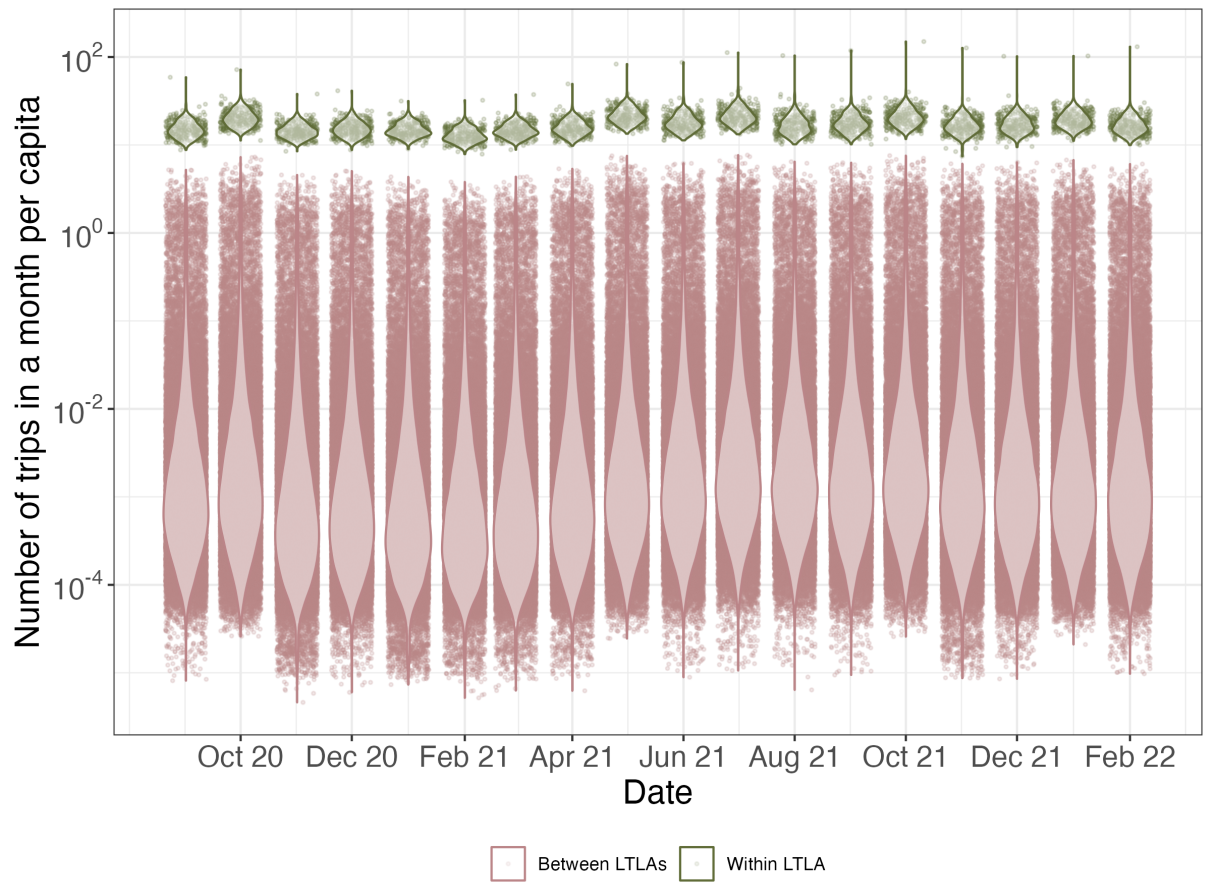


Figure D.2: **Monthly number of trips per capita across LTLAs.** The points represent the observed LTLA specific monthly number of trips per capita. The pink dots refer to average trips between pairs of LTLAs, and green dots refer to trips within an LTLA. The width of each violin represents the density of data points.

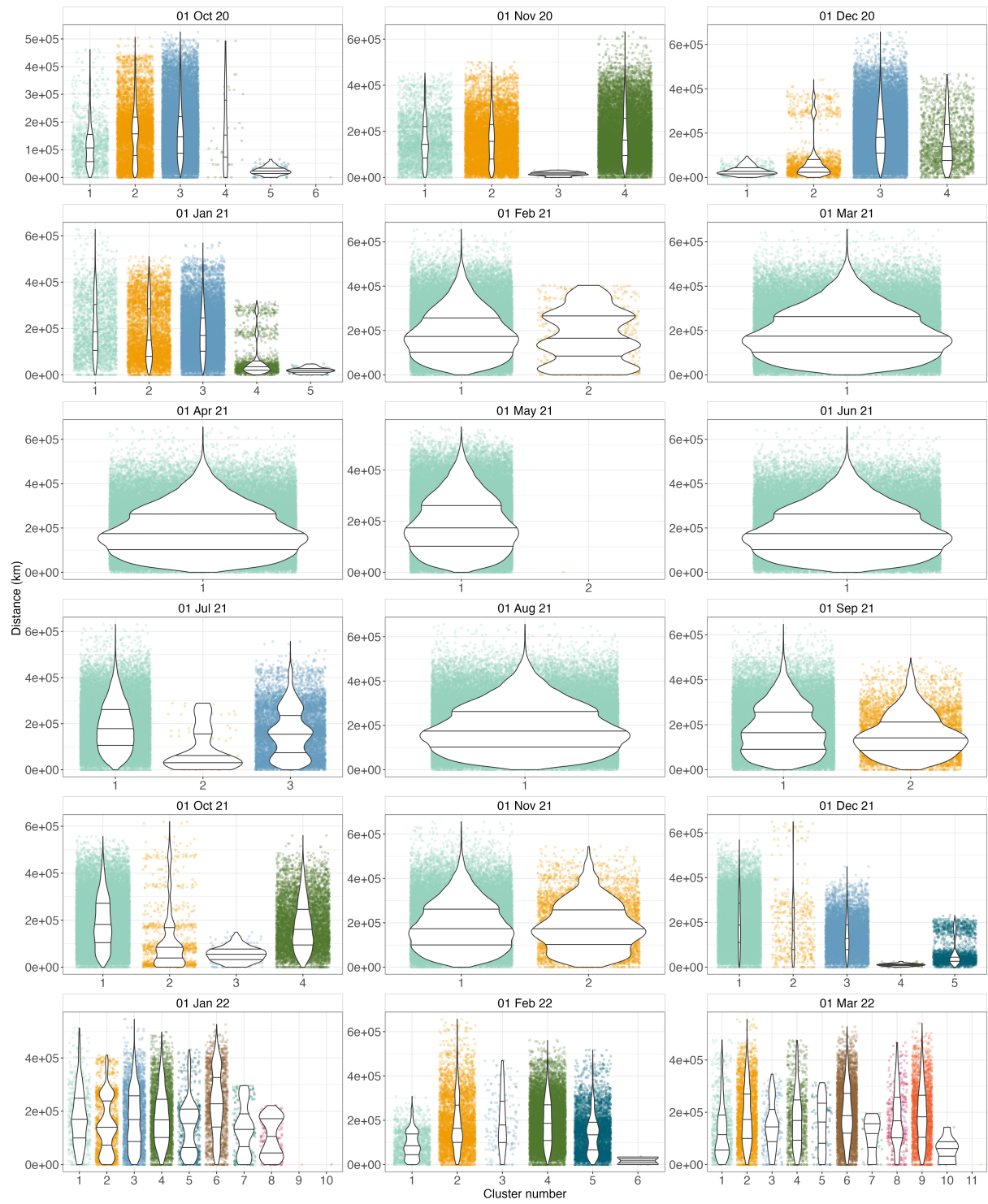


Figure D.3: **Distance between cluster-specific LTLAs.** Distance (in km) between LTLAs assigned to the same cluster for each month. The violins represent the median and the inter-quartile range.

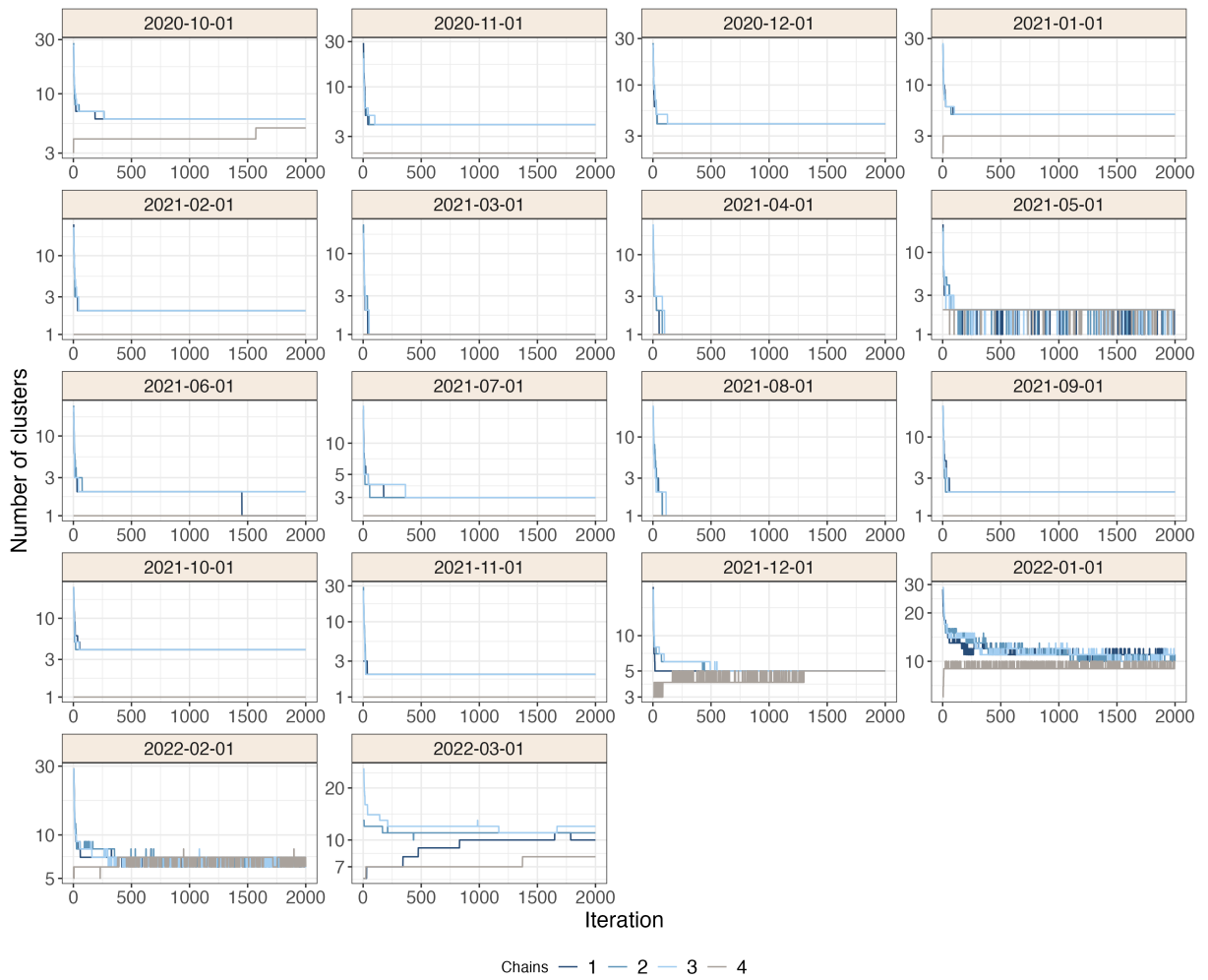


Figure D.4: **Convergence of MCMC chains for clustering.** Monthly number of clusters estimated for each iteration (represented on the x-axis) of the MCMC algorithm. The colors represent the four different chains with different initial values of number of clusters.

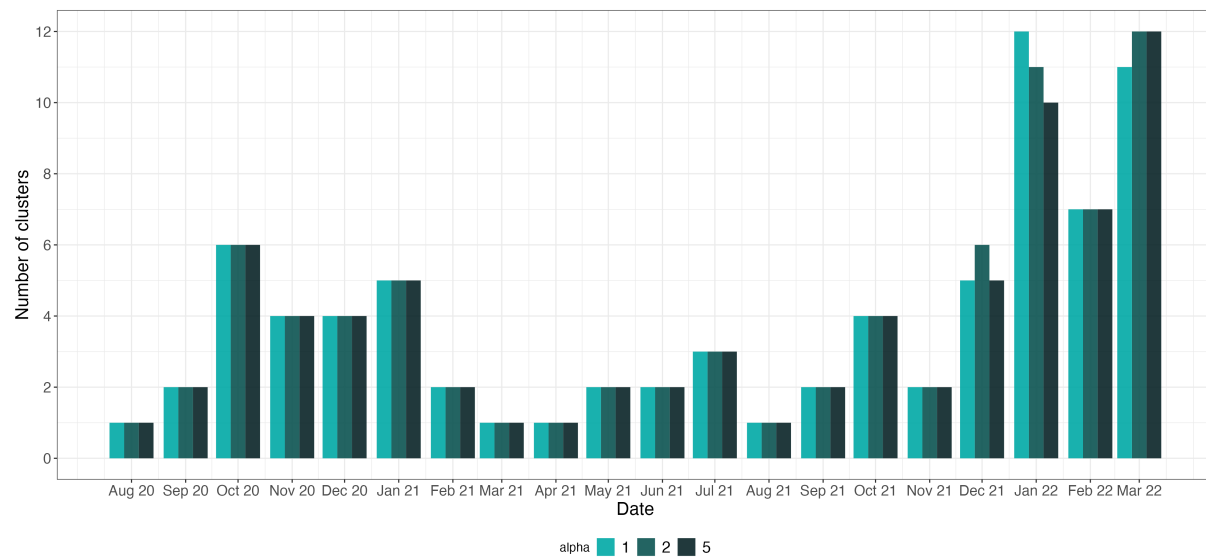


Figure D.5: **Effect of concentration parameter choice on clustering.** The bars represent the monthly number of clusters for different values of the concentration parameter α .



Figure D.6: **Posterior predictive checks of prevalence estimates.** The x-axis corresponds to the week, y-axis corresponds to prevalence, and facets are the cluster number to which LTLAs are assigned in a given month. The yellow points represent LTLA specific observed weekly prevalences, and the green points represent a draw from the posterior distribution of the cluster specific means, to which each LTLA belongs to.

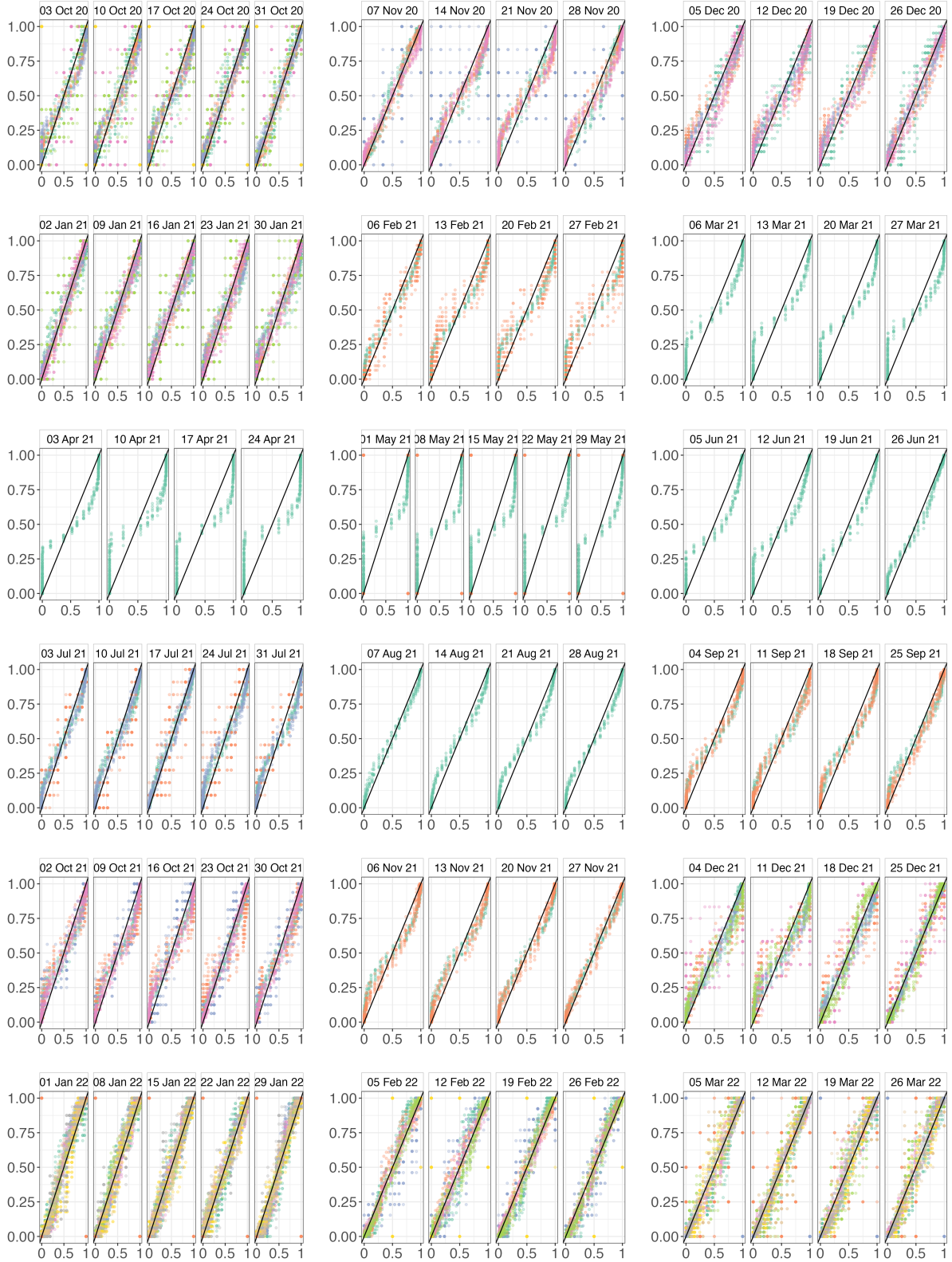


Figure D.7: **Posterior predictive checks, comparing eCDFs.** The x-axis corresponds to the empirical cumulative distribution function (eCDF) of observed prevalences, and y-axis corresponds to the same from simulated prevalences.

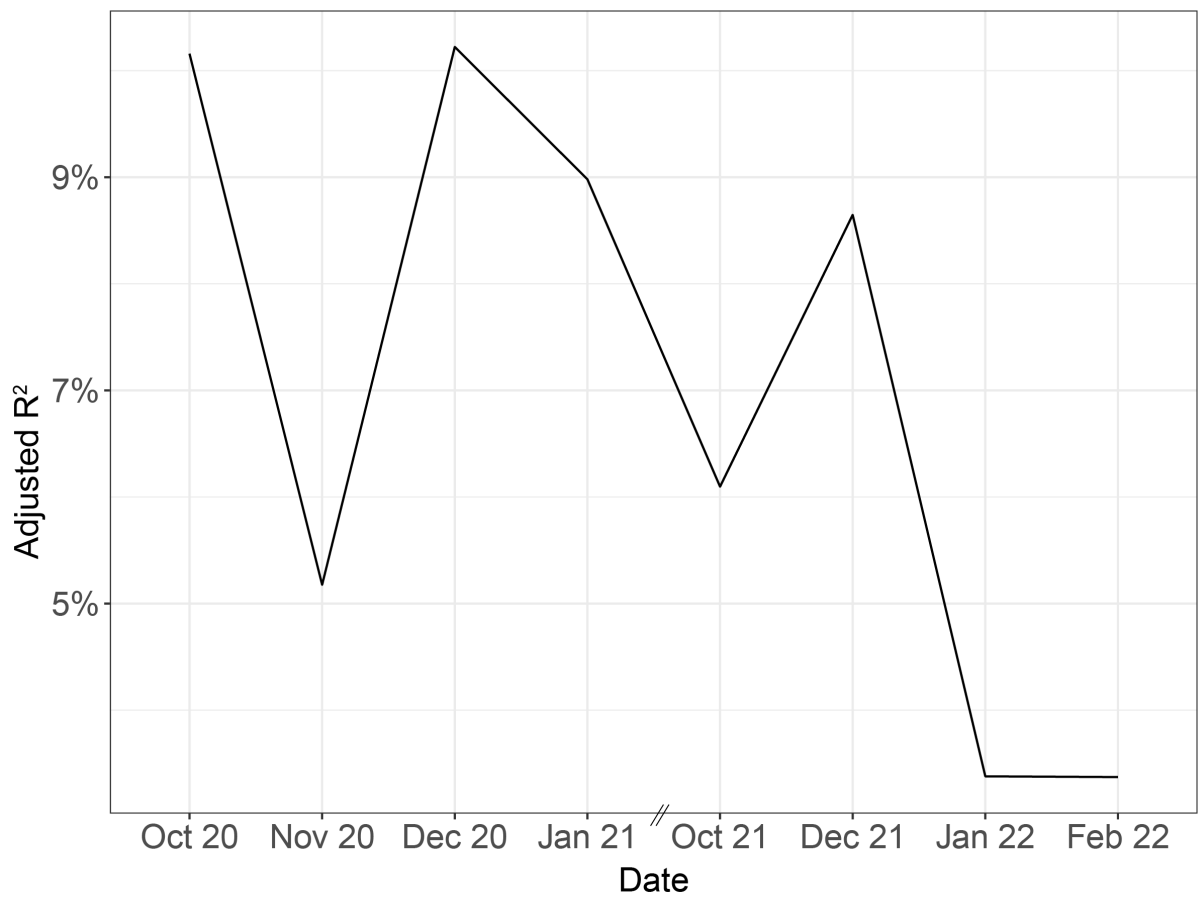


Figure D.8: **Adjusted R^2 from linear regressions.** Adjusted R^2 values for each month where the outcome was probability of co-clustering for pairs of LTLAs and the covariates are described in Methods section 5.3.4.

Variables	Model number
$\log prop64_i - prop64_j $	1
$\log prop64_i - prop64_j ,$ $ MPC_{ii,w} - MPC_{jj,w} $	2
$\log prop64_i - prop64_j ,$ $ MPC_{ii,w} - MPC_{jj,w} ,$ $ \frac{IMD_i}{10} - \frac{IMD_j}{10} $	3
$\log prop64_i - prop64_j ,$ $ MPC_{ii,w} - MPC_{jj,w} ,$ $ \frac{IMD_i}{10} - \frac{IMD_j}{10} ,$ $ PD_i - PD_j $	4
$\log prop64_i - prop64_j ,$ $ MPC_{ii,w} - MPC_{jj,w} ,$ $ \frac{IMD_i}{10} - \frac{IMD_j}{10} ,$ $ PD_i - PD_j ,$ $I(B_{ij})$	5
$\log prop64_i - prop64_j ,$ $ MPC_{ii,w} - MPC_{jj,w} ,$ $ \frac{IMD_i}{10} - \frac{IMD_j}{10} ,$ $ PD_i - PD_j ,$ $I(B_{ij}),$ D_{ij}	6
$\log prop64_i - prop64_j ,$ $ MPC_{ii,w} - MPC_{jj,w} ,$ $ \frac{IMD_i}{10} - \frac{IMD_j}{10} ,$ $ PD_i - PD_j ,$ $I(B_{ij}),$ $D_{ij},$ $MPC_{ij,w}$	7

Table D.1: **Description of models compared** The first column shows which variables were included (as defined in the methods section 5.3.4) in the ANOVA test for model comparison, and the second column represents the number they will be referred by in the subsequent tables.

Month	Model	RSS	SS	F-statistic	p-value
October 20	1	7415.4	-	-	-
	2	7399.2	16.2	111.9	4.12e-26
	3	7027.1	372.1	2567.2	0e+00
	4	7023.7	3.4	23.7	1.1e-06
	5	6984.3	39.3	271.2	9.1e-61
	6	6663.4	320.9	2214.1	0e+00
	7	6661.6	1.8	12.2	4.9e-04
November 20	1	5906.6	-	-	-
	2	5898.6	8.0	65.2	7.0e-16
	3	5736.1	162.4	1316.2	4.1e-284
	4	5736.1	0.01	0.1	0.7
	5	5697.9	38.3	310.0	3.7e-69
	6	5606.8	91.0	737.8	3.5e-161
	7	5601.9	4.9	39.8	2.9e-10
December 20	1	6182.1	-	-	-
	2	6181.9	0.2	1.8	0.2
	3	6181.4	0.4	3.6	0.06
	4	5723.1	458.3	3682.2	0e+00
	5	5689.2	33.9	272.4	5.1e-61
	6	5671	18.2	146.1	1.4e-33
	7	5660.3	10.6	85.4	2.6e-20

Table D.2: **Model comparison for monthly regressions in 2020.** Model number corresponds to the variables included, as specified in Appendix Table D.1. RSS and SS refer to the residual sums of squares and sums of squares for each model.

Month	Model	RSS	SS	F-statistic	p-value
January 21	1	5447.6	-	-	-
	2	5446.7	0.8	7.3	7e-03
	3	5446.6	0.1	1.2	0.2
	4	5240.2	206.4	1808.1	0e+00
	5	5210.3	29.9	261.7	1e-58
	6	5169.2	41.0	359.5	7.6e-80
	7	5156.9	12.6	108.2	2.6e-25
October 21	1	3975.6	-	-	-
	2	3950.9	24.7	303.1	1.2e-67
	3	3946.5	4.4	54.1	2e-13
	4	3776.2	170.3	2093.5	0e+00
	5	3766.6	9.5	117.2	2.7e-27
	6	3764.9	1.8	21.6	3.4e-06
	7	3758.4	6.5	79.8	4.4e-19
December 21	1	5976	-	-	-
	2	5955.9	20.1	166.3	5.5e-38
	3	5938.8	17	141	1.8e-32
	4	5735.1	203.8	1685.2	0e+00
	5	5675	60.1	496.9	1.7e-109
	6	5565.1	109.8	908.2	1.3e-197
	7	5558.8	6.3	52	5.7e-13

Table D.3: **Model comparison for monthly regressions in 2021.** Model number corresponds to the variables included, as specified in Appendix Table D.1. RSS and SS refer to the residual sums of squares and sums of squares for each model.

Month	Model	RSS	SS	F-statistic	p-value
January 22	1	4065.5	-	-	-
	2	4061.4	4.1	47.1	6.7e-12
	3	4047.9	3.5	154.3	2.2e-35
	4	4047.7	0.2	2.3	0.1
	5	4034.8	12.9	147	8.8e-34
	6	4034.3	0.5	5.5	0.001
	7	4030	4.3	49	2.5e-12
February 22	1	4954	-	-	-
	2	4943.5	10.4	99.3	2.2e-23
	3	4943.5	0.002	0.002	0.08
	4	4868.5	75	713.4	5.8e-156
	5	4861	7.6	72.1	2.1e-17
	6	4833.5	27.5	261.3	1.3e-58
	7	4832.1	1.4	1.3	2.6e-04

Table D.4: **Model comparison for monthly regressions in 2022.** Model number corresponds to the variables included, as specified in Appendix Table D.1. RSS and SS refer to the residual sums of squares and sums of squares for each model.

Chapter 6

Discussion

6.1 Summary

In this thesis, I presented four research chapters that described role of spatial and sociodemographic variation on SARS-CoV-2 transmission and public health interventions in England between October 2020 and March 2022.

In chapter 2, I combined phylogeographic results from my collaborators with statistical methods to describe the drivers of transmission of the Alpha and Delta VOCs describing their initial spread throughout the country. I extended methods for estimating growth rates and found that Alpha spread across the UK, driven by mobility patterns from Kent and Greater London, whereas Delta's emergence occurred within a diverse landscape of easing restrictions, where local mobility played an important role in its establishment.

In chapter 3, I presented the differences in the PCR tests performed and LFD tests performed and reported over time across a range of sociodemographic groups. We found a higher prevalence in deprived areas where the number of PCR tests performed was nearly double that of the least deprived areas. However, LFD test reporting was lower in the most deprived areas, almost half compared to least deprived areas, even after the rollout of free asymptomatic testing. We then introduced a method which combined mass testing data with random community surveillance surveys to estimate prevalence at fine spatial scales and by sociodemographic groups. This approach highlighted higher infection prevalence in Asian or Asian British communities during the Alpha and Omicron BA.1 waves. We also found that secondary school-aged children experienced the highest infection rates during these periods. Additionally, we developed a novel method to estimate incidence and found that the national testing programme captured 26% – 40% of the

infections.

In chapter 4, I presented the biases in testing from the method presented in the previous chapter. During the invasion phases of the Delta and Omicron VOCs, the PCR testing bias in the least deprived populations was roughly half of that in the least, suggesting that people in the least deprived areas were more likely to perform PCR tests regardless of their infection status. Using confirmatory PCR data (LFD positive results followed by a PCR test within three days), we estimated false positivity rates, and developed novel methods to estimate the sensitivity, and specificity of the LFD tests which aligned closely with the literature. Additionally, we modelled the likelihood of a positive confirmatory PCR test relative to the time elapsed since the LFD positive result, finding potential delays in reporting positive LFD tests among Black populations.

Lastly in chapter 5, using the finescale prevalence estimates from chapter 3, I introduced Bayesian nonparametric methods to identify transmission clusters, i.e. groups of areas that had similar levels of transmission, and examined how these clustering patterns evolved over time, influenced by factors such as mobility and spatial proximity. I describe the inference process of the Dirichlet process mixture model and highlight the computation challenges associated with these flexible models.

This thesis introduces and extends novel methods from the literature to enhance our understanding of infectious disease transmission and improve the monitoring of public health interventions, such as testing uptake across different groups.

In the following sections, I outline potential areas for future research.

6.2 Future directions

6.2.1 Statistical, mechanistic and causal models

Mechanistic versus statistical methods

Mechanistic models and statistical methods serve different but complementary purposes in the study of infectious diseases. Statistical methods are primarily concerned with identifying relationships between variables, or the predictive power of a collection of variables on outcomes. For example, regression methods can help us identify correlations between prevalence and incidence or other epidemiological factors. These models are data-driven and excel at forecasting and identifying patterns based on existing data, often without requiring an explicit formulation of the

underlying processes. In chapters 2 and 5 we used statistical models to estimate growth rates, identify clusters, and to examine the relationship between mobility patterns and the spread of VOCs, as well as the spatial synchrony of transmission.

On the other hand, mechanistic models are grounded in the underlying biological, ecological, or social mechanisms that drive disease dynamics. These models attempt to formally define the processes governing the transmission dynamics of a pathogen within a population. They are particularly valuable for testing hypotheses about the mechanisms through which variables influence outcomes, such as how changes in human behavior or environmental conditions might impact disease spread. In chapter 3, we explicitly modelled our beliefs about the bias structure in mass testing data to estimate the prevalence of SARS-CoV-2 at fine spatial scales and across sociodemographic groups. Similarly, in chapter 4, we mechanistically defined the relationship between incidence and prevalence to estimate incidence at these fine scales. Additionally, we utilized data from confirmatory PCR tests to clearly outline the underlying mechanisms, enabling us to estimate the specificity and sensitivity of LFD tests, as well as to analyze how PCR detectability varied among ethnic groups and changed over time.

By incorporating our beliefs about how a system works, mechanistic models provide a framework to explore the “how” of disease transmission, complementing the “what” identified by statistical methods. Both statistical and mechanistic approaches are essential for a comprehensive understanding of disease dynamics. When used together, these approaches can guide us toward a deeper understanding of the drivers of disease spread.

COVID-19 and causal models

Causal inference is critical not only for understanding relationships between variables but also for identifying the underlying causes of these relationships. Moving forward, a key area of development should be the integration of statistical, mechanistic, and causal models to create a more robust framework for understanding and controlling infectious diseases. Recent studies during the COVID-19 pandemic effectively utilized causal inference methods to gain deeper insights into transmission dynamics and the impact of interventions. For example, a study used the quasi-experimental design of matched-neighbors to assess the NHS app’s role in reducing COVID-19 transmission [159]. Through mediation analysis, another study highlighted the Simpson’s Paradox in COVID-19 case fatality rates between China and Italy, illustrating how age-related factors can distort data interpretations [160]. Additionally, these researchers emphasised the need to

clearly articulate causal pathways before analysing observational data, which can help to identify selection biases and prevent misleading conclusions regarding vaccine effectiveness. Moving forward, incorporating causal inference methods, from quasi-experimental designs including but not limited to instrumental variables, propensity score matching, difference-in-differences, and regression discontinuity designs can further enhance our understanding of infectious disease dynamics and improve the reliability of our findings.

Causal inference in infectious disease modelling

Despite these promising developments, there are significant challenges in applying existing causal inference literature to the field of infectious disease modeling. In traditional epidemiology, study designs like cohort, cross-sectional, case-control, and clinical trials are frequently used. However, infectious disease epidemiology requires a different approach, as multiple studies are often needed to address the diverse questions surrounding the origin, spread, and control of pathogens. Complicating this further is the fact that we lack a robust theoretical foundation to apply traditional causal inference methods, which are commonly used in other areas of epidemiology. The challenge of establishing causality in infectious disease epidemiology is compounded by the ethical constraints around conducting controlled experiments. While vaccine efficacy studies offer a controlled setting through randomized trials, they are the exception rather than the rule, and even these studies face limitations. For example, they often focus on direct protection rather than exploring broader population-level effects like herd immunity. Beyond vaccines, other non-pharmaceutical interventions such as social distancing for infected individuals are typically evaluated through observational studies or natural experiments, making it difficult to isolate causal relationships. These ethical and logistical constraints necessitate innovative approaches to infer causality without relying on traditional epidemiological designs. This makes it necessary to develop new methods that can accurately identify causal pathways and quantify the causal effects of different transmission drivers.

Two key areas of future research stand out as essential for advancing causal inference in infectious disease epidemiology. The first is addressing the assumption of independence, which is often violated in the context of infectious diseases due to the interconnected nature of disease transmission. In infectious diseases, an individual's risk of infection is not independent of others in their community or environment; instead, it is heavily influenced by interactions with infected individuals, making the transmission process highly dependent on social networks, proximity, and

behaviour. This interconnectedness means that changes in one individual’s infection status can directly alter the risk for others. As a result of this interdependence, spillover effects where the treatment or infection of one individual affects others are common in infectious disease systems. A clear example of this is herd immunity, where the treatment or vaccination of a portion of the population reduces the overall risk of infection for those who remain untreated, illustrating how interconnectedness drives population-level impacts from individual interventions.

The second key area for future research lies in addressing the complexity introduced by feedback loops in infectious disease dynamics. Observational data already pose difficulties for causal inference, and this complexity increases when exposures change over time. Moreover, the feedback loops inherent in infectious disease dynamics further complicate causal analysis, especially given the requirement that the *cause must precede its effect*. Mobility, as discussed in this thesis, plays a central role in pathogen spread, but its influence is bidirectional, highlighting the need for future research to better understand the reciprocal effects in transmission dynamics. I explore this in more detail in the section 6.2.3 below.

Together, these two areas—violations of independence and feedback loops—are likely to drive significant advancements in causal inference within infectious disease epidemiology. Mechanistic models provide a valuable approach to address these challenges, as they are inherently causal in nature. Agent-based models (ABMs), for instance, simulate how individual behaviors contribute to population-level trends, offering insights into the causal pathways of transmission. For example, one of the earliest ABMs demonstrated how individual preferences for same-race neighbors led to racial segregation, showing the power of these models in capturing complex social dynamics [161]. Applied to infectious diseases, ABMs can empirically examine the feedback loops between individual behaviors and pathogen spread, making them particularly useful for studying the dynamics of transmission in populations. By integrating individual actions with population-level outcomes, these models provide a powerful tool for exploring the intricate relationships driving infectious disease transmission.

Lastly, interdisciplinary collaboration with social scientists and other experts will be crucial in addressing the “why” of disease dynamics, and integrating social, behavioral, and environmental factors into our models. This interdisciplinary approach can ultimately lead to more effective, evidence-based interventions and policies, tailored to the complex realities of disease transmission in different populations and contexts.

6.2.2 Sociodemographic factors in epidemiological models

Modelling

While the COVID-19 pandemic saw a boost in innovative modelling solutions, various authors have also called attention to integrating social aspects into infectious disease modelling as a necessary step in our field of research [133, 134, 72]. There have been some studies during the pandemic that have incorporated socio-economic factors to calibrate the more commonly used age- and spatially-structured mathematical models.

One study showed that an eviction crisis stemming from job losses in Philadelphia could lead to a surge in COVID-19 cases, particularly affecting individuals in living in poorer neighborhoods with higher proportions of racial/ethnic minorities [162]. This research underscores the interplay between economic instability and health outcomes, suggesting that public health interventions must consider socio-economic vulnerability of individuals.

Another study analysed fine-scale mobility data from various locations such as restaurants, grocery stores, and religious establishments across ten metropolitan cities in the United States. By combining this mobility data with census block groups, the authors found that areas frequently visited by disadvantaged racial and socioeconomic groups experienced higher transmission rates of SARS-CoV-2 [163]. They also modelled different reopening strategies to minimize the infection burden in these vulnerable populations, highlighting the importance of tailored public health measures that account for disparities in the population.

Another study incorporated heterogeneity in susceptibility and exposure to SARS-CoV-2 infection among different racial and ethnic groups in New York [76]. By explicitly integrating socioeconomic status into their mathematical framework, the researchers revealed that disparities among these groups led to unequal distributions of infections. This finding emphasised that a nuanced understanding of social determinants is essential for accurately modelling disease dynamics, estimating epidemiological parameters by sub-groups, and informing effective public health strategies.

Collectively, these studies illustrate that integrating socioeconomic factors into infectious disease modeling framework can provide a more comprehensive understanding of transmission dynamics and can inform targeted interventions that address health and economic disparities.

Filling the data gap

Despite a few studies during the COVID-19 pandemic, incorporating sociodemographic factors into the framework of mechanistic mathematical models remains largely under-explored and presents a promising direction for future research. However, to make meaningful progress in this area, we must also address the critical issue of data representation. Most of the work in our field of research relies heavily on data, and accurately capturing and accounting for biases within it is crucial. Ensuring that our data are representative is vital, particularly in resource-limited settings where data gaps are most pronounced. As artificial intelligence and computational methods continue to advance, particularly with the integration of multiple data sources, it becomes increasingly important to be mindful of the underlying biases in the data used to inform public health decisions [164].

In chapter 3 and 4, I presented methods to debias prevalence estimates by combining high-quality, low-volume surveys with mass testing data. This approach allows us to leverage the strengths of both data sources, providing more accurate epidemiological estimates even at fine spatial and sociodemographic scales. Extending this work could involve designing cost-effective surveys that maintain representativeness while being feasible to implement in settings with limited resources.

6.2.3 Modelling human behaviour

Building on the findings in Chapters 2 and 5, which highlight the role of human behavior (measured through mobility) in the spatial expansion of VOCs and spatial synchrony patterns of transmission, future research should delve deeper into understanding the *why* behind these behaviors. Specifically, it is crucial to explore why some individuals can modify their behaviors, how human behavior changes in response to perceived risk, and how these changes, in turn, affect disease dynamics [165, 166].

Future work could focus on modeling the dynamic feedback loop between how individuals react to an epidemic and how these reactions influence the spread of the disease. This would involve accounting for differences in individuals' ability to modify their behaviors, which is a complex function of several factors, particularly in the context of COVID-19, as outlined below:

- Perception of risk – influenced by media, government announcements, policies, and the risk perception within an individual's contact network;

- Ability to modify contact patterns – such as implementing stringent measures like working from home and isolation;
- Willingness to comply with preventative measures – influenced by beliefs about the benefits of mask-wearing, vaccine uptake hesitancy due to perceived adverse events, and rational opposition as vaccine uptake in the community increases and perceived risk decreases [167];
- Adherence fatigue – over time, compliance with preventative measures may wane, affecting the overall effectiveness of interventions.

By modeling these factors, we can gain insights into how heterogeneity in micro-level dynamics (individual behaviors) impacts macro-level disease dynamics. A key focus should be on understanding how much of this impact is mediated through individuals' perception of disease risk versus their (in)ability to modify behavior. The former could be targeted by effective communication [168] from authorities, while the latter points to deeper structural inequalities in society [169]. This research could guide the design of more equitable and effective public health interventions.

Bibliography

- [1] Moritz U G Kraemer, Simon I Hay, David M Pigott, David L Smith, G R William Wint, and Nick Golding. Progress and Challenges in Infectious Disease Cartography. *Trends in Parasitology*, 32(1):19–29, 2016.
- [2] Lisa F. Berkman, Ichiro Kawachi, and M. Maria Glymour. *Social Epidemiology*. Oxford University Press, 2014.
- [3] Mohsen Naghavi, Tomislav Mestrovic, Authia Gray, Anna Gershberg Hayoon, Lucien R Swetschinski, Gisela Robles Aguilar, Nicole Davis Weaver, Kevin S Ikuta, Erin Chung, Eve E Wool, Chieh Han, Daniel T Araki, Samuel B Albertson, Rose Bender, Greg Bertolacci, Annie J Browne, Ben S Cooper, Matthew W Cunningham, Christiane Dolecek, Matthew Doxey, Susanna J Dunachie, Sama Ghoba, Georgina Haines-Woodhouse, Simon I Hay, Rebecca L Hsu, Kenneth C Iregbu, Hmwe H Kyu, Jorge R Ledesma, Jianing Ma, Catrin E Moore, Jonathan F Mosser, Vincent Mougin, Pirouz Naghavi, Amanda Novotney, Victor Daniel Rosenthal, Benn Sartorius, Andy Stergachis, Christopher Troeger, Avina Vongpradith, Magdalene K Walters, Han Yong Wunrow, and Christopher J L Murray. Global burden associated with 85 pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Infectious Diseases*, 24(8):868–895, 2024.
- [4] Nuno R Faria, Andrew Rambaut, Marc A Suchard, Guy Baele, Trevor Bedford, Melissa J Ward, Andrew J Tatem, João D Sousa, Nimalan Arinaminpathy, Jacques Pépin, David Posada, Martine Peeters, Oliver G Pybus, and Philippe Lemey. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61, 2014.
- [5] UNAIDS. Fact sheet 2024. Global HIV statistics, 2024. Available at: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf.

