

Human-environment interactions on infectious disease risk in Brazil

A thesis submitted for the degree of

Doctor of Philosophy

Sabrina Luyao Li (Sabrina.li@ouce.ox.ac.uk)

Keble College

University of Oxford

Michaelmas 2021

Abstract

Emerging and re-emerging infectious diseases pose a major threat to global health, yet little is known about their complex socio-ecological pathways that impact population health disproportionately at fine spatial scales. This is common in low-and middle-income countries with stark inequalities such as Brazil, which faces unprecedented changes in land-use and climate, and the burden of several epidemics, including the Zika virus, yellow fever virus, and SARS-CoV-2 responsible for the COVID-19 pandemic. This thesis presents research conducted on these diseases to advance knowledge on the relationships of ecological-epidemiological factors in natural and socio-institutional settings, which shape differential population risk to infectious diseases in Brazil.

Guided by the theoretical frameworks of disease ecology and landscape epidemiology, this thesis uses large open-access datasets, GIS, and statistical modelling to explore the geographic spread of Zika virus, the emergence of yellow fever virus in the natural environment, differential risk to COVID-19 spread in the built environment, and population response to COVID-19 via immunisation. Forest cover and temperature fostered the environmental suitability of yellow fever, while race, income, and healthcare systems contributed to inequalities in access to COVID-19 testing, infection, and death. Mass vaccination campaigns illustrated in the context of COVID-19 showed that vaccine coverage varies by area, age, race, sex, and socioeconomic status. In these settings, population interconnectivity via settlement, mobility, and behaviour significantly influenced disproportionate geographic exposure to disease risk.

Findings and implications from this research can be used to inform public health policies and interventions in Brazil. Future research should continue to explore the integration of large open-access datasets for understanding geographical inequalities in disease risk.

Acknowledgements

This thesis was the result of a three year and two-month journey with many people that supported me along the way. I am very grateful to my supervisors Janey Messina and Oliver Pybus, as well as Danny Dorling, for encouraging and supporting me in all my academic endeavours. I've learned so much from you all and I am eternally grateful for all the academic opportunities and resources that you have all provided me. I want to thank many of my research collaborators within the UK-Brazil Centre for Arbovirus Discovery, Diagnosis, Genomics, and Epidemiology (CADDE) and the Oxford Martin School. Special thanks go to Nuno Faria, Moritz Kraemer, Alex Zarebski, Rafael Pereira, Andreza de Souza Santos, and Carlos Prete Jr. for helping to shape my research interests and understanding of Brazil and infectious diseases. I'd also like to thank my viva examiners, Lisa Wedding and Mike Emch, for your helpful insights and for giving inspiration for another paper,

I want to acknowledge the Canadian Social Sciences and Humanities Council, Oxford Martin school, Oxford School of Geography and the Environment, and Keble College for their financial support and resources that made my DPhil possible.

Despite doing most of my DPhil in lockdown and during the COVID-19 pandemic, I felt very supported by my friends, old and new, at Oxford and abroad. Thank you all for the new and lasting friendships and for the walks, chats, and dinner parties.

Lastly, to my parents – ma and ba. Thank you for your endless encouragement and the countless sacrifices that you've made during the past 29 years to support me. I would not have done this without you both.

Table of Contents

Abstract ..... *ii*

Acknowledgements ..... *iii*

Thesis Outline ..... *vi*

1 Introduction *1*

1.1 Background **1**

1.2 Research aim and significance **4**

1.3 References **7**

2 Literature review *9*

2.1 Background on Brazil **9**

2.1.1 Geography 9
2.1.2 Demographics 11
2.1.3 Government and healthcare 12
2.1.4 Arboviral diseases 13
2.1.5 Coronavirus disease 2019 (COVID-19) 14
2.1.6 History of vaccination in Brazil 15
2.1.7 COVID-19 vaccination strategy 16
2.1.8 COVID-19 vaccine hesitancy and confidence 16

2.2 Theoretical frameworks **17**

2.2.1 Disease ecology 17
2.2.2 Landscape epidemiology 19

2.3 Methodology **20**

2.3.1 Recent advances in infectious disease mapping 20
2.3.2 Modelling the geographical spread of infectious disease: A case study on Zika virus 21

2.4 References **32**

3 The physical environment and yellow fever virus emergence *36*

4 Differential risk to COVID-19 in the built environment *70*

4.1 Differential access to diagnosis in the Metropolitan Region of São Paulo **70**

4.1.1 Introduction 70
4.1.2 Data and methods 71
4.1.3 Results 73
4.1.4 Discussion 76
4.1.5 References 77

4.2 Differential vulnerability to hospitalisation and death in São Paulo state **79**

5 Population response to COVID-19 via mass immunisation *91*

6 Discussion *112*

6.1 Chapter Summary **112**

6.2 Directions for Future Research **120**

6.3 References **121**

Appendices *122*

Statement of Contributions **123**

Chapter 3 Supplementary Information **126**
Chapter 4.1 Supplementary Information **139**
Chapter 4.2 Supplementary Information **143**
Chapter 5 Supplementary Information **168**

Thesis Outline

This thesis comprises of six chapters. Except for the introductory and concluding chapters, each chapter will comprise of one or more research articles that have been published in peer-reviewed journals, in submission, or will be submitted. A paper route was selected due to the variety of case studies conducted, where they are connected by the theme and research aims of this thesis. Some articles are outputs from a collaborative effort with members from the UK-Brazil Centre for Arbovirus Discovery, Diagnosis, Genomics, and Epidemiology (CADDE) and other international collaborators. While full author contributions are provided at the beginning and end each article, I have also outlined my own contributions below.

Chapter 1 begins with the introduction, which outlines the background context of this thesis and its research aims and objectives.

Chapter 2 is a literature review, which begins with background context on the study area and the diseases explored in this thesis. Next, an overview of the theoretical frameworks of disease ecology and landscape epidemiology, which were adopted to guide this research, is provided. This chapter concludes with details on the methodology, which is based on spatial modelling techniques for predicting and evaluating infectious disease risk. Here, I present a review paper that I first-authored, which explores the application of the chosen methodology to investigative the geographic spread of Zika virus, the last infectious disease epidemic in Brazil that was declared as a Public Health Emergency of International Concern by the World Health Organization (WHO). While this paper focuses on Zika, the spatial modelling methods featured in this publication have been applied to Chapters 3 and 4.

Paper 1: **Li, S.L.,** Messina, J.P., Pybus, O.G., Kraemer, M.U.G., Gardner, L. (2021). A review of models applied to the geographic spread of Zika virus. Transactions of the Royal Society of Tropical Medicine and Hygiene, DOI: https://doi.org/10.1093/trstmh/trab009

Chapter 3 investigates how physical environmental factors and anthropogenic changes contribute to the re-emergence of yellow fever virus, an arboviral disease that is endemic to Brazil and have caused sporadic outbreaks in recent years. This chapter consists of a research paper that I first authored, which explores the ecological niche of sylvatic (rural/jungle) mosquito vectors known to transmit yellow fever virus by mapping the spatial distribution of their environmental suitability at 1 x 1km resolution, and identifies the areas at risk of transmission and spillover into human populations.

Paper 2: **Li, S. L. ,** Acosta A.L., Hill, S.C., Brady, O.J., … Messina, J.P., et al. (2021). Mapping environmental suitability of Haemagogus and Sabethes to understand sylvatic transmission risk of yellow fever virus in Brazil. *PLOS Neglected Tropical Diseases - Accepted.*

Chapter 4 explores how human populations interact with the social and built environments to facilitate differential infectious disease risk. Specifically, human dimensions of the recent COVID-19 epidemic in Brazil from a social determinants of health perspective will be discussed through two publications. The first paper discusses the epidemiological characteristics of the early phase of the COVID-19 epidemic in Brazil, and is part of a wider collaborative effort with CADDE where I contributed as a joint first author. Specifically, I led the geospatial analysis that assessed geographical inequalities in access to COVID-19 testing and its linkages to income in the greater metropolitan of São Paulo, where the first COVID-19 cases were detected in Brazil. This analysis will be featured as a paper excerpt in this thesis. The second paper, which I am the leading joint first author, focuses on São Paulo state and investigates how racial and socio-institutional inequalities have shaped differential risk to COVID-19 hospitalisation and death. These two papers complement each

other, as they illustrate geographical inequalities in risk during the early and later phase of the COVID-19 epidemic, respectively, in one of the most populous states in Brazil.

Paper 3: de Souza, W. M.*, Buss, L. F.*, da Silva Candido, D.*, Carrera, J. P.*, **Li, S.*,** Zarebski, A., ... & Faria, N. R. (2020). pidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nature Human Behaviour **4,** 856–865 (2020). DOI: https://doi.org/10.1038/s41562-020-0928-4
*joint first authors

Paper 4**: Li, S. L.*,** Pereira, R. H.*, Prete, C. A.*, Zarebski, A. E., Emanuel, L., Alves, P. J., ... & Messina, J. P. (2021). Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil. BMJ Global Health 2021;0:e004959. DOI: http://dx.doi.org/10.1136/bmjgh-2021-004959
*joint first authors

Chapter 5 draws on findings from the previous chapter to investigate population response to mitigate infectious disease risk through immunisation, while considering the complex interplay between socio-economic and institutional factors. It investigates the geographical variation in vaccination coverage of the COVID-19 vaccine, and potential factors that influence vaccine access and uptake. This paper has been submitted to a scientific journal to be considered for publication.

Paper 5: **Li, S. L**.*, Prete, C. A.*, Zarebski, A. E.*.. & Messina, J. P. (2021). Socio-demographic inequalities in COVID-19 vaccination in Brazil. *Submitted.*
*joint first authors

Chapter 6 is the concluding chapter. Drawing on findings from the previous Zika epidemic, the yellow fever endemic, and the ongoing COVID-19 pandemic, this chapter discusses the overall findings and their implications. It also summarises the limitations and of the research presented in this thesis and outline recommendations and knowledge gaps for future research. The appendices include the supplementary information that complement the papers that I authored and featured in the chapters.