- [6] R Brosch, S V Gordon, M Marmiesse, P Brodin, C Buchrieser, K Eiglmeier, T Garnier, C Gutierrez, G Hewinson, K Kremer, L M Parsons, A S Pym, S Samper, D van Soolingen, and S T Cole. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences*, 99(6):3684–3689, 2002.
- [7] World Health Organisation. Global tuberculosis report. Technical report, 2023. Available at: <https://iris.who.int/bitstream/handle/10665/373828/9789240083851-eng.pdf?sequence=1>.
- [8] Jesse B Bump and Ifeyinwa Aniebo. Colonialism, malaria, and the decolonization of global health. *PLOS Global Public Health*, 2(9):e0000936, 2022.
- [9] Francis E G Cox. History of the discovery of the malaria parasites and their vectors. *Parasites & Vectors*, 3(1):5, 2010.
- [10] Priya Venkatesan. The 2023 WHO World malaria report [News article]. *The Lancet Microbe*, 5(3):e214, 2024.
- [11] World health Organisation. State of inequality: HIV, tuberculosis and malaria. 2021. Available at: <https://www.who.int/publications/i/item/9789240039445>.
- [12] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rongmeng Jiang, Zhancheng Gao, Qi Jin, Jianwei Wang, and Bin Cao. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506, 2020.
- [13] World Health Organisation. COVID-19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum, 2020. Available at: [https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-\(pheic\)-global-research-and-innovation-forum](https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum).
- [14] Centers for Disease Control and Prevention. CDC Museum COVID-19 Timeline. Available at: <https://www.cdc.gov/museum/timeline/covid19.html#>.

- [15] World Health Organisation. Rolling updates on coronavirus disease (COVID-19), 2020. Available at: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen>.
- [16] Louis du Plessis, John T McCrone, Alexander E Zarebski, Verity Hill, Christopher Ruis, Bernardo Gutierrez, Jayna Raghvani, Jordan Ashworth, Rachel Colquhoun, Thomas R Connor, Nuno R Faria, Ben Jackson, Nicholas J Loman, Áine O’Toole, Samuel M Nicholls, Kris V Parag, Emily Scher, Tetyana I Vasylyeva, Erik M Volz, Alexander Watts, Isaac I Bogoch, Kamran Khan, David M Aanensen, Moritz U G Kraemer, Andrew Rambaut, and Oliver G Pybus. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, 371(6530):708 LP – 712, 2021.
- [17] World Health Organisation. Updated working definitions and primary actions for SARS-CoV-2 variants, 2023. Available at: <https://www.who.int/publications/m/item/updated-working-definitions-and-primary-actions-for--sars-cov-2-variants>.
- [18] Moritz U G Kraemer, Verity Hill, Christopher Ruis, Simon Dellicour, Sumali Bajaj, John T McCrone, Guy Baele, Kris V Parag, Anya Lindström Battle, Bernardo Gutierrez, Ben Jackson, Rachel Colquhoun, Áine O’Toole, Brennan Klein, Alessandro Vespignani, null null, Erik Volz, Nuno R Faria, David M Aanensen, Nicholas J Loman, Louis du Plessis, Simon Cauchemez, Andrew Rambaut, Samuel V Scarpino, and Oliver G Pybus. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science*, 373(6557):889–895, 2021.
- [19] John T McCrone, Verity Hill, Sumali Bajaj, Rosario Evans Pena, Ben C Lambert, Rhys Inward, Samir Bhatt, Erik Volz, Christopher Ruis, Simon Dellicour, Guy Baele, Alexander E Zarebski, Adam Sadilek, Neo Wu, Aaron Schneider, Xiang Ji, Jayna Raghvani, Ben Jackson, Rachel Colquhoun, Áine O’Toole, Thomas P Peacock, Kate Twohig, Simon Thelwall, Gavin Dabrera, Richard Myers, Nuno R Faria, Carmen Huber, Isaac I Bogoch, Kamran Khan, Louis du Plessis, Jeffrey C Barrett, David M Aanensen, Wendy S Barclay, Meera Chand, Thomas Connor, Nicholas J Loman, Marc A Suchard, Oliver G Pybus, Andrew Rambaut, Moritz U G Kraemer, and The COVID-19 genomics U K (COG-UK) consortium. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. *Nature*, 610:154–160, 2022.

- [20] Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, 19(3):141–154, 2021.
- [21] Hannah E Davis, Lisa McCorkell, Julia Moore Vogel, and Eric J Topol. Long COVID: major findings, mechanisms and recommendations. *Nature Reviews Microbiology*, 21(3):133–146, 2023.
- [22] Our World in Data. Coronavirus (COVID-19) Vaccinations, 2024. Available at: <https://ourworldindata.org/covid-vaccinations>.
- [23] Nicolò Gozzi, Matteo Chinazzi, Natalie E Dean, Ira M Longini Jr, M Elizabeth Halloran, Nicola Perra, and Alessandro Vespignani. Estimating the impact of COVID-19 vaccine inequities: a modeling study. *Nature Communications*, 14(1):3272, 2023.
- [24] T V Padma. COVID vaccines to reach poorest countries in 2023 — despite recent pledges [News article]. 2021. Available at: <https://www.nature.com/articles/d41586-021-01762-w>.
- [25] World Health Organisation. COVAX, 2023. Available at: <https://www.who.int/initiatives/act-accelerator/covax>.
- [26] Krause Philip R, Fleming Thomas R, Longini Ira M, Peto Richard, Briand Sylvie, Heymann David L, Beral Valerie, Snape Matthew D, Rees Helen, Ropero Alba-Maria, Balicer Ran D, Cramer Jakob P, Muñoz-Fontela César, Gruber Marion, Gaspar Rogério, Singh Jerome A, Subbarao Kanta, Van Kerkhove Maria D, Swaminathan Soumya, Ryan Michael J, and Henao-Restrepo Ana-Maria. SARS-CoV-2 Variants and Vaccines. *New England Journal of Medicine*, 385(2):179–186, 2021.
- [27] Blavatnik School of Government. COVID-19 Government Response Tracker, 2023. Available at: <https://www.bsg.ox.ac.uk/research/covid-19-government-response-tracker>.
- [28] Gonzalo E Mena, Pamela P Martinez, Ayesha S Mahmud, Pablo A Marquet, Caroline O Buckee, and Mauricio Santillana. Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *Science*, 372(6545):eabg5298, 2021.

- [29] Joakim A Weill, Matthieu Stigler, Olivier Deschenes, and Michael R Springborn. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proceedings of the National Academy of Sciences*, 117(33):19658–19660, 2020.
- [30] Giulia Pullano, Eugenio Valdano, Nicola Scarpa, Stefania Rubrichi, and Vittoria Colizza. Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study. *The Lancet Digital Health*, 2(12):e638–e649, 2020.
- [31] Department of Health & Social Care. Coronavirus (COVID-19) Scaling up our testing programmes Department of Health and Social Care [Policy paper]. 2020. Available at: <https://www.gov.uk/government/publications/coronavirus-covid-19-scaling-up-testing-programmes/coronavirus-covid-19-scaling-up-our-testing-programmes>.
- [32] Department of Health and Social Care. Government launches NHS Test and Trace service [Press release], 2020. Available at: <https://www.gov.uk/government/news/government-launches-nhs-test-and-trace-service>.
- [33] Office of National Statistics. Deaths involving COVID-19 by local area and socioeconomic deprivation Statistical bulletins [Technical report], 2020. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19bylocalareasanddeprivation/previousReleases>.
- [34] Office of National Statistics. Updating ethnic and religious contrasts in deaths involving the coronavirus (COVID-19), England: 24 January 2020 to 23 November 2022. 2023. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/updatingethniccontrastsindeathsinvolvingthecoronaviruscovid19englandandwales/24january2020to23november2022>.
- [35] Public Health England. COVID-19: review of disparities in risks and outcomes [Research and analysis], 2020. Available at: <https://www.gov.uk/government/publications/covid-19-review-of-disparities-in-risks-and-outcomes>.

- [36] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.
- [37] M.E.J. Woolhouse, C Dye, J.-F. Etard, T Smith, J.D. Charlwood, G.P. Garnett, P Hagan, J.L.K. Hii, P.D. Ndhlovu, R.J. Quinnell, C.H. Watts, S.K. Chandiwana, and R.M. Anderson. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences*, 94(1):338–342, 1997.
- [38] Kimberly L VanderWaal and Vanessa O Ezenwa. Heterogeneity in pathogen transmission: mechanisms and methodology. *Functional Ecology*, 30(10):1606–1622, 2016.
- [39] Stephen J Chapman and Adrian V S Hill. Human genetic susceptibility to infectious disease. *Nature Reviews Genetics*, 13(3):175–188, 2012.
- [40] A C Allison. Protection afforded by sickle-cell trait against subtertian malareal infection. *BMJ*, 1(4857):290–294, 1954.
- [41] Dami A Collier, Isabella A T M Ferreira, Prasanti Kotagiri, Rawlings P Datir, Eleanor Y Lim, Emma Touizer, Bo Meng, Adam Abdullahi, Stephen Baker, Gordon Dougan, Christoph Hess, Nathalie Kingston, Paul J Lehner, Paul A Lyons, Nicholas J Matheeson, Willem H Owehand, Caroline Saunders, Charlotte Summers, James E D Thaventhiran, Mark Toshner, Michael P Weekes, Patrick Maxwell, Ashley Shaw, Ashlea Bucke, Jo Calder, Laura Canna, Jason Domingo, Anne Elmer, Stewart Fuller, Julie Harris, Sarah Hewitt, Jane Kennet, Sherly Jose, Jenny Kourampa, Anne Meadows, Criona O’Brien, Jane Price, Cherry Publico, Rebecca Rastall, Carla Ribeiro, Jane Rowlands, Valentina Ruffolo, Hugo Tordesillas, Ben Bullman, Benjamin J Dunmore, Stuart Fawke, Stefan Gräf, Josh Hodgson, Christopher Huang, Kelvin Hunter, Emma Jones, Ekaterina Legchenko, Cecilia Matara, Jennifer Martin, Federica Mescia, Ciara O’Donnell, Linda Pointon, Nicole Pond, Joy Shih, Rachel Sutcliffe, Tobias Tilly, Carmen Treacy, Zhen Tong, Jennifer Wood, Marta Wylot, Laura Bergamaschi, Ariana Betancourt, Georgie Bower, Chiara Cossetti, Aloka De Sa, Madeline Epping, Stuart Fawke, Nick Gleadall, Richard Grenfell, Andrew Hinch, Oisín Huhn, Sarah Jackson, Isobel Jarvis, Ben Krishna, Daniel Lewis, Joe Marsden, Francesca Nice, Georgina Okecha, Ommar Omarjee, Marianne Perera, Martin Potts, Nathan Richoz, Veronika Romashova, Natalia Savinykh Yarkoni, Rahul Sharma, Luca Stefanucci, Jonathan Stephens, Mateusz Strezlecki, Lori Turner, Eckart M D. D. De Bie,

Katherine Bunclark, Masa Josipovic, Michael Mackay, Alice Michael, Sabrina Rossi, Mayurun Selvan, Sarah Spencer, Cissy Yong, Ali Ansaripour, Alice Michael, Lucy Mwaura, Caroline Patterson, Gary Polwarth, Petra Polgarova, Giovanni di Stefano, Codie Fahey, Rachel Michel, Sze-How Bong, Jerome D Coudert, Elaine Holmes, John Allison, Helen Butcher, Daniela Caputo, Debbie Clapham-Riley, Eleanor Dewhurst, Anita Furlong, Barbara Graves, Jennifer Gray, Tasmin Ivers, Mary Kasanicki, Emma Le Gresley, Rachel Linger, Sarah Meloy, Francesca Muldoon, Nigel Ovington, Sofia Papadia, Isabel Phelan, Hannah Stark, Kathleen E Stirrups, Paul Townsend, Neil Walker, Jennifer Webster, Anne Elmer, Nathalie Kingston, Barbara Graves, Emma Le Gresley, Daniela Caputo, Laura Bergamaschi, Kenneth G C Smith, John R Bradley, Lourdes Ceron-Gutierrez, Paulina Cortes-Acevedo, Gabriela Barcenas-Morales, Michelle A Linterman, Laura E McCoy, Chris Davis, Emma Thomson, Paul A Lyons, Eoin McKinney, Rainer Doffinger, Mark Wills, Ravindra K Gupta, The CITIID-NIHR BioResource COVID-19 Collaboration, Principal Investigators, C R F Nurses, Volunteer Research, Sample Logistics, Sample Processing Acquisition, Data, Clinical Data Collection, Royal Papworth Hospital ICU, Addenbrooke's Hospital ICU, Cambridge Trust, Peterborough Foundation, ANPC Therapeutics, Centre for Molecular Medicine, Innovative, and NIHR BioResource4. Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. *Nature*, 596(7872):417–422, 2021.

- [42] G A Poland, I G Ovsyannikova, R M Jacobson, and D I Smith. Heterogeneity in Vaccine Immune Response: The Role of Immunogenetics and the Emerging Field of Vaccinomics. *Clinical Pharmacology & Therapeutics*, 82(6):653–664, 2007.
- [43] Roy M Anderson and Robert M May. Vaccination and herd immunity to infectious diseases. *Nature*, 318(6044):323–329, 1985.
- [44] Takis Panagiotopoulos, Abi Berger, Ioanna Antoniadou, and Eleni Valassi-Adam. Increase in congenital rubella occurrence after immunisation in Greece: retrospective survey and systematic review. *BMJ*, 319(7223):1462 LP – 1467, 1999.
- [45] C J E Metcalf, C Cohen, J Lessler, J M McAnerney, G M Ntshoe, A Puren, P Klepac, A Tatem, B T Grenfell, and O N Bjørnstad. Implications of spatially heterogeneous

- vaccination coverage for the risk of congenital rubella syndrome in South Africa. *Journal of The Royal Society Interface*, 10(78):20120756, 2013.
- [46] D J Earn, P Rohani, and B T Grenfell. Persistence, chaos and synchrony in ecology and epidemiology. *Proceedings. Biological sciences*, 265(1390):7–10, 1998.
- [47] J O Lloyd-Smith, S J Schreiber, P E Kopp, and W M Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005.
- [48] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W John Edmunds. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLOS Medicine*, 5(3):e74, 2008.
- [49] Colin J Worby, Sandra S Chaves, Jacco Wallinga, Marc Lipsitch, Lyn Finelli, and Edward Goldstein. On the relative role of different age groups in influenza epidemics. *Epidemics*, 13:10–16, 2015.
- [50] Lulla Opatowski, Christophe Fraser, Jamie Griffin, Eric de Silva, Maria D Van Kerkhove, Emily J Lyons, Simon Cauchemez, and Neil M Ferguson. Transmission Characteristics of the 2009 H1N1 Influenza Pandemic: Comparison of 8 Southern Hemisphere Countries. *PLOS Pathogens*, 7(9):e1002225, 2011.
- [51] T HOUSE and M J KEELING. Household structure and infectious disease transmission. *Epidemiology and Infection*, 137(5):654–661, 2009.
- [52] Benjamin Bolker and Bryan Grenfell. Space, Persistence and Dynamics of Measles Epidemics. *Philosophical Transactions: Biological Sciences*, 348(1325):309–320, 1995.
- [53] T J Hagenaars, C A Donnelly, and N M Ferguson. Spatial heterogeneity and the persistence of infectious diseases. *Journal of Theoretical Biology*, 229(3):349–359, 2004.
- [54] Moritz U G Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, null null, Louis du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, John S Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G Pybus, and Samuel V Scarpino. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490):493–497, 2020.

- [55] Joseph L.-H. Tsui, John T McCrone, Ben Lambert, Sumali Bajaj, Rhys P D Inward, Paolo Bosetti, Rosario Evans Pena, Houriiyah Tegally, Verity Hill, Alexander E Zarebski, Thomas P Peacock, Luyang Liu, Neo Wu, Megan Davis, Isaac I Bogoch, Kamran Khan, Meaghan Kall, Nurin Iwani Binti Abdul Aziz, Rachel Colquhoun, Áine O’Toole, Ben Jackson, Abhishek Dasgupta, Eduan Wilkinson, Tulio de Oliveira, The COVID-19 Genomics U K (COG-UK) consortium[¶], Thomas R Connor, Nicholas J Loman, Vittoria Colizza, Christophe Fraser, Erik Volz, Xiang Ji, Bernardo Gutierrez, Meera Chand, Simon Dellicour, Simon Cauchemez, Jayna Raghvani, Marc A Suchard, Philippe Lemey, Andrew Rambaut, Oliver G Pybus, and Moritz U G Kraemer. Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science*, 381(6655):336–343, 2023.
- [56] Henrik Salje, Justin Lessler, Kishor Kumar Paul, Andrew S Azman, M Waliur Rahman, Mahmudur Rahman, Derek Cummings, Emily S Gurley, and Simon Cauchemez. How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. *Proceedings of the National Academy of Sciences*, 113(47):13420–13425, 2016.
- [57] R M Anderson, G F Medley, R M May, and A M Johnson. A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA journal of mathematics applied in medicine and biology*, 3(4):229–263, 1986.
- [58] Simon Cauchemez, Achuyt Bhattarai, Tiffany L Marchbanks, Ryan P Fagan, Stephen Ostroff, Neil M Ferguson, and David Sverdlow. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108(7):2825–2830, 2011.
- [59] Yuguo Li, Ignatius T S Yu, Pengcheng Xu, J H W Lee, Tze Wai Wong, Peng Lim Ooi, and Adrian C Sleight. Predicting Super Spreading Events during the 2003 Severe Acute Respiratory Syndrome Epidemics in Hong Kong and Singapore. *American Journal of Epidemiology*, 160(8):719–728, 2004.
- [60] Dillon C Adam, Peng Wu, Jessica Y Wong, Eric H Y Lau, Tim K Tsang, Simon Cauchemez, Gabriel M Leung, and Benjamin J Cowling. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Medicine*, 26(11):1714–1719, 2020.

- [61] Tobias M Holden, Reese A K Richardson, Philip Arevalo, Wayne A Duffus, Manuela Runge, Elena Whitney, Leslie Wise, Ngozi O Ezike, Sarah Patrick, Sarah Cobey, and Jaline Gerardin. Geographic and demographic heterogeneity of SARS-CoV-2 diagnostic testing in Illinois, USA, March to December 2020. *BMC Public Health*, 21(1):1105, 2021.
- [62] Calistus N Ngonghala, Mateusz M Pluciński, Megan B Murray, Paul E Farmer, Christopher B Barrett, Donald C Keenan, and Matthew H Bonds. Poverty, Disease, and the Ecology of Complex Systems. *PLOS Biology*, 12(4):e1001827, 2014.
- [63] Ronan F Arthur, Emily S Gurley, Henrik Salje, Laura S P Bloomfield, and James H Jones. Contact structure, mobility, environmental impact and behaviour: the importance of social forces to infectious disease dynamics and disease ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1719):20160454, 2017.
- [64] Kris V. Parag, Robin N. Thompson, and Christl A. Donnelly. Are Epidemic Growth Rates More Informative than Reproduction Numbers? *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement 1):S5–S15, 2022.
- [65] George Nicholson, Briec Lehmann, Tullia Padellini, Koen B Pouwels, Radka Jersakova, James Lomax, Ruairidh E King, Ann-Marie Mallon, Peter J Diggle, Sylvia Richardson, Marta Blangiardo, and Chris Holmes. Improving local prevalence estimates of SARS-CoV-2 infections using a causal debiasing framework. *Nature Microbiology*, 7(1):97–107, 2022.
- [66] Sam Abbott and Sebastian Funk. Estimating epidemiological quantities from repeated cross-sectional prevalence measurements. *medRxiv*, page 2022.03.29.22273101, 2022.
- [67] Sam Abbott and Sebastian Funk. Real-time epidemiological estimates from ONS Community Infection Survey data, 2023. Available at: <https://epiforecasts.io/inc2prev/report>.
- [68] Office of National Statistics. Coronavirus (COVID-19) Infection Survey: methods and further information, 2023. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveyypilotmethodsandfurtherinformation#:~:text=Current%20method%20for%20calculating%20incidence&text=We%20estimate%20the%20time%20between,for%20different%20times%20between%20visits>.

- [69] Duncan J Watts, Roby Muhamad, Daniel C Medina, and Peter S Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences*, 102(32):11157–11162, 2005.
- [70] Laura Alessandretti, Ulf Aslak, and Sune Lehmann. The scales of human mobility. *Nature*, 587(7834):402–407, 2020.
- [71] Bryan Grenfell and John Harwood. (Meta)population dynamics of infectious diseases. *Trends in Ecology & Evolution*, 12(10):395–399, 1997.
- [72] Jon Zelner, Nina B Masters, Ramya Naraharisetti, Sanyu A Mojola, Merlin Chowkwanyun, and Ryan Malosh. There are no equal opportunity infectors: Epidemiological modelers must rethink our approach to inequality in infection risk. *PLOS Computational Biology*, 18(2):e1009795, 2022.
- [73] Yelena Rozenfeld, Jennifer Beam, Haley Maier, Whitney Haggerson, Karen Boudreau, Jamie Carlson, and Rhonda Medows. A model of disparities: risk factors associated with COVID-19 infection. *International Journal for Equity in Health*, 19(1):126, 2020.
- [74] Jonathan Jay, Jacob Bor, Elaine O Nsoesie, Sarah K Lipson, David K Jones, Sandro Galea, and Julia Raifman. Neighbourhood income and physical distancing during the COVID-19 pandemic in the United States. *Nature Human Behaviour*, 4(12):1294–1302, 2020.
- [75] Marco Dueñas, Mercedes Campi, and Luis E Olmos. Changes in mobility and socioeconomic conditions during the COVID-19 outbreak. *Humanities and Social Sciences Communications*, 8(1):101, 2021.
- [76] Kevin C Ma, Tigist F Menkir, Stephen Kissler, Yonatan H Grad, and Marc Lipsitch. Modeling the impact of racial and ethnic disparities on COVID-19 epidemic dynamics. *eLife*, 10:e66601, 2021.
- [77] EY-Oxford Health Analytics Consortium. Evaluation of the national COVID-19 testing programme in England between October 2020 and March 2022 [Technical report], 2023. Available at: https://assets.ey.com/content/dam/ey-sites/ey-com/en_uk/resources/evaluation-of-covid-19-testing-in-england-eoha-conference-edition.pdf.

- [78] Christopher Dye. *The Population Biology of Tuberculosis*. Princeton University Press, 2015.
- [79] Sunetra Gupta. Evolution of pathogen virulence. *EMBO reports*, 24(8):e57611, 2023.
- [80] Erik Volz, Swapnil Mishra, Meera Chand, Jeffrey C Barrett, Robert Johnson, Lily Geidelberg, Wes R Hinsley, Daniel J Laydon, Gavin Dabrera, Áine O’Toole, Robert Amato, Manon Ragonnet-Cronin, Ian Harrison, Ben Jackson, Cristina V Ariani, Olivia Boyd, Nicholas J Loman, John T McCrone, Sónia Gonçalves, David Jorgensen, Richard Myers, Verity Hill, David K Jackson, Katy Gaythorpe, Natalie Groves, John Sillitoe, Dominic P Kwiatkowski, Cherian Koshy, Amy Ash, Emma Wise, Nathan Moore, Matilde Mori, Nick Cortes, Jessica Lynch, Stephen Kidd, Derek J Fairley, Tanya Curran, James P McKenna, Helen Adams, Christophe Fraser, Tanya Golubchik, David Bonsall, Mohammed O Hassan-Ibrahim, Cassandra S Malone, Benjamin J Cogger, Michelle Wantoch, Nicola Reynolds, Ben Warne, Joshua Maksimovic, Karla Spellman, Kathryn McCluggage, Michaela John, Robert Beer, Safiah Afifi, Sian Morgan, Angela Marchbank, Anna Price, Christine Kitchen, Huw Gulliver, Ian Merrick, Joel Southgate, Martyn Guest, Robert Munn, Trudy Workman, Thomas R Connor, William Fuller, Catherine Bresner, Luke B Snell, Amita Patel, Themoula Charalampous, Gaia Nebbia, Rahul Batra, Jonathan Edgeworth, Samuel C Robson, Angela H Beckett, David M Aanensen, Anthony P Underwood, Corin A Yeats, Khalil Abudahab, Ben E W Taylor, Mirko Menegazzo, Gemma Clark, Wendy Smith, Manjinder Khakh, Vicki M Fleming, Michelle M Lister, Hannah C Howson-Wells, Louise Berry, Tim Boswell, Amelia Joseph, Iona Willingham, Carl Jones, Christopher Holmes, Paul Bird, Thomas Helmer, Karlie Fallon, Julian Tang, Veena Raviprakash, Sharon Campbell, Nicola Sheriff, Victoria Blakey, Lesley-Anne Williams, Matthew W Loose, Nadine Holmes, Christopher Moore, Matthew Carlile, Victoria Wright, Fei Sang, Johnny Debebe, Francesc Coll, Adrian W Signell, Gilberto Betancor, Harry D Wilson, Sahar Eldirdiri, Anita Kenyon, Thomas Davis, Oliver G Pybus, Louis du Plessis, Alex E Zarebski, Jayna Raghwani, Moritz U G Kraemer, Sarah Francois, Stephen W Attwood, Tetyana I Vasylyeva, Marina Escalera Zamudio, Bernardo Gutierrez, M Estee Torok, William L Hamilton, Ian G Goodfellow, Grant Hall, Aminu S Jahun, Yasmin Chaudhry, Myra Hosmillo, Malte L Pinckert, Iliana Georgana, Samuel Moses, Hannah Lowe, Luke Bedford, Jonathan Moore, Susanne Stonehouse, Chloe L Fisher, Ali R Awan, John BoYes, Judith Breuer, Kathryn Ann Harris, Ju-

lianne Rose Brown, Divya Shah, Laura Atkinson, Jack C D Lee, Nathaniel Storey, Flavia Flaviani, Adela Alcolea-Medina, Rebecca Williams, Gabrielle Vernet, Michael R Chapman, Lisa J Levett, Judith Heaney, Wendy Chatterton, Monika Pusok, Li Xu-McCrae, Darren L Smith, Matthew Bashton, Gregory R Young, Alison Holmes, Paul Anthony Randell, Alison Cox, Pinglawathee Madona, Frances Bolt, James Price, Siddharth Mookerjee, Manon Ragonnet-Cronin, Fabricia F Nascimento, David Jorgensen, Igor Siveroni, Rob Johnson, Olivia Boyd, Lily Geidelberg, Erik M Volz, Aileen Rowan, Graham P Taylor, Katherine L Smollett, Nicholas J Loman, Joshua Quick, Claire McMurray, Joanne Stockton, Sam Nicholls, Will Rowe, Radoslaw Poplawski, Alan McNally, Rocio T Martinez Nunez, Jenifer Mason, Trevor I Robinson, Elaine O'Toole, Joanne Watts, Cassie Breen, Angela Cowell, Graciela Sluga, Nicholas W Machin, Shazaad S Y Ahmad, Ryan P George, Fenella Halstead, Venkat Sivaprakasam, Wendy Hogsden, Chris J Illingworth, Chris Jackson, Emma C Thomson, James G Shepherd, Patawee Asamaphan, Marc O Niebel, Kathy K Li, Rajiv N Shah, Natasha G Jesudason, Lily Tong, Alice Broos, Daniel Mair, Jenna Nichols, Stephen N Carmichael, Kyriaki Nomikou, Elihu Aranday-Cortes, Natasha Johnson, Igor Starinskij, Ana da Silva Filipe, David L Robertson, Richard J Orton, Joseph Hughes, Sreenu Vattipally, Joshua B Singer, Seema Nickbakhsh, Antony D Hale, Louissa R Macfarlane-Smith, Katherine L Harper, Holli Carden, Yusri Taha, Brendan A I Payne, Shirelle Burton-Fanning, Sheila Waugh, Jennifer Collins, Gary Eltringham, Steven Rushton, Sarah O'Brien, Amanda Bradley, Alasdair Maclean, Guy Mollett, Rachel Blacow, Kate E Templeton, Martin P McHugh, Rebecca Dewar, Elizabeth Wastenge, Samir Dervisevic, Rachael Stanley, Emma J Meader, Lindsay Coupland, Louise Smith, Clive Graham, Edward Barton, Debra Padgett, Garren Scott, Emma Swindells, Jane Greenaway, Andrew Nelson, Clare M McCann, Wen C Yew, Monique Andersson, Timothy Peto, Anita Justice, David Eyre, Derrick Crook, Tim J Sloan, Nichola Duckworth, Sarah Walsh, Anoop J Chauhan, Sharon Glaysher, Kelly Bicknell, Sarah Wyllie, Scott Elliott, Allyson Lloyd, Robert Impey, Nick Levene, Lynn Monaghan, Declan T Bradley, Tim Wyatt, Elias Allara, Clare Pearson, Husam Osman, Andrew Bosworth, Esther Robinson, Peter Muir, Ian B Vipond, Richard Hopes, Hannah M Pymont, Stephanie Hutchings, Martin D Curran, Surendra Parmar, Angie Lackenby, Tamyo Mbisa, Steven Platt, Shahjahan Miah, David Bibby, Carmen Manso, and The COVID-19 Genomics U K (COG-UK) consortium. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, 593(7858):266–269,

2021.

- [81] World Health Organisation. Tracking SARS-CoV-2 variants. Available at: <https://www.who.int/activities/tracking-SARS-CoV-2-variants>.
- [82] A O'Toole, V Hill, O G Pybus, A Watts, I I Bogoch, K Khan, J P Messina, COVID-19 Genomics UK (COG-UK) consortium, Network for Genomic Surveillance in South Africa (NGS-SA), Brazil-UK CADDE Genomic Network, H Tegally, R R Lessells, J Giandhari, S Pillay, K A Tumedi, G Nyepetsi, M Kebabonye, M Matsheka, M Mine, S Tokajian, H Hassan, T Salloum, G Merhi, J Koweyes, J L Geoghegan, J de Ligt, X Ren, M Storey, N E Freed, C Pattabiraman, P Prasad, A S Desai, R Vasanthapuram, T F Schulz, L Steinbrück, T Stadler, Swiss Viollier Sequencing Consortium, A Parisi, A Bianco, D de Viedma, S Buenestado-Serrano, V Borges, J Isidro, S Duarte, J P Gomes, N S Zuckerman, M Mandelboim, O Mor, T Seemann, A Arnott, J Draper, M Gall, W Rawlinson, I Deveson, S Schlebusch, J McMahon, L Leong, C K Lim, M Chironna, D Loconsole, A Bal, L Josset, E Holmes, K St. George, E Lasek-Nesselquist, R S Sikkema, B Oude Munnink, M Koopmans, M Brytting, V Sudha rani, S Pavani, T Smura, A Heim, S Kurkela, M Umair, M Salman, B Bartolini, M Rueca, C Drosten, T Wolff, O Silander, D Eggink, C Reusken, H Vennema, A Park, C Carrington, N Sahadeo, M Carr, G Gonzalez, SEARCH Alliance San Diego, National Virus Reference Laboratory, SeqCOVID-Spain, Danish Covid-19 Genome Consortium (DCGC), Communicable Diseases Genomic Network (CDGN), Dutch National SARS-CoV-2 surveillance program, Division of Emerging Infectious Diseases (KDCA), T de Oliveira, N Faria, A Rambaut, and M U G Kraemer. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. *Wellcome Open Research*, 6(121), 2021.
- [83] Nicholas G Davies, Christopher I Jarvis, Kevin van Zandvoort, Samuel Clifford, Fiona Yueqian Sun, Sebastian Funk, Graham Medley, Yalda Jafari, Sophie R Meakin, Rachel Lowe, Matthew Quaife, Naomi R Waterlow, Rosalind M Eggo, Jiayao Lei, Mihaly Koltai, Fabienne Krauer, Damien C Tully, James D Munday, Alicia Showering, Anna M Foss, Kiesha Prem, Stefan Flasche, Adam J Kucharski, Sam Abbott, Billy J Quilty, Thibaut Jombart, Alicia Rosello, Gwenan M Knight, Mark Jit, Yang Liu, Jack Williams, Joel Hellewell, Kathleen O'Reilly, Yung-Wai Desmond Chan, Timothy W Russell, Simon R Procter, Akira Endo, Emily S Nightingale, Nikos I Bosse, C Julian Villabona-Arenas, Frank G Sandmann,

- Amy Gimma, Kaja Abbas, William Waites, Katherine E Atkins, Rosanna C Barnard, Petra Klepac, Hamish P Gibbs, Carl A B Pearson, Oliver Brady, W John Edmunds, Nicholas P Jewell, Karla Diaz-Ordaz, Ruth H Keogh, and CMMID COVID-19 Working Group. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*, 593(7858):270–274, 2021.
- [84] Jennifer Brown, Esme Kirk-Wade, Carl Baker, and Sarah Barber. Coronavirus: A history of English lockdown laws, 2021. Available at: <https://researchbriefings.files.parliament.uk/documents/CBP-9068/CBP-9068.pdf>.
- [85] Gayathri Vaidyanathan. Coronavirus variants are spreading in India — what scientists know so far [News article]. *Nature*, 593:321–322, 2021.
- [86] Public Health England. Investigation of SARS-CoV-2 variants of concern: technical briefings [Research and analysis], 2020. Available at: <https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201>.
- [87] UK Cabinet Office. COVID-19 Response - Spring 2021 (Summary) [Guidance], 2021. Available at: <https://www.gov.uk/government/publications/covid-19-response-spring-2021/covid-19-response-spring-2021-summary>.
- [88] Office of National Statistics. Coronavirus (COVID-19) Infection Survey. Available at: <https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/covid19infectionsurvey>.
- [89] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.
- [90] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, v2.26.1, 2021. Available at: <https://mc-stan.org>.
- [91] Nicholas G Davies, Sam Abbott, Rosanna C Barnard, Christopher I Jarvis, Adam J Kucharski, James D Munday, Carl A B Pearson, Timothy W Russell, Damien C Tully, Alex D Washburne, Tom Wenseleers, Amy Gimma, William Waites, Kerry L M Wong, Kevin van Zandvoort, Justin D Silverman, CMMID COVID-19 Working Group^{1‡}, COVID-19 Genomics U K (COG-UK) Consortium[‡], Karla Diaz-Ordaz, Ruth Keogh, Rosalind M

- Eggo, Sebastian Funk, Mark Jit, Katherine E Atkins, and W John Edmunds. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372(6538):eabg3055, 2021.
- [92] Sang Woo Park, Benjamin M Bolker, Sebastian Funk, C Jessica E Metcalf, Joshua S Weitz, Bryan T Grenfell, and Jonathan Dushoff. The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants. *Journal of The Royal Society Interface*, 19(191):20220173, 2022.
- [93] William S Hart, Sam Abbott, Akira Endo, Joel Hellewell, Elizabeth Miller, Nick Andrews, Philip K Maini, Sebastian Funk, and Robin N Thompson. Inference of the SARS-CoV-2 generation time using UK household data. *eLife*, 11:e70767, 2022.
- [94] Sumali Bajaj, Siyu Chen, Richard Creswell, Reshania Naidoo, Joseph L.-H. Tsui, Olumide Kolade, George Nicholson, Briec Lehmann, James A Hay, Moritz U G Kraemer, Ricardo Aguas, Christl A Donnelly, Tom Fowler, Susan Hopkins, Liberty Cantrell, Prabin Dahal, Lisa J White, Kasia Stepniewska, Merryn Voysey, Ben Lambert, and the EY-Oxford Health Analytics Consortium. Understanding COVID-19 testing behaviour in England through a sociodemographic lens. *medRxiv*, page 2023.10.26.23297608, 2023.
- [95] Department of Health & Social Care. Technical report on the COVID-19 pandemic in the UK [Independent report]. Technical report, 2022. Available at: <https://www.gov.uk/government/publications/technical-report-on-the-covid-19-pandemic-in-the-uk/chapter-6-testing#timeline-of-testing>.
- [96] Paul Elliott, Matthew Whitaker, David Tang, Oliver Eales, Nicholas Steyn, Barbara Bodinier, Haowei Wang, Joshua Elliott, Christina Atchison, Deborah Ashby, Wendy Barclay, Graham Taylor, Ara Darzi, Graham S Cooke, Helen Ward, Christl A Donnelly, Steven Riley, and Marc Chadeau-Hyam. Design and Implementation of a National SARS-CoV-2 Monitoring Program in England: REACT-1 Study. *American Journal of Public Health*, 113(5):545–554, 2023.
- [97] Office of National Statistics. English indices of deprivation 2019. Technical report. Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019>.

- [98] Office of National Statistics. Population characteristics research tables, 2019. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationcharacteristicsresearchtables>.
- [99] Office of National Statistics. Research report on population estimates by characteristics, 2017. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/researchreportonpopulationestimatesbycharacteristics>.
- [100] Office of National Statistics. Estimates of the population for the UK, England, Wales, Scotland and Northern Ireland-Mid 2020 edition. 2021. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland>.
- [101] Andrew Rambaut, Edward C Holmes, Áine O’Toole, Verity Hill, John T McCrone, Christopher Ruis, Louis du Plessis, and Oliver G Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5(11):1403–1407, 2020.
- [102] Wellcome Sanger Institute. COVID–19 Genomic Surveillance, 2023. Available at: <https://covid19.sanger.ac.uk/downloads>.
- [103] Joel Hellewell, Timothy W Russell, Rebecca Matthews, Abigail Severn, Sajida Adam, Louise Enfield, Angela McBride, Kathleen Gärtner, Sarah Edwards, Fabiana Lorencatto, Susan Michie, Ed Manley, Maryam Shahmanesh, Hinal Lukha, Paulina Prymas, Hazel McBain, Robert Shortman, Leigh Wood, Claudia Davies, Bethany Williams, Emilie Sanchez, Daniel Frampton, Matthew Byott, Stavroula M Paraskevopoulou, Elise Crayton, Carly Meyer, Nina Vora, Triantafylia Gkouleli, Andrea Stoltenberg, Veronica Ranieri, Tom Byrne, Dan Lewer, Andrew Hayward, Richard Gilson, Naomi Walker, Aaron Ferron, Aaron Sait, Abhinay Ramaprasad, Abigail Perrin, Adam Sateriale, Adrienne E Sullivan, Aileen Nelson, Akshay Madoo, Alana Burrell, Aleksandra Pajak, Alessandra Gaiba, Alice Rossi, Alida Avola, Alison Dibbs, Alison Taylor-Beadling, Alize Proust, Almaz Huseynova, Amar Pabari, Amelia Edwards, Amy Strange, Ana Cardoso, Ana Agua-Doce, Ana Perez Caballero, Anabel Guedan, Anastacio King Spert Teixeira, Anastasia Moraiti, Andreas Wack, Andrew Riddell, Andrew Buckton, Andrew Levett, Andrew Rowan, Angela Rodgers,

Ania Kucharska, Anja Schlott, Annachiara Rosa, Annalisa D'Avola, Anne O'Garra, Anthony Gait, Antony Fearn, Beatriz Montaner, Belen Gomez Dominguez, Berta Terré Torras, Beth Hoskins, Bishara Marzook, Bobbi Clayton, Bruno Frederico, Caetano Reis e Sousa, Caitlin Barns-Jenkins, Carlos M Minutti, Caroline Oedekoven, Catharina Wenman, Catherine Lambe, Catherine Moore, Catherine F Houlihan, Charles Swanton, Chelsea Sawyer, Chloe Roustán, Chris Ekin, Christophe Queval, Christopher Earl, Claire Brooks, Claire Walder, Clare Beesley, Claudio Bussi, Clementina Cobolli Gigli, Clinda Puvirasinghe, Cristina Naceur-Lombardelli, Dan Fitz, Daniel M Snell, Dara Davison, David Moore, Davide Zecchin, Deborah Hughes, Deborah Jackson, Dhruva Biswas, Dimitrios Evangelopoulos, Dominique Bonnet, Edel C McNamara, Edina Schweighoffer, Effie Taylor, Efthymios Fidanis, Eleni Nastouli, Elizabeth Horton, Ellen Knuepfer, Emine Hatipoglu, Emma Russell, Emma Ashton, Enzo ZPoirier, Erik Sahai, Fatima Sardar, Faye Bowker, Fernanda Teixeira Subtil, Fiona McKay, Fiona Byrne, Fiona Hackett, Fiona Roberts, Francesca Torelli, Ganka Bineva-Todd, Gavin Kelly, Gee Yen Shin, Genevieve Barr, George Kassiotis, Georgina H Cornish, Gita Mistry, Graeme Hewitt, Graham Clark, Gunes Taylor, Hadija Trojer, Harriet B Taylor, Hector Huerga Encabo, Heledd Davies, Helen M Golding, Hon-Wing Liu, Hui Gong, Jacki Goldman, Jacqueline Hoyle, James Fleming, James I MacRae, James M Polke, James M A Turner, Jane Hughes, Jean D O'Leary, Jernej Ule, Jerome Nicod, Jessica Olsen, Jessica Diring, Jill A Saunders, Jim Aitken, Jimena Perez-Lloret, Joachim M Matz, Joanna Campbell, Joe Shaw, John Matthews, Johnathan Canton, Joshua Hope, Joshua Wright, Joyita Mukherjee, Judith Heaney, Julian AZagalak, Julie A Marczak, Karen Ambrose, Karen Vousden, Katarzyna Sala, Kayleigh Richardson, Kerol Bartolovic, Kevin W Ng, Konstantinos Kousis, Kylie Montgomery, Laura Churchward, Laura Cubitt, Laura Peces-Barba Castano, Laura Reed, Laura E McCoy, Lauren Wynne, Leigh Jones, Liam Gaul, Lisa Levett, Lotte Carr, Louisa Steel, Louise Busby, Louise Howitt, Louise Kiely, Lucia Moreira-Teixeira, Lucia Prieto-Godino, Lucy Jenkins, Luiz Carvalho, Luke Nightingale, Luke Williams, Lyn Healy, Magali S Perrault, Malgorzata Broncel, Marc Pollitt, Marcel Levi, Margaret Crawford, Margaux Silvestre, Maria Greco, Mariam Jamal-Hanjani, Mariana Grobler, Mariana Silva Dos Santos, Mark Johnson, Mary Wu, Matthew Singer, Matthew J Williams, Maximiliano G Gutierrez, Melanie Turner, Melvyn Yap, Michael Howell, Michael Hubank, Michael J Blackman, Michael D Buck, Michele S Y Tan, Michelle Lappin, Mimmi Martensson, Ming Jiang, Ming-Han C

- Tsai, Ming-Shih Hwang, Mint R Htun, Miriam Molina-Arcas, Moira J Spyer, Monica Diaz-Romero, Moritz Treeck, Namita Patel, Natalie Chandler, Neil Osborne, Nick Carter, Nicola O'Reilly, Nigel Peat, Nikhil Faulkner, Nikita Komarov, Nisha Bhardwaj, Nnennaya Kanu, Oana Paun, Ok-Ryul Song, Olga O'Neill, Pablo Romero Clavijo, Patrick Davis, Patrick Toolan-Kerr, Paul Nurse, Paul Kotzampaltiris, Paul C Driscoll, Paul R Grant, Paula Ordonez Suarez, Peter Cherepanov, Peter Ratcliffe, Philip Hobson, Philip A Walker, Pierre Santucci, Qu Chen, Rachael Instrell, Rachel Ambler, Rachel Ulferts, Rachel Zillwood, Raffaella Carzaniga, Rajnika Hirani, Rajvee Shah Punatar, Richard Byrne, Robert Goldstone, Robyn Labrum, Ross Hall, Rowenna Roberts, Roy Poh, Rupert Beale, Rupert Faraway, Sally Cottrell, Sam Barrell, Sam Loughlin, Samuel McCall, Samutheswari Balakrishnan, Sandra Segura-Bayona, Savita Nutan, Selvaraju Veeriah, Shahnaz Bibi, Sharon P Vanloo, Simon Butterworth, Simon Caidan, Solene Debaisieux, Sonia Gandhi, Sophia Ward, Sophie Ridewood, Souradeep Basu, Stacey-Ann Lee, Steinar Halldorsson, Stephanie Nofal, The SAFER Investigators Team, Field Study, and The Crick COVID-19 Consortium. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *BMC Medicine*, 19(1):106, 2021.
- [104] James A Hay, Stephen M Kissler, Joseph R Fauver, Christina Mack, Caroline G Tai, Radhika M Samant, Sarah Connolly, Deverick J Anderson, Gaurav Khullar, Matthew MacKay, Miral Patel, Shannan Kelly, April Manhertz, Isaac Eiter, Daisy Salgado, Tim Baker, Ben Howard, Joel T Dudley, Christopher E Mason, Manoj Nair, Yaoming Huang, John DiFiori, David D Ho, Nathan D Grubaugh, and Yonatan H Grad. Quantifying the impact of immune history and variant on SARS-CoV-2 viral kinetics and infection rebound: A retrospective cohort study. *eLife*, 11:e81849, 2022.
- [105] Younjung Kim, Christl A Donnelly, and Pierre Nouvellet. Drivers of SARS-CoV-2 testing behaviour: a modelling study using nationwide testing data in England. *Nature Communications*, 14(1):2148, 2023.
- [106] Institute of Population Health University of Liverpool. Covid-SMART Asymptomatic Testing Pilot in Liverpool City Region: Quantitative Evaluation. Technical report.
- [107] UK Health Security Agency. Targeted Community Testing: Programme Evaluation - A review of policy implementation and delivery (July 2021 – March 2022) [Confidential

- Internal Document]. Technical report, 2022.
- [108] UK Health Security Agency and Department of Health & Social Care. Response: Public Accounts Committee, Twenty-Third Report of Session 2021-22 Department of Health and Social Care Test and Trace - Update. Technical report, 2021.
- [109] Department for Education. Coronavirus (COVID-19) asymptomatic testing in schools and colleges [Withdrawn 8 July 2021], 2021. Available at: <https://www.gov.uk/government/publications/coronavirus-covid-19-asymptomatic-testing-in-schools-and-colleges/%0Acoronavirus-covid-19-asymptomatic-testing-%0Ain-schools-and-colleges>.
- [110] UK Health Security Agency. Confirmatory PCR tests to be temporarily suspended for positive lateral flow test results [Press release], 2022. Available at: <https://www.gov.uk/government/news/confirmatory-pcr-tests-to-be-temporarily-suspended-for-positive-lateral-flow-test-results>.
- [111] Tullia Padellini, Radka Jersakova, Peter J Diggle, Chris Holmes, Ruairidh E King, Brieuc C L Lehmann, Ann-Marie Mallon, George Nicholson, Sylvia Richardson, and Marta Blangiardo. Time varying association between deprivation, ethnicity and SARS-CoV-2 infections in England: A population-based ecological study. *The Lancet Regional Health – Europe*, 15, 2022.
- [112] Marc Chadeau-Hyam, Oliver Eales, Barbara Bodinier, Haowei Wang, David Haw, Matthew Whitaker, Joshua Elliott, Caroline E Walters, Jakob Jonnerby, Christina Atchison, Peter J Diggle, Andrew J Page, Deborah Ashby, Wendy Barclay, Graham Taylor, Graham Cooke, Helen Ward, Ara Darzi, Christl A Donnelly, and Paul Elliott. 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. *eClinicalMedicine*, 48:101419, 2022.
- [113] Paul Elliott, Oliver Eales, Nicholas Steyn, David Tang, Barbara Bodinier, Haowei Wang, Joshua Elliott, Matthew Whitaker, Christina Atchison, Peter J Diggle, Andrew J Page, Alexander J Trotter, Deborah Ashby, Wendy Barclay, Graham Taylor, Helen Ward, Ara Darzi, Graham S Cooke, Christl A Donnelly, and Marc Chadeau-Hyam. Twin peaks: The Omicron SARS-CoV-2 BA.1 and BA.2 epidemics in England. *Science*, 376(6600):eabq4411, 2023.

- [114] Louise E Smith, Henry W W Potts, Richard Amlôt, Nicola T Fear, Susan Michie, and G James Rubin. Adherence to the test, trace, and isolate system in the UK: results from 37 nationally representative surveys. *BMJ*, 372, 2021.
- [115] Mingwang Shen, Yanni Xiao, Guihua Zhuang, Yan Li, and Lei Zhang. Mass testing—An underexplored strategy for COVID-19 control. *The Innovation*, 2(2):100114, 2021.
- [116] Vinícius Bonetti Franceschi, Andressa Schneiders Santos, Andressa Barreto Glaeser, Janini Cristina Paiz, Gabriel Dickin Caldana, Carem Luana Machado Lessa, Amanda de Menezes Mayer, Julia Gonçalves Küchle, Paulo Ricardo Gazzola Zen, Alvaro Vigo, Ana Trindade Winck, Liane Nanci Rotta, and Claudia Elizabeth Thompson. Population-based prevalence surveys during the Covid-19 pandemic: A systematic review. *Reviews in Medical Virology*, 31(4):e2200, 2021.
- [117] Joshua Elliott, Matthew Whitaker, Barbara Bodinier, Oliver Eales, Steven Riley, Helen Ward, Graham Cooke, Ara Darzi, Marc Chadeau-Hyam, and Paul Elliott. Predictive symptoms for COVID-19 in the community: REACT-1 study of over 1 million people. *PLOS Medicine*, 18(9):e1003777, 2021.
- [118] Veena S Raleigh. Ethnic differences in covid-19 death rates. *BMJ*, 376:o427, 2022.
- [119] Krishnan Bhaskaran, Sebastian Bacon, Stephen J W Evans, Chris J Bates, Christopher T Rentsch, Brian MacKenna, Laurie Tomlinson, Alex J Walker, Anna Schultze, Caroline E Morton, Daniel Grint, Amir Mehrkar, Rosalind M Eggo, Peter Inglesby, Ian J Douglas, Helen I McDonald, Jonathan Cockburn, Elizabeth J Williamson, David Evans, Helen J Curtis, William J Hulme, John Parry, Frank Hester, Sam Harper, David Spiegelhalter, Liam Smeeth, and Ben Goldacre. Factors associated with deaths due to COVID-19 versus other causes: population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. *The Lancet Regional Health – Europe*, 6, 2021.
- [120] Victoria J McGowan and Clare Bambra. COVID-19 mortality and deprivation: pandemic, syndemic, and endemic health inequalities. *The Lancet Public Health*, 7(11):e966–e975, 2022.

- [121] Clare Bamba, Ryan Riordan, John Ford, and Fiona Matthews. The COVID-19 pandemic and health inequalities. *Journal of epidemiology and community health*, 74(11):964–968, 2020.
- [122] Elizabeth J Williamson, Alex J Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E Morton, Helen J Curtis, Amir Mehrkar, David Evans, Peter Inglesby, Jonathan Cockburn, Helen I McDonald, Brian MacKenna, Laurie Tomlinson, Ian J Douglas, Christopher T Rentsch, Rohini Mathur, Angel Y S Wong, Richard Grieve, David Harrison, Harriet Forbes, Anna Schultze, Richard Croker, John Parry, Frank Hester, Sam Harper, Rafael Perera, Stephen J W Evans, Liam Smeeth, and Ben Goldacre. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821):430–436, 2020.
- [123] Karla Romero Starke, David Reissig, Gabriela Petereit-Haack, Stefanie Schmauder, Albert Nienhaus, and Andreas Seidler. The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis. *BMJ Global Health*, 6(12):e006434, 2021.
- [124] UK Health Security Agency. Tackling London’s ongoing COVID-19 health inequalities [Blog], 2021. Available at: <https://ukhsa.blog.gov.uk/2021/02/03/tackling-londons-covid-19-health-inequalities/>.
- [125] Tim Peto, Dominic Affron, Babak Afrough, Anita Agasu, Mark Ainsworth, Alison Allanson, Katherine Allen, Collette Allen, Lorraine Archer, Natasha Ashbridge, Iman Aurfan, Miriam Avery, Ellena Badenoch, Priya Bagga, Rishab Balaji, Ella Baldwin, Sophie Barraclough, Carol Beane, John Bell, Tracy Benford, Susan Bird, Marina Bishop, Angela Bloss, Richard Body, Rosie Boulton, Abbie Bown, Carla Bratten, Chris Bridgeman, Dominic Britton, Tim Brooks, Margaret Broughton-Smith, Pauline Brown, Beverley Buck, Elaine Butcher, Wendy Byrne, Gloria Calderon, Siobhan Campbell, Olivia Carr, Penny Carter, Daniel Carter, Megan Cathrall, Matthew Catton, Jim Chadwick, David Chapman, Kevin K Chau, Tanzina Chaudary, Shaolin Chidavaenzi, Samatha Chilcott, Bea Choi, Hannah Claasen, Simon Clark, Richard Clarke, Dawn Clarke, Richard Clayton, Kayleigh Collins, Rima Colston, James Connolly, Eloise Cook, Marie Corcoran, Ben Corley, Laura Costello, Caroline Coulson, Ant Crook, Derrick W Crook, Silvia D’Arcangelo, Mary-Anne Darby, John Davis, Rosaline de Koning, Pauline Derbyshire, Pam Devall, Mark Dolman,

Natalie Draper, Mark Driver, Sarah Dyas, Emily Eaton, Joy Edwards, Ruth Elderfield, Kate Ellis, Graham Ellis, Sue Elwell, Rachel Evans, Becky Evans, Marion Evans, Ranoromanana Evans, David Eyre, Codie Fahey, Vanessa Fenech, Janet Field, Alice Field, Tom Foord, Tom Fowler, Mollie French, Hannah Fuchs, Jasmine Gan, Joseph Gernon, Geeta Ghadiali, Narindar Ghuman, Kerry Gibbons, Gurvinder Gill, Kate Gilmour, Anika Goel, Sally Gordon, Tillie Graham, Alexander Grassam-Rowe, David Green, Anna Gronert, Tegan Gumsley-Read, Claire Hall, Bassam Hallis, Sally Hammond, Peter Hammond, Beth Hanney, Victoria Hardy, Gabriella Harker, Andrew Harris, May Havinden-Williams, Elena Hazell, Joanne Henry, Kim Hicklin, Kelly Hollier, Ben Holloway, Sarah J Hoosdally, Susan Hopkins, Lucy Hughes, Steve Hurdowar, Sally-Anne Hurford, Joanne Jackman, Harriet Jackson, Ruth Johns, Susan Johnston, Juliet Jones, Tinashe Kanyowa, Katie Keating-Fedders, Sharon Kempson, Iftikhar Khan, Beinn Khulusi, Thomas Knight, Anuradha Krishna, Patrick Lahert, Zoe Lampshire, Daniel Lasserson, Kirsten Lee, Lennard Y W Lee, Arabella Legard, Cristina Leggio, Justin Liu, Teresa Lockett, Christopher Logue, Vanessa Lucas, Sheila F Lumley, Vindhya Maripuri, Des Markham, Emma Marshall, Philippa C Matthews, Sarah Mckee, Deborah F McKee, Neil McLeod, Antoinette McNulty, Freddie Mellor, Rachel Michel, Alex Mighiu, Julie Miller, Zarina Mirza, Heena Mistry, Jane Mitchell, Mika Erik Moeser, Sophie Moore, Akhila Muthuswamy, Daniel Myers, Gemma Nanson, Mike Newbury, Scott Nicol, Harry Nuttall, Jewel Jones Nwanaforo, Louise Oliver, Wendy Osbourne, Jake Osbourne, Ashley Otter, Jodie Owen, Sulaksan Panchalingam, Dimitris Papoulidis, Juan Dobaldo Pavon, Arro Peace, Karen Pearson, Liam Peck, Ashley Pegg, Suzannah Pegler, Helen Permain, Prem Perumal, Leon Peto, Tim E A Peto, Thanh Pham, Hayleah L Pickford, Mark Pinkerton, Michelle Platton, Ashley Price, Emily Protheroe, Hellen Purnell, Lottie Rawden, Sara Read, Charles Reynard, Susan Ridge, Tom G Ritter, James Robinson, Patrick Robinson, Gillian Rodger, Cathy Rowe, Bertie Rowell, Alexandra Rowlands, Sarah Sampson, Kathryn Saunders, Rachel Sayers, Jackie Sears, Richard Sedgewick, Laura Seeney, Amanda Selassie, Lloyd Shail, Jane Shallcross, Lucy Sheppard, Anna Sherkat, Shelha Siddiqui, Alex Sienkiewicz, Lavanya Sinha, Jennifer Smith, Ella Smith, Emma Stanton, Thomas Starkey, Aleksander Stawiarski, Amelia Sterry, Joe Stevens, Mark Stockbridge, Nicole Stoesser, Anila Sukumaran, Angela Sweed, Sami Tatar, Hema Thomas, Carly Tibbins, Sian Tiley, Julie Timmins, Cara Tomas-Smith, Oliver Topping, Elena Turek, Toi Neibler, Kate Trigg-Hogarth, Elizabeth Truelove, Chris Turn-

- bull, David Tyrrell, Alison Vaughan, John Vertannes, Richard Vipond, Linda Wagstaff, Joanne Waldron, Philip Walker, Ann Sarah Walker, Mary Walters, Jenny Y Wang, Ellie Watson, Kate Webberley, Kimerbley Webster, Grace Westland, Ian Wickens, Jane Willcocks, Herika Willis, Stephen Wilson, Barbara Wilson, Louise Woodhead, Deborah Wright, Bindhu Xavier, Fiona Yelnoorkar, Lisa Zeidan, and Rangeni Zinyama. COVID-19: Rapid antigen detection for SARS-CoV-2 by lateral flow assay: A national systematic evaluation of sensitivity and specificity for mass-testing. *eClinicalMedicine*, 36, 2021.
- [126] Department of Health & Social Care. Twice weekly rapid testing to be available to everyone in England [Press release], 0. Available at: <https://www.gov.uk/government/news/twice-weekly-rapid-testing-to-be-available-to-everyone-in-england>.
- [127] Erika Molteni, Carole H Sudre, Liane S Canas, Sunil S Bhopal, Robert C Hughes, Michela Antonelli, Benjamin Murray, Kerstin Kläser, Eric Kerfoot, Liyuan Chen, Jie Deng, Christina Hu, Somesh Selvachandran, Kenneth Read, Joan Capdevila Pujol, Alexander Hammers, Tim D Spector, Sebastien Ourselin, Claire J Steves, Marc Modat, Michael Absoud, and Emma L Duncan. Illness duration and symptom profile in symptomatic UK school-aged children tested for SARS-CoV-2. *The Lancet Child & Adolescent Health*, 5(10):708–718, 2021.
- [128] Dylan A Mistry, Jenny Y Wang, Mika-Erik Moeser, Thomas Starkey, and Lennard Y W Lee. A systematic review of the sensitivity and specificity of lateral flow devices in the detection of SARS-CoV-2. *BMC Infectious Diseases*, 21(1):828, 2021.
- [129] David W Eyre, Matthias Futschik, Sarah Tunkel, Jia Wei, Joanna Cole-Hamilton, Rida Saquib, Nick Germanacos, Andrew R Dodgson, Paul E Klapper, Malur Sudhanva, Chris Kenny, Peter Marks, Edward Blandford, Susan Hopkins, Tim E A Peto, and Tom Fowler. Performance of antigen lateral flow devices in the UK during the alpha, delta, and omicron waves of the SARS-CoV-2 pandemic: a diagnostic and observational study. *The Lancet Infectious Diseases*, 23(8):922–932, 2023.
- [130] Cabinet office. COVID-19 Response: Living with COVID-19. Available at: <https://www.gov.uk/government/publications/covid-19-response-living-with-covid-19/covid-19-response-living-with-covid-19>.

- [131] Fesani Mahmood, Dev Acharya, Kanta Kumar, and Vibhu Paudyal. Impact of COVID-19 pandemic on ethnic minority communities: a qualitative study on the perspectives of ethnic minority community leaders. *BMJ Open*, 11(10):e050584, 2021.
- [132] Chun-Han Lo, Long H Nguyen, David A Drew, Erica T Warner, Amit D Joshi, Mark S Graham, Adjoa Anyane-Yeboah, Fatma M Shebl, Christina M Astley, Jane C Figueiredo, Chuan-Guo Guo, Wenjie Ma, Raaj S Mehta, Sohee Kwon, Mingyang Song, Richard Davies, Joan Capdevila, Carole H Sudre, Jonathan Wolf, Yvette C Cozier, Lynn Rosenberg, Lynne R Wilkens, Christopher A Haiman, Loïc Le Marchand, Julie R Palmer, Tim D Spector, Sebastien Ourselin, Claire J Steves, and Andrew T Chan. Race, ethnicity, community-level socioeconomic factors, and risk of COVID-19 in the United States and the United Kingdom. *eClinicalMedicine*, 38, 2021.
- [133] Michele Tizzoni, Elaine O Nsoesie, Laetitia Gauvin, Márton Karsai, Nicola Perra, and Shweta Bansal. Addressing the socioeconomic divide in computational modeling for infectious diseases. *Nature Communications*, 13(1):2897, 2022.
- [134] Caroline Buckee, Abdisalan Noor, and Lisa Sattenspiel. Thinking clearly about social aspects of infectious disease transmission. *Nature*, 595(7866):205–213, 2021.
- [135] Cécile Viboud, Ottar N Bjørnstad, David L Smith, Lone Simonsen, Mark A Miller, and Bryan T Grenfell. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*, 312(5772):447–451, 2006.
- [136] B T Grenfell, O N Bjørnstad, and J Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.
- [137] Kimiya Gohari, Anoshirvan Kazemnejad, Ali Sheidaei, and Sarah Hajari. Clustering of countries according to the COVID-19 incidence and mortality rates. *BMC Public Health*, 22(1):632, 2022.
- [138] Syeda Amna Rizvi, Muhammad Umair, and Muhammad Aamir Cheema. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons & Fractals*, 151:111240, 2021.

- [139] Banafsheh Sadeghi, Rex C Y Cheung, and Meagan Hanbury. Using hierarchical clustering analysis to evaluate COVID-19 pandemic preparedness and performance in 180 countries in 2020. *BMJ Open*, 11(11):e049844, 2021.
- [140] Marcus R Andrews, Kosuke Tamura, Janae N Best, Joniqua N Ceasar, Kaylin G Batey, Troy A Kearse, Lavell V Allen, Yvonne Baumer, Billy S Collins, Valerie M Mitchell, and Tiffany M Powell-Wiley. Spatial Clustering of County-Level COVID-19 Rates in the U.S. *International Journal of Environmental Research and Public Health*, 18(22), 2021.
- [141] Yee Whye Teh. Dirichlet Process, 2010. Available at: <https://www.stats.ox.ac.uk/~teh/research/npbayes/Teh2010a.pdf>.
- [142] Jim Pitman. *Combinatorial stochastic processes: Ecole d’été de probabilités de saint-flour xxxii-2002*. Springer, 2006.
- [143] Richard Creswell. Flexible models for scientific data using Dirichlet process mixtures. *In preparation*, 2024.
- [144] Radford M Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- [145] Yuelin Li, Elizabeth Schofield, and Mithat Gönen. A tutorial on Dirichlet process mixture modeling. *Journal of Mathematical Psychology*, 91:128–144, 2019.
- [146] Ioana Bouros, Richard Creswell, Sumali Bajaj, Katherine Shepherd, and Ben Lambert. A student’s guide to Dirichlet processes. *In preparation*, 2024.
- [147] Swapnil Mishra, James A Scott, Daniel J Laydon, Harrison Zhu, Neil M Ferguson, Samir Bhatt, Seth Flaxman, and Axel Gandy. A COVID-19 Model for Local Authorities of the United Kingdom. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement 1):S86–S95, 2022.
- [148] Richard Creswell, Martin Robinson, David Gavaghan, Kris V Parag, Chon Lok Lei, and Ben Lambert. A Bayesian nonparametric method for detecting rapid changes in disease transmission. *Journal of Theoretical Biology*, 558:111351, 2023.
- [149] Iwona Hawryluk, Henrique Hoeltgebaum, Swapnil Mishra, Xenia Miscouridou, Ricardo P Schnekenberg, Charles Whittaker, Michaela Vollmer, Seth Flaxman, Samir Bhatt, and

- Thomas A Mellan. Gaussian Process Nowcasting: Application to COVID-19 Mortality Reporting. *arXiv*, 2021.
- [150] S Bhatt, D J Weiss, E Cameron, D Bisanzio, B Mappin, U Dalrymple, K E Battle, C L Moyes, A Henry, P A Eckhoff, E A Wenger, O Briët, M A Penny, T A Smith, A Bennett, J Yukich, T P Eisele, J T Griffin, C A Fergus, M Lynch, F Lindgren, J M Cohen, C L J Murray, D L Smith, S I Hay, R E Cibulskis, and P W Gething. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526(7572):207–211, 2015.
- [151] Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, Osman Sankoh, Monica F Myers, Dylan B George, Thomas Jaenisch, G R William Wint, Cameron P Simmons, Thomas W Scott, Jeremy J Farrar, and Simon I Hay. The global distribution and burden of dengue. *Nature*, 496(7446):504–507, 2013.
- [152] Gabriela B Cybis, Janet S Sinsheimer, Trevor Bedford, Andrew Rambaut, Philippe Lemey, and Marc A Suchard. Bayesian nonparametric clustering in phylogenetics: modeling antigenic evolution in influenza. *Statistics in Medicine*, 37(2):195–206, 2018.
- [153] Qian Wang, Wen Dong, Kun Yang, Zhongda Ren, Dongqing Huang, Peng Zhang, and Jie Wang. Temporal and spatial analysis of COVID-19 transmission in China and its influencing factors. *International Journal of Infectious Diseases*, 105:675–685, 2021.
- [154] Jack Cordes and Marcia C Castro. Spatial analysis of COVID-19 clusters and contextual factors in New York City. *Spatial and Spatio-temporal Epidemiology*, 34:100355, 2020.
- [155] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- [156] Paula C Dias. Sources and sinks in population biology. *Trends in Ecology & Evolution*, 11(8):326–330, 1996.
- [157] Cliff C Kerr, Robyn M Stuart, Dina Mistry, Romesh G Abeysuriya, Katherine Rosenfeld, Gregory R Hart, Rafael C Núñez, Jamie A Cohen, Prashanth Selvaraj, Brittany Hagedorn, Lauren George, Michał Jastrzębski, Amanda S Izzo, Greer Fowler, Anna Palmer,

- Dominic Delport, Nick Scott, Sherrie L Kelly, Caroline S Bennette, Bradley G Wagner, Stewart T Chang, Assaf P Oron, Edward A Wenger, Jasmina Panovska-Griffiths, Michael Famulare, and Daniel J Klein. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, 17(7):1–32, 2021.
- [158] Matthew J Penn, Daniel J Laydon, Joseph Penn, Charles Whittaker, Christian Morgenstern, Oliver Ratmann, Swapnil Mishra, Mikko S Pakkanen, Christl A Donnelly, and Samir Bhatt. Intrinsic randomness in epidemic modelling beyond statistical uncertainty. *Communications Physics*, 6(1):146, 2023.
- [159] Chris Wymant, Luca Ferretti, Daphne Tsallis, Marcos Charalambides, Lucie Abeler-Dörner, David Bonsall, Robert Hinch, Michelle Kendall, Luke Milsom, Matthew Ayres, Chris Holmes, Mark Briers, and Christophe Fraser. The epidemiological impact of the NHS COVID-19 app. *Nature*, 594(7863):408–412, 2021.
- [160] J von Kügelgen, L Gresele, and B Schölkopf. Simpson’s Paradox in COVID-19 Case Fatality Rates: A Mediation Analysis of Age-Related Causal Effects. *IEEE Transactions on Artificial Intelligence*, 2(1):18–27, 2021.
- [161] Thomas C Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186, 1971.
- [162] Anjalika Nande, Justin Sheen, Emma L Walters, Brennan Klein, Matteo Chinazzi, Andrei H Gheorghe, Ben Adlam, Julianna Shinnick, Maria Florencia Tejeda, Samuel V Scarpino, Alessandro Vespignani, Andrew J Greenlee, Daniel Schneider, Michael Z Levy, and Alison L Hill. The effect of eviction moratoria on the transmission of SARS-CoV-2. *Nature Communications*, 12(1):2274, 2021.
- [163] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.
- [164] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

- [165] Zirou Qiu, Baltazar Espinoza, Vitor V Vasconcelos, Chen Chen, Sara M Constantino, Stefani A Crabtree, LuoJun Yang, Anil Vullikanti, Jiangzhuo Chen, Jörgen Weibull, Kaushik Basu, Avinash Dixit, Simon A Levin, and Madhav V Marathe. Understanding the coevolution of mask wearing and epidemics: A network perspective. *Proceedings of the National Academy of Sciences*, 119(26):e2123355119, 2022.
- [166] Sebastian Funk, Marcel Salathé, and Vincent A A Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of The Royal Society Interface*, 7(50):1247–1256, 2010.
- [167] Pierre Geoffard and Tomas Philipson. Disease Eradication: Private versus Public Vaccination. *American Economic Review*, 87(1):222–230, 1997.
- [168] Cornelia Betsch, Lothar H Wieler, and Katrine Habersaat. Monitoring behavioural insights related to COVID-19. *The Lancet*, 395(10232):1255–1256, 2020.
- [169] J A Patel, F B H Nielsen, A A Badiani, S Assi, V A Unadkat, B Patel, R Ravindrane, and H Wardle. Poverty, inequality and COVID-19: the forgotten vulnerable. *Public health*, 183:110–111, 2020.

CORONAVIRUS

Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence

Moritz U. G. Kraemer^{1*}†, Verity Hill²†, Christopher Ruis³†, Simon Dellicour^{4,5}†, Sumali Bajaj¹†, John T. McCrone², Guy Baele⁵, Kris V. Parag⁶, Anya Lindström Battle⁷, Bernardo Gutierrez¹, Ben Jackson², Rachel Colquhoun², Áine O'Toole², Brennan Klein⁸, Alessandro Vespignani⁸, COVID-19 Genomics UK (COG-UK) Consortium†, Erik Volz⁶, Nuno R. Faria^{1,6,9}, David M. Aanensen^{10,11}, Nicholas J. Loman¹², Louis du Plessis¹, Simon Cauchemez¹³, Andrew Rambaut^{2*}, Samuel V. Scarpino^{8,14,15*}, Oliver G. Pybus^{1,16*}

Understanding the causes and consequences of the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants of concern is crucial to pandemic control yet difficult to achieve because they arise in the context of variable human behavior and immunity. We investigated the spatial invasion dynamics of lineage B.1.1.7 by jointly analyzing UK human mobility, virus genomes, and community-based polymerase chain reaction data. We identified a multistage spatial invasion process in which early B.1.1.7 growth rates were associated with mobility and asymmetric lineage export from a dominant source location, enhancing the effects of B.1.1.7's increased intrinsic transmissibility. We further explored how B.1.1.7 spread was shaped by nonpharmaceutical interventions and spatial variation in previous attack rates. Our findings show that careful accounting of the behavioral and epidemiological context within which variants of concern emerge is necessary to interpret correctly their observed relative growth rates.

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) lineage B.1.1.7 expanded rapidly across the United Kingdom (1, 2) in late 2020 and subsequently spread internationally (3, 4). As of 19 January 2021 (date of the most recent sample in our dataset), B.1.1.7 had reached all but five counties of Wales, Scotland, Northern Ireland, and England, with onward transmission in each. Restrictions on international travel were enacted to contain B.1.1.7's spread; however, genomic surveillance has since detected the presence and growth of the lineage in many countries worldwide (4, 5). Analyses of genomic, laboratory, secondary contact, and aggregated epidemiological data estimate higher transmissibility of B.1.1.7 compared with previous SARS-CoV-2 lineages (1, 6–9) and potentially a greater risk of hospitalization (10–13). The spatial heterogeneity of SARS-CoV-2 transmission—and of emerging infectious diseases in general—can have profound effects on the local likelihood and intensity of transmission, final epidemic size, and immunity (14–22). More specifically, estimates of B.1.1.7's increased relative transmissibility declined during its emergence in the UK (7, 9); understanding why this occurred is necessary if we are to respond effectively to future SARS-

CoV-2 variants. We reconstructed and quantified the spatial dynamics of B.1.1.7's emergence and investigated how human mobility and heterogeneity in previous exposure contributed to B.1.1.7's initial spread and evaluation of higher transmissibility.

Spatial expansion and source sink dynamics of B.1.1.7 in the UK

B.1.1.7 can be first detected in COVID-19 Genomics UK Consortium (COG-UK) genome data in Kent on 20 September 2020 and spread quickly across the UK, with each week adding detections in approximately seven new upper-tier local authorities (UTLAs) (Fig. 1, A and B, and table S2). B.1.1.7 was already reported in several UTLAs before the start of the second English lockdown (5 November 2020). By the end of that lockdown (2 December 2020), B.1.1.7 was widespread throughout the UK (Fig. 1, A and B).

The spatial expansion of SARS-CoV-2 lineages [for example, (16, 23)] can be tracked by using data from the UK's national surveillance of SARS-CoV-2 genomes (24). By combining these data with aggregated mobile phone data, we examined the dissemination of B.1.1.7 through human mobility, from its likely location of emergence (Kent and Greater

London) to other UK regions (Fig. 1, D and E, and supplementary materials, materials and methods). Human mobility among UK regions increased at the end of the second English lockdown, from 55 million to 75 million weekly movements (Fig. 1E). Because of its centrality, Greater London exhibits an important connective role in the UK human movement network (Fig. 1D; red lines indicate the week the second lockdown was eased). Compared with that of previous weeks, movements out of Greater London were more frequent and reached more destinations (fig. S1). For each UTLA, we found that the date of first detection of B.1.1.7 is predicted well by human mobility from Kent and Greater London to that UTLA [Pearson's correlation coefficient (r) = -0.73 ; 95% confidence interval (CI): -0.61 , -0.81 ; Akaike information criteria (AIC) = 734] (Fig. 1C) and similarly well by using movements from Kent and Greater London separately (fig. S2). This correlation strengthens through time as new locations of B.1.1.7 detection are added (fig. S3) and is robust to changes in human mobility through time in among-region human movement (Pearson's r = -0.44 ; 95% CI: -0.16 , -0.65 ; P < 0.01; mobility data through 23 January 2021) (materials and methods). Geographic distance from Greater London correlates less strongly with B.1.1.7 arrival times (Pearson's r = 0.60; 95% CI: 0.44 to 0.71; AIC = 763) (fig. S4).

To understand better the spatial dispersal of B.1.1.7 during its emergence, we reconstructed its spread across England using large-scale phylogeographic analysis (25–27). We analyzed 17,716 B.1.1.7 genomes collected between 20 September 2020 and 19 January 2021 (Fig. 2 and fig. S5), collated from polymerase chain reaction (PCR)–positive community samples that represent a random selection of SARS-CoV-2–positive samples (28). These genomes represent ~4% of UK B.1.1.7 cases during the study period [n = 460,510 estimated tests with PCR S-gene target failure (SGTF) between 20 September 2020 and 19 January 2021]. Samples per location (UTLA) and per week in the SGTF and whole-genome datasets are strongly correlated (Pearson's r = 0.69; 95% CI: 0.63 – 0.73; P < 0.001) (fig. S6) (7), making it feasible to reconstruct B.1.1.7 expansion history by using phylogeographic approaches (29).

We identified distinct phases to the emergence of B.1.1.7. Initially, during the second

¹Department of Zoology, University of Oxford, Oxford, UK. ²Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. ³Molecular Immunity Unit, Department of Medicine, Cambridge University, Cambridge, UK. ⁴Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Bruxelles, Belgium. ⁵Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, 3000 Leuven, Belgium. ⁶MRC Centre for Global Infectious Disease Analysis, Jameel Institute for Disease and Emergency Analytics, Imperial College London, London, UK. ⁷Department of Plant Sciences, University of Oxford, Oxford, UK. ⁸Network Science Institute, Northeastern University, Boston, USA. ⁹Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. ¹⁰Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, UK. ¹¹Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK. ¹²Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. ¹³Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, UMR2000, CNRS, Paris, France. ¹⁴Vermont Complex Systems Center, University of Vermont, Burlington, USA. ¹⁵Santa Fe Institute, Santa Fe, USA. ¹⁶Department of Pathobiology and Population Sciences, Royal Veterinary College London, London, UK.

*Corresponding author. Email: a.rambaut@ed.ac.uk (A.R.); moritz.kraemer@zoo.ox.ac.uk (M.U.G.K.); s.scarpino@northeastern.edu (S.V.S.); oliver.pybus@zoo.ox.ac.uk (O.G.P.)

†These authors contributed equally to this work.

‡Consortium members and affiliations are listed in the supplementary materials.

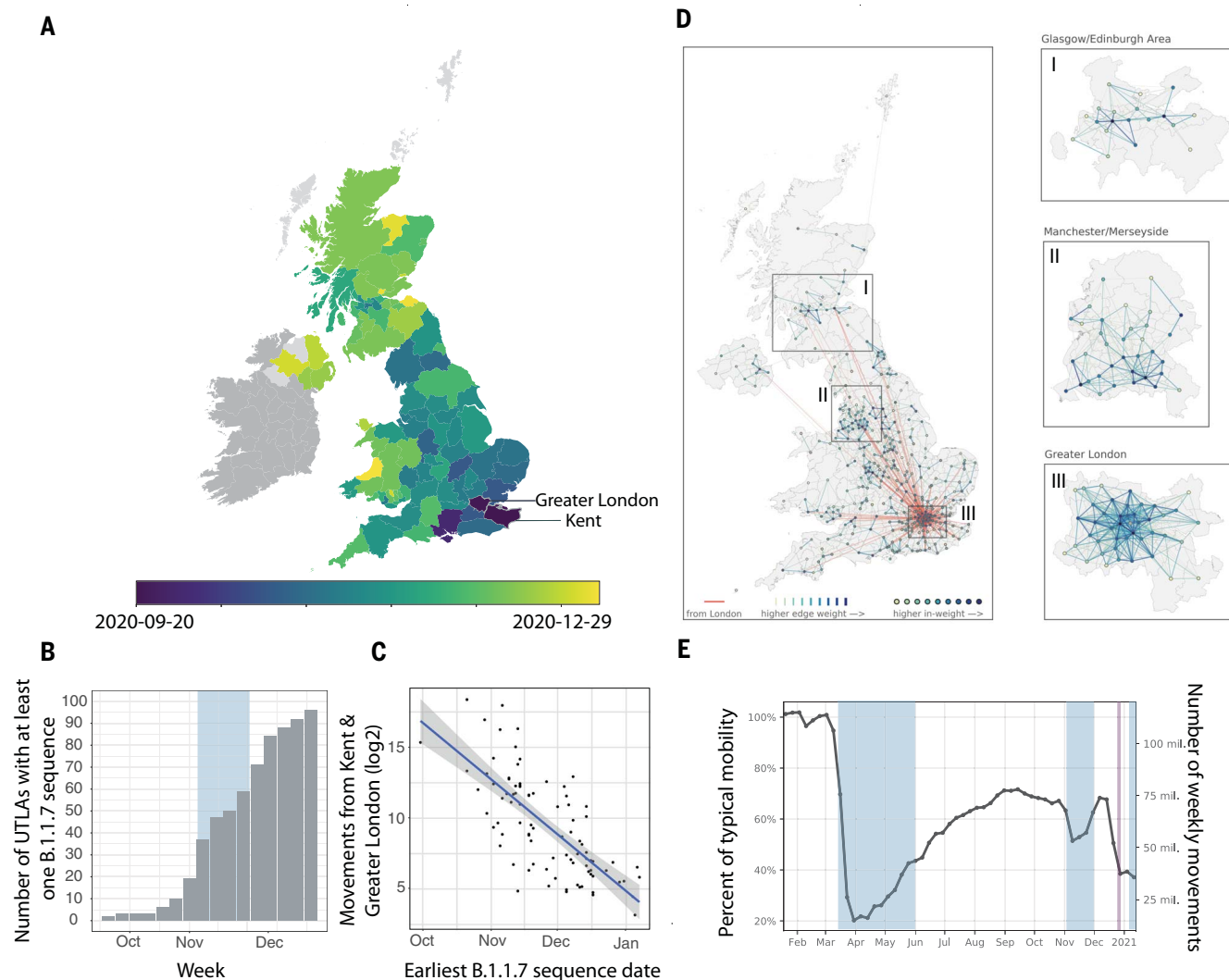


Fig. 1. Human mobility and spatial expansion of B.1.1.7 across the UK.

(A) Map at the UTLA level of arrival dates of lineage B.1.1.7. Darker colors indicate earlier dates, and lighter colors indicate later dates. Arrival time is defined as the earliest sampling date of a B.1.1.7 genomic sequence in each UTLA. (B) Cumulative number of UTLAs in which B.1.1.7 has been detected, in 7-day intervals. The blue shaded area indicates the period of the second lockdown in England. (C) Relationship between the arrival time of B.1.1.7 and estimated number of movements from Kent and London during February 2020 for each UTLA (Pearson's $r = -0.73$; 95% CI: -0.61 , -0.81 ; $P < 0.001$) (materials and methods). (D) Human mobility at the UK local

authority district level (LAD) (table S2) during the epidemiological week 29 November to 5 December 2020. Thicker lines (edges) indicate more movements between regions. Nodes with larger absolute incoming movements are indicated with darker colors. Red lines indicate movements from Greater London. (Insets I, II, and III) Mobility within three UK metropolitan areas. (E) Trends in human mobility across the UK (indicating movements between but not within LADs). The blue shaded areas indicate the period of the first, second, and third lockdown in England. Dark red indicates the timing (20 December 2020) of the Tier 4 restrictions imposed in southeast England, including London (56).

English lockdown, most (71.2%) B.1.1.7 phylogenetic branch movements originated and ended in Greater London or Kent; long-distance dispersal events were relatively infrequent (Figs. 2 and 3). After the lockdown ended, and new cases in London subsequently rose rapidly, observed virus lineage movements from southeast England to other regions increased, and other large cities started to exhibit local transmission (Figs. 2 and 3). This phase of a growing number of exported B.1.1.7 cases from London and environs stabilized in mid-December and coincided with reduced mobility from Greater London (Tier 4 restrictions were announced on 20 December 2020 and entailed a "Stay at home" order, closure of

nonessential shops and hospitality, and strict limitations on household mixing) (Figs. 1E and 2C). However, the total number of B.1.1.7 lineage exports did not immediately decline because the growing number of B.1.1.7 cases in southeast England offset the decline in outward travel (Fig. 2C) (30), indicating a limited effect of delayed action on B.1.1.7 spread from Greater London. Our analysis did not allow us to establish a causal link between nonpharmaceutical interventions (NPIs) and their impact on lineage exportations, so these results should be interpreted with caution.