1 Introduction

1.1 Background

Increasing global connectivity, urbanisation, and climate change have redefined the complex relationship between humans, animals, and their environment in the 21st century. The pandemic era has directed mass attention to emerging and re-emerging pathogens that survive in nature, and their capability to cause infectious diseases in humans through zoonotic spillover. Globally, zoonotic emerging infectious disease risk is elevated in forested tropical regions with rich biodiversity and increasing land-use changes (1). This characteristic profiles Brazil, which continues to experience significant disease burden and public health challenges from infectious diseases with zoonotic origin. In addition to having the third highest number of COVID-19 cases globally (as of November 2021), Brazil is plagued with coexisting sporadic epidemics from vector-borne diseases, which are transmitted to humans through vectors such as arthropods (e.g. mosquitoes), and can cause severe illness and death in humans. Many vector-borne diseases, including well known diseases such as Zika, malaria, and West Nile fever, are caused by only a select few species including *Aedes*, *Anopheles*, and *Culex* (2). In order to prevent the next global health threat, more local evidence is needed on how ecological-epidemiological factors contribute to infectious disease risk in the context of low- and middle-income countries (LMICs) such as Brazil, which are experiencing rapid change in land use. This will help to broaden existing perspectives on human interactions with the natural environment and help identify natural and socio-institutional settings that contribute to infectious disease emergence.

Human exposure to disease risk arises from direct and indirect interactions with the environment containing the animal or human host of the pathogen. Many zoonotic and

vector-borne diseases that infect humans arise from direct human-animal interaction through intentional contact fostered by legal and illegal wildlife trade (3) and land-use change (4). The alteration and repurposing of the natural landscape to urban, agricultural, or anthropogenic ecosystems can modify the spatial distribution and niche of the pathogen, vector, and host, including their behaviour and movement, as well as their community composition (5). Furthermore, land-use change redefines the distribution of socioeconomic and environmental conditions and their mediating effects among human populations. Human contact results in humans being hosts of the pathogen, which may trigger a chain of transmission events when they come into contact with other humans in their surroundings. Disease risk can also develop incidentally through frequent human contact with an environment that contains wildlife capable of hosting pathogens. In the case of mosquito-borne diseases in the Atlantic forest biome of Brazil, the rise in ecotourism has led people to pursue a lifestyle that integrates frequent contact with the natural environment, especially the forests in the Amazonas, which has driven the transmission of viruses from mosquitoes to humans (6). Indirect modification of the natural environment could degrade the original habitats of mosquitoes and non-human primates, which act as vertebrate hosts for certain arboviral diseases such as yellow fever (7). This process has led to mosquito species having a proclivity for migrating towards ecotones between the natural and built environment and between forest patches, leading to dispersion of infectious agents (8).

Due to optimal environmental and ecological conditions that foster mosquito breeding and socioeconomic factors (e.g. poverty, poor housing, etc.) that facilitate human exposure to mosquito bites, the occurrence of arboviral diseases is common in Brazil year-round. Among these diseases, the most widespread mosquito-borne pathogens are the four dengue virus serotypes (DENV 1-4), chikungunya virus (CHIKV), and in recent years, Zika virus (ZIKV); all are transmitted by *Aedes aegypti* mosquitoes that thrive in urban and peri-

urban environments (9). In rural environments, the circulation of yellow fever virus (YFV) maintained by a sylvatic transmission cycle between non-human primates and sylvatic mosquitoes has caused occasional spillover outbreaks resulting in a high number of human cases and deaths (10). Despite arboviral diseases being nationally notifiable in Brazil, their co-endemicity (simultaneous circulation of more than one arbovirus), which is common in many parts of Brazil (9), makes clinical notification complex. Moreover, only a small number of cases notified through public healthcare undergo laboratory confirmation via virus-specific molecular or serological assays.

While Brazil continues to cope with concomitant illnesses caused by arboviruses and other pathogens, the emergence of COVID-19, a respiratory disease caused by severe acute respiratory syndrome coronavirus (SARS-CoV-2), could further complicate clinical diagnosis (11,12). In Brazil, arboviral outbreaks take a seasonal pattern where increased incidence is observed at the beginning of the year between the months of March and April that corresponds to high temperatures and precipitation. In the first quarter of 2020, the increase in the number of dengue cases surpassed the numbers reported in 2019 (13); more than 2.1 million dengue cases were reported during the 2019-early 2020 period in Brazil. However, the number of dengue case reports declined after March 2020, which coincided with the introduction of COVID-19 in Brazil (14). It is likely that case notification has decreased or delayed due to the reallocation of epidemiological surveillance, diagnostic laboratories, and healthcare resources towards addressing COVID-19 epidemic and associated severe health outcomes.

As of October 2021, COVID-19 has resulted in over 21.7 million cases more than 600,000 deaths (world's third highest) in Brazil alone. Despite having a universal healthcare system and a history of managing infectious disease epidemics, Brazil was unprepared for the

epidemic. Conflicting messages on non-pharmaceutical interventions from national and local governments and overall administration of the epidemic have contributed to rapid and widespread infection (15). In January 2021, a new Variant of Concern (VOC) known as the P.1 lineage was detected in Japan from travellers coming from the capital city of the northern Amazonas state, Manaus. It is known to be 1.4 – 2.2 times more transmissible than non-VOC lineages (16). To combat severe disease and death risk, vaccination against COVID-19 started on 17 January with high-risk priority groups using two authorised vaccines, CoronaVac (Sinovac Biotech, China) and Covishield (Oxford-AstraZeneca, U.K.). Brazil experienced an initial slow vaccine roll-out and concomitant vaccine shortage amid the rapid transmission of P.1, which pushed its public health system to a brink of collapse in early 2021 (17).

Even though public health resources have been mobilized for COVID-19, Brazil has limited COVID-19 testing which is mostly reserved for severe cases (14). Given that more than 70% of the Brazilian population depends on public healthcare (18), the current public healthcare system may be overwhelmed by the concurrent high number of arbovirus infections and COVID-19 cases in the long term. In order to understand infectious disease emergence and the factors that facilitate transmission and escalate risk, a better understanding of how humans interact with their changing environments is urgently needed.

1.2 Research aim and significance

Despite significant decrease in the global burden of infectious diseases in the last three decades (19), the complex relationship between human behaviour and the environment has continued to redefine the pathogen-host interactions and their multi-dimensional impact on health and wellbeing. Human disturbances of the natural environment through anthropogenic activities have continuously expanded the reservoirs of zoonotic viruses

closer to human settlements, triggering the spillover of emerging and re-emerging infectious diseases. The increase in global connectivity, alongside the magnitude and frequency of human travel have enabled the dynamic spatial redistribution of local emerging diseases in unprecedented ways. This has been illustrated by the spillover of SARS-CoV-2 from wildlife reservoirs into the local urban environment in China, which was then exported to more than 180 countries as of April, 2021.

Brazil is a multi-faceted and extremely diverse country with characteristics that makes it a suitable study area of interest. Firstly, its climate and the natural environment provide suitable habitats for wildlife vectors, but moreover, ongoing climate and land-use changes have instigated the emergence of new vector-borne diseases. Secondly, despite having the largest economy in Latin America, it is consistently ranked as one of the most unequal countries in the world, with stark racial, social, and health disparities. The coexistence of several continuous arboviral epidemics, including ZIKV and YFV, is now coupled with the recent COVID-19 pandemic, creating complex public health challenges that require novel solutions. In order to address these major challenges, an understanding of the root drivers of disease emergence is needed.

This research aims to estimate and understand geographical inequalities in infectious disease risk in Brazil derived from human-environment interactions. It draws from the theoretical frameworks of human disease ecology and landscape epidemiology combined with spatial epidemiological and GIS methods to understand how human decisions around interactions with the physical, built, and social environments shape infectious disease risk. By following the life course of infectious diseases from “emergence”, “spread”, to “response”, my thesis will draw on case studies from recent infectious disease outbreaks and epidemics

in Brazil to address the following three research questions (RQ) and six corresponding objectives (O):



Figure 1.1 The life trajectory of infectious diseases.

Emergence

RQ1. How does the physical environment contribute to disease emergence and interact with human populations to create suitable conditions for outbreaks?

(O1) Develop data and models to investigate how climate and environmental factors cultivate the spatial distribution and niche of arbovirus vectors, focusing on those known to transmit yellow fever virus.

(O2) Identify human populated areas at risk of transmission and spillover of yellow fever virus.

Spread

RQ2. How do interactions between human populations and their environment facilitate differential susceptibility to infectious disease risk?

(O3) Map the geographical inequalities in population-level COVID-19 risk and identify its connection to socioeconomic status.

(O4) Develop models to examine how population vulnerability to COVID-19 risk varies by behaviour and socio-economic status.

Response

RQ3. How can our understanding of these interactions aid public health response through immunisation to infectious disease epidemics?

(O5) Integrate data on hospitalisations and deaths, vaccination coverage, and socioeconomic data to understand potential COVID-19 vaccine uptake during the first vaccination campaign.

(O6) Conduct analyses to identify the facilitators and barriers of initial COVID-19 immunisation access and uptake.

To understand the human dimensions of infectious disease risk, I utilise various spatial modelling techniques to predict how human-environment dynamics contribute to spread and map the geographical distribution in exposure and risk. Data, methods, findings and data visualisations from this thesis will be shared with researchers, knowledge users, public health officials to facilitate knowledge dissemination and to help develop evidenced-based public health policy.