By combining mobility and SGTF data with estimates of the proportion of the population

testing SARS-CoV-2-positive (materials and methods), we can estimate the frequency of B.1.1.7 export from Greater London to other English regions (Fig. 2C and fig. S7) and explore its role in accelerating the lineage's emergence. Using these combined data sources, we estimate that the number of B.1.1.7 case exports from Greater London rose during November (including during lockdown) from <600 to >12,000 in early December (Fig. 2C, gray curve), reflecting growth in B.1.1.7 infections in Greater London and an increase in human mobility among UK geographic regions across in late November (Fig. 1E). The estimated intensity of B.1.1.7 case exportation from Greater London remained high in December,

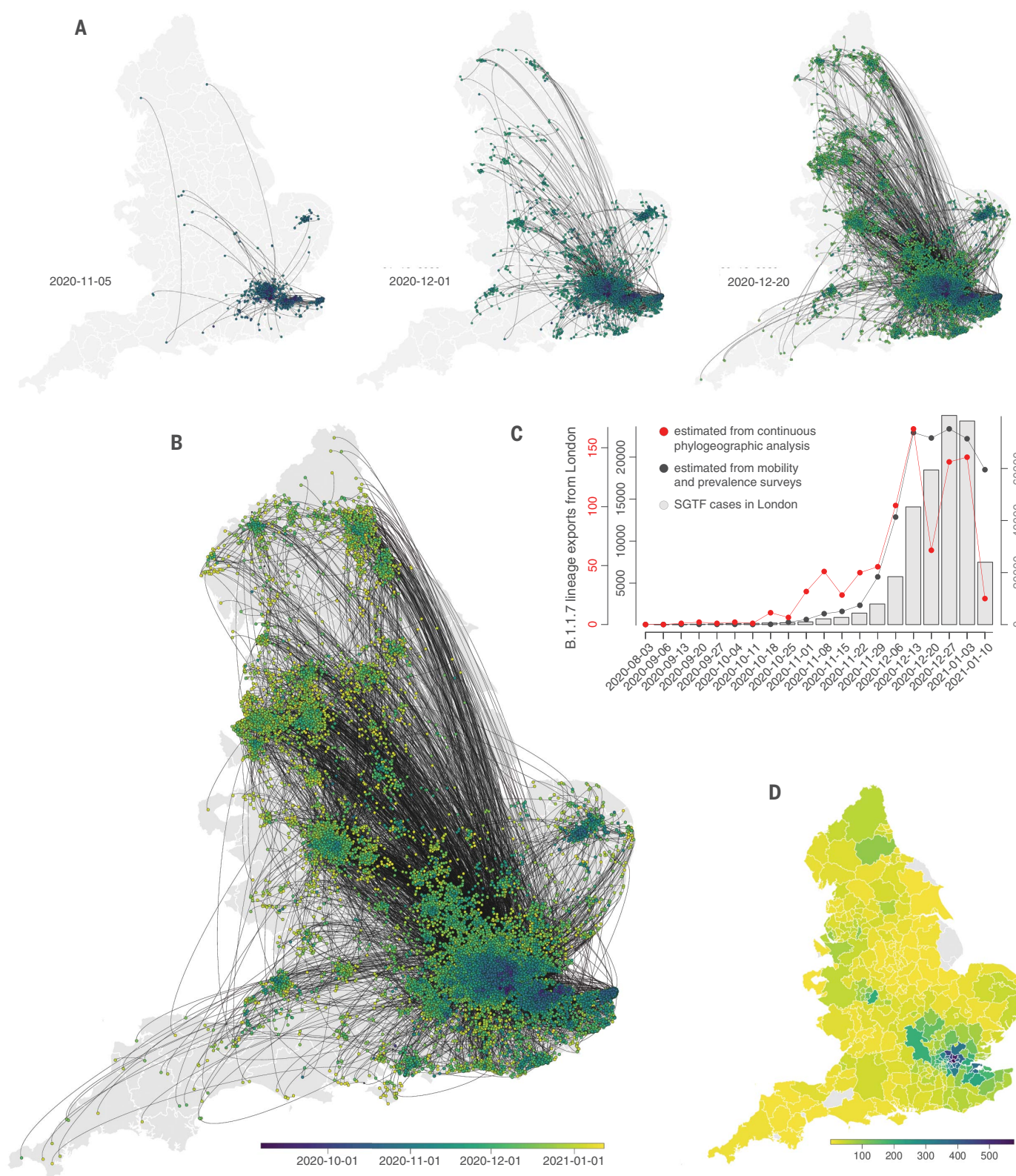


Fig. 2. Spatial emergence dynamics of SARS-CoV-2 lineage B.1.1.7 in England.

(A and B) Continuous phylogeographic reconstruction with phylogeny nodes colored according to their time of occurrence and dispersal direction of phylogeny branches indicated by edge curvature (counterclockwise). From left to right, data to 5 November, 1 December, and 20 December 2020, respectively. (B) Map of the entire recon-

struction, up to 19 January 2021. (C) Estimated number of weekly exports of lineage B.1.1.7 from the Greater London area, inferred from the continuous phylogeographic analysis (red), and estimated from mobility and prevalence survey data (black). (D) Estimated number of cumulative B.1.1.7 introductions inferred from phylogeographic analysis into each administrative area (UTLA) by 12 December 2020.

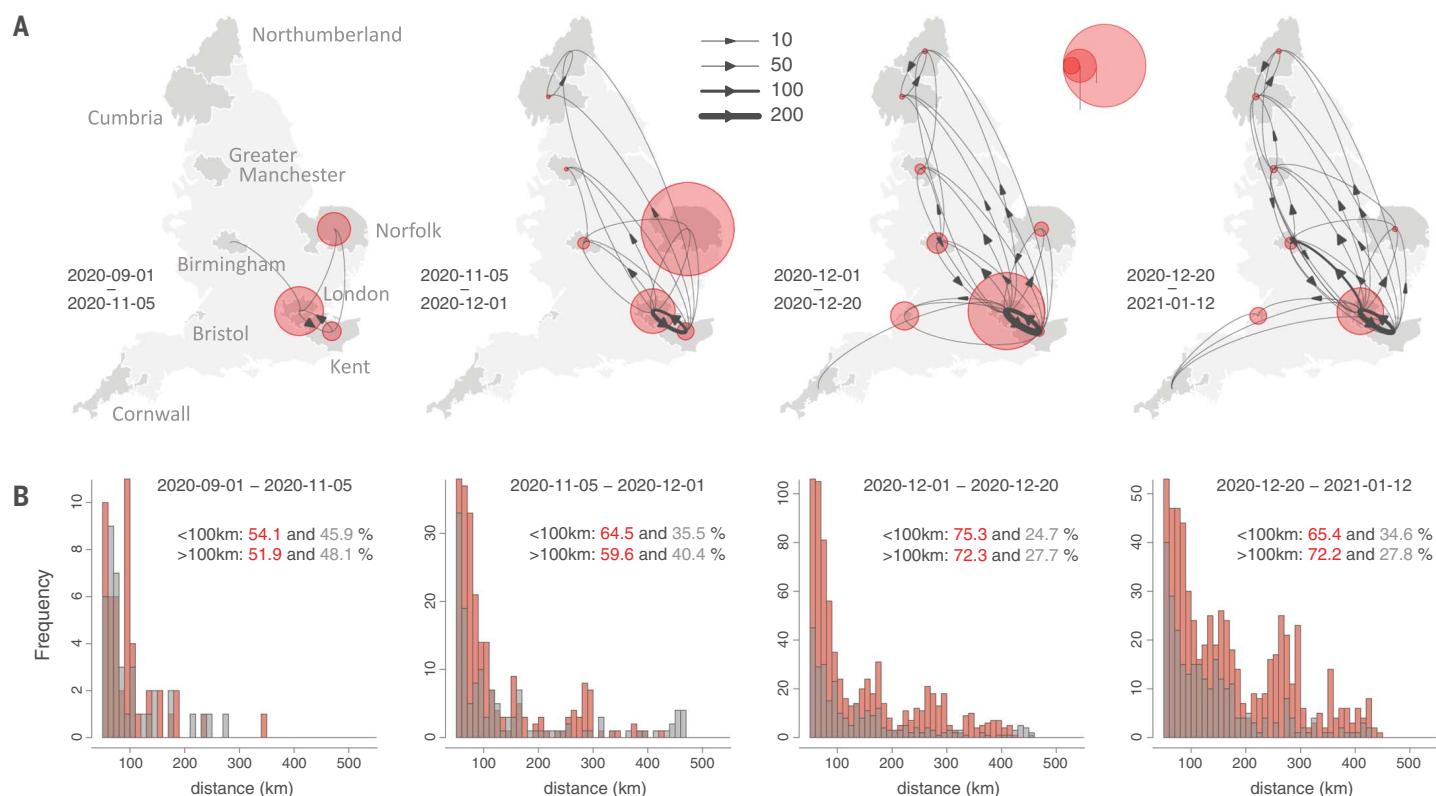


Fig. 3. Spatial structure of B.1.1.7 lineage dispersal in England from phylogeographic reconstruction. (A) Curved arrows and line thicknesses indicate the direction and intensity of B.1.1.7 lineage flows among regions. Red circles indicate, for a given location, the ratio of inferred local movements to inferred importations into that location. Four time periods are shown (left to

right) and roughly correspond to (i) before second lockdown, (ii) second lockdown, (iii) after second lockdown, and (iv) implementation of Tier 4 restrictions in southeast England. (B) Distribution of the geographic distances of phylogenetic lineage movement events (>50 km). Those from Greater London are in red, and those from other locations are in gray.

peaking in mid-December at ~20,000 weekly exports, before declining in early January after the third national lockdown started on 5 January 2021. These estimates (Fig. 2C, gray curve) closely match the trends in lineage B.1.1.7 movement inferred from phylogeographic analysis (Fig. 2C, red curve), cross-validating both data sources (exports estimated by using each method are strongly correlated; Pearson's $r = 0.62$; 95% CI: 0.61 to 0.64; $P < 0.001$) (fig. S8). Lineage exportation events estimated from genomic data are lower from late December onward, possibly owing to reporting lags in genomic data generation and/or delayed care-seeking because of the Christmas holidays (31). Our simple model assumes that nonsymptomatic infectious individuals are equally likely to travel (Fig. 2C, gray line), which may bias our estimates of infectious travellers upward.

B.1.1.7 dispersal dynamics shifted in late December to more bidirectional exchange of phylogenetic lineages in and out of Greater London (Fig. 3), coinciding with rapid growth in B.1.1.7 cases across England (9). Throughout, the weekly number of B.1.1.7 cases in a UTLA was positively associated with the number of B.1.1.7 lineage introductions into that UTLA during that week (Pearson's $r = 0.41$,

0.76, 0.91, and 0.73, for October, November, December, and January, respectively; $P < 0.001$ for all; further analysis is provided in the supplementary materials) (fig. S6). We observed spatial heterogeneity in B.1.1.7 lineage importations; in the phylogeographic analysis, some locations received >500 inferred importations, despite our genomic dataset representing <4% of reported B.1.1.7 cases during the study period (Fig. 2D).

Detailed mapping of the spatial dynamics of SARS-CoV-2 lineages is difficult without comprehensive, well-sampled epidemiological and genomic data (32, 33). However, the COG-UK data enables us to study dissemination trends by comparing inferred B.1.1.7 importations with within-location movements. Greater London (and to some extent Kent) acted as the main exporter of B.1.1.7 lineages to other UTLAs until mid-December 2020 (Fig. 3A). The longest (>100 km) and shortest (<100 km) dispersal events consistently originated from Greater London throughout the study period (Fig. 3B), primarily because of its large epidemic. However, the relative percentage of lineage movements that originated from Greater London approximately halved between September 2020 and January 2021 (table S1).

Spatial heterogeneity in SARS-CoV-2 incidence and B.1.1.7 expansion

Using SGTF PCR-positive tests as a proxy for B.1.1.7 infection (34), we next examined daily growth rates of SARS-CoV-2 cases at the UTLA level for SGTF and non-SGTF cases (excluding case data from 25 to 31 January to account for reporting and testing delays) (materials and methods) (35). Case growth rates immediately after the November 2020 lockdown were highest in regions of southeast England connected to Greater London and/or Kent (fig. S9). Acceleration in SGTF case growth rates in Greater London began in mid-November and preceded acceleration in other regions (Fig. 4B). At the UTLA level, growth rates of SGTF cases were higher than non-SGTF cases (fig. S9), a key observation used to support an increased transmissibility for B.1.1.7 (7, 9).

We added to those findings by quantifying the import of B.1.1.7 cases from London and investigating the association of importation trends with lineage-specific case growth rates (materials and methods). Using our phylogeographic analysis results (Figs. 2 and 3), we found that growth in the rate of B.1.1.7 importation into a lower-tier local authority (LTLA) closely matches the early growth rate of SGTF

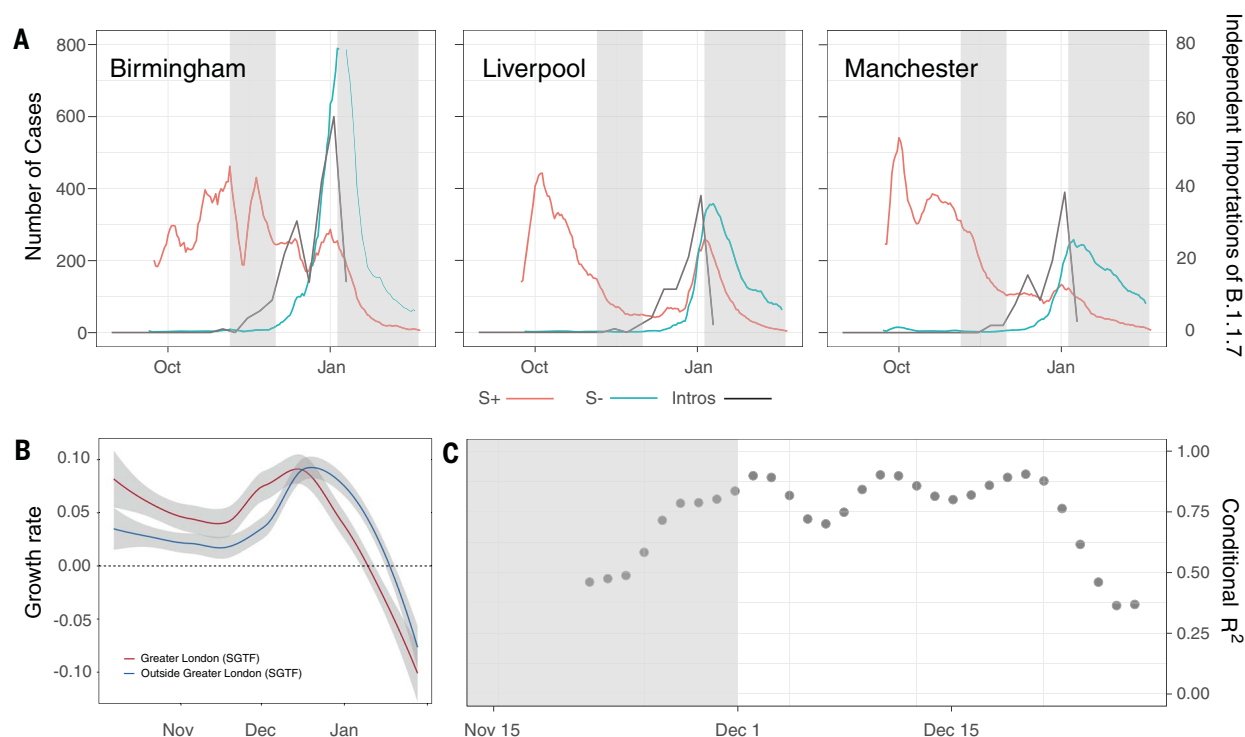


Fig. 4. Case growth rates of B.1.1.7 are correlated with human mobility and attack rates across the UK. (A) Seven-day rolling average number of cases reported that had the SGTF (green) and cases reported with non-SGTF (red) for three selected LTLAs, Birmingham, Liverpool, and Manchester (table S2). The black line indicates the weekly number of independent introductions estimated from the phylogeographic analysis. The gray shaded area indicates the timing of the second (5 November to 1 December) and start of the third (5 January) English lockdown. **(B)** Rolling 7-day average growth rates of SGTF cases in Greater London (red line)

and outside of Greater London (blue line). **(C)** Association between per-region (LTLA) difference between SGTF and non-SGTF case growth rates (corrected to account for differences through time in sampling intensity of SGTF cases) and number of B.1.1.7 importations into that region, as estimated from prevalence surveys and human mobility (gray dots) (Fig. 2C and materials and methods). The gray area shows the time of the second English lockdown. Modeling results for SGTF growth rates are shown in fig. S9, and regression results under different assumptions about the frequency of SGTF are shown in fig. S10.

cases in that LTLA (Birmingham, Liverpool, and Manchester are shown in Fig. 4A). We further calculated the per-region difference between SGTF and non-SGTF case growth rates [the estimated raw additive increase in SGTF growth rate is 0.0715, and the median multiplicative advantage is 1.576, assuming a generation time of 6.5 days, which is qualitatively similar to those reported previously (7, 9), with the caveat that generation times may differ between B.1.1.7 and other lineages (36, 37)] (fig. S11). The degree to which this difference is positively correlated with B.1.1.7 importation rate grew during the latter half of the November lockdown and remained very high [coefficient of determination (R^2) > 0.75] until mid-December, before declining (the trend remains when accounting for uncertainty in the estimated number of infections across Greater London) (Fig. 4C and fig. S12). This result is robust to the data and methods used to estimate per-location B.1.1.7 importation rates (figs. S9 and S10). Accounting for continued export of B.1.1.7 from Greater London and Kent can explain in part why estimates of the growth advantage of B.1.1.7 declined during the second half of December 2020, before the im-

plementation of tighter control measures (Tier 4, 20 December) (7, 9).

Human mobility and prior outbreaks as predictors of B.1.1.7 growth

The epicenter of SARS-CoV-2 transmission in the UK shifted during the November 2020 lockdown: between 1 September and 1 December 2020, ~80% of reported cases were reported outside London and southeast England, whereas those regions accounted for ~40% of all cases during 1 to 7 December. We sought to understand how, in each location, post-lockdown growth rates related to previous attack rates as well as travel inflow to that location. We investigated predictors of the increase in the relative frequency of B.1.1.7 genomes compared with that of other SARS-CoV-2 lineages (Fig. 5A) (7, 9). In a multivariate model, we found that about half of the variation in the increase in B.1.1.7 relative frequency between 2 and 16 December is associated with human mobility from Greater London and attack rates before the November lockdown (Fig. 5, B and C). UTLAs with lower previous attack rates tended to have faster-increasing B.1.1.7 frequencies. We repeated this analysis using SGTF

case frequency data and obtained similar results ($R^2 = 0.57$, $P < 0.001$) (fig. S13). However, neither human mobility nor pre-lockdown attack rate were significant predictors of later changes. Instead, change in the relative frequency of B.1.1.7 genomes after 17 December was best predicted simply by its frequency on that date ($R^2 = 0.13$, $P < 0.01$) (fig. S14), although a model identified through exhaustive search by using Bayesian information criteria (BIC) includes the “frequency of B.1.1.7 on 17 December,” an interaction between arrival time and “frequency of B.1.1.7 on 17 December,” and an interaction between incidence before the November lockdown and mobility from London (BIC 178.467; $R^2 = 0.68$; $P < 0.001$) (fig. S14). Mobility from Greater London remains a significant predictor of B.1.1.7 growth after controlling for population size by means of both a multivariate regression and model-selection by using exhaustive search with both BIC and AIC.

Conclusions, limitations, and future work

We found that the emergence of B.1.1.7 throughout the UK was associated with a high export frequency from a major source location that

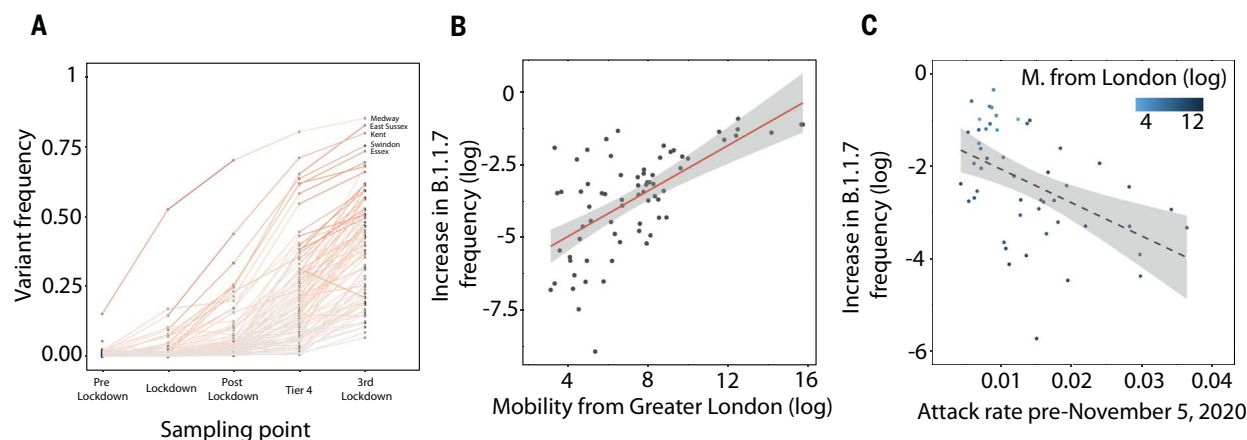


Fig. 5. (A) Frequency of B.1.1.7 at the UTLA level at different sampling times. Pre-lockdown, dates before 5 November; lockdown, 5 to 30 November; post-lockdown, 1 to 15 December; Tier 4, 16 to 31 December; and the most recent sampling point, 1 to 12 January (materials and methods). **(B)** Increase in the frequency of B.1.1.7 sampled genomes between 2 and 16 December

2020 is associated with mobility from Greater London. **(C)** Increase in the frequency of B.1.1.7 sampled genomes at the UTLA level is associated with previous attack rates in each location. Results for equivalent analyses of SGTF data are similar and are provided in the supplementary materials.

was identified only retrospectively. This pattern recapitulates at a national scale the role that international mobility played in the early spread of the SARS-CoV-2 pandemic (38–40). We conclude that the exceptionally rapid spatial spread and early growth rates of lineage B.1.1.7 likely reflect the combined effects of its higher intrinsic transmissibility (1, 7, 9) and the spatial structure of incidence and mobility before, during, and after the second lockdown in England (41).

Understanding what causes a new SARS-CoV-2 lineage to grow and replace preexisting lineages is a complex problem. In addition to virus genetic changes to relevant phenotypes (such as per-contact transmissibility, duration of infectiousness, and immune evasion), lineage replacement dynamics are likely affected by spatiotemporal heterogeneity in incidence, NPIs, prior infection, and among-region mobility (42). The role of the latter may be enhanced in the context of low or declining prevalence, as suggested by the frequency growth of lineage B.1.1.7 in the UK and Europe during summer 2020, which was associated with international travel (43–45). Evidence for the increased intrinsic transmissibility of B.1.1.7 is clear, but estimates have varied considerably [38 to 130% increase (7, 9)]. The growth potential of new SARS-CoV-2 variants will depend also on the average durations of their exposed and infectious phases, as well as their per-contact transmissibility (36). Our results indicate that exportations from a high-incidence epidemic source region raised early location-specific growth rate estimates across the UK (Fig. 4B), and that this effect declined through time. Similar trends have since been observed for lineage B.1.617.2 into the UK, after its importation from high-incidence regions onto

a background of low incidence and lockdown easing. This conclusion is relevant for the interpretation of the current and future estimates of the increased transmissibility of B.1.1.7 (and other variants of concern) in other countries [such as the United States and Denmark (3)]. Further epidemiological and experimental work is needed to discriminate transient demographic factors from the permanent contribution to increased transmissibility conferred by the mutations carried by B.1.1.7.

Although B.1.1.7 was first detected in Kent, UK, and is speculated to have accumulated its mutations during a chronic infection (2), because of the strong correlation between human mobility from those areas and date of B.1.1.7 detection elsewhere our results support the hypothesis that B.1.1.7 originated in Kent or Greater London. Further, our phylogeographic reconstruction shows early lineage dissemination from Kent and Greater London, indicating that B.1.1.7 spread through the UK from one dominant UK source region, as opposed to a large undetected epidemic elsewhere, which would likely have resulted in multiple introductions through international travel (16).

We demonstrate that large-scale and well-sampled genomic surveillance data can reveal the detailed spatial transmission dynamics of individual SARS-CoV-2 lineages and compensate for their comparatively low genetic diversity (46). To achieve a representative genomic sample, we used only samples from population-level testing rather than those from specific outbreak investigations. However, this approach does not fully mitigate reduced representation from populations less likely to seek testing (47), and there is some geographic variation in the proportion of cases sequenced (fig. S15). Greater London consistently has a

higher sampling proportion than other regions throughout the study timeframe. Although sampling biases cannot be wholly eliminated, the selection procedure used here, and our cross-validation between independent data sources (human mobility and SGTF datasets), help to ensure that our conclusions are robust. As SARS-CoV-2 genome sequencing efforts are accelerated worldwide, careful consideration and communication of sampling frameworks are needed to facilitate downstream epidemiological analyses (48). Spatial heterogeneity at the within-city scale was not accounted for in our analysis, consideration of which may further refine our understanding of the mechanisms of lineage emergence and invasion.

Coordinated and unified systems of genomic surveillance are needed worldwide to identify, track, and mitigate the transmission of SARS-CoV-2 variants of concern, including mechanisms to pair virus genomic and contact tracing data. Continuing rises in global incidence will increase the rate generation of viral genetic variation, and the accrual of higher levels of population immunity will create new selective pressures (49), the effects of which on virus evolution are difficult to predict (50–52). It is therefore critical to rapidly and accurately disentangle the contributions of genetic and ecological factors to the emergence of new SARS-CoV-2 variants. Geographic variation in vaccine availability, uptake, and delivery is expected to further contribute to variability in COVID-19 burden and the differential risk of disease resurgence (17, 53, 54), which can be mitigated through increased global access to vaccination and continued transmission control measures (52). Importation of SARS-CoV-2 lineages and variants from areas of high incidence will continue to pose a risk to those

regions that are reducing NPIs after having controlled infection.

REFERENCES AND NOTES

1. M. Chand *et al.*, Investigation of novel SARS-CoV-2 variant Variant of Concern 202012/01 (2021); https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/952490/Variant_of_Concern_VOC_202012_01_Technical_Briefing_4_England.pdf.
2. COVID-19 Genomics Consortium UK (CoG-UK), Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations (2020); <https://virological.org/t/563>.
3. T. Alpert *et al.*, *Cell* **184**, 2595–2604.e13 (2021).
4. A. O'Toole *et al.*, *Wellcome Open Res.* **6**, 121 (2021).
5. N. L. Washington *et al.*, *Cell* **184**, 2587–2594.e7 (2021).
6. H. Vöhringer *et al.*, on behalf of The COVID-19 Genomics UK (CoG-UK) consortium Lineage-specific growth of SARS-CoV-2 B.1.1.7 during the English national lockdown (2020); <https://virological.org/t/lineage-specific-growth-of-sars-cov-2-b-1-1-7-during-the-english-national-lockdown/575>.
7. E. Volz *et al.*, *Nature* **593**, 266–269 (2021).
8. T. C. Jones *et al.*, *Science* **373**, eab5273 (2021).
9. N. G. Davies *et al.*, *Science* **372**, eabg3055 (2021).
10. R. Challen *et al.*, *BMJ* **372**, n579 (2021).
11. Scientific Advisory Group for Emergencies, NERVTAG: Update note on B.1.1.7 severity, 11 February 2021 (2021); www.gov.uk/government/publications/nervtag-update-note-on-b117-severity-11-february-2021.
12. N. G. Davies *et al.*, *Nature* **593**, 270–274 (2021).
13. D. Frampton *et al.*, *Lancet Infect. Dis.* S1473-3099(21)00170-5 (2021).
14. A. Rambaut *et al.*, *Nature* **453**, 615–619 (2008).
15. C. Viboud *et al.*, *Science* **312**, 447–451 (2006).
16. L. du Plessis *et al.*, *Science* **371**, 708–712 (2021).
17. R. M. May, R. M. Anderson, *Math. Biosci.* **72**, 83–111 (1984).
18. D. J. Watts, R. Muhamad, D. C. Medina, P. S. Dodds, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 11157–11162 (2005).
19. C. J. E. Metcalf, C. V. Munayco, G. Chowell, B. T. Grenfell, O. N. Bjørnstad, *J. R. Soc. Interface* **8**, 369–376 (2011).
20. B. D. Dalziel *et al.*, *Science* **362**, 75–79 (2018).
21. M. S. Y. Lau *et al.*, *Nat. Ecol. Evol.* **4**, 934–939 (2020).
22. G. Chowell, L. Sattenspiel, S. Bansal, C. Viboud, *Phys. Life Rev.* **18**, 66–97 (2016).
23. D. S. Candido *et al.*, *Science* **369**, 1255–1260 (2020).
24. COVID-19 Genomics UK (CoG-UK) consortiumcontact@cogconsortium.uk, *Lancet Microbe* **1**, e99–e100 (2020).
25. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
26. O. G. Pybus *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15066–15071 (2012).
27. S. Dellicour *et al.*, *Mol. Biol. Evol.* **38**, 1608–1613 (2021).
28. UK Department of Health and Social Care, COVID-19 testing data: methodology note; www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note.
29. A. Kalkauskas *et al.*, *PLOS Comput. Biol.* **17**, e1008561 (2021).
30. S. Francis, Covid: Christmas Tier-4 heartbreak for Londoners. *BBC* 19 December 2020; www.bbc.co.uk/news/uk-england-london-55380644.
31. Public Health England, Official UK Coronavirus Dashboard; <https://coronavirus.data.gov.uk/details/testing?areaType=nation%26areaName=England>.
32. J. Bahl *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19359–19364 (2011).
33. G. Dudas *et al.*, *Nature* **544**, 309–315 (2017).
34. Public Health England, Investigation of novel SARS-CoV-2 variants of concern (2020); www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201.
35. S. A. Kemp *et al.*, *bioRxiv* [Preprint] 15 December 2020. doi:10.1101/2020.12.14.422555.
36. S. W. Park *et al.*, *bioRxiv* [Preprint] 5 May 2021. doi:10.1101/2021.05.03.21256545.
37. W. S. Hart *et al.*, *medRxiv* [Preprint] 30 May 2021. doi:10.1101/2021.05.27.21257936.
38. M. Chinazzi *et al.*, *Science* **368**, 395–400 (2020).
39. M. Worobey *et al.*, *Science* **370**, 564–570 (2020).
40. A. S. Gonzalez-Reiche *et al.*, *Science* **369**, 297–301 (2020).
41. S. Riley *et al.*, *bioRxiv* [Preprint] 22 January 2021. doi:10.1101/2021.01.20.21250158.
42. N. W. Ruktanonchai *et al.*, *Science* **369**, 1465–1470 (2020).
43. P. Lemey *et al.*, *Nature* **595**, 713–717 (2021).
44. E. Volz *et al.*, *Cell* **184**, 64–75.e11 (2021).
45. E. B. Hodcroft *et al.*, *Nature* **595**, 707–712 (2021).
46. J. Lu *et al.*, *Cell* **181**, 997–1003.e9 (2020).
47. Public Health England, COVID-19: Review of disparities in risks and outcomes (2020); www.gov.uk/government/publications/covid-19-review-of-disparities-in-risks-and-outcomes.
48. M. U. G. Kraemer *et al.*, *Epidemiol. Infect.* **147**, 1–7 (2018).
49. N. R. Faria *et al.*, *Science* **372**, 815–821 (2021).
50. C. M. Saad-Roy *et al.*, *Science* **372**, 363–370 (2021).
51. B. T. Grenfell *et al.*, *Science* **303**, 327–332 (2004).
52. R. N. Thompson, E. M. Hill, J. R. Gog, *Lancet Infect. Dis.* **21**, 913–914 (2021).
53. B. L. Rice *et al.*, *Nat. Med.* **27**, 447–453 (2021).
54. N. B. Masters *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 28506–28514 (2020).
55. V. Hill, CoG-UK/B.1.1.7_spatial_analysis_UK: V1.0.0. Zenodo (2021); doi:10.5281/zenodo.5085521.
56. NHS Foundation Trust, Tier 4 Restrictions for London, Bedfordshire and Luton; www.elft.nhs.uk/News/Tier-4-Restrictions-for-London-Bedfordshire-and-Luton.

ACKNOWLEDGMENTS

We thank all involved in the collection and processing of SARS-CoV-2 testing and genomic data. We also thank Public Health England (PHE) for making anonymized epidemiological data available for this analysis. We thank the Office of National Statistics (ONS) for their effort to publish the Coronavirus (COVID-19) Infection Surveys in real time. **Funding:** V.H. was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) (grant BB/M010996/1). A.R. acknowledges the support of the Wellcome Trust (Collaborators Award 206298/Z/17/Z-ARTIC network) and the European Research Council (grant agreement 725422-ReservoirDOCS). M.U.G.K. acknowledges support from the Branco Weiss Fellowship. M.U.G.K. and S.D. acknowledge support from the European Union's Horizon 2020 project MOOD (grant agreement 874850). O.G.P. and M.U.G.K. acknowledge support from the Oxford Martin School. A.L.B., S.V.S., and M.U.G.K. acknowledge support from the Rockefeller Foundation and Google.org. C.R. was supported by a Fondation Botnar Research Award (Programme grant 6063) and UK Cystic Fibrosis Trust (Innovation Hub Award 001). A.L.B. acknowledges support from the Biotechnology and Biological Sciences Research Council (BBSRC) [grant BB/M011224/1]. S.D. acknowledges support from the Fonds National de la Recherche Scientifique (FNRS, Belgium). G.B. acknowledges support from the Research Foundation–Flanders (Fonds voor Wetenschappelijk Onderzoek–Vlaanderen, GOE1420N and G098321N) and from the Interne Fondsen KU Leuven/Internal Funds KU Leuven under grant agreement C14/18/094. CoG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research and Innovation (UKRI), the National Institute of Health Research (NIHR), and Genome

Research Limited, operating as the Wellcome Sanger Institute. A.O. is supported by the Wellcome Trust Hosts, Pathogens and Global Health Programme (grant grant.203783/Z/16/Z) and Fast Grants (award 2236). S.B. is supported by the Clarendon Scholarship, University of Oxford and NERC DTP (grant NE/S007474/1). N.R.F. acknowledges support from Wellcome Trust and Royal Society (Sir Henry Dale Fellowship: 204311/Z/16/Z) and Medical Research Council–São Paulo Research Foundation CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0). The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission or any of the other funders. **Author contributions:** M.U.G.K., A.R., V.H., C.R., S.D., S.V.S., and O.G.P. conceived and planned the research. M.U.G.K., V.H., A.R., C.R., S.D., S.B., G.B., B.K., A.L.B., S.D., S.G., and S.V.S. analyzed the data. M.U.G.K. and O.G.P. wrote the first draft. All authors contributed to writing and interpreting the results. M.U.G.K., A.R., S.V.S., and O.G.P. jointly supervised this work. **Competing interests:** O.G.P., A.R., and A.O. have undertaken consulting for AstraZeneca relating to the genetic diversity and classification of SARS-CoV-2 lineages. S.V.S. is a paid consultant with Pandefense Advisory and Booz Allen Hamilton; is on the advisory board for BioFire Diagnostics Trend Surveillance, which includes paid consulting; and holds unexercised options in Iliad Biotechnologies. These entities provided no financial support associated with this research; did not have a role in the design of this study; and did not have any role during its execution, analyses, interpretation of the data, and/or decision to submit. **Data and materials availability:** Aggregated epidemiological data used in this study are available from <https://coronavirus.data.gov.uk/details/download>. SARS-CoV-2 infection survey data are available from the Office of National Statistics (ONS) and available at www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/coronaviruscovid19infectionsurveydata. Raw epidemiological SARS-CoV-2 line list data are available from Public Health England (PHE) and aggregated statistics are available via Github. All genomes, phylogenetic trees, and basic metadata are available at the CoG-UK Consortium website (www.cogconsortium.uk/data). The O₂ aggregated, anonymized mobile data insights dataset is not publicly available owing to stringent licensing agreements. Information on the process of requesting access to the O₂ aggregated mobile data insights dataset is available at o2@businesso2.co.uk. The Google COVID-19 Aggregated Mobility Research Dataset is not publicly available owing to stringent licensing agreements. Information on the process of requesting access to the Google mobility data are available from sadilekadam@google.com. Code and data are available on the following GitHub repository https://github.com/CoG-UK/B.1.1.7_spatial_analysis_UK and permanently on Zenodo (55). This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6557/889/suppl/DC1
Materials and Methods
Figs. S1 to S19
Tables S1 to S3
COVID-19 Genomics UK (CoG-UK) Consortium Author List
References (57–72)
MDAR Reproducibility Checklist

16 April 2021; accepted 12 July 2021
Published online 22 July 2021
10.1126/science.abj0113

Context-specific emergence and growth of the SARS-CoV-2 Delta variant

<https://doi.org/10.1038/s41586-022-05200-3>

Received: 10 December 2021

Accepted: 5 August 2022

Published online: 11 August 2022

Open access

 Check for updates

John T. McCrone^{1,29}, Verity Hill^{1,29}, Sumali Bajaj^{2,29}, Rosario Evans Pena^{2,29}, Ben C. Lambert³, Rhys Inward^{2,4}, Samir Bhatt^{4,5}, Erik Volz⁴, Christopher Ruis⁶, Simon Dellicour^{7,8}, Guy Baele⁸, Alexander E. Zarebski², Adam Sadilek⁹, Neo Wu⁹, Aaron Schneider⁹, Xiang Ji¹⁰, Jayna Raghwan², Ben Jackson¹, Rachel Colquhoun¹, Áine O'Toole¹, Thomas P. Peacock^{11,12}, Kate Twohig¹², Simon Thelwall¹², Gavin Dabrera¹², Richard Myers¹², The COVID-19 Genomics UK (COG-UK) Consortium*, Nuno R. Faria^{2,4,13}, Carmen Huber¹⁴, Isaac I. Bogoch^{15,16}, Kamran Khan^{14,16,17}, Louis du Plessis^{2,18,19}, Jeffrey C. Barrett²⁰, David M. Aanensen²¹, Wendy S. Barclay¹¹, Meera Chand¹², Thomas Connor^{22,23,24}, Nicholas J. Loman²⁵, Marc A. Suchard²⁶, Oliver G. Pybus^{2,27,28,30}, Andrew Rambaut^{1,30} & Moritz U. G. Kraemer^{2,28,30}✉

The SARS-CoV-2 Delta (Pango lineage B.1.617.2) variant of concern spread globally, causing resurgences of COVID-19 worldwide^{1,2}. The emergence of the Delta variant in the UK occurred on the background of a heterogeneous landscape of immunity and relaxation of non-pharmaceutical interventions. Here we analyse 52,992 SARS-CoV-2 genomes from England together with 93,649 genomes from the rest of the world to reconstruct the emergence of Delta and quantify its introduction to and regional dissemination across England in the context of changing travel and social restrictions. Using analysis of human movement, contact tracing and virus genomic data, we find that the geographic focus of the expansion of Delta shifted from India to a more global pattern in early May 2021. In England, Delta lineages were introduced more than 1,000 times and spread nationally as non-pharmaceutical interventions were relaxed. We find that hotel quarantine for travellers reduced onward transmission from importations; however, the transmission chains that later dominated the Delta wave in England were seeded before travel restrictions were introduced. Increasing inter-regional travel within England drove the nationwide dissemination of Delta, with some cities receiving more than 2,000 observable lineage introductions from elsewhere. Subsequently, increased levels of local population mixing—and not the number of importations—were associated with the faster relative spread of Delta. The invasion dynamics of Delta depended on spatial heterogeneity in contact patterns, and our findings will inform optimal spatial interventions to reduce the transmission of current and future variants of concern, such as Omicron (Pango lineage B.1.1.529).

The SARS-CoV-2 pandemic has been characterized by the appearance and spread of genetically distinct virus variants that are associated with faster spread than pre-existing lineages. In May 2021, the World Health Organization (WHO) announced a new variant of concern (VOC), designated Delta. Delta became the variant primarily responsible for a wave of transmission and mortality in India in early-to-mid 2021, replacing Alpha (Pango lineage B.1.1.7) and Kappa (Pango lineage B.1.617.1) in the process^{3,4}. Studies indicate that Delta has increased transmissibility⁵, rates of hospitalization⁶ and immune evasion⁷ compared with Alpha^{8,9}, the variant that was previously dominant in many countries. These phenotypes are attributed to a constellation of 30 mutations across the virus genome (Supplementary Table 2) compared to the Wuhan-Hu-1 reference sequence, including: the spike mutation P681R in the furin cleavage site, which is thought to increase the speed and efficiency

with which the virus fuses with host cells^{10,11}; mutation L452R in the receptor-binding domain, which is thought to reduce neutralization by antibodies¹²; and the nucleocapsid mutation R203M, which is thought to increase virion infectivity¹³. Delta disseminated rapidly from India to locations worldwide and has been detected in 174 countries as of 12 April 2022 (https://cov-lineages.org/global_report.html). Delta became the dominant lineage in the UK by mid-May 2021¹⁴, and similar increases in frequency were observed worldwide (for example, ref. 15).

The emergence of Delta in the UK occurred in the context of a heterogeneous landscape of prior immunity (from infection and vaccination) and non-pharmaceutical interventions (NPIs). Here we examine virus genomes generated from a random sample of all COVID-19-positive tests with PCR with reverse transcription (RT-PCR) C_t values greater than 30, collected during community-based testing in England between

A list of affiliations appears at the end of the paper.

12 March 2021 and 15 June 2021. Our data include 52,992 Delta genomes from England with known dates and locations of sampling, representing more than 40% of all positive lateral flow and PCR tests in England during the study period (see Methods and details on case data at <https://coronavirus.data.gov.uk/details/about-data>; an estimated 27% of COVID-19 infections were detected during the study period¹⁶). Using these data, we evaluate the effectiveness of policies in reducing international importations and how they contributed to the establishment and local transmission dynamics of Delta in England. We then investigate, at a high spatial resolution, how human mobility contributed to context-specific growth of Delta in England.

International importations of Delta

To provide a global context for the emergence of Delta in the UK, we first conducted a phylodynamic analysis by uniformly subsampling SARS-CoV-2 Delta genome sequences by collection date between 4 March 2021 and 15 June 2021 ($n = 975$). Details of the origin and spread of Delta within India are uncertain. A substantial increase in genomic surveillance across the country would probably facilitate the study of the emergence and expansion of Delta there, but is outside the scope of this work. To put the UK epidemic into context, we estimated the time of the most recent common ancestor (TMRCA) of Delta globally to be 19 October 2020 (95% highest posterior density (HPD) interval: 6 September 2020 to 29 November 2020). The relative frequency of Delta in India does not appear to increase substantially until March 2021 (Fig. 1), coinciding with a rapid expansion in case numbers there (Extended Data Fig. 1) and a decline in the relative frequency of genomes assigned to Kappa, a sibling lineage of Delta (<https://www.gisaid.org/>). Genomic surveillance in India revealed that several sub-lineages of Delta existed prior to its expansion in March¹⁷ (Fig. 1a,b). This standing diversity is consistent with undetected transmission of Delta in India between late 2020 and March 2021.

We evaluated the global dissemination of Delta from March 2021 by multiplying, for each country, estimated numbers of SARS-CoV-2 cases, relative frequencies of Delta, and relative numbers of outward international passengers (estimated exportation intensity (EEI); Methods). The EEI of Delta increased rapidly during March 2021 and was highest around late April 2021, coinciding with its peak incidence in India (Extended Data Fig. 1). The subsequent rapid spread of Delta in the USA, Russia, UK, Mexico and elsewhere and its decline in India resulted in the former locations becoming main exporters of Delta by June 2021 (Extended Data Fig. 1), corroborating global trends in Delta phylogeography (Fig. 1a) and reported cases (Fig. 1b). Similar patterns of rapidly changing foci of international dissemination were observed for the initial wave of SARS-CoV-2 in 2020¹⁸.