1.3 References

1. Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, Di Marco M, et al. Global hotspots and correlates of emerging zoonotic diseases. Nat Commun. 2017 Oct 24;8(1):1124.
2. WHO. Vector-borne diseases [Internet]. WHO; 2020 [cited 2021 Nov 5]. Available from: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases
3. Bezerra-Santos MA, Mendoza-Roldan JA, Thompson RCA, Dantas-Torres F, Otranto D. Illegal Wildlife Trade: A Gateway to Zoonotic Infectious Diseases. Trends Parasitol. 2021 Mar 1;37(3):181–4.
4. Plowright RK, Reaser JK, Locke H, Woodley SJ, Patz JA, Becker DJ, et al. Land use-induced spillover: a call to action to safeguard environmental, animal, and human health. Lancet Planet Health [Internet]. 2021 Mar 6 [cited 2021 Mar 26]; Available from: https://www.sciencedirect.com/science/article/pii/S2542519621000310
5. Gottdenker NL, Streicker DG, Faust CL, Carroll CR. Anthropogenic Land Use Change and Infectious Diseases: A Review of the Evidence. EcoHealth. 2014 Dec;11(4):619–32.
6. Possas C, Lourenço-de-Oliveira R, Tauil PL, Pinheiro F de P, Pissinatti A, Cunha RV da, et al. Yellow fever outbreak in Brazil: the puzzle of rapid viral spread and challenges for immunisation. Mem Inst Oswaldo Cruz [Internet]. 2018 Sep 3 [cited 2019 Feb 22];113(10). Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02762018001000200&lng=en&tlng=en
7. de Azevedo Fernandes NCC, Guerra JM, Díaz-Delgado J, Cunha MS, Saad L delC., Iglezias SD, et al. Differential Yellow Fever Susceptibility in New World

Nonhuman Primates, Comparison with Humans, and Implications for Surveillance. Emerg Infect Dis. 2021 Jan;27(1):47–56.
8. Chaves LSM, Conn JE, López RVM, Sallum MAM. Abundance of impacted forest patches less than 5 km 2 is a key driver of the incidence of malaria in Amazonian Brazil. Sci Rep. 2018 May 4;8(1):7077.
9. Magalhaes T, Chalegre KDM, Braga C, Foy BD. The Endless Challenges of Arboviral Diseases in Brazil. Trop Med Infect Dis [Internet]. 2020 May 9 [cited 2021 Jan 18];5(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7345859/
10. de Oliveira Figueiredo P, Stoffella-Dutra AG, Barbosa Costa G, Silva de Oliveira J, Dourado Amaral C, Duarte Santos J, et al. Re-Emergence of Yellow Fever in Brazil during 2016–2019: Challenges, Lessons Learned, and Perspectives. Viruses [Internet]. 2020 Oct 30 [cited 2021 Mar 13];12(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7692154/
11. Ribeiro VST, Telles JP, Tuon FF. Arboviral diseases and COVID-19 in Brazil: Concerns regarding climatic, sanitation, and endemic scenario. J Med Virol. 2020;92(11):2390–1.
12. Silva SJR da, Magalhães JJF de, Pena L. Simultaneous Circulation of DENV, CHIKV, ZIKV and SARS-CoV-2 in Brazil: an Inconvenient Truth. One Health. 2021 Jun 1;12:100205.
13. Lorenz C, Dias Bocewicz AC, Corrêa de Azevedo Marques C, Reis Santana LM, Chiaravalloti-Neto F, Alves Gomes AH, et al. Have measures against COVID-19 helped to reduce dengue cases in Brazil? Travel Med Infect Dis. 2020;37:101827.
14. de Souza WM, Buss LF, Candido D da S, Carrera J-P, Li S, Zarebski AE, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat Hum Behav. 2020 Aug;4(8):856–65.
15. Castro MC, Kim S, Barberia L, Ribeiro AF, Gurzenda S, Ribeiro KB, et al. Spatiotemporal pattern of COVID-19 spread in Brazil. Science [Internet]. 2021 Apr 14 [cited 2021 Apr 14]; Available from: https://science.sciencemag.org/content/early/2021/04/13/science.abh1558
16. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science [Internet]. 2021 Apr 14 [cited 2021 Apr 14]; Available from: https://science.sciencemag.org/content/early/2021/04/13/science.abh2644
17. Taylor L. Covid-19: Brazil's hospitals close to collapse as cases reach record high. BMJ. 2021 Mar 23;372:n800.
18. WHO. Flawed but fair: Brazil's health system reaches out to the poor [Internet]. World Health Organization; 2008 [cited 2021 Mar 26]. Available from: https://www.who.int/bulletin/volumes/86/4/08-030408/en/
19. Roser M, Ritchie H. Burden of Disease [Internet]. OurWorldInData.org; 2016. Available from: https://ourworldindata.org/burden-of-disease

2 Literature review

2.1 Background on Brazil

2.1.1 Geography

Brazil is situated on the eastern coastline of South America along the Atlantic Ocean (14.23° S, 51.93° W) sharing geographic borders with every South American country except Chile and Ecuador (Figure 2.1.1). It is the largest country in South America, and has a land area of 8.52 million km^2. The Brazilian landscape is mainly tropical to subtropical and includes wetlands, savannas, plateaus, and low mountains. Brazil has six major ecosystems. Situated in the north is the Amazon River basin, which consists of the world's largest river system and rainforest. The Cerrado is a vast savanna region that covers most of central region and experiences a strong dry season. Towards the northwest is the Pantanal, which encompasses the world's largest tropical wetlands and flooded grasslands, while the northeast is the Caatinga, which comprises of small thorny scrublands. From the Northeast to the South along the coast is the highly biodiverse Atlantic Forest, which has experienced 55-70 thousand km^2 of forest loss due to deforestation in recent years (1). The far South is the fertile Pampas lowlands. With eight drainage basins that all drain into the Atlantic Ocean, Brazil has one of the most extensive river systems in the world. The Amazon and Tocantins-Araguaia basins account for majority of the total drainage area.

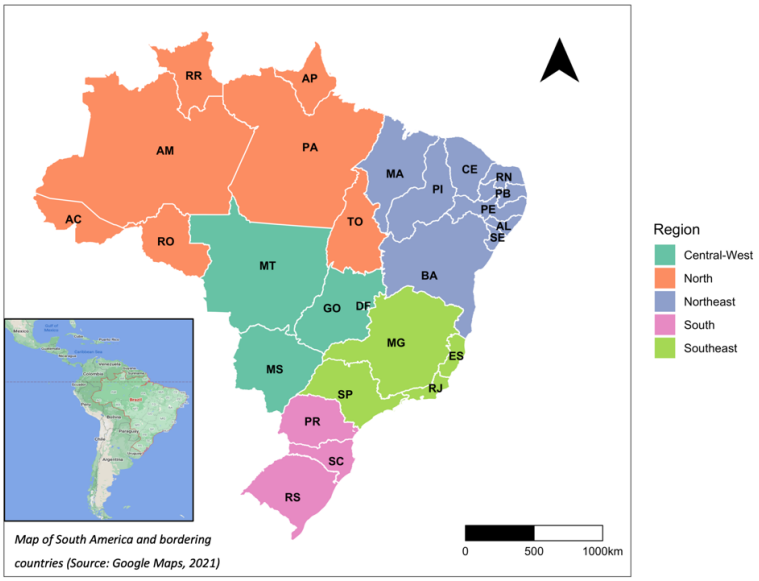


Figure 2.1.1. Map of Brazil illustrating regions and the 26 states. Central-West: Mato Grosso do Sul (MS), Mato Grosso (MT), Goías (GO), Distrito Federal/Federal District (DF). North: Rondônia (RO), Amazonas (AM), Acre (AC), Roraima (RR), Pará (PA), Amapá (AP), Tocantins (TO). Northeast: Maranhão (MA), Piauí (PI), Ceará (CE), Rio Grande do Norte (RN), Paraíba (PB), Pernambuco (PE), Alagoas (AL), Sergipe (SE), Bahia (BA). South: Rio Grande do Sul (RS), Santa Catarina (SC), Paraná (PR). Southeast: São Paulo (SP), Minas Gerais (MG), Rio de Janeiro (RJ), Espírito Santo (ES).

Brazil has humid tropical to subtropical climate, except for semi-arid regions which experiences droughts in the Northeast. Land surface temperatures during the summer months (November to April) are uniform and ranges between 23 ℃ – 29 ℃, with temperatures exceeding 38℃ in hotter regions. During the winter months (May to October), temperature can range between 13℃ – 23℃ though it varies greatly by region. Most of Brazil receives 1,000 mm – 1,800 mm of precipitation annually, where rainfall is heavier in the Amazon basin and the Southeastern region along the Atlantic coast. The Brazilian highlands, which include the eastern, southern, and central regions, covers half of the country's land area and receives most of the precipitation during the summer months in the form of heavy downpours. Given diverse climate and drainage conditions, there exist a range of vegetation types in Brazil. Tropical soil conditions make the land extremely arable and enable the production of a variety of crops.

2.1.2 Demographics

Brazil is home to more than 212.9 million people (2020 estimate) and is the sixth most populated country in the world. According to the 2019 World Population Prospects published by the United Nations (2), more than 69% of the country's population are between the ages of 15 to 64 years, while 9.5% of the population was 65 years older (UN). The geographical distribution of population is uneven, with high population density concentrated along the coast mainly in the Northeastern and Southeastern regions, and lowly populated areas within the interior of the Amazon Basin. São Paulo is the largest urban agglomeration in the country and the state is home to 45.9 million people (2019 estimate, 22% of Brazil's population), where more than 75% of population resides in the Metropolitan Region of São Paulo.

The population of Brazil has a very diverse racial and ethnic composition. The majority of Brazilians trace their ancestry to early European settlers from Portugal, Sub-Saharan Africans, and indigenous groups in Brazil. During the post-colonial era in the 19th century, Brazil also received significant immigration from other countries within Europe, Lebanon, and East Asian countries such as Japan. The Brazilian Institute of Geography and Statistics (IBGE) classifies race into five categories, mainly White, Pardo (mixed ancestry), Black, Asian, and Indigenous. According to the 2010 IBGE census, over 47% of the population is White and are identified as European Brazilians, while 43% of the population are identified as Pardo, which refer to multiracial Brazilians. Next, 7.6% of the population are identified as Black, while 1.1% and 0.4% are identified as Asian and Indigenous, respectively. Due to racism, classism, and stark differences in socioeconomic status, some Pardos are likely to self-declare as White or Black if they belong to a higher social class.

Extreme inequalities exist between regions, urban and rural populations, socioeconomic groups, and racial groups in Brazil's society. Regional polarisation present at the start of the 20th century has created clear-cut divides between regions, leading to disparities present between the wealthy and urbanised Southeast and the poor Northeast (3). About 20% of its population lives in poverty and an estimated 12 million people live in overcrowded housing and neighbourhoods called "favelas" (4). Moreover, recent statistics published by the IBGE in 2018 found that 40.8% of the employed population are considered informal workers, defined as those lacking a formal contract and labour protections that can help cushion temporary unemployment should it occur. Almost half of informal workers are identified as Black or Pardo and income disparities exist between men and women (5).

2.1.3 Government and healthcare

Since 1889, Brazil has been a federal republic that underwent major political changes, alternating between authoritarian and short-lived democratic governments in the 20th century. A democratic government was restored in the mid-1980's. Its autonomous government consists of three levels in descending rank: a federal government, 26 states and a federal district, and 5,563 municipalities. The Brazilian Health Reform movement occurred during democracy restoration and was driven by civil society. Its political ideology was to promote the acknowledgement that health is a social and political issue that needs to be publicly addressed (6).

In 1988, this led to the recognition of health as a constitutional right and a state duty in Brazil, where care has been provided via both public and private services. Brazil has Universal Health Care and thus healthcare is provided free of charge to everyone through the Unified Health System (Sistema Único de Saúde, SUS). This system is decentralised and regionalised, meaning that states and municipalities manage their own local hospitals,

services, and disease surveillance (7). The system covers free access to all health services from primary care to specialists and have helped to overcome past public health crises such as the Zika epidemic and yellow fever outbreaks (8). However, the implementation of SUS has been hindered by competing support for private health services and health services in developed areas (9). Moreover, SUS suffers from chronic underfunding, which has been exacerbated by the 2014-2016 economic recession, leading to disparities in access and coverage to health services (10). While 25% of the population have access to private care through employment insurance, there has been a decline in private health coverage (11). Moreover, in recent years, there has been an increase in out-of-pocket expenditures for medicine (12).

The COVID-19 pandemic has further aggravated public healthcare. Since the start of the epidemic in Brazil, conflicting messages sent by the President and local leaders have led to widespread confusion about staying at home and going to work, with President Bolsonaro driven by a political agenda that has consistently downplayed the severity of the pandemic (13). As a result, the rapid spread of COVID-19 in March 2021, further driven by the emergence of the P.1 variant (14), has led to the occupation of more than 80% of intensive care beds by COVID-19 patients in 24 out of 26 states, causing the “biggest health system collapse in Brazil’s history” (15). As of November 5, 2021, 55.2% of the Brazilian population has been vaccinated with at least one dose of the COVID-19 vaccine (16). Section 1.1.7 and 1.1.8 will elaborate further on the COVID-19 vaccination efforts in Brazil.

2.1.4 Arboviral diseases

Arboviruses are zoonotic viruses that are naturally maintained in mosquitoes, ticks, midges, and other insects, which feed on humans and causes illnesses, posing a serious threat to public health in Brazil. The main arboviruses causing epidemics belong to the families

Flaviviridae and *Togaviridae* (17). In Brazil, arboviruses of major public health importance and are endemic include dengue (DENV), Zika (ZIKV), chikungunya (CHIK), along with Mayaro (MAYV), Oropouche (OROV), and yellow fever (YFV) (18). The number of arboviruses has increased in parallel with deforestation, climate change, migration, and poor sanitation (19). Concurrent prevalence of several arboviral diseases in Brazil have created concerns regarding the co-circulation of arboviruses and the likelihood of consecutive or simultaneous infection (20). By April 2020, the number of notified arbovirus cases from dengue, Zika, and chikungunya amassed over 660,000 (21).

Arboviral diseases are nationally notifiable diseases in Brazil, however only a small proportion of cases are notified by the public healthcare system due to limited laboratory resources, which are required to perform serological assays for case confirmation (22). Accurate diagnosis is difficult due to cross-reactivity between arboviruses during serology testing. Moreover, uneven distribution in public health resources across states also results in reporting inconsistencies. In this thesis, the spread of Zika virus will be discussed in Chapter 2.3.2, while Chapter 3 will focus on the emergence of yellow fever virus in Brazil.

2.1.5 Coronavirus disease 2019 (COVID-19)

Coronavirus disease (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which has zoonotic origin traced to bats (23). It was first detected at Huanan Seafood Market in Wuhan, China in December 2019 (24), and is spread between humans in close contact with respiratory droplets from coughs and sneezes. WHO declared COVID-19 a Public Health Emergency of International Concern on 30 January, 2020 and a pandemic on 11 March, 2020.

As of November 5, 2021, there have been more than 21.8 million confirmed cases and 609,000 deaths in Brazil, making Brazil the third country in the world behind the United States and India with the most number of cases. The first confirmed COVID-19 cases in South America were detected in travellers from Italy in São Paulo on 25 February, 2020 (25) and was then spread to every state in Brazil by 21 March, 2020. The severity of the epidemic in Brazil was significantly driven by politics and the government response (13). While Brazil's mayors and governors implemented non-pharmaceutical interventions (NPIs) to curb the spread of COVID-19, these actions have been criticised by President Bolsonaro, who favoured the opening of businesses and the easing of social isolation guidelines, calling the pandemic "just a little flu" (13). The speed and intensity of COVID-19 spread across Brazil has illustrated the lack of a prompt and coordinated response (26). In January 2021, a surge of hospitalisations associated with COVID-19 was witnessed in Manaus city, Amazonas, (27), instigating an investigation that led to the detection of the P.1 variant (28).

2.1.6 History of vaccination in Brazil

Founded in 1973, the *Programa Nacional de Imunizações do Brasil* (PNI - Brazilian National Immunization Program) administers free vaccination to the Brazilian population under the coordination of the Ministry of Health. Vaccines provided through the PNI are exclusively funded by public resources. Similar to the Brazilian Unified Health System (SUS), PNI adopts decentralized governance at each government level and the prioritisation of universal and equitable care for all target groups (29). Across all 5500 municipalities in Brazil, PNI can reach 50,000 immunization rooms during vaccination campaigns where 114, 101 health professionals can be deployed. Citizens can access vaccines at any time of the year through one of the 10,000 Basic Health Units (Fonesca et al., 2021). Vaccines offering protection over 20 disease have been offered at different stages of the life course including 15 vaccines for children, 9 for adolescents, and 5 for

adults and the elderly. Approximately 75% of vaccines used in the program come from public laboratories (30).

2.1.7 COVID-19 vaccination strategy

The first COVID-19 mass vaccination campaign begun on 17 January 2021. Compared to other Latin American and middle-income countries, Brazil started its rollout a month later, but managed to catch up quickly. The initial strategy for immunization was aligned with the WHO SAGE values framework for the allocation and prioritization of COVID-19 vaccination (31) the Ministry of Health established 29 priority groups, which included health professionals, the elderly population, people with comorbidities, and in-person workers at risk due to increased exposure. However, these categories did not define specifically who qualifies for each group (e.g. who qualifies as a healthcare worker) and thus local governments had to establish and modify their own priority lists. This resulted in alterations where some states included particular working professionals at the expense of excluding patients with comorbidities (32).

2.1.8 COVID-19 vaccine hesitancy and confidence

COVID-19 vaccine hesitancy in Brazil is associated with perception of the efficacy, safety, and adverse effects of the vaccine (33,34). A cross-national study that conducted surveys in late 2020 found that Brazil was ranked seventh among 32 countries with the highest willingness to take a COVID-19 vaccine (35). This was corroborated by a later study, which found via a survey of more than 170,000 participants in all regions of Brazil, that about 89.5% were willing to vaccinate. Those that did not intend to vaccinate accounted for 2.5% while 8% were unsure about vaccinating with a particular vaccine (36).

Another national survey (n=2,771) conducted in late 2020 focused on politics and COVID-19 found that the vaccine's country of origin strongly affected the likelihood of vaccination (37). It found that Brazilians were less inclined to vaccinate if they were told that the vaccine was developed in another country, such as China and Russia, compared to the US and UK. Furthermore, the survey also found that those in support of President Bolsonaro, who is a strong critic of vaccination, were also less likely to vaccinate and showed stronger resistance towards the vaccine from China than those who view him negatively.

2.2 Theoretical frameworks

This thesis will draw upon two key theoretical frameworks, namely the theory of human ecology of disease (38,39) and landscape epidemiology (40). These frameworks will be combined to address the interactions between the pathogen and host, while accounting for its interdependency with environment, behaviour, and population characteristics.

2.2.1 Disease ecology

The World Health Organization (WHO) defines health as a "state of complete physical, mental and social well-being and not merely the absence of disease or infirmity". This resonates with the work of medical doctor Ralph Audy, who conceptualised human health as measurable adaptability from the "physical, chemical, infectious, psychological, and social" hazards rather than as the "absence of disease". Applying medical ecology to medical geography, Audy argues that ill health from exposure to infectious diseases is driven by both human behaviour and the environment, which includes the social and natural environments and their interactions (41). Building on this notion, medical geographer Melinda Meade defined the human ecology of disease framework based on an ecosystem where the susceptibility to a disease is driven by a triad formed by behaviour, population,

and environment (Figure 2.2.1). These factors interact in different ways across space to produce or prevent the transmission of a disease (39,42).