To evaluate the temporal dynamics of Delta importation into England and to reconstruct its subsequent local spread, we conducted a travel history-aware Bayesian phylogeographic analysis¹⁹ of 93,649 Delta sequences, from GISAID and COVID-19 Genomics UK Consortium (COG-UK), which accounts in part for the phylogenetic uncertainty inherent in SARS-CoV-2 phylogenies¹⁸. To render the analysis tractable, we split the full tree into three independent subtrees (Fig. 1a) prior to phylogeographic analysis. Virus genomes were generated from around 40–60% of all positive cases in England during the emergence of Delta between March and May 2021²⁰ (Fig. 1c) and combined with metadata on the locations (at the upper tier local authority (UTLA) level), enabling us to trace the introduction of the virus and characterize its spread at a high spatio-temporal resolution²⁰.

We estimated a minimum of 1,458 (95% HPD 1,398–1,513) separate international introductions of Delta into England, with approximately half inferred to have originated from India (posterior mean 56.5%; 95% HPD 53.7%–59.1%). We found that the majority of Delta genomes in England can be traced back to introductions that are inferred to have occurred prior to the implementation of a mandatory hotel quarantine

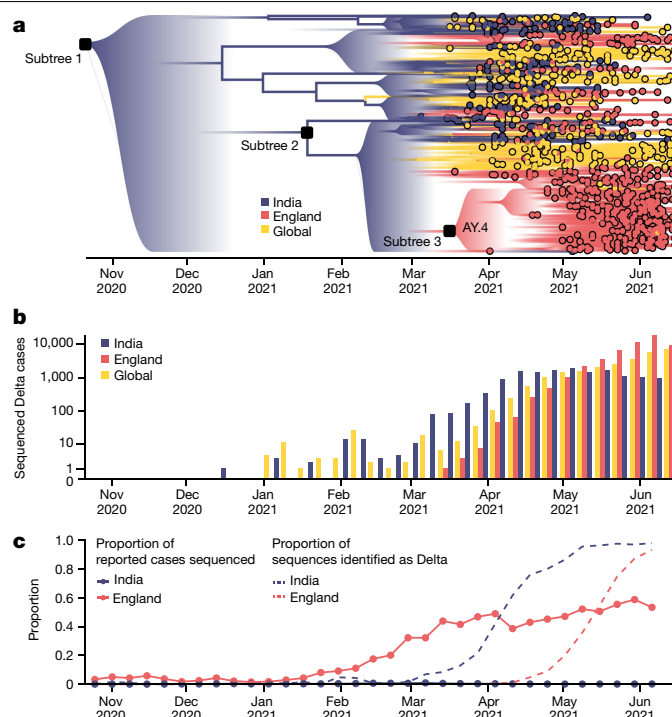


Fig. 1 | The emergence and rapid geographic expansion of Delta.

a, Time-calibrated phylogenetic reconstruction of Delta based on 1,000 sequences subsampled from 93,649 sequences from 100 countries (52,992 from England). The tree was split into three subtrees (with $n = 28,783$, 28,715 and 36,151 sequences, respectively) prior to full analysis. The roots of these three subtrees, and of lineage AY.4 are labelled with black squares. Lineage colours represent the inferred countries and/or regions where transmission occurred. **b**, The number of sequenced cases of Delta per week in India, England and globally, where 'global' refers all countries other than England and India. **c**, The proportion of sequenced, reported positive cases in India and England (solid lines, $n = 52,992$ sequences from England, corresponding to 84% of all sequences from the UK during the study period) and the proportion of sequenced cases classified as Delta in India and England (dashed lines).

for people arriving from India on 23 April 2021 (posterior mean 84.3%; 95% HPD 77.8–90.4%). During this period, 90.0% of introductions are inferred to have originated from India (95% HPD 86.5–93.1%). These inferred importation dynamics closely match data on individual travel histories obtained from infected incoming international passengers (origin–destination travel histories are available for 1.4% of genomes; $n = 770$) (Fig. 2b and Extended Data Fig. 2).

The high variation in sampling intensity among countries (specifically, the higher sampling intensity in England than in other countries) means that the true number of importations into England is probably much larger than that inferred from phylogeographic analysis alone (Fig. 1b,c; see the related discussion in the context of the first wave in the UK¹⁸). For example, the AY.4 lineage (Fig. 1a) comprises 42,445 sequences and was probably imported to England many times. We investigated AY.4 by pairing genomic data with contact tracing data collated by Public Health England (now the UK Health Security Agency). During the study period we found 61 sequenced cases with AY.4 had a travel history from India and 140 had a travel history from elsewhere, similar to the time-varying importation dynamics seen across the entire dataset (Fig. 2a and Extended Data Fig. 2). Thus, sampling heterogeneity means that the number of importations estimated from phylogenetic analysis represents a lower bound on the true number¹⁸.

To investigate the importation of Delta into England specifically, and to cross-validate the results above using independent data, we use the estimated importation intensity (EII) of Delta to England over time^{18,21}.

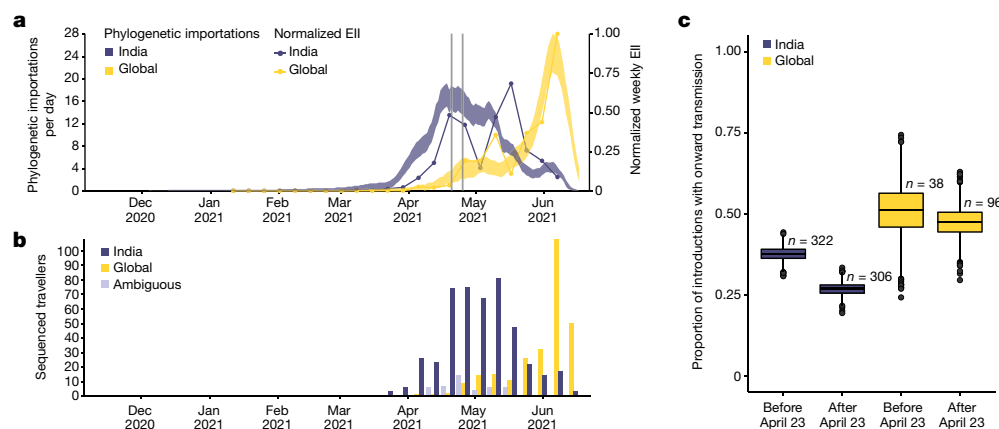


Fig. 2 | Timing of importations of Delta into England. **a**, The estimated daily number of importations of Delta from India (blue shaded area) and other countries (yellow shaded area), inferred from phylogenetic analysis. Shaded areas show 95% HPDs of the estimate. Blue and yellow lines show the EII of Delta, obtained by combining data on human movements, cases and prevalence of Delta, normalized to the same scale as the phylogenetic estimates. Grey vertical lines show the timing of the announcement of travel restrictions from India to England (18 April 2021) and their implementation on 23 April 2021. **b**, Temporal distribution of genome sequences from cases with a known travel history from India (blue) and other countries (yellow). Isolates with recent

travel to both India and other countries are considered ambiguous (grey). **c**, The proportion of all virus introductions that show evidence of onward transmission in the UK, estimated separately for weeks before and after the implementation of hotel quarantine (23 April 2021) and stratified by the location of origin (India, blue; other countries, yellow). The box plot displays the median, with lower and upper hinges representing the 25th and 75th percentiles of each group. Whiskers extend to the most extreme data points no more than 1.5 times the interquartile range beyond each hinge. The number of observations in each group is annotated above each box.

The EII is a metric of Delta importation that represents trends in the number of Delta cases arriving in the country, irrespective of whether or not those cases result in local transmission. This is distinct from the phylogenetic analysis above, which better captures trends in the number of Delta introductions that did lead to forward transmission in England. The EII combines (1) weekly reported cases, (2) weekly prevalence of Delta genomes, and (3) weekly aggregate human mobility (inferred from mobile phone data) into England through direct connections (Fig. 2a; see refs.^{18,21} for related approaches). The EII from India increased rapidly in April 2021 following the rise in cases in India, and remained high until the end of May 2021. EII correlates strongly with the inferred importations from genomic data but is weaker for imports from India after the implementation of hotel quarantine (Extended Data Fig. 3). We then estimate, from genomic data, the proportion of inferred importations that led to observed onward transmission in the community (defined as at least one ancestral node in England), stratified by location of origin in the three weeks pre- and post-implementation of hotel quarantine. We find that pre-quarantine, 37.7% (95% HPD 34.0%–41.7%) of importations from India led to observable onward transmission. After the implementation, the fraction of importations from India leading to observed onward transmission dropped to 26.9% (95% HPD 23.0%–30.2%) (Fig. 2c). For comparison, the proportion of introductions from other locations leading to onward transmission did not change during the two periods (around 50%) (Fig. 2c). The decrease in onward transmission is most apparent in importations associated with travel history, which suggest the trend is driven by the implementation of hotel quarantine and not temporal biases in lineage detection (Extended Data Fig. 3). Even though we observe that the implementation of hotel quarantine was effective in reducing onward transmission, substantial importation had already occurred before its implementation and additional introductions from other countries probably further accelerated the spread of Delta in England from May onwards (Fig. 2a).

There are several reasons why some importations led to onward transmission within England after the implementation of hotel quarantine for arriving travellers: (1) a separate terminal for arrivals from mandatory quarantine countries was not opened at the UK's largest airport (London Heathrow) until 1 June 2021²², so arriving passengers may have mixed with others before entering mandatory quarantine;

(2) individuals may have become infectious and transmitted the virus only after leaving quarantine, either owing to an unusually long latent period or within-group transmission during the quarantine period; (3) individuals may have infected others on a connecting flight where the connecting airport did not require hotel quarantine; (4) there were exemptions to hotel quarantine that may have led to onward transmission in the community²³.

Lineage dynamics of Delta in England

Importations of Delta occurred on a background of relaxation of social distancing in England: on 12 April 2021, outdoor dining and non-essential retail reopened, and on 17 May 2021, restrictions on indoor dining and international travel were relaxed²⁴. The relative frequency of Delta genomes in England increased rapidly during May and the number of reported COVID-19 cases subsequently increased²⁵ (Fig. 1c). Initially, Delta transmission clusters were concentrated in the North West region of England and were commonly associated with returning travellers²⁶. We sought to reconstruct the dispersal dynamics of independently imported Delta transmission lineages within England, in the context of changing NPIs.

We analysed all identified Delta transmission lineages in England and inferred their history of dissemination among subnational regions (UTLAs). Sequence sampling was highly representative of reported cases at the UTLA level (Extended Data Fig. 4), making possible the reconstruction of virus movements across England using continuous phylogeography approaches²⁷. We observe high heterogeneity among UTLAs in the numbers of Delta introductions from other English regions (Fig. 3a), with Lancashire and Greater Manchester each receiving more than 2,000 estimated independent introductions and Torbay receiving only 9. The majority ($n = 11,960$) of Delta sequences in England belong to a single transmission lineage (lineage I) (Fig. 3d), which was sampled mostly in Greater Manchester and Lancashire, and we observe many short-range lineage movements among UTLAs in these areas (Fig. 3a). Greater London also received many Delta cases from elsewhere in England (Fig. 3a), as expected, given its population size and connectedness to other metropolitan areas²¹. Transmission lineages II and III each comprised 3,000–4,000 genomes; lineage II is

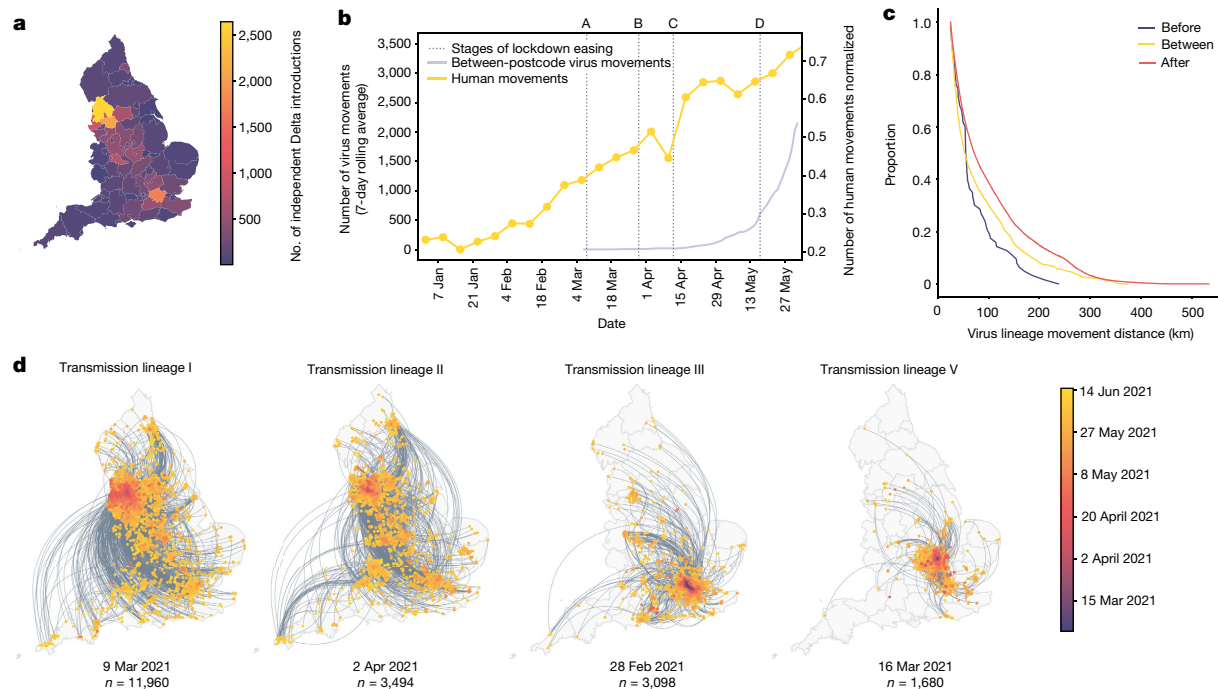


Fig. 3 | Introductions and regional dynamics of Delta transmission lineages. **a**, The number of independent introductions per UTLA in England derived from continuous phylogeographic analysis of all Delta transmission lineages with more than five sequences. **b**, Trends in aggregate intra- and inter-UTLA mobility normalized to pre-pandemic levels (yellow) and the number of virus lineage movements between postcode districts. Letters denote stages of lockdown easing: A, schools re-open and limited outdoor mixing between households is permitted (8 March 2021); B, ‘stay at home’ directive is lifted, more outdoor mixing (up to six people from two households) is allowed (29 March 2021); C, non-essential retail, holiday lets and campsites re-open and outdoor dining is permitted (12 April 2021); D, indoor hospitality

re-opens and indoor mixing is permitted (17 May 2021). **c**, The proportion of virus lineage movements between postcodes more than 25 km apart: the y axis denotes the proportion of movements that are less than or equal to the value on the x axis. This is shown for movements before lockdown easing (C (blue)), between C and D (yellow) and D (red). C and D are defined in **b**. **d**, Virus lineage movements inferred by continuous phylogeographic analysis for four example large transmission lineages (transmission lineages IV, VI and VII are shown in Extended Data Fig. 5). The direction of lineage movement is anticlockwise, and dots represent the start and end points of movements, coloured by inferred date. The size of each lineage and its inferred TMCA date are shown below each map. Distance kernels for each lineage are shown in Extended Data Fig. 7.

distributed across multiple urban areas (especially in the North West), whereas the latter is focused in Greater London and the South East (Fig. 3d). We also highlight transmission lineage V (Fig. 3d), originally centred in Bedfordshire, the location of one of the first Delta outbreaks in England that was subject to surge testing²⁸ (Extended Data Fig. 5).

In early May 2021, the number of virus lineage movements among locations accelerated (Fig. 3b and Extended Data Fig. 5), showing that increase in Delta frequency (Fig. 1c) was associated with regional dissemination. This spread occurred on the background of relaxing NPIs and increased mixing (between mid-January and June 2021, mobility in England increased from 20% to 70% of its pre-pandemic level, and estimated mean daily contacts increased²⁹ from approximately 2 to 5). By contrast, the initial wave of SARS-CoV-2 introductions to the UK in spring 2020 occurred during a period of increasing travel and social restrictions¹⁸. In general, we find that as NPIs were progressively relaxed over time, long-range viral lineage movements comprised an increasing proportion of all movements (Fig. 3c).

For the seven largest Delta transmission lineages in England (I–VII) we observed approximately three times more exports from Greater Manchester than from Greater London. Further, we observe that Bolton, Blackburn, Salford, Bury and Greater Manchester had on average higher than expected numbers of exportations for their population sizes (Extended Data Fig. 6). This difference matches early epidemiological data: the largest and earliest Delta outbreaks were in North West England (on 21 May, Bolton had 452 cases per 100,000 population and Greater London had 21.6 cases per 100,000 population) (<https://coronavirus.data.gov.uk/>; Methods). Introductions of Delta into other, smaller urban areas also spread rapidly (for example, transmission

lineage V) (Fig. 3d) and were important for the propagation of the variant across England. We observe a spatial structure of the seven largest lineages; the frequency of viral movements declined with the distance away from the origin location but we also observe a second peak at around 260 km (similar to the distance between Greater London and Greater Manchester) (Extended Data Fig. 7). Although North West England was a focus of early Delta transmission, the Delta epidemic in England derived from many successful independent international importations. Each of the main Delta transmission lineages in England grew at a similar rate (Extended Data Fig. 8). By contrast, the Alpha variant expanded across the UK from a single origin in South East England²¹. The spatial expansion of Delta transmission lineages plateaued after early June, when most UTLAs had established Delta transmission and the relative frequency of Delta genomes in England had exceeded 90% (<https://covid19.sanger.ac.uk/lineages/raw>).

Although Scotland, Wales or Northern Ireland were not included here, case count data suggest that cities in England (<https://coronavirus.data.gov.uk/details/download>) were the main source of the expanding Delta epidemic in the UK; owing to this source-sink structure we do not anticipate that omitting these countries substantially affects our reconstruction of epidemic dynamics in England (of the Delta genomes available before 15 June 2021, 57,592 were from England, 9,738 were from Scotland, 1,067 were from Wales and 325 were from Northern Ireland).

Factors contributing to the growth of Delta

Regional and international heterogeneity in incidence, vaccination and human mobility determine the dynamics of infectious diseases³⁰,

including those of SARS-CoV-2^{18,31,32}. We used a combination of epidemiological, aggregate human mobility and genomic data to test whether relaxation of NPIs, virus importations and vaccination rates correlate with local Delta growth rates. To do so, we developed a hierarchical Bayesian model to estimate the effect of these factors on the weekly relative growth of Delta (that is, the weekly change in the observed proportion of Delta genomes on a log odds scale³³) at the UTLA level for England. Models for estimating the increase in transmissibility of new variants are typically based on increases in relative frequency^{1,33} but rarely take into account other potential confounding factors, such as variation in population behaviour, vaccination rates or numbers of independent virus introductions.

In general, growth rates varied widely across locations and weeks in England (Fig. 4a and Extended Data Fig. 10). Our model estimates that the most important tested predictor of the variation in growth of Delta (relative to Alpha) across UTLAs in England was within-UTLA mixing (that is, relative changes in weekly within-UTLA human mobility, compared with the pre-pandemic period) (Fig. 4a,b and Supplementary Tables 5 and 6). The importance of within-UTLA mobility as a factor during the emergence of a new variant (until Delta prevalence reached 85% (25%–75% quantiles: 78%–96%)) is unsurprising, as pre-emptive restrictions on movement and social mixing slow the emergence of new pathogens or variants³⁴ (see counterfactual scenarios in Extended Data Fig. 10); the cost/benefit ratio of such restrictions will of course depend on the specific context of variant emergence. The relaxation of NPIs therefore increased both within- and among-region transmission (see Fig. 3c). Other European countries did not observe such a rapid increase in Delta relative frequency during May 2021 (<https://www.gisaid.org/>); possible reasons for this difference are (1) during this period, levels of mobility and mixing (both local and regional) were lower in those countries and/or (2) those countries potentially received fewer international importations of Delta (86,489 passengers flew from India to the UK between March and June, whereas 43,515 flew to Germany, and 16,688 flew to France, during the same period).

Model fit did not improve when including weekly numbers of independent viral introductions estimated from genomic data or vaccination rates (Supplementary Table 6). We refrained from translating estimates of the growth rate of Delta relative frequency into differences in the reproduction number, as this is sensitive to assumptions about the generation time of the variant, which is also influenced by NPIs and immunity³⁵. Further studies should consider estimating the generation times of VOCs in specific contexts of immunity, NPIs and household structure³⁶ to accurately translate relative growth rates into R_t .

Discussion, limitations and future work

We find that growing epidemics of SARS-CoV-2 Delta worldwide led to a wave of importations of the VOC into England, initially from India, and later from other countries. These importations found fertile ground as they arrived in a context of easing social restrictions, and consequently expanded rapidly across England. Much transmission occurred in unvaccinated and younger populations (<https://coronavirus.data.gov.uk/details/download>), and high levels of Delta transmission within the UK led to onward dissemination of the variant to other countries (see for example, ref. ³⁷). By pairing the phylogenetic results with contact tracing data we conclude that hotel quarantine measures were effective in reducing onward transmission of imported Delta cases in England. However, after 21 May 2021, we found that levels of local social mixing in England—and not numbers of importations—were associated with faster relative growth of Delta. At that point, the independently introduced transmission lineages grew at a similar pace; details of their geographic distribution and expansion will support future work defining the optimal spatial interventions to reduce transmission of VOCs in England.

Compared with the Alpha variant, which arose and spread from a single location in South East England¹⁸, the expansion of the Delta

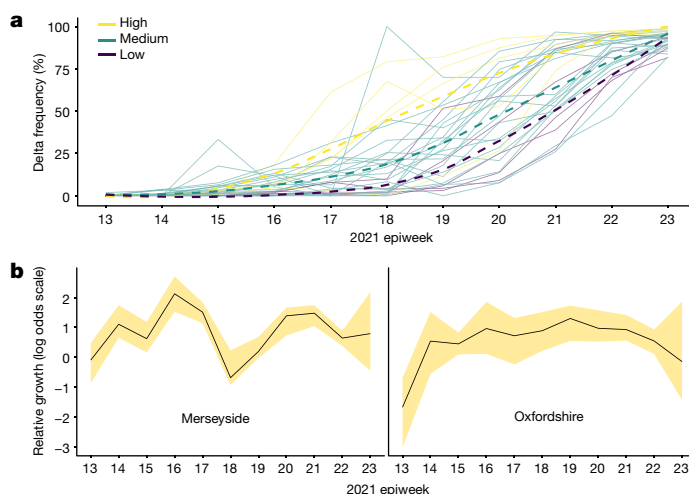


Fig. 4 | Variation in Delta growth rates across UTLAs in England. a, The increase in Delta frequency compared with Alpha at the UTLA level. UTLAs are coloured according to the level of average within-UTLA mobility: high, the five UTLAs with the highest within-location mobility; low, the five UTLAs with the lowest within-location mobility; medium, the remaining UTLAs. Solid lines show data for given UTLAs; dashed lines show LOESS curves fit to the data for each mobility category. **b**, Examples of weekly growth (the solid line corresponds to posterior medians) of UTLAs with high (left) and low (right) within-UTLA mobility. The shaded regions represent the corresponding 95% Bayesian credible intervals (2.5th and 97.5th quantiles of the posterior distribution). In **a**, data are shown only for UTLAs with 500 or more sequenced samples.

variant was predominantly owing to exports from the North West (Fig. 3). Analysis of the Alpha variant and of the first wave of SARS-CoV-2 in spring 2020 suggested that Greater London had a substantial role in spreading SARS-CoV-2 across England, as expected, given it is the largest city in England by far, and is highly connected by road, rail and air to other locations. However, Greater London was less important in the spread of Delta, even after Delta had become established there. This indicates the importance of founder effects; where a VOC first becomes established within a country may have a strong effect on subnational spatial dissemination, and this information is useful for planning localized interventions.

Furthermore, although there are intrinsic differences in transmissibility between VOCs, the role of NPIs and levels of immunity from prior infection or vaccination also affect their dynamics. After the start of the first UK national lockdown during the first wave of infections in 2020, lineage movements were severely curtailed and most lineages went extinct¹⁸; by contrast, the viral movement of Delta lineages increased after the relaxation of NPIs, accompanied with a subsequent rise in positive cases (Fig. 3). For the most recent VOC, Omicron, NPIs have remained relatively stable throughout England, and the increase in cases of the Omicron subvariant BA.2 in more rural areas in the South West in February and March 2022 has been speculated to be a result of lower infection rates there during the previous Omicron BA.1 wave³⁸ (December 2021–January 2022). Therefore, the effect of seeding location, immunity from previous waves or vaccination and NPI changes all contribute to the large and continued spatial heterogeneity in the spread of VOCs.

The undetected genetic diversity and uneven sampling of Delta in India make the precise estimation of the number of importations to England difficult to achieve from genetic data alone²⁷ (Extended Data Fig. 9). However, our phylogenetic estimates correlate strongly with estimates derived from independent data on case incidence, Delta prevalence and arriving travellers (EII) (Methods and Fig. 2c) during the period before quarantine policies were announced. Fortunately,

additional contact tracing data from public health agencies enabled us to overcome the limitations inherent in the unevenly sampled global virus genomic dataset and provide additional confidence in our findings.

Our statistical analysis shows that higher Delta growth rates were positively associated with levels of local mixing in England. The existence and magnitude of future NPIs needed to reduce the healthcare burden of future VOCs to sustainable levels will depend on the local levels of population immunity (from vaccination and prior infection). Future work should focus on identifying the factors that are most conducive to spread in particular contexts (for example, high versus low NPI regimes and across levels of population immunity³⁹) so that responses can be planned accordingly. This will require a better characterization of the distribution and variation of infectiousness over time, and an understanding of the virus generation time in different behavioural contexts⁴⁰—for example, among individuals who are vaccinated or unvaccinated and/or those who have had previous exposure to SARS-CoV-2 (including knowledge of the lineage or variant). To do so effectively will require investments in large-scale and coordinated serological studies⁴¹, especially for VOCs with the ability to evade immunity.

Even though reporting of case numbers, virus genomic surveillance, sampling strategies and mobile phone penetration differ across the world, our estimates can still provide qualitative insights into the trends in the source locations and the rates of international importation. Including estimates of probable importations in disease surveillance programmes may help support public health decision making⁴², and further improvements in these estimates can be achieved when global health surveillance systems are more integrated, and investments in data generation and capacity are linked directly to paired genomic–epidemiological analytical pipelines.

The detail with which we document the spatial invasion process of Delta in England provides an opportunity to re-examine how more spatially targeted interventions can support COVID-19 control in the future. Our work highlights the relative importance of local (within country) behavioural and mobility changes in determining the speed at which Delta spread in England; such changes will probably be more important than international travel restrictions during the emergence of future variants. Globally coordinated data and analytical pipelines that capture heterogeneity in virus circulation, immunity and policy responses will be necessary to produce the insights necessary to curb the spread of emerging infectious diseases and new variants. However, they can only be successful when integrated into a public health framework that can respond and adapt rapidly to public health threats during their emergence.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-022-05200-3>.

- Vöhringer, H. S. et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021).
- Earnest, R. et al. Comparative transmissibility of SARS-CoV-2 variants Delta and Alpha in New England, USA. *Cell Rep. Med.* **3**, 100583 (2022).
- Kupferschmidt, K. & Wadman, M. Delta variant triggers dangerous new phase in the pandemic. *Science* <https://www.sciencemag.org/news/2021/06/delta-variant-triggers-dangerous-new-phase-pandemic> (2021).
- Vaidyanathan, G. Coronavirus variants are spreading in India—what scientists know so far. *Nature* **593**, 321–322 (2021).
- Elliott, P. et al. Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science* **374**, eab19551 (2021).
- Twohig, K. A. et al. Hospital admission and emergency care attendance risk for SARS-CoV-2 Delta (B.1.617.2) compared with Alpha (B.1.1.7) variants of concern: a cohort study. *Lancet Infect. Dis.* **22**, 35–42 (2021).
- Lucas, C. et al. Impact of circulating SARS-CoV-2 variants on mRNA vaccine-induced immunity. *Nature* **600**, 523–529 (2021).
- Challen, R. et al. Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.06.05.21258365> (2021).
- Mishra, S. et al. Changing composition of SARS-CoV-2 lineages and rise of Delta variant in England. *eClinicalMedicine* **39**, 101064 (2021).
- Papa, G. et al. Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell–cell fusion. *PLoS Pathog.* **17**, e1009246 (2021).
- Ilcova, P. et al. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021).
- Cherian, S. et al. SARS-CoV-2 spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. *Microorganisms* **9**, 1542 (2021).
- Syed, A. M. et al. Rapid assessment of SARS-CoV-2 evolved variants using virus-like particles. *Science* **374**, 1625–1632 (2021).
- Investigation of SARS-CoV-2 Variants of Concern: Technical Briefings; <https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20210201> (Public Health England, 2020).
- Bericht zu Virusvarianten von SARS-CoV-2 in Deutschland, 9 June 2021; https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Bericht_VOC_2021-06-09 (Robert Koch Institut, 2021).
- Abbott, S. & Funk, S. Estimating epidemiological quantities from repeated cross-sectional prevalence measurements. Preprint at *medRxiv* <https://doi.org/10.1101/2022.03.29.22273101> (2022).
- Dhar, M. S. et al. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science* **374**, 995–999 (2021).
- du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
- Lemey, P. et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **11**, 5110 (2020).
- SARS-CoV-2 Variants of Concern and Variants Under Investigation in England, Technical briefing 17; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1001354/Variants_of_Concern_VOC_Technical_Briefing_17.pdf (Public Health England, 2021).
- Kraemer, M. U. G. et al. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021).
- Covid-19: red list arrivals terminal opens at Heathrow Airport. *BBC News* <https://www.bbc.co.uk/news/business-57310148> (2021).
- Booking and Staying in a Quarantine Hotel If You've Been in a Red List Country <https://www.gov.uk/guidance/booking-and-staying-in-a-quarantine-hotel-when-you-arrive-in-england> (Department for Transport and Department of Health and Social Care, 2021).
- COVID-19 Response—Spring 2021 (Summary); <https://www.gov.uk/government/publications/covid-19-response-spring-2021/covid-19-response-spring-2021-summary> (UK Cabinet Office, 2021).
- Willis, R. Y. A. Coronavirus (COVID-19) Infection Survey, Characteristics of People Testing Positive for COVID-19, UK: 3 November 2021; <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveycharacteristicsofpeopletestingpositiveforcovid19uk/3november2021> (Office for National Statistics, 2021).
- Ferguson, N. M. B.1.617.2 Transmission in England: Risk Factors and Transmission Advantage <https://www.gov.uk/government/publications/imperial-college-london-delta-b16172-transmission-in-england-risk-factors-and-transmission-advantage-1-june-2021> (Imperial College London, 2021).
- Kalkauskas, A. et al. Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Comput. Biol.* **17**, e1008561 (2021).
- Covid: Surge testing in Bedford due to Indian variant. *BBC News* <https://www.bbc.co.uk/news/uk-england-beds-bucks-herts-57151534> (2021).
- Jarvis, C. I. et al. CoMix study—Social Contact Survey in the UK; <https://cmmid.github.io/topics/covid19/comix-reports.html> (2020).
- Grenfell, B. T., Bjørnstad, O. N. & Kappey, J. Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414**, 716–723 (2001).
- Lemey, P. et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* **595**, 713–717 (2021).
- Hodcroft, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595**, 707–712 (2021).
- Volz, E. et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
- Tian, H. et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638–642 (2020).
- Park, S. W. et al. Roles of generation-interval distributions in shaping relative epidemic strength, speed, and control of new SARS-CoV-2 variants. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.03.21256545> (2021).
- Kraemer, M. U. G. et al. Monitoring key epidemiological parameters of SARS-CoV-2 transmission. *Nat. Med.* **27**, 1854–1855 (2021).
- Relatório de Situação Sobre Diversidade Genética do Novo Coronavírus SARS-CoV-2 em Portugal—20-07-2021; <http://www.insa.min-saude.pt/relatorio-de-situacao-sobre-diversidade-genetica-do-novo-coronavirus-sars-cov-2-em-portugal-20-07-2021/> (INSA, 2021).
- Elliott, P. et al. Twin peaks: the Omicron SARS-CoV-2 BA.1 and BA.2 epidemics in England. *Science* **376**, eabq441 (2022).
- Madhi, S. A. et al. Population immunity and Covid-19 severity with Omicron variant in South Africa. *N. Engl. J. Med.* **386**, 1314–1326 (2022).
- Ali, S. T. et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* **369**, 1106–1109 (2020).
- Mina, M. J. et al. A Global Immunological Observatory to meet a time of pandemics. *eLife* **9**, e58989 (2020).
- Bastani, H. et al. Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature* **599**, 108–113 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

¹Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. ²Department of Zoology, University of Oxford, Oxford, UK. ³College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK. ⁴MRC Centre of Global Infectious Disease Analysis, Jameel Institute for Disease and Emergency Analytics, Imperial College London, London, UK. ⁵Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark. ⁶Molecular Immunity Unit, Department of Medicine, Cambridge University, Cambridge, UK. ⁷Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Bruxelles, Belgium. ⁸Department of Microbiology, Immunology and

Transplantation, Rega Institute, KU Leuven, Leuven, Belgium. ⁹Google, Mountain View, CA, USA. ¹⁰Department of Mathematics, School of Science and Engineering, Tulane University, New Orleans, LA, USA. ¹¹Department of Infectious Disease, Imperial College London, London, UK. ¹²UK Health Security Agency, London, UK. ¹³Instituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil. ¹⁴BlueDot, Toronto, Ontario, Canada. ¹⁵Divisions of Internal Medicine and Infectious Diseases, Toronto General Hospital, University Health Network, Toronto, Ontario, Canada. ¹⁶Department of Medicine, Division of Infectious Diseases, University of Toronto, Toronto, Ontario, Canada. ¹⁷Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada. ¹⁸Department of Biosystems Science and Engineering, ETH Zurich, Zurich, Switzerland. ¹⁹Swiss Institute of Bioinformatics, Lausanne, Switzerland. ²⁰Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. ²¹Centre for Genomic Pathogen Surveillance, Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. ²²Pathogen Genomics Unit, Public Health Wales NHS Trust, Cardiff, UK. ²³School of Biosciences, The Sir Martin Evans Building, Cardiff University, Cardiff, UK. ²⁴Quadram Institute, Norwich, UK. ²⁵Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. ²⁶Departments of Biostatistics, Biomathematics and Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. ²⁷Department of Pathobiology and Population Sciences, Royal Veterinary College London, London, UK. ²⁸Pandemic Sciences Institute, University of Oxford, Oxford, UK. ²⁹These authors contributed equally: John T. McCrone, Verity Hill, Sumali Bajaj, Rosario Evans Pena. ³⁰These authors jointly supervised this work: Oliver G. Pybus, Andrew Rambaut, Moritz U. G. Kraemer. *A full list of members and their affiliations appears in the Supplementary Information. ✉e-mail: opybus@rvc.ac.uk; a.rambaut@ed.ac.uk; moritz.kraemer@biology.ox.ac.uk

Methods

Genomic data

International (non-UK) sequences were downloaded from GISAID on 15 September 2021 and combined with sequences from England taken as part of community surveillance (pillar 2) available from COG-UK as of September 2021. Each week, each of the pillar 2 testing laboratories selected a number of 96-well plates proportional to the fraction of all testing done at the laboratory, for sequencing. Even though the instruction was that these should be selected randomly, we cannot exclude the possibility of some level of error. However, at the scale at which the COG-UK consortium is operating, we do not anticipate that this affects the results of the study. Sequences were processed and aligned as part of the daily datapipe analysis managed by CLIMB on behalf of COG-UK. Duplicate and environmental sequences, as well as those with impossible or incomplete collection dates, were removed. All sequences were aligned to the reference Wuhan-Hu-1 (GenBank accession MN908947.3) with minimap2 and samples with less than 93% coverage were discarded. Scorpio (<https://github.com/cov-lineages/scorpio>) was run as part of Pangolin⁴³, and sequences containing the Delta VOC constellation of mutations were kept for further analysis.

Problematic sites in the resulting alignment were masked prior to phylogenetic inference and isolates with known sequence artefacts were removed (see <https://github.com/COG-UK/Delta-analysis> for details). Additionally, mutations in the Delta VOC have caused widespread amplicon dropout of amplicon 72 in the commonly used ARTIC primer scheme (<https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye>) before the introduction of version 4 of the primer scheme. To avoid spurious phylogenetic associations based on differential treatment of amplicon dropout with COG-UK and across the globe, we masked sites 2142–21990, which represent the region solely covered by amplicon 72 and are not overlapped by neighbouring amplicons. Delta sequences from India were highly heterogeneous in space (Extended Data Fig. 9).

Phylogenetic analyses

To provide an overview of the global expansion of Delta (Fig. 1a), we analysed a subset of 1,000 Delta genomes sampled evenly through time. To minimize the effect of incorrectly reported collection dates, we restricted our analysis to samples where the lag between sample collection date and GISAID submission date is less than four weeks. To further ensure only the highest quality samples were included, we built an maximum likelihood tree using iqtree2⁴⁴, rooted with Wuhan-Hu-1 (GenBank accession MN908947.3) as an outgroup, and used TreeTime⁴⁵ to remove tips lying beyond two interquartile ranges from the regression of time against root-to-tip distance. This analysis resulted in a final dataset of 975 samples. The temporal tree estimated by treeTime was used as a starting tree in the following Bayesian analysis with slight modifications to randomly resolve polytomies. Two chains of 100 million states were run using BEAST v1.10.4⁴⁶ with sampling every 20,000 states. Both chains were combined with the first 10 million states removed for burnin. We used a HKY + Γ substitution model⁴⁷, a flexible Skygrid coalescent prior⁴⁸ with grid points every two weeks⁴⁵, and an asymmetric, discrete phylogeographic model with samples assigned to Indian, English and global locales. Preliminary analysis showed very little temporal signal in the data, which is unsurprising given the relatively slow evolutionary rate of SARS-CoV-2 and the short study period. Therefore, in all analyses the evolutionary rate was fixed to 7.5×10^{-4} substitutions per site, as estimated in ref. ¹⁸. Convergence was assessed using Tracer v1.7⁴⁹.

The goal of our phylogenetic analysis was to accurately and efficiently describe importation dynamics into England, without sacrificing the dense sampling needed to reconstruct internal spread at a high resolution. Owing to the large size of the required dataset, we followed a similar phylogenetic approach to that used in ref. ¹⁸. First, an

approximately maximum likelihood phylogeny was built using a JC69 substitution model in FastTree⁵⁰, and rooted on Wuhan-Hu-1 (GenBank accession MN908947.3), a high quality Pango lineage B sample from 2019-12-26, as an outgroup. Internal branches representing less than one substitution were collapsed to polytomies. This tree was then split into three subtrees of roughly equal size (Fig. 1a) (28,783, 28,715 and 36,151 tips). As above, TreeTime⁴⁵ was then used to remove temporal outliers, generate a starting time tree, and estimate the number of mutations along each branch. For subtree an empirical distribution of time trees was estimated independently using a recently implemented model in BEAST v1.10⁴⁶ (commit:d1a45) which replaces the substitution model in classical analyses. In brief, in this approach the likelihood of the number of mutations along each branch was calculated from a Poisson distribution with mean equal to the evolutionary rate multiplied by the length of the branch in time⁵¹. In this approach, the standard topological tree search is constrained to operators that sample node heights and resolutions of polytomies present in the substitution tree.

For each subtree, 50 MCMC chains of 40 million iterations were run, sampling trees every 2 million states with the first 20 million states removed as burnin, resulting in datasets of 514–520 empirical trees. The analyses were run using a flexible Skygrid coalescent prior⁴⁸ with grid points every two weeks⁴⁵. Model convergence and proper statistical mixing were verified in Tracer v1.7⁴⁹.

The empirical trees sets estimated above were used to reconstruct importations into England under an asymmetric discrete phylogeographic model. Taxa were split into three locations: England, India and global, with the global state representing all countries other than England and India. We used the recently developed travel-aware phylogenetic model available in BEAST v1.10⁴⁹ to better inform the transition rates in the reconstructed phylogeography. ‘Travel history’ nodes were placed 1 week before isolates from England with known travel history. Where such travel included both India and other countries, ambiguous non-UK states were used. We ran eight chains of 625,000 states, sampling every 2,250 states and with the first 62,500 states removed as burnin, resulting in a total of 1,998 trees sampled from the posterior distribution. Introductions were defined as nodes inferred to be in England with parents in either India or the catch-all global location. The date of importation was assumed to be half-way between such a node and its parent.

Following the importation analysis, the seven largest importations (those with >1,500 sequences, $n = 25,983$) were selected, as well as all importations with five or more sequences, from a representative tree from the posterior set with the same number of total importations as the posterior median. Within this analysis, only sequences with unambiguous postcode districts were used, resulting in a dataset of 25,139 sequences for the seven largest transmission lineages and 24,411 across 280 smaller lineages, which were extracted from the master COG-UK alignment, described in ‘Genomic data’ above. Within those postcode districts, we assigned random coordinates to each sequence, as the continuous phylogeographic analysis does not permit identical values. This was achieved using geographical data from⁵². We then reconstructed the geographic movement of nodes on a fixed tree (pruned from the overall maximum clade credibility (MCC) tree) in BEAST v1.10⁴⁶, using a relaxed random walk model⁵³, and a Cauchy distribution to account for among-branch heterogeneity in dispersal velocity. Large lineages were inferred independently, and all small lineages were inferred in a single run, with the shared parameters for likelihood, precision, and covariance of coordinates, but independent estimates of diffusion rate and trait likelihood. Following this run, 22 small introductions were removed due to their chains not converging to the same posterior. An MCC tree was then generated using TreeAnnotator⁴⁶ to summarize the posterior tree distribution for all lineages. Visualizations were made using a custom Python script. XML files were generated using beastgen.py (<https://github.com/ViralVerity/beastgenpy>) and can be found along with data processing and visualization scripts on GitHub.

Article

For the export analyses we compare Greater London to Greater Manchester which consists of the UTLAs Salford, Trafford, Stockport, Oldham, Bolton, Tameside, Bury, Rochdale, Wigan and Manchester.

State-level incidence data from India. State-level COVID-19 case count data were extracted from <https://api.covid19india.org/csv/latest/states.csv>.