Environment includes the physical, social, and built, and are concerned specifically with elements that shape the habitat of populations. This includes but are not limited to factors such as climate, social connections, and infrastructure-based services such as health care. Population factors such as age, sex, and demographic related composition and structure defines who will be exposed and the duration and severity of the health impact. Human behaviour is defined by sense of perceived normality at the individual level, which influences human actions to modify exposure and risk of disease. Socioeconomic status, social institutions, and beliefs are examples of behavioural actions and structure that can influence individual exposure to disease (39). These dimensions interact in complex and dynamic ways to shape disease exposure and ill health. Recently, health geographers have argued for engagement with biosocial conceptualisations - the integration of a biological perspective with respect to social relations (43). For instance, sex is a biological characteristic that is always entangled with gendered social relations (44). Since the onset of the COVID-19 pandemic, the increasing adoption of longitudinal data on exposure to infectious diseases could enable the analysis of long-term imprints of exposure and connect individual life-course to their behaviours and environment at the population level.

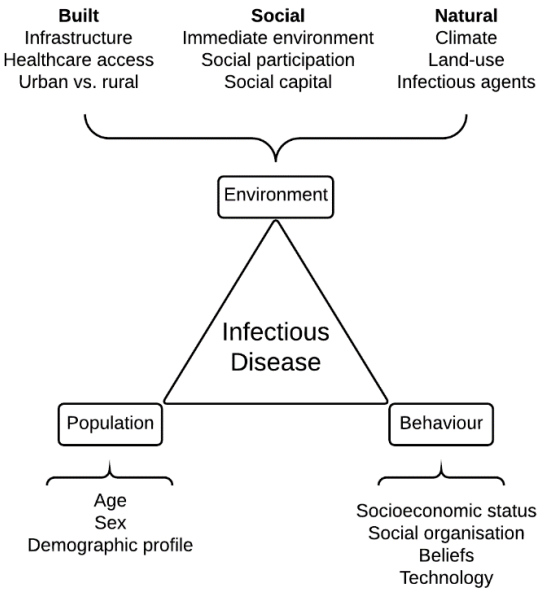
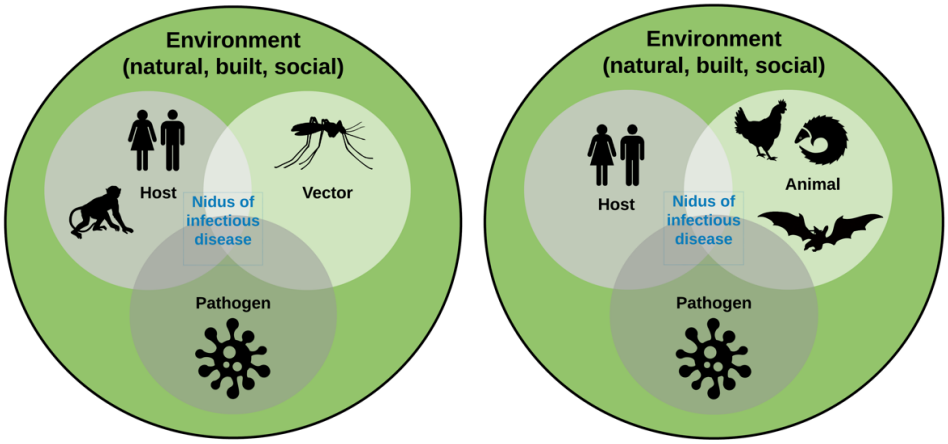


Figure 2.2.1. The human ecology of disease framework defined by the three dimensions – environment, population, and behaviour (39).

2.2.2 Landscape epidemiology

Similar to disease ecology, the concept of landscape epidemiology defines a pathogen-driven infection shaped by the intersection of competent pathogens, vectors, and hosts within a permissive and dynamic environment (45,46). The concept of a “foci” was first conceptualised by medical geographer Evgenly Pavlovsky when he analysed the emergence of zoonotic infectious diseases with respect to the associations between animal and insect life and elements of the natural landscape such as vegetation, soil, and precipitation (40). Figure 2.2.2 illustrates the foci of zoonotic and vector-borne disease transmission in the context of the landscape epidemiology framework. The nidus of disease transmission is subject to alteration by the virus, the vector or animal (e.g. mosquitoes, bats), humans, and in the case of vector-borne diseases, sometimes animal hosts. The geographic variation of this interaction between host, vector or animal, and pathogen can be quantified using various computing tools across various spatiotemporal scales (47). Given recent technological advances in geographic information systems (GIS) and

sophisticated spatial statistical models, landscape epidemiology can now be applied analytically to study the geographical patterns of infectious diseases, while linking it with factors from the natural, built, and social environment (48). This can also be combined with phylogeographic data, which can provide “genetic footprints” of past disease transmission events, giving valuable insight on the evolution of a virus across time and space (49–51).



(a) (b)

Figure 2.2.2. The presence of infectious agents is defined by the vector or animal, host, and pathogen in (a) vector-borne and (b) zoonotic disease transmission (40).

2.3 Methodology

2.3.1 Recent advances in infectious disease mapping

The usage of cartography to investigate the origin and spread of infectious diseases has advanced significantly since the days of John Snow’s 1854 dot map for tracing cholera cases in London. Nowadays, the collection of spatial big data at individual-level resolution can be processed using sophisticated geographic information systems (GIS) to map, predict, and evaluate spatial and temporal patterns. They can be made into easily-shareable interactive web-based tools that present information in near-real time. This has

revolutionised quantitative health geography and our knowledge of how space and place shape people's health outcomes (52).

The COVID-19 pandemic has further redefined the application of GIS for rapid dissemination of crucial information on infectious disease outbreaks. In late January 2020 when the COVID-19 epidemic was still in its infancy, researchers from John Hopkins University (JHU) created a COVID-19 dashboard that mapped open-access data on confirmed infections, fatalities, and recoveries at the state level (https://coronavirus.jhu.edu/map.html) for the general public. Users were also presented with visualisations detailing the progression of case counts over time, the sources of the data, and the most recent update made to the dashboard. While the JHU dashboard presented the most up-to-date data on the pandemic, researchers at the University of Oxford, including myself, proposed an alternative crowd-sourcing approach by utilising line lists tabulating individual case-information including age, sex, symptoms, travel information, and the GPS coordinates of the location where the case was reported (53). This information was then mapped at the finest administrative division unit on the Healthmap (https://www.healthmap.org/covid-19/) dashboard in real time. Our open-access centralised repository has enabled researchers to model the various phases of the epidemic in many countries (54) and as of April 2021, are still used as a data source for COVID-19 modelling studies.

2.3.2 Modelling the geographical spread of infectious disease: A case study on Zika virus

The following paper published in the Transactions of the Royal Society of Tropical Medicine and Hygiene is a literature review summarising quantitative modelling methods adopted to predict and infer the geographic spread of Zika virus. Prior to the COVID-19 pandemic, the

2016 Zika epidemic was the last Public Health Emergency of International Concern (PHEIC) that significantly impacted Brazil. The WHO defines PHEIC as “an extraordinary event which is determined to constitute a public health risk to other States through the international spread of disease and to potentially require a coordinated international response". Thus, there are still lessons to be learnt from the research conducted surrounding that epidemic, especially the modelling methodologies adopted to produce findings that were used to aid public health response.

Despite the focus on Zika, the methods discussed here have been applied to studies in Chapters 2 and 3. This review highlights the various modelling methods adopted to characterise spatial spread, and classifies them by type, aim, application, scale, and objectives. The effectiveness of integrating socio-demographic and environmental data, epidemiological data, and phylogeographic findings in these models are also discussed. Furthermore, the application of data and methods used to characterise human movement, which can be adopted to infer the geographic spread of disease, is also compared and summarised. Lastly, various techniques used for model validation are assessed.

Trans R Soc Trop Med Hyg 2021; **115**: 956–964
doi:10.1093/trstmh/trab009 Advance Access publication 11 February 2021

REVIEW ARTICLE

A review of models applied to the geographic spread of Zika virus

Sabrina L. Li [a,*], Jane P. Messina[a,b], Oliver G. Pybus[c], Moritz U. G. Kraemer[c], and Lauren Gardner[d,e]

[a]School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK; [b]School of Global and Area Studies, University of Oxford, 12 Bevington Road, Oxford, OX2 6LH, UK; [c]Department of Zoology, University of Oxford, 11a Mansfield Rd, Oxford, OX1 3SZ, UK; [d]Department of Civil and Systems Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, 21218-2682, USA; [e]Center for Systems Science and Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, 21218-2682, USA

*Corresponding author: Tel.: +44 01865 285070; E-mail: lisabrinaly@gmail.com

Received 29 June 2020; revised 13 December 2020; editorial decision 5 January 2021; accepted 8 January 2021

In recent years, Zika virus (ZIKV) has expanded its geographic range and in 2015–2016 caused a substantial epidemic linked to a surge in developmental and neurological complications in newborns. Mathematical models are powerful tools for assessing ZIKV spread and can reveal important information for preventing future outbreaks. We reviewed the literature and retrieved modelling studies that were developed to understand the spatial epidemiology of ZIKV spread and risk. We classified studies by type, scale, aim and applications and discussed their characteristics, strengths and limitations. We examined the main objectives of these models and evaluated the effectiveness of integrating epidemiological and phylogeographic data, along with socioenvironmental risk factors that are known to contribute to vector–human transmission. We also assessed the promising application of human mobility data as a real-time indicator of ZIKV spread. Lastly, we summarised model validation methods used in studies to ensure accuracy in models and modelled outcomes. Models are helpful for understanding ZIKV spread and their characteristics should be carefully considered when developing future modelling studies to improve arbovirus surveillance.

Keywords: arbovirus, mobility, surveillance, vector-borne, ZIKV

Introduction

Zika virus (ZIKV) is transmitted to humans via the bite of infected *Aedes* sp. mosquitoes. Since the virus was first isolated in 1947, ZIKV has circulated for decades without causing large reported outbreaks or severe disease.[1] The first known outbreak occurred on Yap Island in the Western Pacific in 2007, but no hospitalisations or deaths were reported.[2] ZIKV then reached French Polynesia and the South Pacific in 2013 and was introduced to Brazil in late 2015, likely through international air travel.[3] The Zika epidemic in Brazil received global attention when a strong epidemiological link was established with an increase in cases of microcephaly.[4] Subsequently ZIKV was declared a Public Health Emergency of International Concern by the WHO[5] in 2016. Since its introduction in the Americas, more than 5.8 million Zika cases have been reported, as of December 2020,[6] and more than 7452 cases were reported in 2020 alone.[7] As of July 2019, 87 countries and territories reported cases of the ZIKV lineage that spread from French Polynesia.[8] No vaccines exist for preventing ZIKV infections, thus prevention depends solely on the deployment of effective vector control measures and public health interventions.

Mathematical models can help to determine when and where future ZIKV outbreaks may occur and can retrospectively elucidate the effects of risk factors that facilitate spread. Here, we reviewed and compared models that have been adopted for investigating the geographic spread of ZIKV. We described their shared characteristics and evaluated the integration of various data sources for improving ZIKV surveillance. Finally, we discussed how studies have performed model validation to ensure robustness.

Box 1. Review search strategy

We conducted searches in PubMed and Web of Science for all published studies, as of 18 November 2020. We filtered studies using a combination of search strings (Figure 1) as follows:

- Zika OR ZIKV
- Spatial OR spatio-temporal OR geographic OR map OR spread OR dispersal OR transmission

© The Author(s) 2021. Published by Oxford University Press on behalf of Royal Society of Tropical Medicine and Hygiene. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

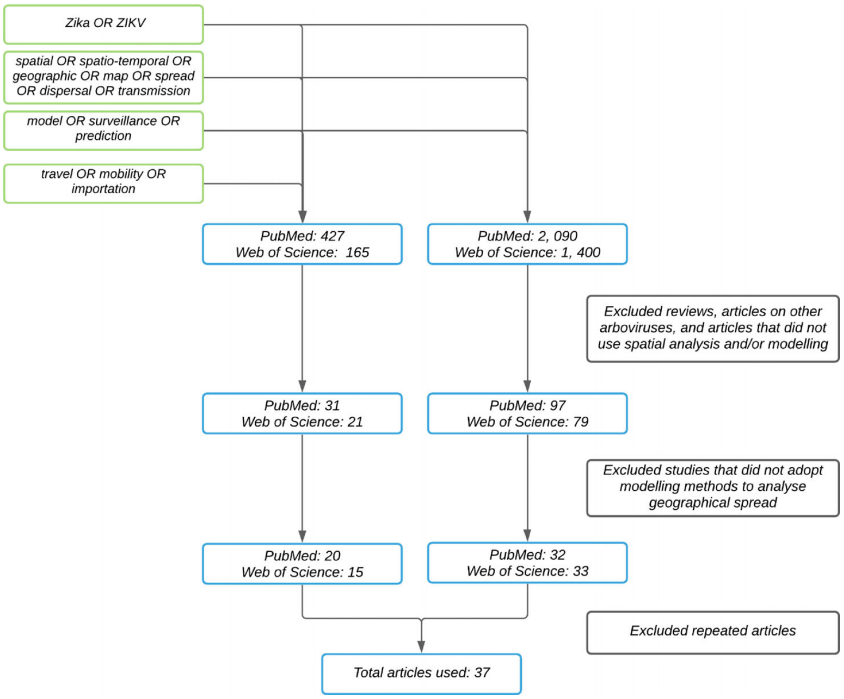


Figure 1. Flowchart illustrating the process of searching and retrieving the articles used in the review.

- travel OR mobility OR importation
- model OR surveillance OR prediction

We then filtered studies using two exclusion criteria. First, we examined article abstracts and excluded articles that were (1) identified as reviews, (2) focused on other arboviruses and (3) did not use spatial analysis. Second, for the remaining articles, we evaluated the study content and excluded studies that did not adopt a modelling approach to look at the geographical spread of Zika. In total, we retained 37 studies that met our inclusion criteria.

Characteristics of Zika modelling studies

We categorised studies by model type (stochastic or deterministic), aim, application, temporal and spatial scale. This is illustrated in Figure 2 as an alluvial diagram which shows how studies are connected by these categories. We compared model features and identified whether each model incorporated surveillance data and factors on human mobility, socioeconomic conditions, population, environmental conditions, mosquito suitability and mosquito-to-human transmission. We summarised this information using a network diagram (Figure 3), where each study is represented by a pie chart that illustrates the model features incorporated by that study. The thickness of each link represents the number of features shared between two studies and the colours of the links show how studies are clustered based on shared features.

Model type

Stochastic, deterministic and phylogeographic methods have all been adopted to investigate ZIKV spread (Figure 2). Deterministic models use a set of input parameters, often from experimental findings or from the literature, to generate estimates on epidemiological characteristics that quantify spread,[9–13] whereas stochastic models predict or infer retrospectively the process (or parameters) of spread using methods based on statistical theory.[14–24] Deterministic models are effective for estimating mosquito–human interactions that facilitate infection[9–11,15,23,25,26] but they are unable to account for the inherent stochastic nature of disease transmission.[10,12,13] The majority of existing studies use stochastic models, which can flexibly integrate high-resolution information about environmental conditions and vector competence and thereby model fine-scale spatial heterogeneities in disease spread.[14,17,24,27] However, stochastic models may rely heavily on the availability of sufficiently large and detailed surveillance data, which may not be available.[17,24,28]

Several studies have adopted phylogeographic methods, which have been increasingly adopted to rapidly assess the introduction and circulation of the viruses, including ZIKV. Phylogeographic modelling analyses can estimate the frequency and

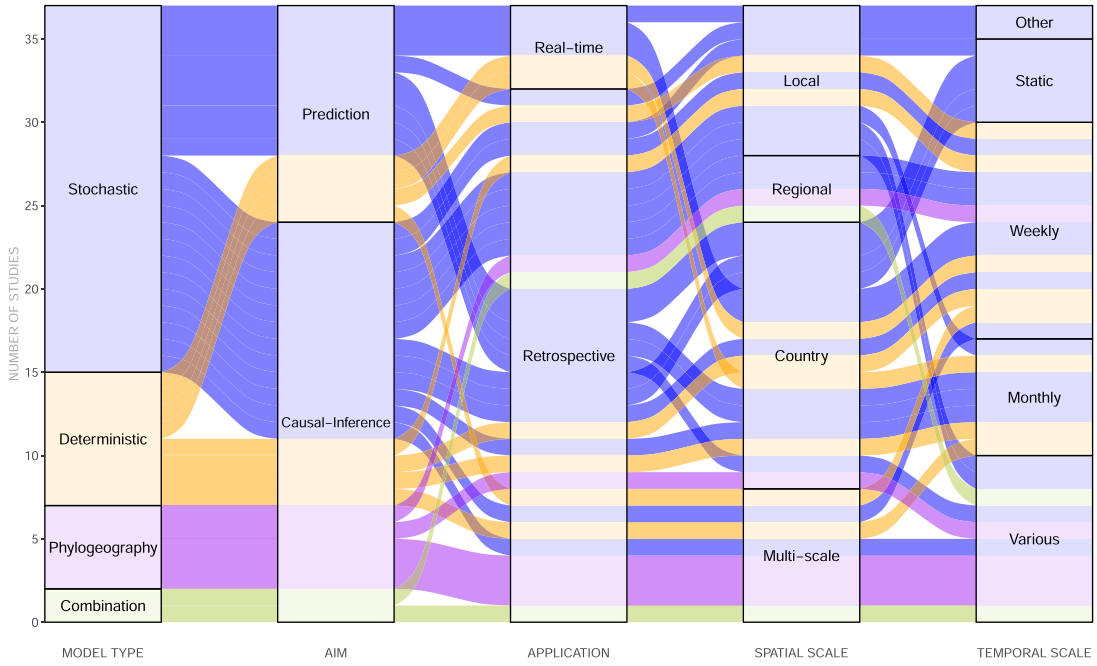


Figure 2. Characteristics of reviewed studies, summarised by model type, aim, application and scale.

sources of ZIKV introduction into specific locations and new regions. Examples of this include studies of ZIKV in Latin America and the Caribbean[3,29,30] and the USA.[31,32] Some of these studies have adopted stochastic models in combination with phylogeography to predict ZIKV spread.[3,29,33] However, it is worth noting that phylogeographic inferences may be sensitive to variation in sampling among locations.