Incidence data from England. COVID-19 case count data for each Local Tier Local Authority were downloaded via <https://coronavirus.data.gov.uk/details/download>.

Travel history data

Four sources of data were compiled to provide the travel history for laboratory-confirmed cases, depending on availability for each individual case: (1) public health passenger locator forms are required for entry into the UK; (2) routine public health contact tracing data including UK Health Security Agency Second Generation Surveillance System (SGSS)⁵⁴; (3) COVID-19 test requests with reported travel associations and (4) responses to additional telephone interviews for cases.

Covariate processing for statistical analyses

COVID-19 case count and vaccination data for the UK. COVID-19 case count data and cumulative vaccination data were downloaded by UTLA from 30 January 2020 to 28 July 2021 by specimen and dosage date, respectively, via <https://coronavirus.data.gov.uk/details/download>. These data include positive laboratory-based PCR tests and positive lateral flow tests, but do not include tests where the lateral flow test was positive and PCR follow-up tests were negative (further details at <https://coronavirus.data.gov.uk/details/about-data>). The COVID-19 case count at the UK country level was calculated by aggregating case data on the UTLA level. Additionally, to match the genomic data, the COVID-19 case count and vaccination data for some UTLAs were aggregated under an area code made up of these multiple UTLAs (see Supplementary Table 3). All entries with the recently discontinued area code E10000002 were assigned the new area code E06000060.

UK population data. UTLA-level 2020-mid-year population size estimates were downloaded from <https://www.ons.gov.uk/people-populationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland>. Population size data were used to calculate the proportion of the population that was partially or fully vaccinated in a location.

Global population data. Country-level population size estimates for the year 2021 were downloaded from https://data.worldbank.org/indicator/SP.POP.TOTL?name_desc=false.

Aggregated and anonymised human mobility data. We used the Google COVID-19 Aggregated Mobility Research Dataset^{31,55}, which contains anonymized relative mobility flows aggregated over users who have turned on the 'location history' setting, which is turned off by default. This is similar to the data used to show how busy certain types of places are in Google Maps, helping identify when a local business tends to be the most crowded. The mobility flux is aggregated per week, between pairs of approximately 5 km² cells worldwide, and for the purpose of this study further aggregated for LTLAs in the UK (<https://geoportal.statistics.gov.uk/datasets/lower-tier-local-authority-to-upper-tier-local-authority-december-2016-lookup-in-england-and-wales/explore>) and to the country level (<https://gadm.org/>) for all other countries for the time period of 29 October 2020 to 6 June 2021.

To produce this dataset, machine learning is applied to log data to automatically segment it into semantic trips. To provide strong privacy

guarantees⁵⁶, all trips were anonymized and aggregated using a differentially private mechanism to aggregate flows over time (see <https://policies.google.com/technologies/anonymization>). This research is done on the resulting heavily aggregated and differentially private data. No individual user data was ever manually inspected; only heavily aggregated flows of large populations were handled. All anonymized trips are processed in aggregate to extract their origin and destination location and time. For example, if users travelled from location a to location b within time interval t , the corresponding cell (a, b, t) in the tensor would be $n \pm \text{err}$, where err is Laplacian noise. The automated Laplace mechanism adds random noise drawn from a zero-mean Laplacian distribution and yields (ϵ, δ) -differential privacy guarantee of $\epsilon = 0.66$ and $\delta = 2.1 \times 10^{-29}$ per metric. Specifically, for each week W and each location pair (A, B) , we compute the number of unique users who took a trip from location A to location B during week W . To each of these metrics, we add Laplace noise from a zero-mean distribution of scale $1/0.66$. We then remove all metrics for which the noisy number of users is lower than 100, following the process described⁵⁶, and publish the rest. This yields that each metric we publish satisfies (ϵ, δ) -differential privacy with values defined above. The parameter ϵ controls the noise intensity in terms of its variance, while δ represents the deviation from pure ϵ -privacy. The closer they are to zero, the stronger the privacy guarantees.

These results should be interpreted in light of several important limitations. First, the Google mobility data is limited to smartphone users who have opted into Google's location history feature, which is off by default. These data may not be representative of the population as whole, and furthermore their representativeness may vary by location. Importantly, these limited data are only viewed through the lens of differential privacy algorithms, specifically designed to protect user anonymity and obscure fine detail. Moreover, comparisons across rather than within locations are only descriptive since these regions can differ in substantial ways.

Flight data. We used data from the International Air Transport Association (<https://bluedot.global/>) on the monthly number of confirmed passengers on flights (direct and indirect) from India to all other countries from January 2021 to June 2021.

Estimated importation intensity. We estimated the weekly importation intensity of the Delta variant for each destination location at the weekly level using the human mobility, GISAID and COG-UK genomic data and COVID-19 case data. An importation intensity value was calculated for each international movement by multiplying the proportion of Delta in the location of origin, the total number of new weekly reported COVID-19 cases and the movement intensity between each origin location and the destination location. We then aggregated all importation intensity values by week and destination location to obtain the EII.

Estimated exportation intensity. We estimated the exportation intensity of the Delta variant for each location of origin at the weekly level using aggregated human mobility, genomic and case count data. An exportation intensity value was calculated for each international movement by multiplying the proportion of Delta in the country of origin, the total number of new weekly reported cases and the movement intensity between the country of origin and the destination country. We then aggregated all importation intensity values by week and origin location to obtain the EEI.

Estimated local human mobility intensity. To obtain an estimate of the intensity of human mobility within a location, we calculated a 'relative self-mobility' value indicating the intensity of mobility within a location (where the origin and destination of the trips are the same) as a percent of the highest recorded of movement within this location in our mobility data during the time period from 22 March 2020 to 6 June 2021 using the human mobility data described above.

New Delta lineage introductions. Daily new lineage introductions into the UK by UTLA were obtained from the continuous phylogenetic analysis described above. The data were aggregated by week and UTLA.

Statistical modelling of Delta growth

Data pre-processing: we kept data starting from the 13th (week commencing 28th March 2021) epidemiological week. These dates are referred to as baseline elsewhere in the main text. We excluded weeks after the first time 95% of samples were observed to be Delta in each UTLA because, after this point, we can no longer estimate the relative growth rates reliably since Delta is effectively fixed in the population. Finally, we kept only those UTLAs which had data on Delta for at least 9 weeks (which are not required to be consecutive). In the final dataset, we had 683 observations (across 64 UTLAs with approximately 11 weeks of non-missing data on average for each) (Supplementary Table 8).

Model. In what follows, we model the dynamics of Delta penetration within a UTLA. Here, we model how the number of Delta samples per UTLA (i), varies over time (t) (here measured in weeks). The background transmission conditions driving the observed number of Delta samples in a given UTLA may be similar to other UTLAs within the same region. We model this variation hierarchically and index variables at the UTLA level by $i[j]$ to indicate that UTLA i is nested within (the overarching) NUTS1 unit j . We use a binomial sampling distribution to model the number of Delta samples $Z_t^{i[j]}$,

$$Z_t^{i[j]} \sim \text{binomial}(Y_t^{i[j]}, p_t^{i[j]}),$$

where $Y_t^{i[j]}$ is the total number of sequenced samples, and $0 \leq p_t^{i[j]} \leq 1$ is the corresponding proportion of Delta samples in subregion i in week t . We then transform this probability, so that it is on the (unconstrained) logit scale:

$$\phi_t^{i[j]} = \text{logit}(p_t^{i[j]}).$$

A key quantity of interest is the relative growth in the proportion of Delta on the logit (that is, log odds) scale, which we estimate weekly and is denoted by $\rho_t^{i[j]}$, where

$$\rho_t^{i[j]} = \phi_t^{i[j]} - \phi_{t-1}^{i[j]}.$$

Relative growth for each UTLA is modelled spatially as depending hierarchically on its containing region, j . It is also assumed to depend on UTLA-specific covariates:

$$\rho_t^{i[j]} = \rho_t^j + \beta' x_t^{i[j]} + \delta_t^{i[j]},$$

where ρ_t^j is a NUTS1-region-level growth trend, $x_t^{i[j]}$ is a vector of covariates, and $\delta_t^{i[j]}$ is a UTLA- and week-specific term representing the deviation from the region-level growth. To account for temporal autocorrelation in the relative growth rate, a given region's relative growth is assumed to follow a random walk centred around its relative growth in the previous week:

$$\rho_t^j \sim \text{normal}(\rho_{t-1}^j, \sigma_2).$$

To assess the importance of covariates, we compared the predictive performance of models which included different sets of covariates. All covariates were standardized by subtracting the mean and dividing by the standard deviation. Since the cumulative proportions vaccinated (considering either the cumulative proportion vaccinated with a 1st or 2nd dose) increased monotonically throughout the time period of observation, we included the UTLA-level mean of these variables

in our regressions: that is, we used time-invariant regressors. We did so to avoid the risk of spurious association due to both Delta and proportions vaccinated growing coincidentally. Covariates were chosen as important predictors if including them in the model improved the model fit on a hold-out set comprising the last two weeks of data for each UTLA. Our best model included within-UTLA mobility and time since baseline, which outperformed the model where we included no covariates (Supplementary Table 6). A model including the cumulative proportion vaccinated with a second dose, time since baseline and within-UTLA mobility also outperformed the no covariate model. However, the improvement in prediction accuracy was minimal, and this was the only model outperforming the no covariate model which included vaccinations, so we do not take this as strong evidence of the importance of vaccination in slowing Delta growth.

We estimated our model in a Bayesian framework and chose priors (Supplementary Table 9) so that a wide range of possible Delta proportions were possible yet were centred on low values in the absence of further information: our prior predictive distributions in Extended Data Fig. 11 illustrate these characteristics.

The computations were done using R and Stan using four parallel chains with 50,000 to 60,000 iterations (depending on the model), half of which were discarded as warm-up iterations; the chains were subsequently thinned by a factor of 10. In all cases, MCMC sampling was diagnosed as converged with $\hat{R} < 1.01$, and bulk and tail effective sample sizes > 400 for all parameters. For 6 of 15 models used for model comparison (which included neither the no covariate model nor the best fit model), there remained 2 out of 4,410 parameters which had $\hat{R} > 1.01$ or had a tail effective sample size below 400; in all cases, the bulk effective sample sizes exceeded 400. In these models, the last two weeks were held-out from each UTLA to perform out of sample predictions, resulting in a smaller dataset, which likely explains the difficulty in obtaining convergence with 50,000 iterations.

Our model outputted two sets of key quantities: the weekly relative growth rate of Delta over time ($\rho_t^{i[j]}$) and the estimated 'effect' of a variable on Delta growth (β). To determine the implications of the effect sizes, we computed the estimated proportion of Delta samples when the covariates took factual versus counterfactual values. We considered counterfactual scenarios for within-UTLA mobility, holding all other covariates at their factual values. The counterfactual scenario we considered was:

- Minimum mobility (within-UTLA mobility = 0)
- Maximum mobility (within-UTLA mobility = 1)

The results of these counterfactual simulations are shown in Extended Data Fig. 10.

Simulation and model robustness: to test model parameter identifiability, we performed inference on simulated data. We fixed the parameters and simulated from the model to create hypothetical data (with 5 regions, each with 6 sub-regions (that is, UTLAs) and 15 time points). We then used these data to estimate the known parameters. We were reasonably able to recover our parameters, and the model converged with $\hat{R} < 1.01$, bulk and tail effective sample sizes > 400 after 20,000 iterations, discarding 10,000 warm-up iterations and thinning by a factor of 10 (Extended Data Fig. 12 and Supplementary Table 7).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

UK genome sequences used were generated by the COVID-19 Genomics UK consortium (COG-UK, <https://www.cogconsortium.uk/>). Data linking COG-IDs to location have been removed to protect privacy, however if you require this data please visit <https://www.cogconsortium.uk/contact/> for information on accessing consortium-only data.

The Google COVID-19 Aggregated Mobility Research Dataset used for this study is available with permission from Google LLC. Shapefiles for county-level analyses in the UK are openly accessible via the Global Administrative Database (<https://gadm.org/>).

Code availability

Code to reproduce the statistical analyses on Delta growth can be found at <https://github.com/sumalibajaj/Delta-Statistical-analysis-share>. The code and accession ids of sequences used to run the phylogenetic analysis as well as a GISAIID acknowledgment table are available at <https://github.com/COG-UK/Delta-analysis>.

43. O'Toole, Á. et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).
44. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
45. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
46. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
47. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
48. Gill, M. S. et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
49. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
50. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
51. Zuckerkandl, E. & Pauling, L. B. in *Horizons in Biochemistry* (eds Kasha, M. & Pullman, B.) 189–225 (Academic Press, 1962).
52. Pope, A. *GB Postcode Area, Sector, District* <https://doi.org/10.7488/ds/1947> (Univ. of Edinburgh, 2017).
53. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
54. *SGSS and CHES Data—NHS Digital*; <https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/data-provision-notices-dpns/sgss-and-sari-watch-data> (NHS, 2021).
55. Kraemer, M. U. G. et al. Mapping global variation in human mobility. *Nat Hum Behav* **4**, 800–810 (2020).
56. Wilson, R. J. et al. Differentially private SQL with bounded user contribution. Preprint at *arXiv* <https://arxiv.org/abs/1909.01917> (2019).

Acknowledgements COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research and Innovation (UKRI), the National Institute of Health Research (NIHR) (grant code: MC_PC_19027), and Genome Research Limited, operating as the

Wellcome Sanger Institute. M.U.G.K. acknowledges support from a Branco Weiss Fellowship, Reuben College Oxford, Google.org, the Foreign, Commonwealth and Development Office and Wellcome (225288/Z/22/Z), and The Rockefeller Foundation. S.D. and M.U.G.K. acknowledge support from the European Union Horizon 2020 project MOOD (grant agreement number 874850). O.G.P., M.U.G.K., L.d.P. and A.E.Z. acknowledge support from the Oxford Martin School. V.H. was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) (grant number BB/M010996/1). S.D. is supported by the Fonds National de la Recherche Scientifique (FNRS) (Belgium). J.T.M., R.C. and A.R. acknowledge support from the Wellcome Trust (Collaborators Award 206298/Z/17/Z—ARTIC network). A.R. is also supported by the European Research Council (grant agreement number 725422—ReservoirDOCS) and Bill and Melinda Gates Foundation (OPP1175094—HIV-PANGEA II). C.R. was supported by a Fondation Botnar Research Award (programme grant 6063). G.B. acknowledges support from the Research Foundation—Flanders (Fonds voor Wetenschappelijk Onderzoek—Vlaanderen) (GOE1420N and G098321N) and from the Interne Fondsen KU Leuven (Internal Funds KU Leuven) under grant agreement C14/18/094. A.O. is supported by the Wellcome Trust Hosts, Pathogens and Global Health Programme (grant number 203783/Z16/Z) and Fast Grants (award number 2236). S. Bajaj is supported by the Clarendon Scholarship, University of Oxford and NERC DTP (grant number NE/S007474/1). M.A.S. acknowledges support from US National Institutes of Health grant R01 AI153044. X.J. acknowledges support from US National Institutes of Health grant U19 AI135995. T.P.P. and W.S.B. acknowledge support from the G2P-UK National Virology Consortium funded by the MRC (MR/W005611/1). I.I.B. is supported by the Canadian Institutes of Health Research (grant 02179-000). N.R.F. acknowledges support from the Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z), Bill and Melinda Gates Foundation (INV-034540), the Medical Research Council-Sao Paulo Research Foundation (FAPESP) CADDE partnership award (MR/SO195/1 and FAPESP 18/14389-0) and the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1). E.V. acknowledges support from the Wellcome Trust (220885/Z/20/Z). The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission or any of the other funders.

Author contributions J.T.M., V.H., S. Bajaj, O.G.P., A.R. and M.U.G.K. conceived and planned the research. J.T.M., V.H., S. Bajaj, R.E.P., R.I., C.R., C.H., I.I.B., O.G.P., A.R. and M.U.G.K. analysed and processed the data. S. Bajaj performed statistical epidemiological analyses. B.C.L., M.U.G.K., E.V., S. Bhatt, S.D., G.B., X.J. and M.A.S. advised on statistical methodology. J.T.M., V.H., B.J., R.C., A.O., N.J.L., D.M.A. and A.R. designed and implemented genomic data processing pipelines. J.T.M., V.H., O.G.P., S. Bajaj, R.E.P., A.R. and M.U.G.K. wrote the first draft of the manuscript. All authors contributed to editing and interpreting the results. A.R., O.G.P. and M.U.G.K. jointly supervised the work.

Competing interests The authors declare no competing interests.

Additional information

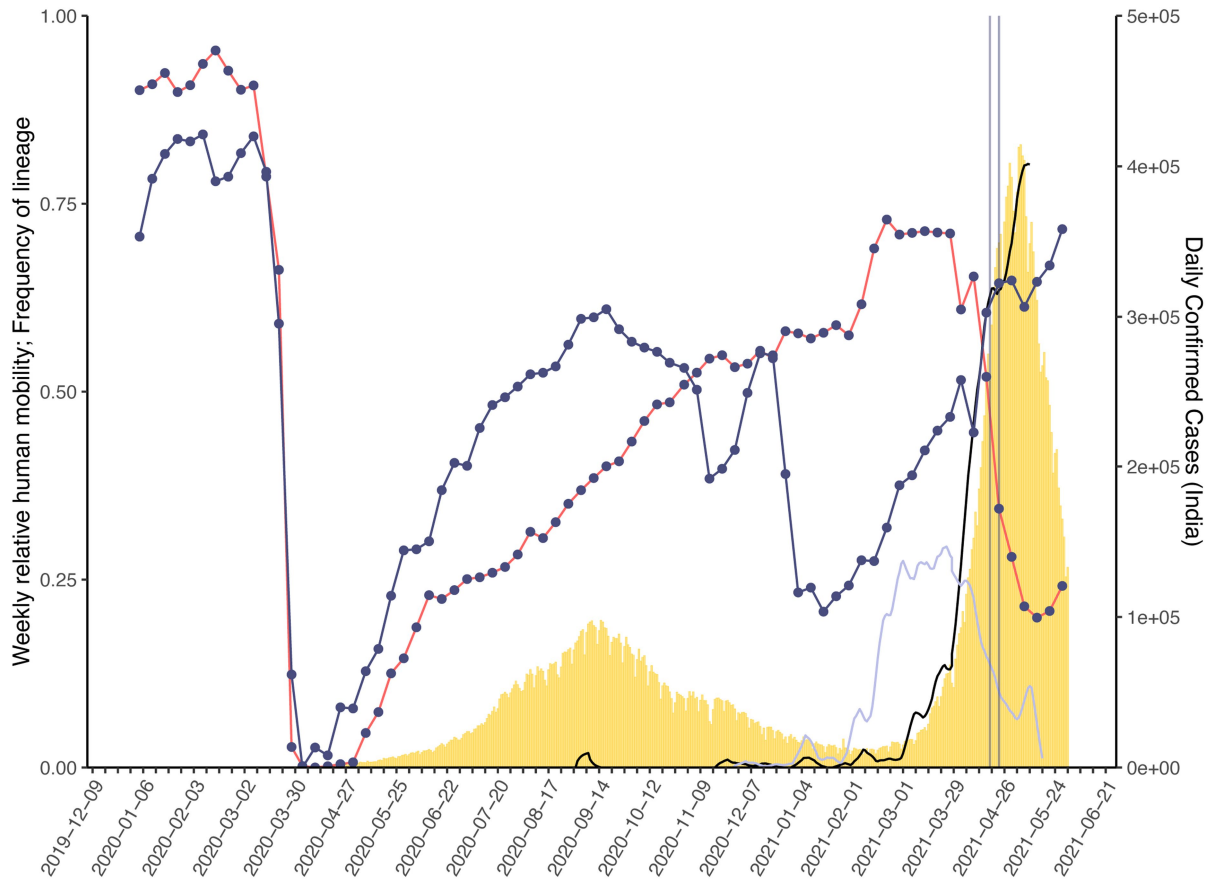
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-022-05200-3>.

Correspondence and requests for materials should be addressed to Oliver G. Pybus, Andrew Rambaut or Moritz U. G. Kraemer.

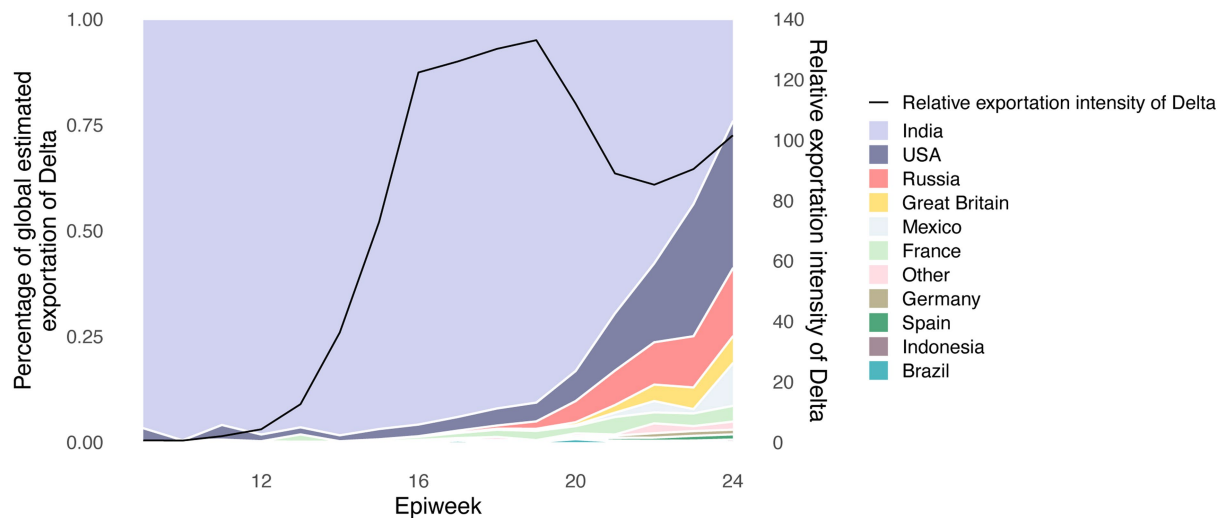
Peer review information *Nature* thanks Christopher Illingworth, Tommy Tsan-Yuk Lam and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

A.

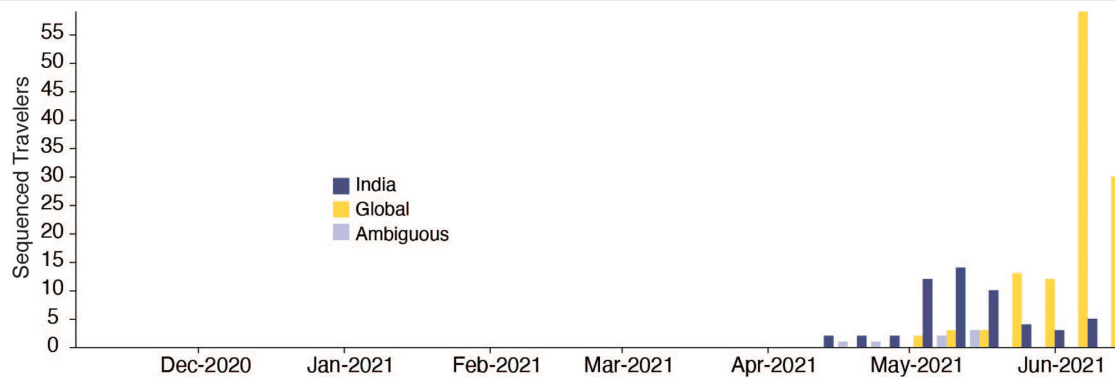


B.

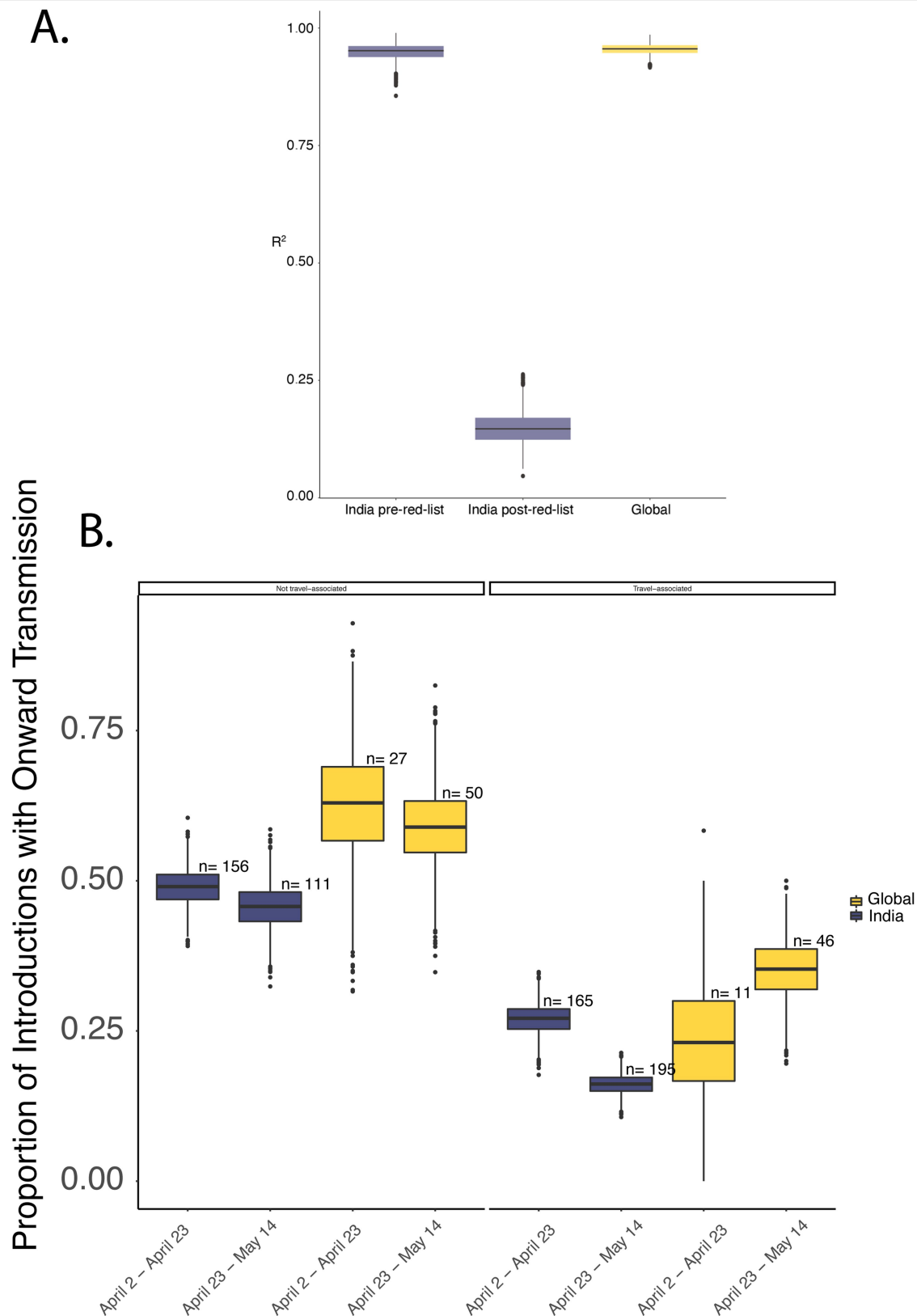


Extended Data Fig. 1 | SARS-CoV-2 movement dynamics. A) Daily number of reported SARS-CoV-2 cases (yellow bars, right hand axis) in India. Weekly human movements in England, relative to the maximum in England (dark blue line, left hand axis, Methods), and in India, relative to the maximum in India (red line, left hand axis, Methods). Proportion of genomes in India that are assigned to lineages B.1.617.2 (black line, no points) and B.1.617.1 (light blue line, no points)

(left hand axis). First vertical line represents the announcement of the quarantine policy for arrivals of travellers from India to England (17 March 2021) and the second vertical line represents the date of implementation (23 March 2021). **B)** Proportion of weekly Estimated Exportation Intensity (EEI) of Delta by country. See Methods for details of calculation (left y-axis). The black line represents the total EEI by week (right y-axis).

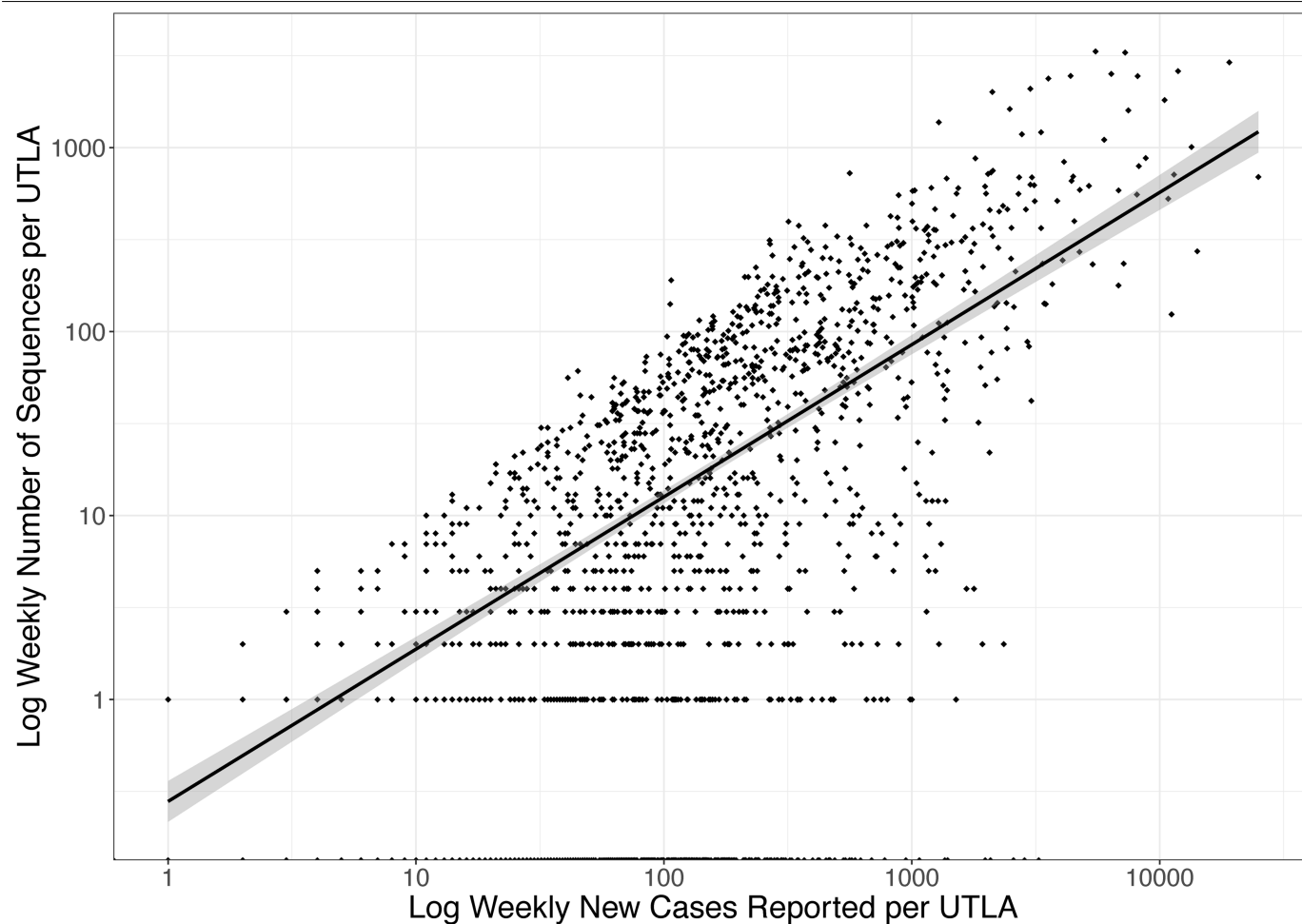


Extended Data Fig. 2 | Travel history of Delta importations. Temporal distribution of genomic isolates from the AY.4 sublineage with travel history, by the likely location of exposure.



Extended Data Fig. 3 | Changing importation dynamics. A) R^2 (coefficient of determination) between estimated number of importations from the phylogenetic analysis and the Estimated Importation Intensity (EII) (Fig. 2a). The R^2 is calculated separately for India (blue) before and after hotel quarantine was introduced, and for other countries (yellow), whilst also accounting for phylogenetic uncertainty. **B)** Proportion of singletons vs non-singletons stratified by non-travel associated clusters and travel associated clusters and

origin locations for the 3 weeks before and after the implementation of the hotel quarantine. The box plot displays the median with lower and upper hinges representing the 25th and 75th percentiles of each group. Whiskers extend to the most extreme data points no more than 1.5 times the interquartile range beyond each hinge. Figure 3a includes each of the trees in the posterior sample ($n = 1998$ for each box). The number of observations in each group in B is annotated above each box.



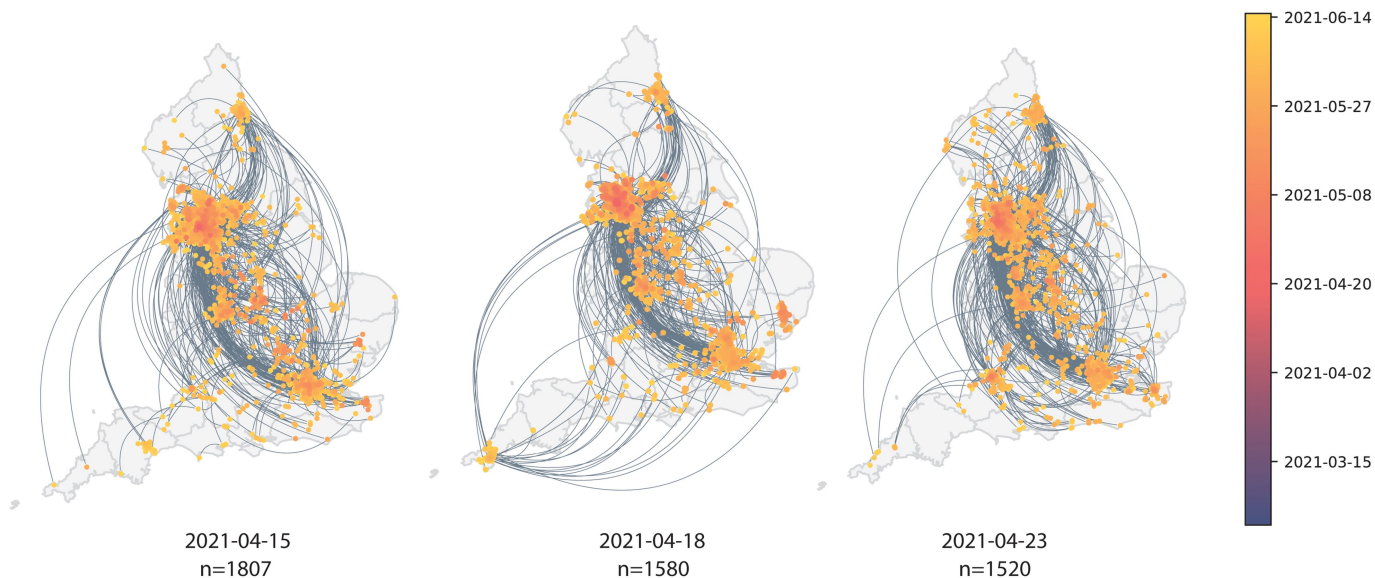
Extended Data Fig. 4 | Representativeness of SARS-CoV-2 Delta genomes. Correlation between the number of weekly Delta sequences and the number of weekly Delta cases each UTLA (Pearson's $r = 0.68$, 95% CI: 0.65–0.71). Shaded region represents 95% CI.

A.

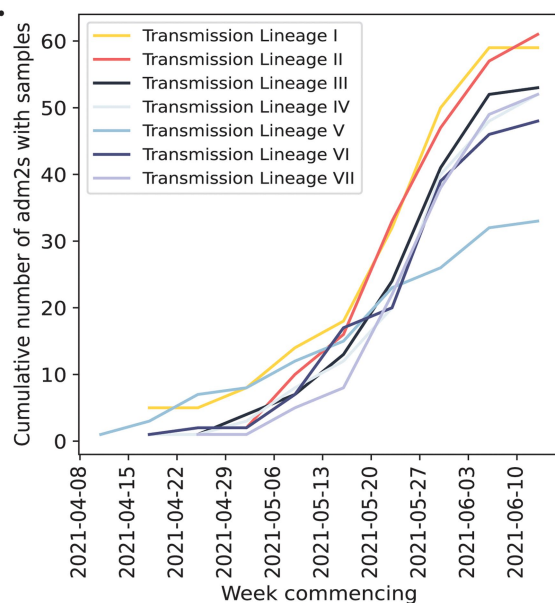
Transmission lineage IV

Transmission lineage VI

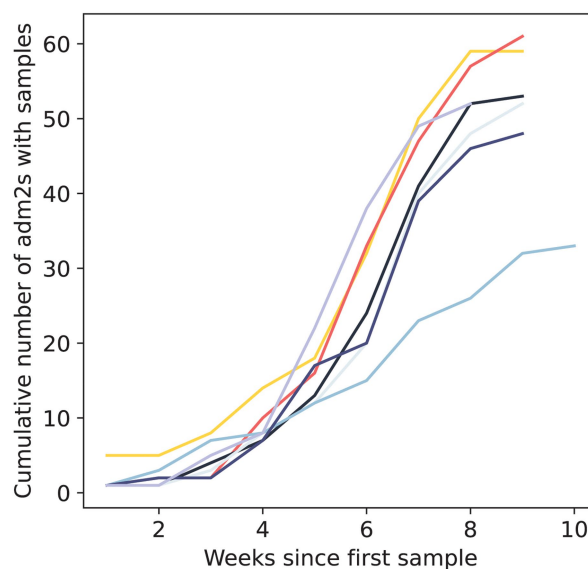
Transmission lineage VII



B.

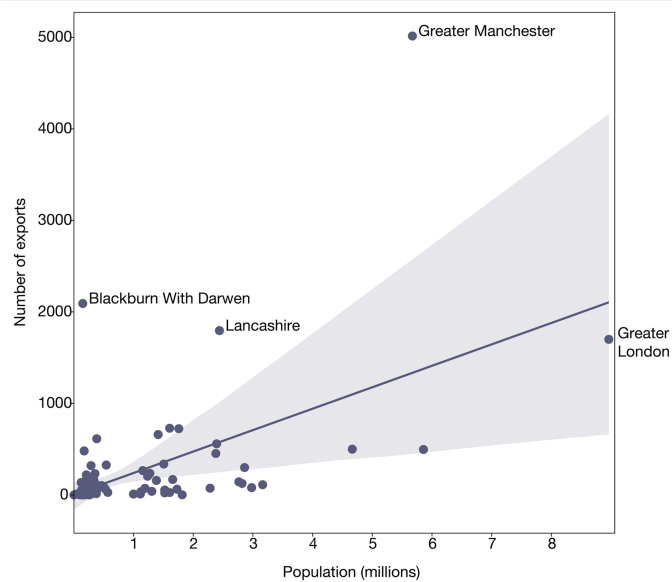


C.

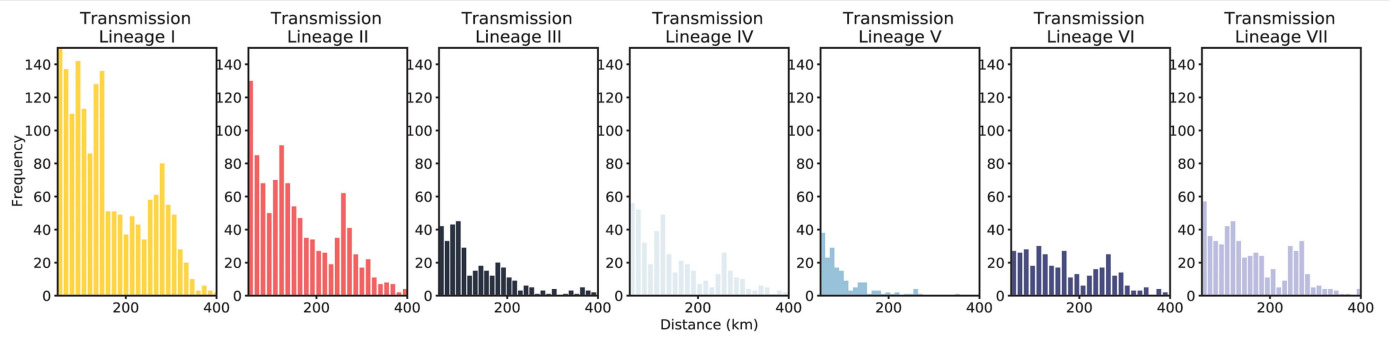


Extended Data Fig. 5 | Internal dissemination of Delta. A) Maps showing virus movements inferred using continuous phylogeographic analysis for the fourth, sixth and seventh largest transmission lineages. Direction of movement

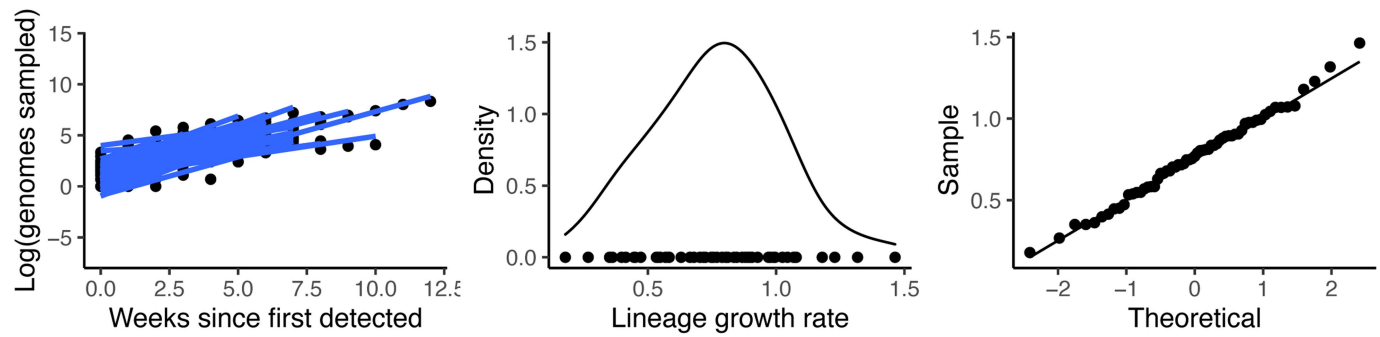
is anti-clockwise, and dots are coloured by date. Cumulative number of UTLAs that the five largest Delta transmission lineages are sampled in absolute (B) and relative (C) time.