Model aim and application

We distinguished the modelling studies we reviewed according to two primary study objectives: causal-inference and prediction. The majority of studies are causal-inference (Figure 2), which aim to assess the contribution and interaction of factors in predicting ZIKV transmission.[22,26,34,35] Additionally, they aim to identify the geographical origins of virus spread and epidemic history using phylogeographic methods.[3,30–32] Predictive studies can either be stochastic or deterministic and may aim to estimate Zika incidence,[14,24,36] importation of cases,[11,12,14] distribution of risk[18,21,37–39] and transmission potential.[15,23] Both predictive and causal-inference models can be applied retrospectively to assess disease spread, which is the main application identified among the studies we reviewed. Predictive models can also be used to track disease spread with potential real-time applications.[13–15,20]

Spatial and temporal scale

The geographic spread of ZIKV has been assessed extensively at the country and local level, with time intervals ranging from weekly to monthly (Figure 2). In addition to conducting analysis at the global level,[9,13,18,21,38,40] studies have focused on the Latin American and Caribbean region,[20,28–30,33,36,41] Africa and Asia-Pacific, Oceania, Europe[12,23] and on specific countries such as the USA,[11,15,27,31] Colombia,[16,19,34,42] Brazil[3,10,37,43] and Singapore.[22,44] Multi-scale analysis[29,44] and models that integrate data at different temporal and spatial scales[27,40,43] can help to infer risk factors or make accurate predictions on geographic spread. This is especially useful for countries with passive surveillance systems that experience long delays in releasing official case data. However, aggregating data at different scales across a region with high spatial heterogeneity could result in inaccurate estimates of potential ZIKV risk.[16]

ZIKV modelling objectives

Among the studies that we reviewed, we identified a set of common objectives that the authors aimed to achieve, which are elaborated below. Studies with similar objectives are also shown to be in the same clusters depicted in Figure 3.

Mapping geographical distribution and risk

Prior to the availability of epidemiological data, mapping the environmental suitability of ZIKV can elucidate the potential distribution of exposure and risk of infection. This requires an understanding of the geographical distribution of *Aedes* sp. mosquito presence and the competence of vectors in transmitting ZIKV based on environmental factors, which can be inferred from mapping

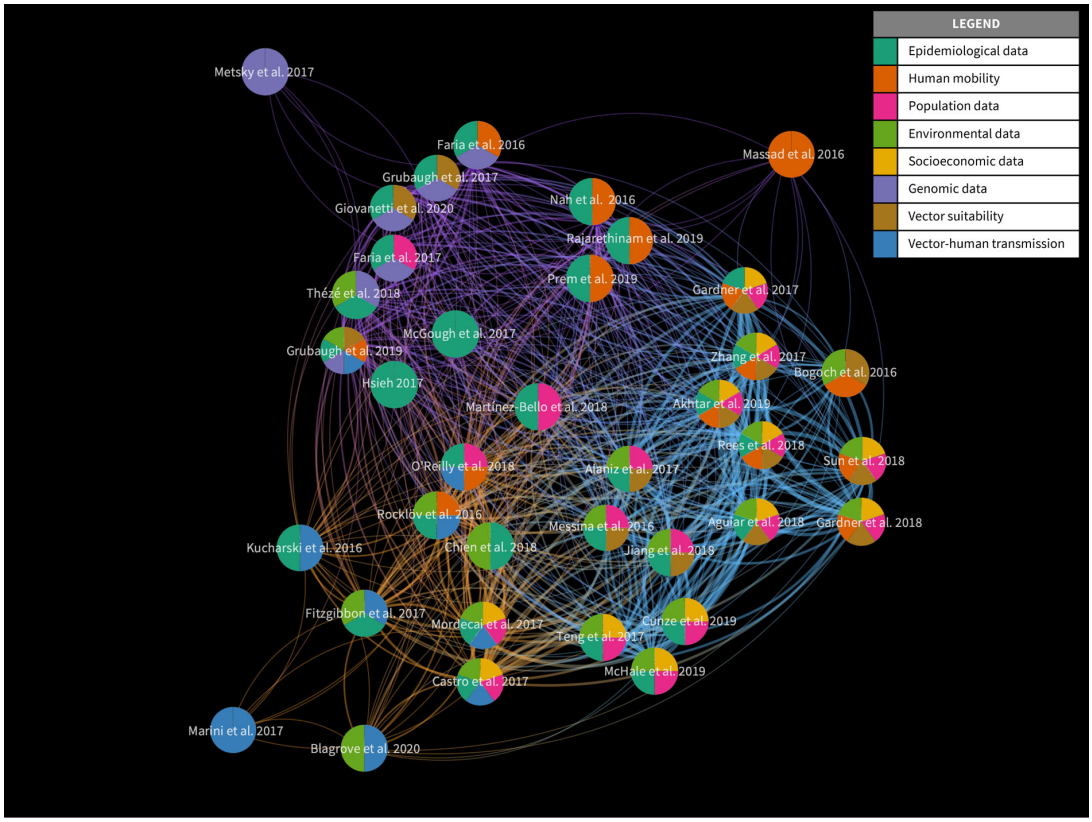


Figure 3. Network of reviewed studies, illustrating how studies are related to each other based on shared features. The pie chart at each node summarises the features incorporated in that particular study. Visualisation created using VOS viewer (v. 1.6.13).

the mosquito's habitat suitability using ecological niche models.[21] One of the earlier works by Messina et al.[21] used ensemble boosted regression trees to map global environmental suitability for ZIKV. Combining this model output with disease occurrence, once it becomes available, can help to predict potential population exposure,[38] transmission risk[41] and risk factors[35] at both regional and global scales.

In particular, integrating airline data and socioeconomic factors can help to characterise travel-related spread[39] and the effects of socioeconomic factors on transmission.[37] More recently, machine learning models, such as propagation neural network, gradient boosting machine and random forests, have also been applied to identify regions at risk globally.[18] Similarly, spatio-temporal modelling using Bayesian inference can also be used to identify areas of high risk at the local level[19,42] and the influence of travel history on risk.[44]

Predicting local and imported infections

The frequency and risk of ZIKV infection, including local and imported infections, has been predicted using several methods. The rate of infection per population and region have been predicted using a stochastic spatial model,[24] a dynamic neural network model[14] and a deterministic model.[36] A SEIR (susceptible, exposed, infectious, recovered) model framework has also been adopted to assess infection dynamics by quantifying vector–human interactions[25] and the likelihood of sustaining mosquito-borne transmission.[15] Furthermore, risk models developed by Gardner et al.[40] that considered the vector competence of both *Aedes aegypti* and *Aedes Albopictus* have further revealed the risk of importation into new regions. Once the number of imported infections is known, the time and risk of importation from travellers can be used to understand the dynamics of local transmission[13,27] and to dynamically model epidemic risk in real time.[15] Exportation and importation risk profiles can also be generated to understand the risk of spread between geographical regions.[17]

The availability of near real-time data sources, such as Google searches, Twitter microblogs and ProMed (https://promedmail.org/), has also been leveraged to make timely predictions of ZIKV infections before official case data become available. Predictions of weekly ZIKV cases have been made 1–3 wk ahead of publication date in Latin American countries.[20] The accuracy and geographic coverage of 'big' mobile and internet data are highly dependent on human–computer interactions and participation dynamics, thus reliability and biases from these sources may be limited.

Quantifying transmission using R_0

The basic reproduction number (R_0) is a key metric used to define the capability of an infectious agent to proliferate[45] and is commonly adopted in studies to understand transmission in populations. This number can vary across time (R_t), as an estimate determined during an epidemic driven by human contact will differ as it transitions to an endemic state when the population has achieved advanced control efforts or herd immunity.

At the start of an epidemic, when public health interventions and population immunity are absent, R_0 can be adopted as an effective metric to understand rapid spread. This is heavily dependent on the vector capacity of the mosquito, which is driven by temperature, the extrinsic incubation period of ZIKV, mosquito biting rate and vector abundance in a location.[15,23] Given the optimal conditions for human–vector transmission, integrating data on local and regional travel can help assess transmission dynamics within and between populations[15,23,28] and between human and vector to dissipate or sustain epidemics.[10] Furthermore, by identifying and modelling *Aedes* sp. competence and transmission intensity under various temperature ranges, a range of R_0 estimates by region can be derived to understand the distribution of transmission duration and risk during recent and future climate scenarios.[9,26]

Reconstructing transmission pathways using phylogeography

Phylogeographic analyses can elucidate timely information on the rapid evolution and spatial spread of viral pathogens. Evolutionary trees combined with genetic and geolocation data have helped to determine the initial date of ZIKV circulation in the Americas, highlighting that local circulation began months before the first confirmed case.[32] While it was costly and time-consuming to sequence large numbers of virus genomes in the past, more accessible and timely methods of pathogen genome sequencing are now available using portable genomic technologies. For example, nanopore instruments such as the MinION[46] device can be rapidly and effectively adopted in remote and resource-low settings. For instance, ZIKV samples collected from the 2016 outbreaks in Brazil were generated within 48 h.[33]

Genomic data have been combined with epidemiological data to detect unreported outbreaks and to understand whether a virus is driven by importation or local transmission. By analysing viral genomes and epidemiological data, Faria et al.[33] found that ZIKV was introduced to northeast Brazil as early as February 2014 prior to circulation in Brazil and detection in the Americas. Grubaugh et al.[29] combined genomic data, passenger air travel information and local and travel-related infections to detect unreported outbreaks, such as one in Cuba that occurred a year after peak transmission in neighbouring islands. Applying genomic data to fill in gaps existing in epidemiological data has significant potential; however, given limited quantity and quality of travel data, combining travel surveillance in a joint framework remains a challenge that merits further exploration.

Identifying risk factors and drivers of spread

Human mobility, climate change, urbanisation and socioeconomic disparities can drive variations in the geographic spread of ZIKV as well as outbreak intensity. Global human mobility, which has been increasing at unprecedented rates due to tourism, trade and migration, is a key determinant of global arbovirus distribution.[47] The role of human mobility as a key predictor of ZIKV spread has been investigated by several studies, particularly when it concerns the risk of spread due to international travel[13,17,23,29,39] and migration.[10,14,22,34,36] Types of data and models used to quantify human mobility for Zika modelling are discussed further in the next section.