Extended Data Fig. 6 | Drivers of exportations. Number of estimated exportations from phylogeographic analysis and population size at the UTLA level (Pearson's $r = 0.54$, 95% CI: 0.38–0.68, p -value < 0.005). Grey shaded region represents 95% CI.

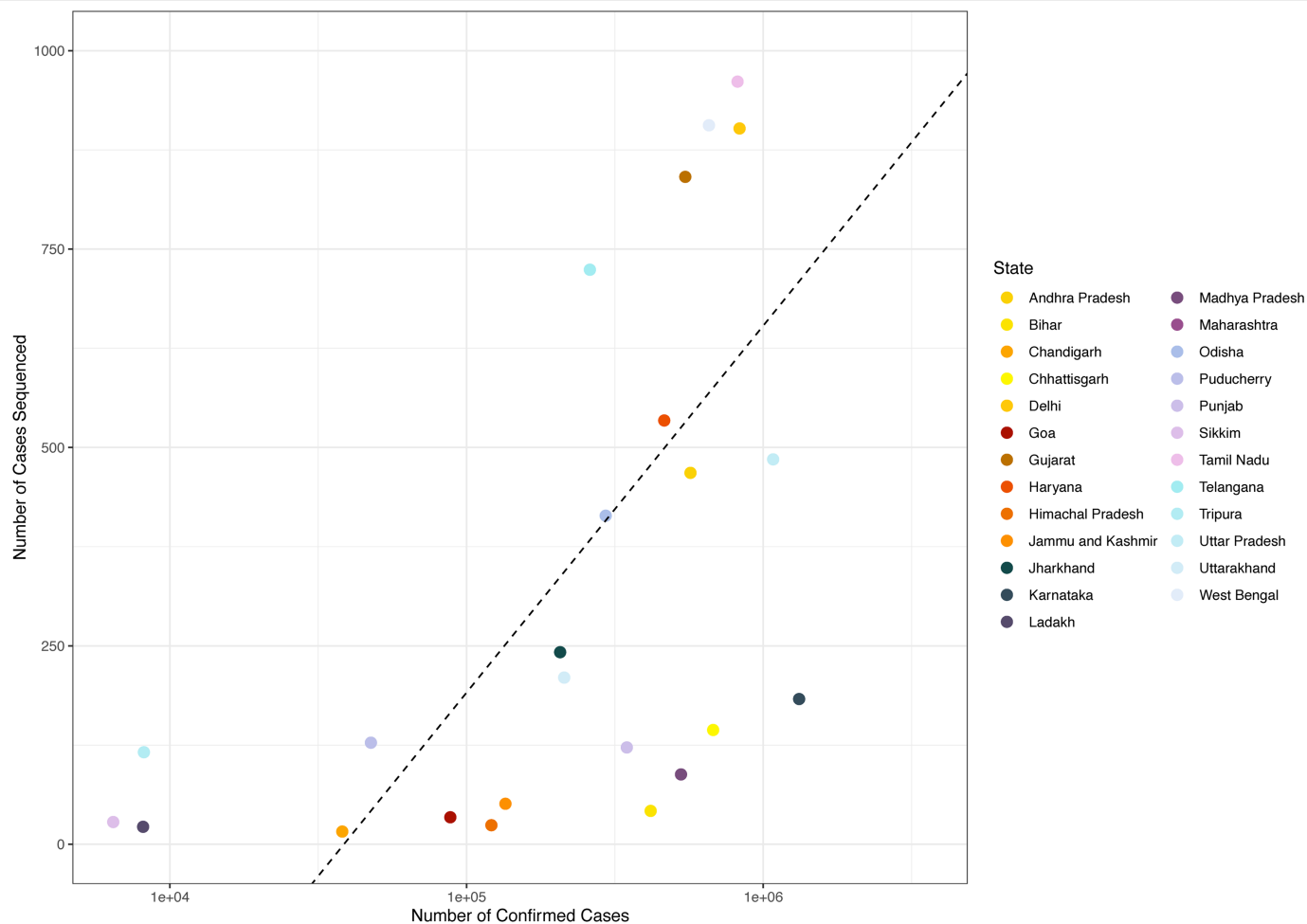


Extended Data Fig. 7 | Distance of viral movements. Histograms of the distance of viral movements over 50km for each of the largest seven Delta transmission lineages in England.



Extended Data Fig. 8 | Weekly lineage growth rates. Growth of transmission lineages in England for lineages observed for at least 3 weeks and with >100 genomes sampled in total. **A)** The log number of weekly sampled genomes per transmission lineage plotted over time. Lines represent a linear fit (assuming

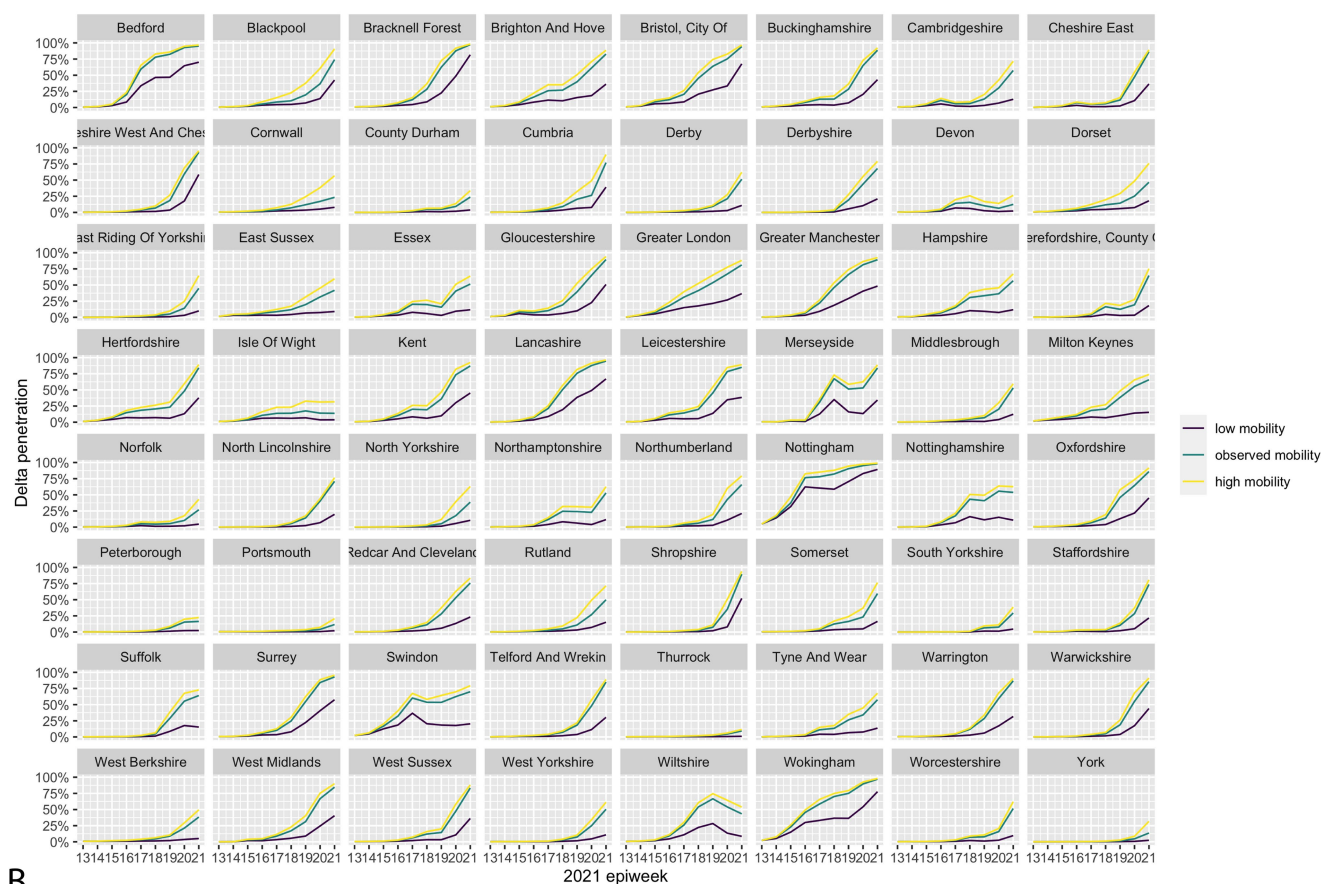
exponential growth). **B)** Distribution of growth rates (slopes in A). **C)** Quantile plot comparing the observed quantiles in the growth rate distribution to theoretical quantiles from a normal distribution.



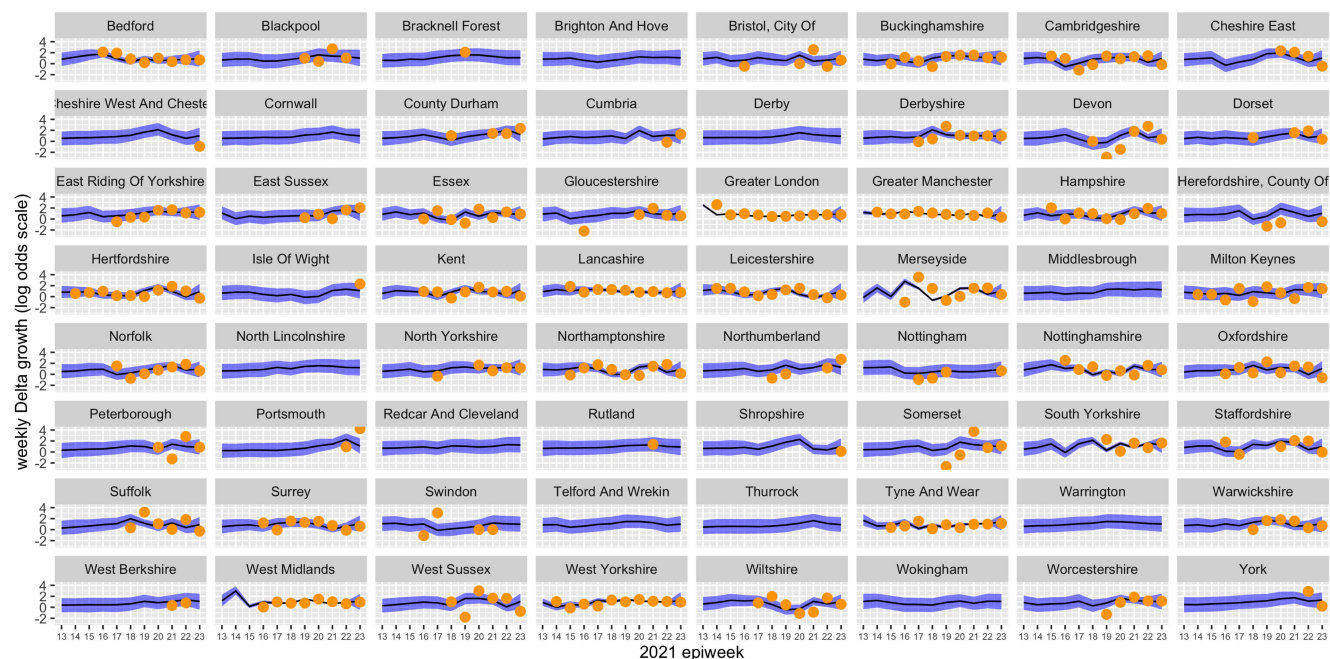
Extended Data Fig. 9 | Representativeness of SARS-CoV-2 Delta genomes in India. Scatter plot showing the number of confirmed cases per state in India vs. the number of cases sequenced in that state in India between 28th of November

2020 to the 16th of May 2021. In states above the line more than the mean number of cases were sequenced.

A.



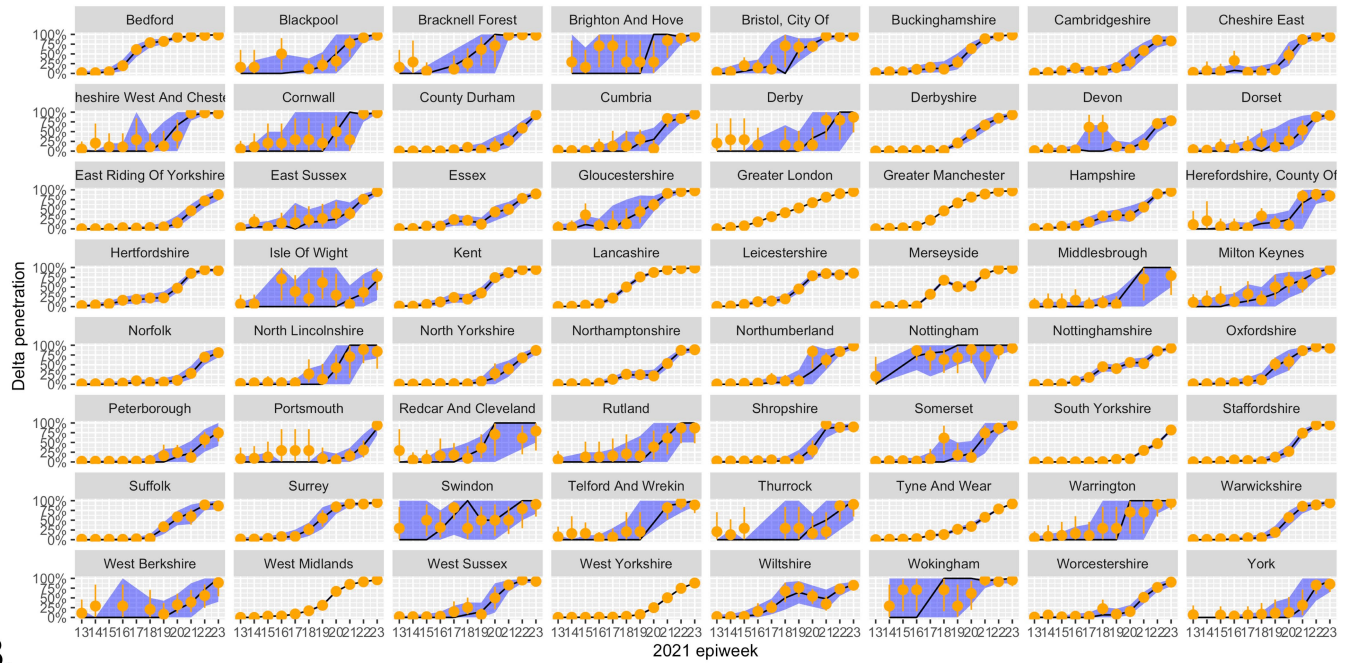
B.



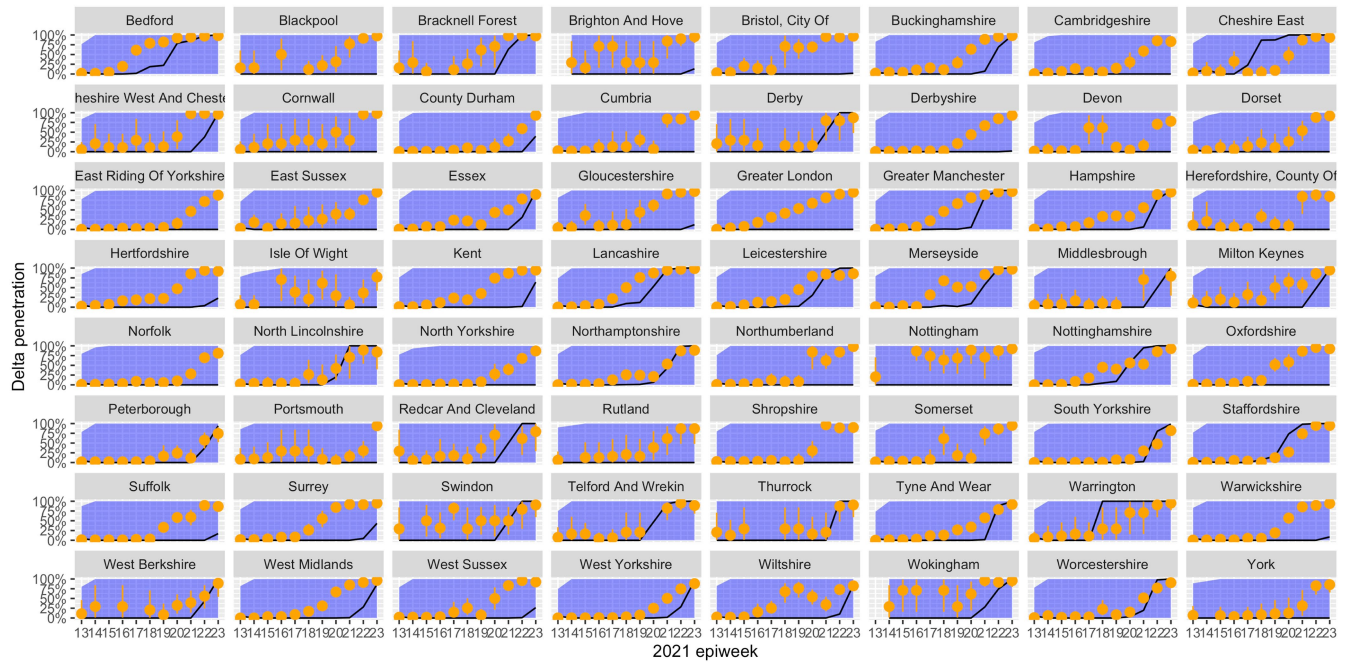
Extended Data Fig. 10 | Counterfactual scenarios. A) Estimated and observed proportions of Delta variant samples across UTLAs, for various counterfactual scenarios: minimum (purple) and maximum relative within UTLA mobility (yellow), observed within UTLA mobility (light blue). All lines represent median posterior estimates using the model with epiweek and within UTLA mobility as covariates which was fit to data from 2021 epiweeks 13–21.

B) Time-varying relative growth of Delta (on the log odds scale). The shaded regions represent the corresponding 95% Bayesian credible intervals. The orange points indicate the raw growth rates calculated directly from the data (note, there are missing values in these data due to the presence of dates when no samples were taken or when the frequency of Delta remained at 0% or 100%).

A.

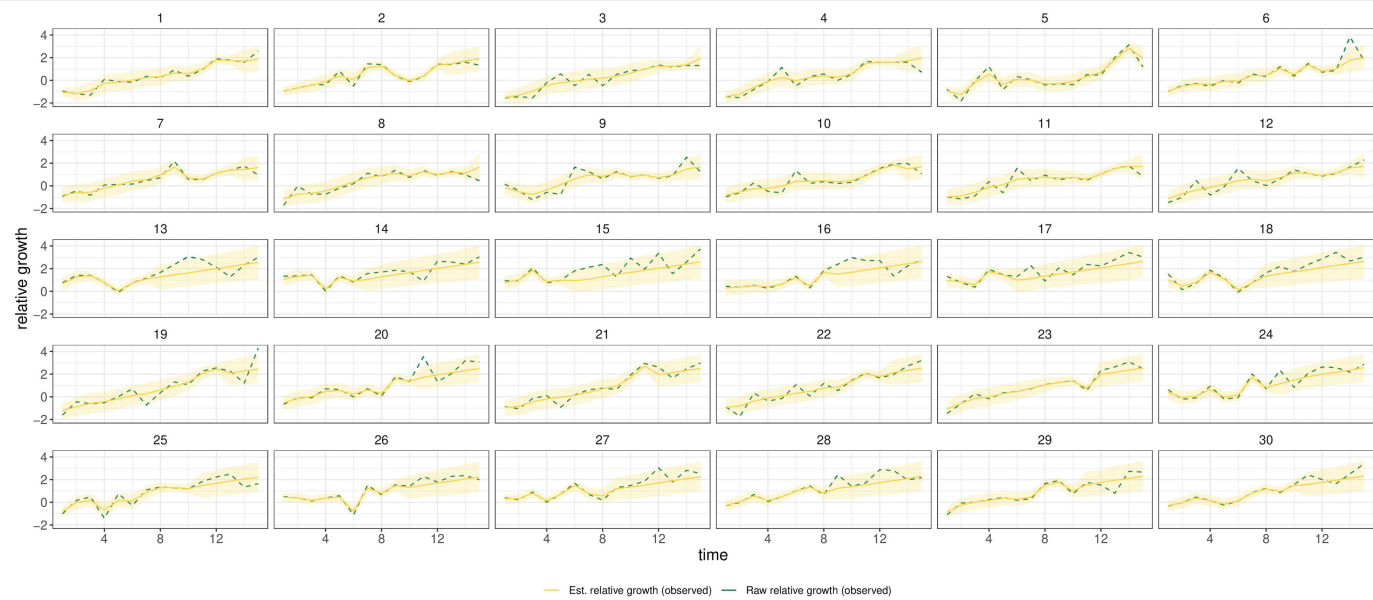


B.



Extended Data Fig. 11 | Model evaluation. A) Posterior and **B)** prior predictive simulations for the Delta frequency. In both cases, the orange point-ranges display the observed data (points indicate posterior medians and whiskers indicate 2.5%–97.5% posterior quantiles assuming a uniform prior for each

independent data point). The blue shaded region represents in (A) the posterior simulated 2.5%–97.5% quantiles and in (B) the equivalent prior simulated quantiles; lines indicate median simulation values.



Extended Data Fig. 12 | Simulation model. Simulation comparing known vs estimated relative growth rates (see Methods) for hypothetical locations.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.

Data analysis

Code to partially reproduce the statistical analyses on Delta growth can be found here: <https://github.com/sumalibajaj/Delta-Statistical-analysis-share>. The code and accession ids of sequences used to run the phylogenetic analysis as well as an GISAID acknowledgment table are available here: <https://github.com/COG-UK/Delta-analysis>. Explicit details about the analyses and their dependencies can be found in the repositories listed above. Details regarding COGUK genomic processing pipeline can be found at <https://github.com/COG-UK/datapipe>. In particular these analyses use the following software.
minimap2v2.17
FastTreev2.1.10
BEASTv1.10.4 and BEASTv1.10.4 (commit:d1a45)
Tracer v1.7

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

UK genome sequences used were generated by the COVID-19 Genomics UK consortium (COG-UK, <https://www.cogconsortium.uk/>). Data linking COG-IDs to location have been removed to protect privacy, however if you require this data please visit <https://www.cogconsortium.uk/contact/> for information on accessing consortium-only data. The Google COVID-19 Aggregated Mobility Research Dataset used for this study is available with permission from Google LLC. Code to reproduce the statistical analyses on Delta growth can be found here: <https://github.com/sumalibajaj/Delta-Statistical-analysis-share>. The code and accession ids of sequences used to run the phylogenetic analysis as well as an GISAID acknowledgment table are available here: <https://github.com/COG-UK/Delta-analysis>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Our study characterizes the importation and subsequent spread of the SARS-CoV-2 Delta variant of concern in England. We use the UK genomic data produced by COG-UK to highlight the temporal and spatial dynamics of imported transmission lineages, and novel epidemiological models to determine factors that led to the growth of Delta across both the UK and US.
Research sample	The phylogenetic analysis used UK sequences produced by COG-UK and international sequences shared on GISAID. The UK data set consists of pillar 2 English samples taken as part of the national testing infrastructure for community surveillance. These samples were chosen to as they best represent a random sample of lineages circulating in the general English public at the time of study. Only samples identified as the Delta variant of concern were included in this study. UK genome sequences used were generated by the COVID-19 Genomics UK consortium (COG-UK, https://www.cogconsortium.uk/)
Sampling strategy	No subsampling strategy was used in this study. All sequences identified as Delta were included. This resulted in a dataset of over 90,000 sequences. This is the largest phylogenetic analysis of it's kind, and used all available pillar 2 English samples generated by COGUK. This was done to generate the best approximation of circulating lineages in England at the time of study. The density of the data allows for detailed modeling of geographic spread in England.
Data collection	English samples were taken and sequenced by COG-UK affiliated partners across the UK and shared according to the COG-UK data sharing agreement. Post-processing of genomic data was done as part of the daily COG-UK 'datapipeline/phylopipe' hosted on CLIMB. Four sources of data were compiled to provide the travel history for laboratory confirmed cases, depending on availability for each individual case: (1) public health passenger locator forms are required for entry into the UK; (2) routine public health contact tracing data including UK Health Security Agency Second Generation Surveillance System (SGSS)89, (3) COVID-19 test requests with reported travel associations and (4) responses to additional telephone interviews for cases. This data was processed by authors in PHE and shared with others under the appropriate data sharing agreement. We used two human mobility datasets, one a country flight dataset from IATA and another from the Google Mobility Research Dataset which was aggregated to the country level and UTLA level in the UK and state level in the USA.
Timing and spatial scale	All SARS-CoV-2 non-UK genomes in GISAID and pillar 2 samples from England that were identified as Delta up until June 15, 2021 were included in this study. The spatial resolution of the global analyses were at the country level and for the detailed phylogeographic analyses at the postcode level.
Data exclusions	Sequences with known quality flaws as well as those that failed to pass standard quality measures (e.g. those with low sequence coverage and temporal outliers) were removed from the study.
Reproducibility	This study is based on epidemiological data and subsequent modeling. All code to reproduce the results and what data we are able to share has been shared to make reproducing such analysis in this and other settings possible.
Randomization	Randomization was not relevant in this study, as it was based on epidemiological observations in the general public and there were no experimental treatments applied.
Blinding	Blinding was not relevant to this study design, as it was based on epidemiological observations in the general public and there were no experimental treatments applied.
Did the study involve field work?	☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Understanding COVID-19 testing behaviour in England through a sociodemographic lens

Sumali Bajaj^{1,*†}, Siyu Chen^{2,*}, Richard Creswell^{3,*}, Reshania Naidoo^{4,5,*}, Joseph L.-H. Tsui^{1,*}, Olumide Kolade⁶, George Nicholson^{7,8}, Brieuc Lehmann^{8,9}, James A Hay², Moritz U. G. Kraemer^{1,10}, Ricardo Aguas⁵, Christl A. Donnelly^{11,12}, Tom Fowler⁶, Susan Hopkins⁶, Liberty Cantrell¹³, Prabin Dahal¹⁴, Lisa J. White¹, Kasia Stepniowska¹⁴, Merryn Voysey¹³, Ben Lambert^{11,15,†} and the EY-Oxford Health Analytics Consortium^{**}

1. Department of Biology, University of Oxford, Oxford, UK
2. Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK
3. Department of Computer Science, University of Oxford, Oxford, UK
4. Ernst & Young (EY) UKI Health Sciences and Wellness, London, UK
5. Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK
6. UK Health Security Agency, UK
7. Department of Statistics, University of Oxford, Oxford, UK
8. The Alan Turing Institute and Royal Statistical Society Health Data Lab, London, UK
9. Department of Statistical Science, University College London, London, UK
10. Pandemic Sciences Institute, University of Oxford, Oxford, UK
11. Department of Statistics, University of Oxford, Oxford, UK
12. MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK
13. Oxford Vaccine Group, Department of Paediatrics, University of Oxford, Oxford, UK
14. Infectious Diseases Data Observatory, Nuffield Department of Medicine, University of Oxford, UK
15. College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

* Contributed equally.

** The EY-Oxford Health Analytics Consortium membership list is at the end of the article.

† Corresponding authors: sumali.bajaj@biology.ox.ac.uk, ben.lambert@stats.ox.ac.uk

Abstract:

Background:

Understanding underlying mechanisms of heterogeneity in test-seeking and reporting behaviour can help to protect the vulnerable and guide equity-driven interventions. Using COVID-19 testing data for England and data from community prevalence surveillance surveys (REACT-1 and ONS-CIS) from October 2020 to March 2022, we investigated the relationship between sociodemographic factors and testing behaviours in England.

Methods

We used mass testing data for lateral flow device (LFD; data for 290 million tests performed and reported) and polymerase chain reaction (PCR) (data for 107 million tests performed and returned from the laboratory) tests made available for the general public, provided by date, self-reported age and ethnicity at lower tier local authority (LTLA) level. Using a mechanistic causal model to debias the PCR testing data, we obtained estimates of weekly SARS-CoV-2 prevalence by self-reported ethnic groups and age groups for LTLAs in England. This approach to debiasing the PCR (or LFD) testing data also estimated a testing bias parameter defined as the odds of testing in infected versus not infected individuals, which would be close to zero if the likelihood of test seeking (or seeking and reporting) was the same regardless of infection status. Using confirmatory PCR data, we estimated false positivity rates, sensitivity, specificity, and the rate of decline in detection probability by PCR by sociodemographic groups. We also estimated the daily incidence allowing us to determine the fraction of cases captured by the testing programme.

Findings

From March 2021 onwards, individuals in the most deprived regions reported approximately half as many LFD tests per-capita than those in the least deprived areas (Median ratio [Inter quartile range, IQR]: 0.50 [0.44, 0.54]). During October 2020 – June 2021, PCR testing patterns were in the opposite direction (Median ratio [IQR]: 1.8 [1.7, 1.9]). Infection prevalences in Asian or Asian British communities were considerably higher than those of other ethnic groups during the Alpha and Omicron BA.1 waves. Our estimates indicate that the England COVID-19 testing program detected 26% - 40% of all cases (including asymptomatic cases) over the study period with no consistent differences by deprivation levels or ethnic groups.

PCR testing biases were generally higher than for LFDs, which was in line with the general policy of symptomatic and asymptomatic use of these tests. During the invasion phases of the Delta and Omicron variants of concern, the PCR testing bias in the most deprived populations was roughly double (ratio: 2.2 and 2.7 respectively) that in the least. We also determined that ethnic minorities and older individuals were less likely to use confirmatory PCR tests through most of the pandemic and that there was possibly a longer delay in reporting a positive LFD test in the Black populations.

Interpretation

Differences in testing behaviours across sociodemographic groups may be reflective of the relatively higher costs of self-isolation to vulnerable populations, differences in test accessibility, digital literacy, and differing perception about the utility of tests and risks posed by infection. Our work shows how mass testing data can be used in conjunction with surveillance surveys to identify gaps in the uptake of public health interventions at fine scale levels and by sociodemographic groups. It provides a framework for monitoring local interventions and yields valuable lessons for policy makers in ensuring an equitable response to future pandemics.

Funding

UK Health Security Agency.

Introduction

More unequal societies have poorer health outcomes¹. The COVID-19 pandemic likely further exacerbated the impact of existing societal inequalities, yet our understanding of the impact of the pandemic across different parts of society remains poor. Understanding the underlying causes of heterogeneity in infection risk and disease burden is essential to protect the vulnerable and to guide equity-driven interventions in real-time.

In England, testing for SARS-CoV-2 played an important role during the COVID-19 pandemic, allowing the government to monitor disease transmission and forming the cornerstone of self-isolation policies. However, it is unclear whether testing was delivered and taken up equitably across society²⁻⁶. The UK government rolled-out mass testing in response to the COVID-19 pandemic, with the initial testing response starting in March 2020⁷. The testing programme component of NHS Test and Trace (NHSTT) played an integral role through its various testing services. Since May 2020, PCR tests were available for testing swabs from suspected COVID-19 cases; LFD tests for COVID-19 were rapidly developed and evaluated for use between Dec 2020 and April 2021.^{8,9} LFD testing was first rolled-out to key populations, for example, those working in healthcare, and subsequently to the wider population through a universal testing offer. The universal testing offer was a major component of the national testing programme, with a key aim to expand asymptomatic LFD testing and was delivered via various channels, with rollout to the general public from 9 April 2021¹⁰.

We had unprecedented access to COVID-19 testing data for England as part of a UK Health Security Agency (UKHSA) commissioned retrospective evaluation of England's testing programme over an evaluation period of October 2020 – March 2022⁹. Here, we present analyses of these data which quantify biases in test uptake and reporting patterns across different sociodemographic groups, including those based on age, self-reported ethnic group and a measure of deprivation. We also estimate SARS-CoV-2 infection prevalence over time and space for sociodemographic groups, which additionally allows us to estimate the reach of the national testing programme. Our analysis focusses only on the Pillar 2 testing data, which includes PCR and LFD tests made available to the general public (not including tests made available to individuals with clinical need, and health and care workers through Pillar 1).

Methods

Data

The testing data were provided by UKHSA in the context of a retrospective evaluation of the delivery of England's COVID-19 testing programme and its impact on public health outcomes¹¹. Our analysis period covered 1st October 2020 - 30th March 2022. LFD data relied on self-reporting, meaning that our analysis focused on self-reported results. In contrast, the PCR data are results from registered and submitted samples of individuals. These Pillar 2 data were provided by date by a range of categories (including self-reported age and ethnicity) and at lower tier local authority (LTLA) level. Individuals self-reported their ethnicities according to five groupings: “Asian or Asian British” (abbreviated throughout as “Asian ethnic groups” or “Asian people”), “Black; African; Black British or Caribbean” (abbreviated as “Black ethnic groups” or “Black people”), “White”, “Mixed or Multiple ethnic groups”, and “Other ethnic group”; the Office for National Statistics (ONS) also publishes population estimates for these groups, which we used. We present results for the first three ethnic groups in the main section since the population sizes for these were considerably higher; in the supplementary materials, we report results for the remaining categories. We used publicly available data on population size estimates by LTLAs, LTLA and ethnic groups, and age groups, as well as deprivation indices/levels.

We used weekly LTLA-level and weekly age group specific community surveillance random sampling survey data from the REal-time Assessment of Community Transmission (REACT-1) study which performed PCR tests of cross-sectional random samples of individuals approximately every month from 1st May 2020 to 31st March 2022¹². We also used publicly available national weekly prevalence estimates from the ONS COVID-19 Infection Survey (CIS) which repeatedly tested a sample of people¹³. ONS prevalence estimates were used only to estimate national incidence (see Methods below).

Variant periods

The following were the time periods during which the Alpha, Delta and Omicron variants of concern (VOCs) were dominant in England i.e. when the relative frequency of each VOC exceeded 90% nationally (see Figure 1A, Figure S1):

- Alpha: 30th January 2021 to 24th April 2021
- Delta: 5th June 2021 to 4th December 2021
- Omicron: 1st January 2022 to 14th May 2022 (but we ended our Omicron period in March 2022 since this was the end of our study period)

Prevalence estimates and testing bias

Community surveillance random sampling survey testing programmes such as REACT-1 and ONS CIS provide a gold-standard measure of infection prevalence; they have large sample sizes nationally and can produce precise estimates of prevalence at this level. At combinations of fine geographic or other scales (LTLA level and sociodemographic groups), the effective sample size is much reduced, particularly because prevalence was generally low (<10%). Conversely, at finer scales, there is a wealth of Pillar 2 PCR and LFD tests conducted, but the naïve prevalence estimates derived from these tests may be biased. This is because PCR tests were used for symptomatic testing or as part of test-to-release protocols¹⁴, and even though LFD tests were recommended for asymptomatic testing, they likely suffer from reporting bias (unlike PCR). We implement and extend the debiasing methodology of Nicholson et al.¹⁵ which in addition makes use of large volumes of Pillar 2 PCR and LFD tests. This method uses swab results from the REACT-1 study to correct for biases in mass PCR testing data and thus produces fine scale PCR swab positivity rate estimates at a higher precision than is possible just from the community surveillance survey. For ease of readability, we use we refer to PCR swab positivity rate as prevalence. The method estimates testing bias at a coarse scale where there are high numbers of both community surveillance survey tests and mass tests. These biases are then propagated to the nested finer spatial scales to estimate prevalence, making use of the sufficiently large volume but biased mass testing data at fine scales. The testing biases estimated represent the log odds of testing in infected versus not infected individuals. For example, if the probability of performing a PCR test was the same regardless of whether people were infected, then the biases would equal zero ($= \log(1)$). So, a large bias corresponds to a higher chance of an infected individual performing a PCR test compared with an uninfected individual, which would result in overestimated prevalence using the biased testing data. We also estimate the LFD test reporting bias using data on reported LFD tests at the coarse and fine scales. We term this as bias in context of using these data naively for surveillance purposes; however, we recognise that the testing programme was primarily a public health intervention rather than a surveillance programme and did not aim to use PCR or LFD testing data to determine SARS-CoV-2 prevalence.

Using the debiasing method described above, we generate two novel sets of prevalence estimates for each LTLA: one set, stratified by ethnic groups (Figure S2); another, by age group (Figure S3). We also use the method to estimate the LTLA specific prevalence (Figure S4) and testing bias for deprivation levels (Index of Multiple Deprivation IMD).

False positivity and test specificity and sensitivity

We collated a subset of our Pillar 2 testing data comprising all reported LFD positive results which were followed up with a PCR test taken within three days of the reported positive. These data allowed us to estimate the false positive rate for LFD tests, assuming PCR tests were a gold standard. Using the total number of people reporting LFD positives, we can additionally determine the probability that individuals reporting an LFD positive result subsequently took a PCR test, as was generally mandated throughout the period under observation, with two exceptions: from January to March 2021 and after January 2022¹⁶ when confirmatory PCR testing was temporarily suspended given the high COVID-19 prevalence meaning LFD tests likely had low false discovery rates. Using our estimates of prevalence and false positivity rates, we estimated LFD test specificity and sensitivity across different age and ethnic groups throughout the pandemic.

Modelling the decay in PCR detection rate after reporting an LFD positive

Using the data described above, we modelled how the probability of detecting SARS-CoV-2 infection by PCR after reporting a positive LFD result declined according to the time elapsing between reporting the LFD test and subsequent testing by PCR. We determined this decay rate for the different ethnic groups for monthly time windows throughout the study period.

Estimating SARS-CoV-2 incidence and the case detection ratio from prevalence time series

An uptick in cases may signal future increases in hospitalisations and deaths. However, case data are imperfect, being influenced by changes in both the true incidence of infection and changes in testing / test reporting behaviour. We estimated daily incidence in a Bayesian paradigm, assuming the following probabilistic relationship between infections and prevalence:

$$\theta_t \sim N(\sum_{\tau=1}^{\infty} \omega_{\tau} i_{t-\tau}, \sigma),$$

where θ_t denotes prevalence on day t and i_t is the incidence on the same day; $0 \leq \omega_{\tau} \leq 1$ is the probability of being PCR positive given an infection which began τ days ago and $N(\mu, \sigma)$ represents a normal distribution with mean μ and standard deviation σ .

We assume that within 15-day rolling blocks, the reported PCR and LFD positive counts for the Pillar 2 mass testing data at time t were a to-be-estimated constant fraction of the true infection incidence, given by the following relationship:

$$i_t = \frac{\text{reported cases}_t}{\text{case detection ratio}_t},$$

with case detection ratio _{t} corresponding to the 15-day block that time t corresponds to. We describe our priors in supplementary materials.

The availability of two prevalence datasets (the REACT-1 surveys and the UK ONS at the national level only) made it possible for us to infer the true incidence of infection in England and to compare this with the reported cases derived from the national mass testing programme. We used the LTLA-specific prevalences estimated using Pillar 2 PCR data and REACT-1 data as described above to estimate true incidence at the LTLA level and then aggregated it to get true national incidence estimates. Using the ONS prevalence estimates for England, we estimated another set of national incidences. We report time series of the average national case detection ratios derived from REACT-1 and ONS-CIS datasets defined as the fraction of PCR and LFD positive tests in the Pillar 2 dataset compared to the estimated true incidence, by LTLA-specific sociodemographic groups. Details of all methods are in the Supplementary Materials.

Results

Trends in testing uptake and reporting

Over our study period, a total of 290 million LFD tests were reported with 2% of these results positive; and 107 million PCR tests were registered with 13% of tests positive (Figure S5). Of the 7 million positive LFD tests reported, 2.4 million were confirmed by a PCR test with 68% of these on the same day (Table 1 and Figure S6), and 89% returning a positive result.

From October 2020 – June 2021, individuals in the most deprived areas (IMD=1) (Figure 1B) performed almost double (Median ratio [Inter quartile range, IQR]: 1.8 [1.7, 1.9]) the number of PCR tests per capita compared to those in the least deprived (IMD=9) (Figure 1E), and we speculate that this is due to the higher prevalence in deprived areas (Figure S7A) during this period. Linear regression models for the PCR tests per capita suggest that differences in prevalence may explain some of the variation in testing across deprivation levels (Figure S7B, C). But myriad other factors likely also influenced testing, including individuals' age, their occupation, their risk perception and their vaccination status, and a more comprehensive approach would be necessary to probe further the observed differences in testing levels¹⁷. LFD test reporting displayed the opposite and more pronounced trend, where those in the most deprived areas reported the fewest LFD tests per capita, which held even after the roll-out of free asymptomatic testing on 9th April 2021 (Figure 1A). From March 2021, the median number of LFD tests reported per capita in the most deprived areas was roughly half that in the least deprived (Median ratio [IQR]: 0.50 [0.44, 0.54]) despite more deprived areas having the highest prevalence [Figure S7A] – a finding qualitatively similar to the results of the

Liverpool COVID-19 study¹⁸. This may be partly explained by the disproportionately low digital access and resources to report LFDs, as well as lack of adequate support to isolate if testing positive. Those in Black ethnic groups had generally higher PCR testing coverages throughout the analysis period (Figure 1G). Since our prevalence estimates do not indicate a consistently higher prevalence in Black ethnic groups (Figure 2A), it is possible that this reflects a greater risk of symptomatic disease, access to tests through workplace, and could also be partly explained by the ramping up of targeted community testing^{19,20}.

The roll-out of the universal testing offer sought to expand access to LFDs, but it corresponds with a separation in LFD testing patterns across ethnic groups (Figure 1H), and, subsequently, per capita LFD tests reported by White people were consistently higher than in other ethnic groups (from April 2021 per capita LFD tests were 34% (IQR: 22%, 44%) higher in White people versus Black people and 47% (IQR: 42%, 55%) higher in White people versus Asian people). Our estimates of prevalence in White people were similar to or lower than for other ethnic groups (Figure 2A) and similarly so for LFD positivity (Figure S8) suggesting that they performed and reported LFDs more frequently with asymptomatic or mild disease.

Working age populations had consistently the highest PCR testing coverages throughout the evaluation period (Figure 1I and Figure S9B) followed by that in the oldest group (those aged 75+). PCR testing coverages in secondary school-aged children (12-17) were more volatile, likely driven by school terms and the high variability in prevalence in these groups (Figure 2B). LFD test reporting trends were similar to those of PCR test seeking (Figure 1J), albeit with a large spike in LFD testing in school-aged children after the roll-out of asymptomatic testing in schools in March 2021²¹ (Figure 1A).

While the results described above are aggregated at the national level, there was variability between the sociodemographic groups at the LTLA level (Figure S11). This variation may be attributed to limited sample sizes, especially when we disaggregated the tests by self-reported ethnic groups in LTLAs.

From 11th January 2022 onwards, confirmatory PCR tests were no longer required following a positive LFD test result due to the high rates of COVID-19 across the UK¹⁶, which explains the sharp decline in PCR tests in Figures 1E, G & I.

Prevalence, false positivity and PCR positivity rate

During the Alpha wave, the SARS-CoV-2 prevalence in Asian people was considerably higher than for Black people which, in turn, was higher than for White people (Figure 2A), as was found previously²². During the Omicron BA.1 and BA.2 variant waves, the prevalence in Asian groups remained higher than for others, and this pattern was particularly striking in the North-West, West Midlands, and Yorkshire and the Humber regions (Figure S12). During the Omicron wave, the prevalence in White ethnic groups was higher than in Black ethnic groups. It is unclear what drove these complex prevalence patterns which remained generally consistent across regions of England, but natural and vaccine-induced immunity along with differences in mixing behaviour are possible explanations^{23–25}.

SARS-CoV-2 prevalence in secondary school-aged children surged much higher than in other age groups during the VOC waves (Figure 2B), as has been reported previously^{26,27}, likely driven by the higher contact rates of this age group. Prevalence in the oldest age group (75+) was generally lower than for others, although in early-mid 2022, it sharply increased and coalesced with the others, which may have been partly due to the perceived lower risk of severe disease from the Omicron variant, increased social mixing, and the generally high levels of vaccine coverage in this group^{25,27}.

The proportion of PCR tests which were confirmatory varied substantially throughout the evaluation period and according to self-reported ethnicity and age group (Figure 2C&D). Generally, this proportion increased over time, and White people and secondary school-aged children typically had considerably higher proportions of PCRs which were confirmatory than other groups. In addition to asymptomatic testing in schools from March 2021 (Figure 1A), the lower risk of symptomatic disease²⁸ and higher prevalence (Figure 2B) in this age group may partly explain their higher rates of confirmatory PCRs (Figure 2D).

Between October 2020 – December 2021, 60% reported LFD positives were followed up with confirmatory PCRs (Figure S6B). From December 2020 – December 2021, White populations were 1.45 times (IQR: 1.34, 1.51) more likely than Black populations to take PCR tests following a reported LFD positive and 1.33 times (IQR: 1.29, 1.47) more likely than Asian people (Figure 2E). Adherence to confirmatory testing policy was generally similar across age groups apart from those aged 75+ during the period between the end of the Alpha wave and the start of the Delta wave (Figure 2F). Figure 2G shows that from March – May 2021, the false positive rate (proportion of negative confirmatory PCRs) for Black populations was roughly 1.3 times (IQR: 1.07, 1.61) that in White people and roughly 1.85 times (IQR: 1.66, 2.00) that in Asian populations. From May – June 2021, the false positive rate was substantially elevated in those aged 75+ (Figure 2H). At least transiently then, there were times when there were substantial differences across ethnic groups in the risk of isolating unnecessarily if following the governmental policy to self-isolate. False positive rates are inversely related to disease prevalence, with the sensitivity and specificity of the test determining the exact relationship (see Supplementary Materials). This allowed us to estimate the sensitivity (range of central estimates: 66-72%) and specificity (99.5-99.9%) of LFDs by ethnicity and age group which we report in Table S5. While our posterior estimates had wide uncertainty intervals, the medians of these test characteristics were similar between age groups and differed across ethnic groups. These were in line with previous literature estimates^{29,30}.

We also considered how the probability of a positive confirmatory PCR test declined with time elapsing between when the LFD positive was reported and the PCR test subsequently taken. We found that this rate of decay was considerably higher in Black people during the Alpha wave and generally remained above the other groups throughout the study period (Figure 3A), which could indicate that individuals in this ethnic group tended to test and report their positive LFD test later in their infection. Lower values of this parameter later in the study period may indicate that infections were caught earlier by LFD tests or may reflect a longer period of PCR detectability for later VOCs.

Trends in testing biases

Our analysis indicates that infected individuals were much more likely to be tested by PCR than those uninfected. The testing biases for LFDs were considerably lower than for PCRs, meaning symptoms played a lesser role in shaping testing behaviours for LFDs. These results for both PCR and LFD tests are reflective of national testing policy throughout the study period. Throughout the analysis period, the relationship between testing biases (Figure 4C&D) and prevalence (Figure 2A&B) was not strong, but during the Delta wave and Omicron waves, there were visible increases in PCR testing biases (Figure 4C). In February 2022, the government removed the legal requirement for those in England to self-isolate following a positive test result and for close contacts who were fully vaccinated to conduct daily tests³¹. It is possible that this policy change prompted a shift towards increased usage of LFDs for symptomatic testing, which could explain the increasing bias in the use of LFDs in infected individuals (Figure 4C, D).

We found that deprivation and age were associated with testing biases on average, however the uncertainty intervals overlapped across deprivation levels (Figure S13), although the age specific patterns were more distinct (Figure S14). During the invasion phases of the Delta and Omicron VOCs, the most deprived populations were disproportionately more likely to test themselves via PCR only when infected versus when uninfected, and the same pattern held for LFD test-then-reporting. For example, in June 2021, at the start of the Delta wave (the second shaded box in Figure 4A), the average odds of PCR testing in the most-deprived areas were $\exp(3.21) \approx 24$ times higher in individuals with infection compared with individuals without infection; at the same time, the corresponding figure was only ≈ 11 in the least-deprived areas. A range of hypotheses could explain these behavioural differences, such as differential costs of self-isolation, differences in test accessibility and different views about the risks posed by infection, but our data cannot probe these. From mid-February 2022, individuals in areas with different levels of deprivation had more similar PCR test-seeking behaviour. This was likely due in part to the lifting of the legal requirement to self-isolate³¹.

Starting with the Alpha wave, the infection status for those aged 75+ played a lesser role in determining whether individuals tested themselves via PCR than for younger age groups (Figure 4B).

This could be because older individuals were more at risk of severe disease, and they tested themselves more frequently by PCR due to caution¹⁷. It could also be because older people generally have more comorbidities, some of which may produce symptoms similar to COVID-19 causing them to seek PCR tests. We also found that the PCR testing bias was particularly high in school aged children (12-17 years) groups during March –June 2021 (Figure 4B) when LFD was extensively being used (corresponding to a low bias in LFD in Figure 4D)²¹. LFD test reporting biases were more similar across age groups, although the oldest age group had a higher bias early on during the analysis period (Figure 4D).

Incidence and case detection ratio trends

By comparing the estimated true underlying incidence (using REACT-1 and ONS estimates) to reported cases (defined as PCR and LFD positive tests), we found that the national testing programme detected between 26%-40% of all cases across the entire evaluation period (see Methods), although this was likely an overestimate because of potential double counting of LFD and confirmatory PCR tests in the Pillar 2 data as well as multiple tests by the same person. This ratio varied substantially over time and increased until January 2022 after which PCR tests were no longer mandatory following a positive LFD test¹⁶. While the differences in the REACT-1 and ONS derived ratios (Figure 5A) are likely explained by the differences in their sampling strategies (and prevalence estimates), there were no notable persistent differences in case detection ratios according to the deprivation levels of LTLAs (Figure 5B). Throughout much of 2021, Black ethnic groups had higher rates of case detection (Figure 5C), and it is unclear what underpins this relationship, but targeted community testing could have played a role^{19,20}.

Discussion

Our work reveals the intricate and dynamic nature of COVID-19 testing patterns in England throughout our study period (October 2020 - March 2022), illustrating strident differences in test use and reporting between sociodemographic groups. These differences complicate and obfuscate the real-time assessment of transmission across non-traditionally reported groups, and our prevalence estimates point towards an unequal burden of infection across them.

Importantly, these insights were based on the PCR testing programme available to the population and the universal testing offer for LFD tests. However, the LFD tests were also accessible to individuals through a variety of other routes such as school, university, workplace or adult social care settings and we do not distinguish between these. In addition to these two, the targeted community testing service, established in July 2021, aimed to reach disproportionately impacted people and those underserved by the national universal testing offer^{19,20}. These target populations included those in areas of social economic deprivation; ethnic minority groups; those in high-risk occupations; individuals experiencing homelessness or who were sleeping rough; migrants; asylum seekers; refugees; and Gypsy, Roma and Traveller communities. This service was designed and delivered by local authorities and their partners which aimed to more directly reach disproportionately impacted groups. Unfortunately, while tests from this service were included in the data we analysed, targeted testing data were not separately labelled, and we could therefore not summarise this service's data in isolation.

We considered three categories of sociodemographics in this work: self-reported ethnicities of individuals testing or reporting test results; their self-reported age; and a grouping of individuals according to the index of multiple deprivation of the LTLA where each test was taken (PCR) or taken and reported (LFD). We chose these categories of analysis opportunistically according to criteria recorded in the testing data. These are suboptimal because they pool people according to somewhat arbitrary characteristics meaning they lack operational use; if it is known that self-reported Black individuals have a higher burden of infection, how should public health authorities actually respond? This group encompasses a huge diversity of different groups who live and interact in different communities. Existing categorisations also miss undocumented and presumably vulnerable individuals, such as undocumented migrants. Furthermore, there is compelling evidence suggesting a connection between belonging to an ethnic minority group and residing in economically disadvantaged areas, leading to an unequal burden of COVID-19³². While our current study examines these factors in isolation, a potential future expansion could explore their combined impact²². Social

scientists and local community leaders are better placed to understand these nuances, and interdisciplinary collaborations between these and those working in public health should be prioritised^{33,34}.

The CORSAIR study, conducted between March 2020 and January 2021, found that financial hardship, IMD, lower socioeconomic status and having a dependent child in the household were associated with poorer recognition of symptoms and a lower likelihood of booking a test³⁵. Major barriers to testing for people in deprived areas were the financial and practical implications of having to self-isolate following a positive COVID-19 test¹⁸, and those from ethnic minority backgrounds were disproportionately affected by this³⁶. These barriers were also acute for those working in sectors which effectively closed during the pandemic, and women and young people were more likely to fall into this group^{37,38}. The Test and Trace Support Payment scheme³⁹ aimed to provide a financial safety net for those required to self-isolate, but many in need had difficulties accessing the scheme⁴⁰.

The PCR and LFD data from Pillar 2 which we explore offer complementary viewpoints of testing uptake and behaviour since PCR test results reflect test uptake (as the taking of a test resulted in the data being logged failing any mistakes in data recording) whereas LFD test data were available only if individuals self-reported their results. We find that the use (and for LFD, reporting of) these two tests were different throughout the study period, with PCR test use heavily concentrating in those infected and reported LFD positivity giving a more representative view of underlying prevalence; both of these are supportive of the general testing policy in England throughout the period. However, only 15% of LFD tests distributed were reported⁹, and, unfortunately, data on subnational test distribution were unavailable, which prevented granular analysis of LFD uptake and reporting.

During infectious disease epidemics, testing can guide individual- and community-level interventions, for example by allowing self-isolation following exposure, and a return to work following a negative test result. Community surveillance surveys such as the world-leading REACT-1 and ONS-CIS surveys rolled-out in England and the UK, respectively, used representative samples of individuals to produce estimates of SARS-CoV-2 infection prevalence, while adjusting for selection bias due to non-response^{41,42}. Ours and previous work¹⁵ has shown how these high-quality, low-volume surveys can be combined with mass testing data to determine fine scale estimates of disease prevalence for various spatial scales or sociodemographic groups. These data could then be used to determine granular estimates of disease transmission, such as the time-varying reproduction number, R_t . Such nowcasting approaches could be deployed in real-time in future pandemics to improve disease surveillance, particularly to probe transmission in minority risk groups and spatial areas for which community surveillance random sampling surveys in isolation are insufficiently powered to estimate accurately.

The EY-Oxford Health Analytics Consortium membership list

Ricardo Aguas, Ma'ayan Amswych, Billie Andersen-Waine, Sumali Bajaj, Kweku Bimpong, Adam Bodley, Liberty Cantrell, Siyu Chen, Richard Creswell, Prabin Dahal, Sophie Dickinson, Sabine Dittrich, Tracy Evans, Angus Ferguson-Lewis, Caroline Franco, Bo Gao, Rachel Hounsell, Muhammad Kasim, Claire Keene, Ben Lambert, Umar Mahmood, Melinda Mills, Ainura Moldokmatova, Sassy Molyneux, Reshania Naidoo, Randolph Ngwafor Anye, Jared Norman, Wirichada Pan-Ngum, Sarah Pinto-Duschinsky, Sunil Pokharel, Anastasiia Polner, Katarzyna Przybylska, Emily Rowe, Sompob Saralamba, Rima Shretta, Sheetal Silal, Joseph Tsui, Merryn Voysey, Marta Wanat, Lisa White, Gulsen Yenidogan

Acknowledgements

The authors would like to express gratitude to Oliver Munn, Sarah Tunkel, Nick Sharp, Sariyu Shoge and Olutoye Olatunbosun of UKHSA for sponsoring this research and enabling access to the data used in this study. We thank the REACT investigators for their assistance and for access to data from the REACT-1 study. REACT-1 was funded by the Department of Health and Social Care. This study was funded by Secretary of State for Health and Social Care acting as part of the Crown through the UK Health Security Agency (UKHSA), reference number C80260/PRO5331. S.B. is supported by the Clarendon Scholarship and St. Edmund Hall College, University of Oxford and NERC DTP [grant number NE/S007474/1]. C.A.D. is supported by the UK National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emerging and Zoonotic Infections in partnership with Public Health England (PHE), Grant Number: HPRU200907.

References

1. Marmot, M. The health gap: the challenge of an unequal world. *Lancet* **386**, 2442–2444 (2015).
2. Public Health England. *COVID-19: review of disparities in risks and outcomes*. (2020).
3. Bambra, C., Riordan, R., Ford, J. & Matthews, F. The COVID-19 pandemic and health inequalities. *J. Epidemiol. Community Health* **74**, 964–968 (2020).
4. Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).
5. Starke, K. R. *et al.* The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis. *BMJ Glob. Heal.* **6**, e006434 (2021).
6. UK Health Security Agency. Tackling London’s ongoing COVID-19 health inequalities. <https://ukhsa.blog.gov.uk/2021/02/03/tackling-londons-covid-19-health-inequalities/> (2021).
7. Department of Health & Social Care. Coronavirus (COVID-19) Scaling up our testing programmes Department of Health and Social Care [Policy paper]. (2020).
8. Peto, T. *et al.* COVID-19: Rapid antigen detection for SARS-CoV-2 by lateral flow assay: A national systematic evaluation of sensitivity and specificity for mass-testing. *eClinicalMedicine* **36**, (2021).
9. EY-Oxford Health Analytics Consortium. *Evaluation of the national COVID-19 testing programme in England between October 2020 and March 2022*. https://assets.ey.com/content/dam/ey-sites/ey-com/en_uk/resources/evaluation-of-covid-19-testing-in-england-eoha-conference-edition.pdf (2023).
10. Department of Health & Social Care. Twice weekly rapid testing to be available to everyone in England [Press release]. <https://www.gov.uk/government/news/twice-weekly-rapid-testing-to-be-available-to-everyone-in-england>.
11. Naidoo, R. *et al.* A multistage mixed-methods evaluation of the UKHSA testing response during the COVID-19 pandemic in England. *medRxiv* 2022.10.27.22281604 (2022) doi:10.1101/2022.10.27.22281604.
12. 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 (2023).
13. Office of National Statistics. Coronavirus (COVID-19) Infection Survey. <https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/covid19infectionsurvey>.
14. Department of Health & Social Care. *Technical report on the COVID-19 pandemic in the UK*. <https://www.gov.uk/government/publications/technical-report-on-the-covid-19-pandemic-in-the-uk/chapter-6-testing#timeline-of-testing> (2022).
15. Nicholson, G. *et al.* Improving local prevalence estimates of SARS-CoV-2 infections using a causal debiasing framework. *Nat. Microbiol.* **7**, 97–107 (2022).
16. UK Health Security Agency. Confirmatory PCR tests to be temporarily suspended for positive lateral flow test results [Press release]. <https://www.gov.uk/government/news/confirmatory-pcr-tests-to-be-temporarily-suspended-for-positive-lateral-flow-test-results>.
17. Kim, Y., Donnelly, C. A. & Nouvellet, P. Drivers of SARS-CoV-2 testing behaviour: a modelling study using nationwide testing data in England. *Nat. Commun.* **14**, 2148 (2023).
18. Institute of Population Health University of Liverpool. *Covid-SMART Asymptomatic Testing Pilot in Liverpool City Region: Quantitative Evaluation*. https://www.liverpool.ac.uk/media/livacuk/coronavirus/Liverpool_City_Region_Covid_SMA_RT_Evaluation-Feb.pdf.
19. UK Health Security Agency. [confidential internal document] *Targeted Community Testing: Programme Evaluation - A review of policy implementation and delivery (July 2021 – March 2022)*. (2022).
20. UK Health Security Agency & Department of Health & Social Care. *Response: Public Accounts Committee, Twenty-Third Report of Session 2021-22 Department of Health and Social Care Test and Trace - Update*. <https://committees.parliament.uk/publications/8425/documents/85514/default/> (2021).
21. Department for Education. Coronavirus (COVID-19) asymptomatic testing in schools and

- colleges [Updated 30 March 2021; withdrawn 8 July 2021] 2021 30 September 2022].
<https://www.gov.uk/government/publications/coronavirus-covid-19-asymptomatic-testing-in-schools-and-colleges>.
22. Padellini, T. *et al.* Time varying association between deprivation, ethnicity and SARS-CoV-2 infections in England: A population-based ecological study. *Lancet Reg. Heal. – Eur.* **15**, (2022).
23. Kraemer, M. U. G. *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* (80-.). **373**, 889–895 (2021).
24. McCrone, J. T. *et al.* Context-specific emergence and growth of the SARS-CoV-2 Delta variant. *Nature* (2022) doi:10.1038/s41586-022-05200-3.
25. Tsui, J. L.-H. *et al.* Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science* (80-.). **381**, 336–343 (2023).
26. Chadeau-Hyam, M. *et al.* 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. *eClinicalMedicine* **48**, 101419 (2022).
27. Elliott, P. *et al.* Twin peaks: The Omicron SARS-CoV-2 BA.1 and BA.2 epidemics in England. *Science* (80-.). **376**, eabq4411 (2023).
28. Molteni, E. *et al.* Illness duration and symptom profile in symptomatic UK school-aged children tested for SARS-CoV-2. *Lancet Child Adolesc. Heal.* **5**, 708–718 (2021).
29. Mistry, D. A., Wang, J. Y., Moeser, M.-E., Starkey, T. & Lee, L. Y. W. A systematic review of the sensitivity and specificity of lateral flow devices in the detection of SARS-CoV-2. *BMC Infect. Dis.* **21**, 828 (2021).
30. Eyre, D. W. *et al.* Performance of antigen lateral flow devices in the UK during the alpha, delta, and omicron waves of the SARS-CoV-2 pandemic: a diagnostic and observational study. *Lancet Infect. Dis.* **23**, 922–932 (2023).
31. Cabinet office. COVID-19 Response: Living with COVID-19 [Guidance].
<https://www.gov.uk/government/publications/covid-19-response-living-with-covid-19/covid-19-response-living-with-covid-19>.
32. Lo, C.-H. *et al.* Race, ethnicity, community-level socioeconomic factors, and risk of COVID-19 in the United States and the United Kingdom. *eClinicalMedicine* **38**, (2021).
33. Tizzoni, M. *et al.* Addressing the socioeconomic divide in computational modeling for infectious diseases. *Nat. Commun.* **13**, 2897 (2022).
34. Buckee, C., Noor, A. & Sattenspiel, L. Thinking clearly about social aspects of infectious disease transmission. *Nature* **595**, 205–213 (2021).
35. UK Health Security Agency. *Targeted Community Testing: Programme Evaluation - A review of policy implementation and delivery (July 2021 – March 2022) [Confidential Internal Document]*. (2022).
36. Mahmood, F., Acharya, D., Kumar, K. & Paudyal, V. Impact of COVID-19 pandemic on ethnic minority communities: a qualitative study on the perspectives of ethnic minority community leaders. *BMJ Open* **11**, e050584 (2021).
37. UK Health Security Agency. *London Borough of Havering Isolation Outreach Pilot Evaluation [Confidential Internal Document]*.
38. House of Commons. *Unequal impact? Coronavirus and the gendered economic impact. Fifth Report of Session 2019–21*.
<https://committees.parliament.uk/publications/4597/documents/46478/default/> (2021).
39. UK Health Security Agency. Claiming financial support under the Test and Trace Support Payment scheme. (2022).
40. BBC. Covid: How many people get self-isolation payments? (2021).
41. Elliott, J. *et al.* Predictive symptoms for COVID-19 in the community: REACT-1 study of over 1 million people. *PLOS Med.* **18**, e1003777 (2021).
42. Office of National Statistics. Coronavirus (COVID-19) Infection Survey: quality and methodology information (QMI). (2023).

Tables

	Number of tests	Number of positive results (%)
PCR tests performed*		
With LTLA	106,986,026	14,021,052 (13·1)
With ethnicity	106,831,682	13,933,106 (13·0)
With age	103,681,935	13,554,671 (13·1)
LFD tests reported*		
With LTLA	290,462,387	7,007,391 (2·4)
With ethnicity	290,227,237	6,991,160 (2·4)
With age	286,525,766	6,441,966 (2·2)
Confirmatory PCR tests	Number of LFD tests	Number of PCR positive results (%)
LFD positives followed by PCR	2,328,414	2,069,391 (88·9)
Within same day	1,598,213	1,416,229 (88·6)
After one day	513,567	461,159 (89·8)
After two days	209,143	185,511 (88·7)
After three days	7491	6492 (86·7)

Table 1: Summary of Pillar 2 mass testing data for England between 1st October 2020 and 31st March 2022. * includes tests where there were data with information on LTLAs or ethnicity or age.

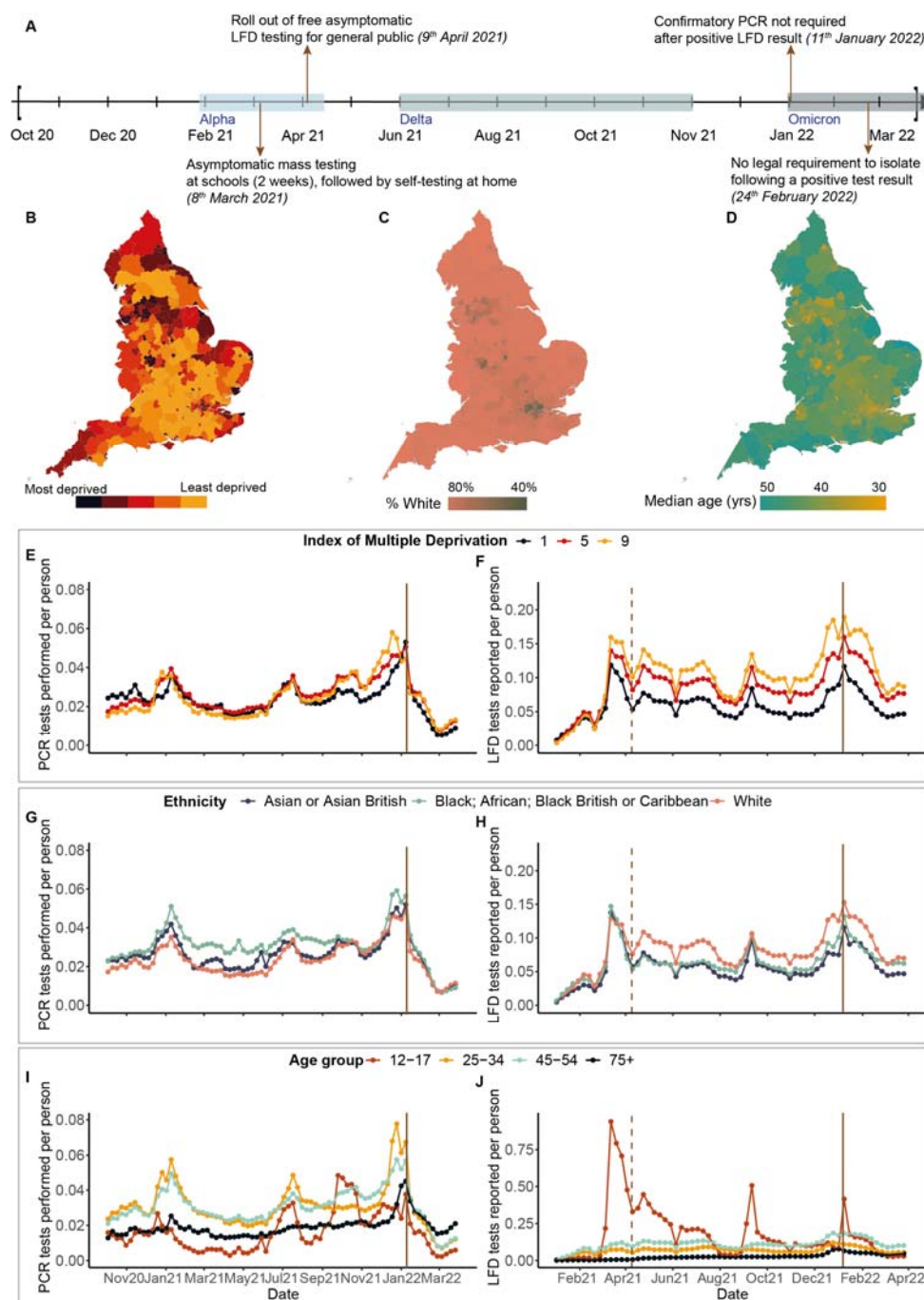


Figure 1: (A) Timeline of the Alpha, Delta, and Omicron variants of concern and other important dates mentioned in our study (B) Spatial distribution of deprivation across England according to the rank of population weighted average of scores for the LSOAs in a given LTLA. (C) Proportion of White people across LTLAs. (D) Median age across LTLAs of England. (E, F) Weekly number of PCR, LFD tests per person by deprivation levels. All LTLAs were assigned deprivation levels from 1 to 9. (G, H) Weekly number of PCR, LFD tests per person by three ethnic groups (see Figure S10 for the other two ethnic groups). (I, J) Weekly number of PCR, LFD tests per person by four age groups (see Figure S9 for the other age groups). The solid vertical line corresponds to the date (11th January 2022) after which confirmatory PCR tests were no longer required following a positive LFD test, and the vertical dashed line corresponds to the date (9th April 2021) when free asymptomatic testing was made available to all individuals in England. See Figure S11 for variation in tests per person by LTLAs and sociodemographic groups.

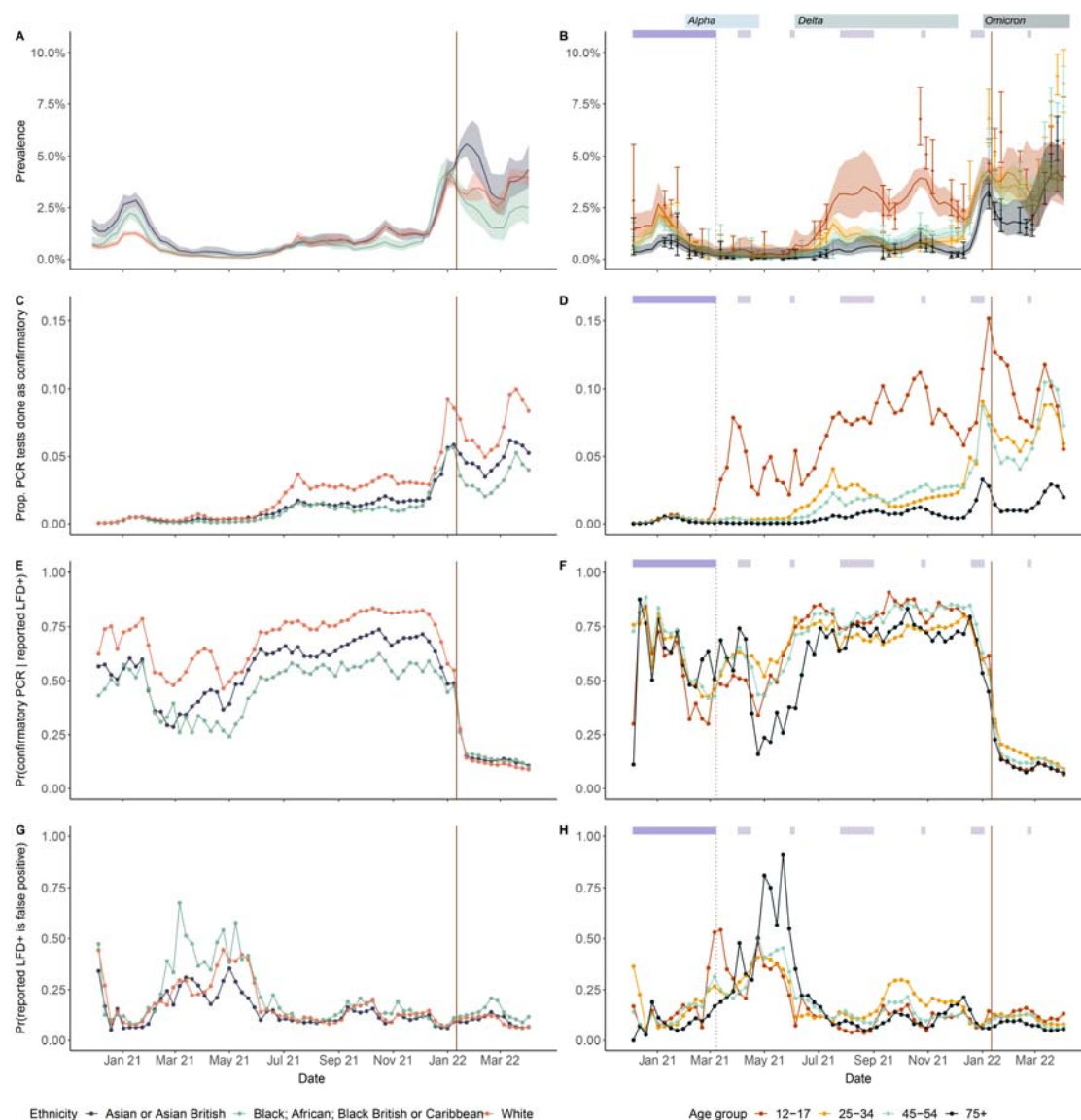


Figure 2: (A) Aggregated weekly national prevalence estimates for ethnic groups. This was the population-weighted average of posterior median estimates (solid lines) and 95% central credible intervals (shaded region) of ethnic group-specific debiased prevalences in LTLAs. (B) Aggregated weekly national prevalence estimates for age groups and 95% central credible intervals (solid and shaded areas respectively). National prevalence and 95% confidence intervals for age groups estimated by REACT-1 are shown as points and whiskers. (C, D) Proportion of PCR tests performed as confirmatory tests by ethnic and age groups. (E, F) Proportion of reported LFD positive tests followed by PCR confirmatory tests within 72 hours among different ethnic and age groups. (G, H) The probability of testing negative with a confirmatory PCR test after reporting a positive LFD test, among different ethnic and age groups, conditional on testing positive with an LFD, reporting the positive LFD test and seeking a confirmatory PCR test. The solid vertical line corresponds to the date (11th January 2022) after which confirmatory PCR tests were no longer required following a positive LFD test, and the vertical dotted line corresponds to the date (8th March 2021) when onsite asymptomatic testing was made available in schools. Restrictions on face-to-face education and school holidays are indicated by the dark and light purple rectangles, respectively. The blue rectangles represent the time periods of Alpha, Delta, and Omicron VOCs as defined in Methods. The date axis starts in December 2020, when LFD tests started being reported.

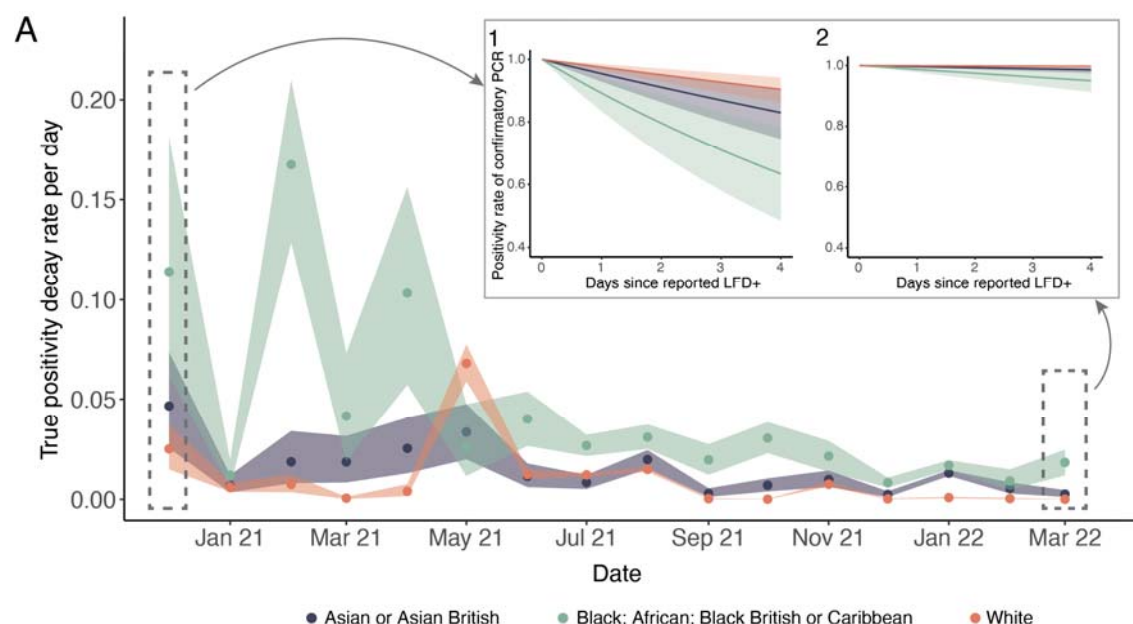


Figure 3: (A) Rate of decline in the probability of detecting a SARS-CoV-2 infection by PCR within 3 days from reporting a positive LFD test over the study period by ethnic groups. Solid dots (and shaded areas) in the main figures are the posterior median estimates (and 50% central credible intervals) of ethnic group specific decay rates. The inset plots translates what the rate estimates in A correspond to in terms of the rate of decline in probability of testing positive by PCR subsequent to reporting of an LFD positive (with 50% credible intervals) for the first (1: December 2020) and last month in our study (2: March 2022).

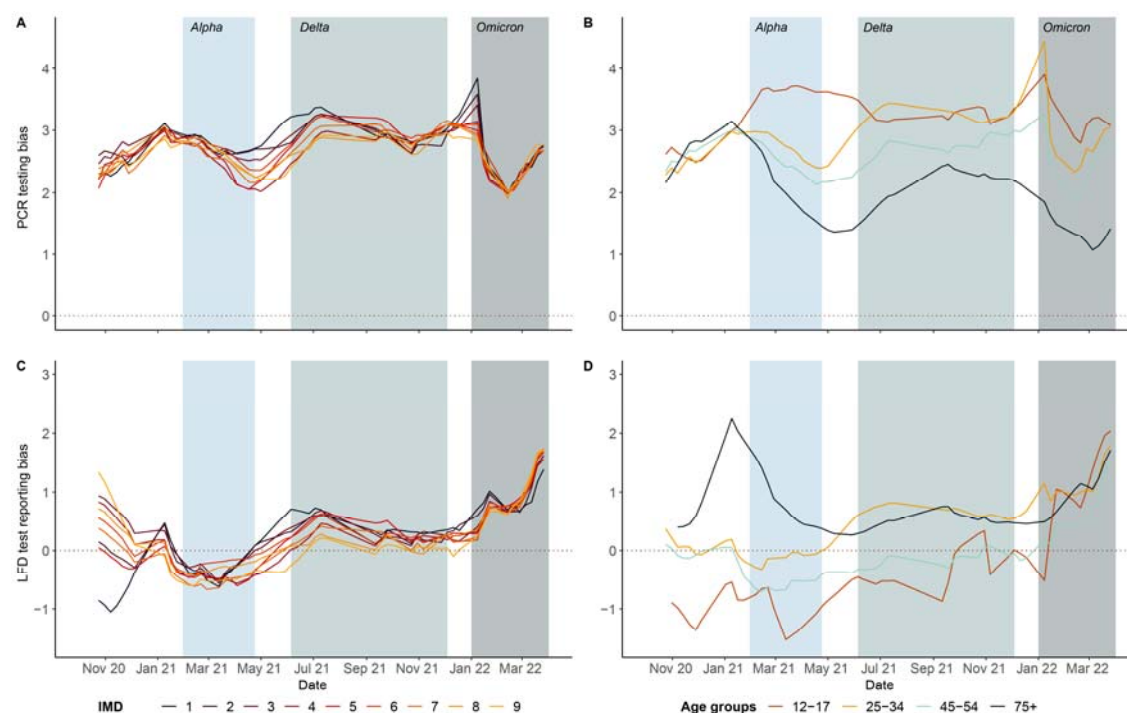


Figure 4: (A, B) Estimated mean PCR testing bias (log odds of taking a PCR test in infected versus uninfected subpopulations) by deprivation levels and age groups in England over time. (C, D) Estimated mean LFD test reporting bias in deprivation levels and age groups. 1 and 9 corresponding to most deprived and least deprived levels respectively. Shaded areas from left to right, correspond to the time periods of Alpha, Delta, and Omicron VOCs respectively (defined in Methods). Testing biases with uncertainty intervals and other age groups are provided in Figure S13, S14.

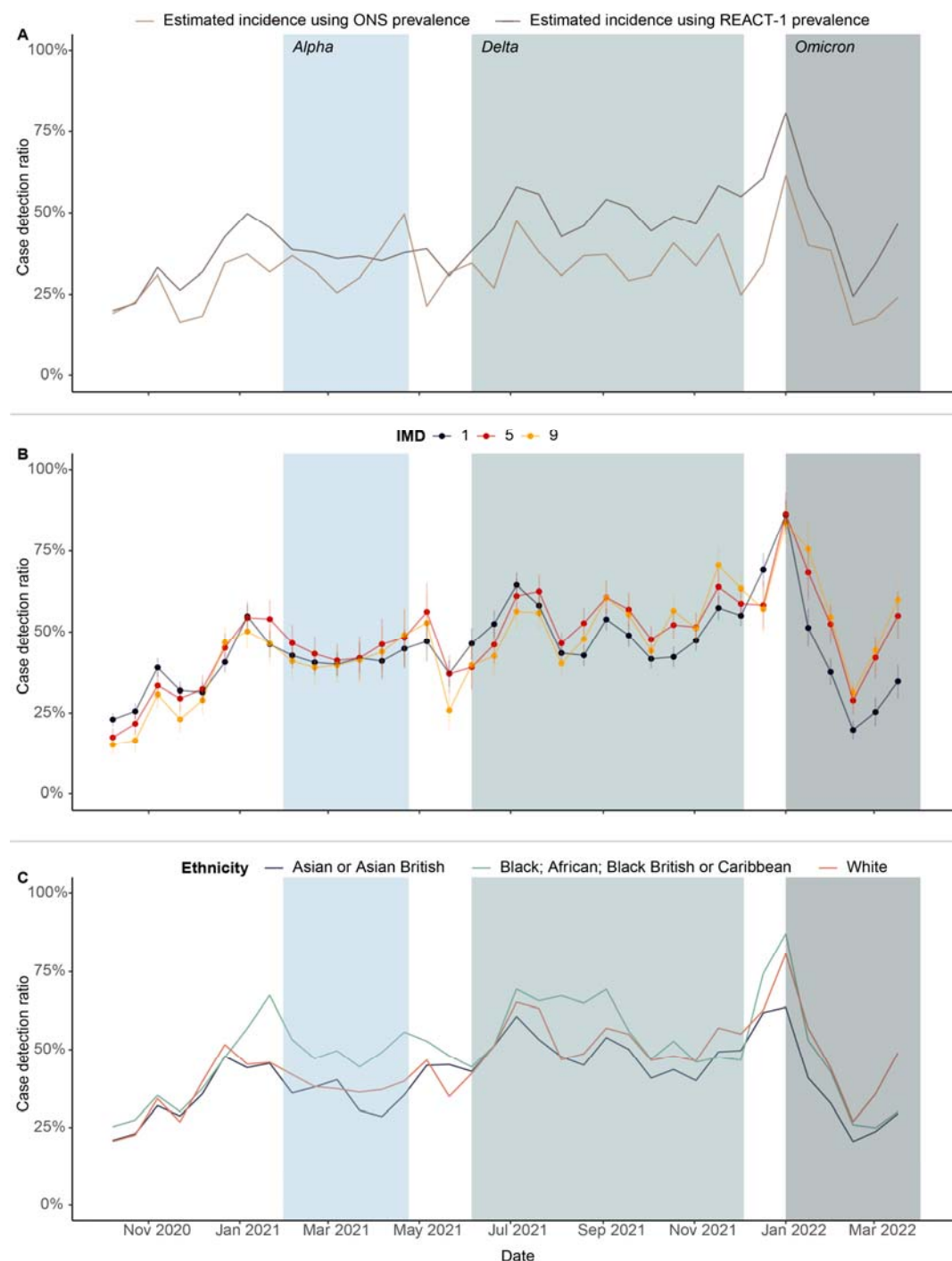


Figure 5: (A) Bi-monthly national estimated case detection ratio over time, using PCR and LFD positive and estimated incidence. The dark and light brown curves represent ratios using REACT-1 and ONS prevalence data, respectively. (B) Case detection ratio over time by deprivation levels (IMD). The dots (and error bars) represent the coefficient estimates (and standard error) of a linear regression with LTLA specific case detection ratio as the dependent variable and IMD levels as predictors for each month. Here IMD level 1 (black) and 9 (yellow) refer to the most- and least-deprived areas, respectively. (C) Bi-monthly national case detection ratio tests by ethnic groups. Shaded areas from left to right, correspond to the time periods of Alpha, Delta, and Omicron VOCs respectively (defined in Methods).