Climate factors, coupled with travel, can further drive changes in vector competence, particularly in locations with similar climate conditions that are conducive to *Aedes* habitat suitability. Temperature is a strong determinant of *A. aegypti* competence in transmitting Zika[48] and has been incorporated in both deterministic and stochastic models to account for its role in driving Zika infections.[16,23,24,26,34,37,43] Other factors such as precipitation and relative humidity also play an important role in fostering mosquito growth, which is conducive to Zika transmission.[16,18,21,34,39]

The role of socioeconomic factors in facilitating ZIKV spread remains inconclusive, as it varies by factor and spatial scale. While a strong negative association between gross domestic product and ZIKV transmission was found among countries[17,35] this remains contested at the subnational level.[37] Household conditions, namely, access to air conditioning, sanitation, piped water and garbage collection, were found to be linked with ZIKV spread and risk.[24,37] Rees et al.[34] found a negative relationship between poverty level and ZIKV detection for Colombia, but noted that this could be linked to under-reporting in poor areas due to limited healthcare access. Typically, urban settings have high landscape heterogeneity, providing optimal conditions for *A. aegypti* mosquitoes to thrive and foster arbovirus transmission to humans.[49,50] Urbanisation has high predictive power for estimating ZIKV risk,[18,37] along with population density, which is a significant risk factor.[17,35]

Variables relating to healthcare infrastructure have rarely been included as explanatory variables in existing ZIKV modelling studies. Gardner et al.[17] used data on the number of hospital beds and number of physicians per 10 000 people to represent the distribution of healthcare infrastructure in their model. They found that these variables did not contribute significantly to country-level geographic spread and local transmission of ZIKV. Future studies should focus on incorporating indicators of healthcare access into their models to understand their role in predicting spread.

Using human mobility to model ZIKV spread

Quantifying mobility using travel data

Human mobility is a key driver of ZIKV spread and integrating this variable in spatial models can provide a valuable insight into spread within and across populations in near real time. Information on human movement has traditionally been collected from household travel surveys and censuses[51]; recently, such data have been acquired at higher resolutions from new digital sources. Table 1 summarises the strengths and weaknesses of

Table 1. Indicators of human mobility for modelling arbovirus spread

Indicator of human mobility	Data source	Open access (Yes/No)	Spatial scale	Source
Passenger air travel	Global flights network (OpenFlights)	Yes	Global	13*
	International air transport association (IATA)	No	Global	39*;17*; 29*;24*; 40*; 14*;23*;12*;3*; 29*
	Official aviation guide (OAG)	No	Global	24*
Call data records	Operators and private companies	No	Local, regional	57;22*;59
Night-time lights brightness	Defense meteorological satellite program (DMSP)	Yes	Global coverage at 1 km x 1 km resolution	60
Global positioning system trackers	GPS data loggers used to track individual movement	No	Local (e.g. neighbourhood level)	61; 62; 63
	Google location history (GLH)	No		
	Cell phone towers	No		
Travel surveys	Government agencies	Yes	Local, regional	51
Infrastructure data	Road networks	Yes	Global	34*; 64
	Public transport contactless cards	No	Local	44*

*Studies that examined human mobility and its effects on ZIKV spread.

existing human mobility indicators that have been used to track the spread of ZIKV and similar arboviruses. A more comprehensive list of data used to quantify human mobility for analysing disease and health risks has been reviewed elsewhere.[52]

Data on passenger flights between locations can help to track ZIKV spread,[29,39,40] but limit our understanding of movement to the origin and destination airports and rarely considers connections made between flights and travel made at the individual level.[53] Call data records from mobile phones offer large coverage of populations and areas and can be leveraged to measure individual-level movement in near real time.[22,52] Access to call data records is currently restricted to research groups and access is granted via a negotiated agreement with the operator, which makes their widespread use challenging. The release of personal individual details via mobile phone records to third parties is a privacy concern, thus censored via data aggregation and often provided as a small sample. Given these restrictions, ongoing research has outlined ways to improve accessibility,[54] while initiatives such as the Open Algorithms project (https://www.opalproject.org) and FlowKit (https://flowkit.xyz) are working towards scaling the privacy-conscientious use of call data for research. Moreover, modelling efforts towards understanding the COVID-19 pandemic have further opened up these data sources to researchers.

Quantifying mobility using mathematical models

When trip-level data are not available, human mobility can be quantified using models such as the gravity model of migration, a deterministic model that assumes that population movement between locations is proportional to some power of the population sizes of the origin and destination locations.[55] O'Reilly et al.[36] used a gravity model to assess the international spread of ZIKV in the Americas and found this model to be effective in fitting the data. However, assumptions of the gravity model, such as the lack of theoretical guidance for fitting empirical data, the requirement of existing traffic data to fit parameters and issues concerning modelling the flux in travel between two locations[56] make it difficult to characterise complex travel behaviour across a large region.

To address these limitations, Simini et al.[56] developed the radiation model, which is parameter-free and is based on a stochastic process that only requires information on population distribution. This model can be applied to estimate movement patterns in areas that lack mobility data and its predictions have generally aligned with observed mobility patterns, including long-term migration and population diffusion between areas. This method was widely adopted to estimate population movement between affected cities in Latin American countries during the 2015–2017 ZIKV outbreaks[36] and similar arbovirus outbreaks.[57,58] When information on population travel is limited, radiation models can effectively infer human mobility behaviour during an outbreak to estimate the magnitude of spread.

Model validation

A good model for assessing disease spread is often validated for its accuracy, ensuring that the model, combined with assumptions, results in a sufficiently accurate representation of reality.

Studies of ZIKV that aimed to forecast the number of infections or R_0 at various spatial and temporal scales have validated their models by comparing the projected output with surveillance data not directly used to calibrate the model[15,22–24] or by adopting statistical analysis to understand whether the predicted relationship was consistent with the data.[15,26,34] While this is a simple way to validate model outputs, it may result in high bias if limited data were used. When the goal is to find the most suitable model among a selection of models, studies that modelled disease risk at the local and regional level have taken a criterion-based approach by comparing the Akaike Information Criterion (AIC)[16,19] and the Deviance Information Criterion (DIC) when using Bayesian inference.[42,44]

When large amounts of data are available, k-fold cross-validation, a method for testing performance for machine learning models, and area-under-the-curve (AUC), a measure of predictive accuracy, can be adopted as quantitative diagnostic tools. Several studies that have explored ZIKV spread at regional and global levels have adopted k-fold cross-validation to train their models,[18,21,34,36,38] particularly studies that have adopted an ecological niche modelling approach. Subsequently, AUC has been adopted to evaluate the predictive performance of these models.[18,21,34] A subset of randomly selected data can be used as a test sample and the AUC can be applied to assess predictive accuracy[37]; when multiple models are present, AUCs can be compared.[13] AUC can also be combined with the receiving operating characteristic curve analysis to further examine a model's predictive ability.[14,35]

Conclusions

Mathematical models can enhance surveillance and help identify potential risk factors that drive ZIKV spread. We reviewed studies that investigated the geographical spread of ZIKV, discussed common model features and examined the role of various transmission risk factors. We highlighted the value of adopting novel data sources that characterise human mobility to monitor transmission in real time, as well as the potential integration of such data with environmental and socioeconomic data phylogeographic methods. We also summarised model validation strategies and recommend their implementation. Our review provides an overview of model characteristics that future studies should consider when modelling the geographic spread of ZIKV and other arboviruses to prevent future outbreaks.

Author contributions: SLL, JPM, MUGK and LG conceived and designed the study. SLL collated and summarised the literature and wrote the manuscript. All the authors read, revised and approved the final version of the manuscript.

Funding: SLL is supported by the Oxford Martin School Programme on Pandemic Genomics and a Canadian Social Sciences and Humanities Research Council (SSHRC) Doctoral Fellowship. MUGK is supported by The Branco Weiss Fellowship, acknowledges funding from a Google.org and the European Union Horizon 2020 project MOOD (#874850). The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

Competing interests: None declared.

Ethical approval: Not required.

Data availability: All search filters used in this review are stated in the text.

References

1 Musso D, Gubler DJ. Zika virus. Clin Microbiol Rev. 2016;29:487–524.

2 Duffy MR, Chen T-H, Hancock WT, et al. Zika virus outbreak on Yap Island, Federated States of Micronesia. New Engl J Med. 2009;360:2536–43.

3 Faria NR, Azevedo R do S da S, Kraemer MUG, et al. Zika virus in the Americas: Early epidemiological and genetic findings. Science. 2016;352:345–9.

4 Mlakar J, Korva M, Tul N, et al. Zika virus associated with microcephaly. New Engl J Med. 2016;374:951–8.

5 WHO. WHO statement on the first meeting of the International Health Regulations (2005). 2016. Available at https://www.who.int/en/news-room/detail/01-02-2016-who-statement-on-the-first-meeting-of-the-international-health-regulations-(2005)-(ihr-2005)-emergency-committee-on-zika-virus-and-observed-increase-in-neurological-disorders-and-neonatal-malformations. Accessed December 13, 2020.

6 PAHO. Cases of Zika Virus Disease. 2020. Available at https://www.paho.org/data/index.php/en/mnu-topics/zika/524-zika-weekly-en.html. Accessed December 13, 2020.

7 Pan American Health Organization/World Health Organization. Epidemiological Update Dengue and other Arboviruses. Washington, D.C.; 2020. https://www.paho.org/en/documents/epidemiological-update-dengue-and-other-arboviruses-10-june-2020.

8 WHO. Zika Epidemiology Update. 2019. Available at https://www.who.int/emergencies/diseases/zika/zika-epidemiology-update-july-2019.pdf?ua=1. Accessed June 29, 2020.

9 Blagrove MSC, Caminade C, Diggle PJ, et al. Potential for Zika virus transmission by mosquitoes in temperate climates. Proc Royal Soc B Biol Sci. 2020;287:20200119.

10 Fitzgibbon WE, Morgan JJ, Webb GF. An outbreak vector-host epidemic model with spatial structure: the 2015–2016 Zika outbreak in Rio De Janeiro. Theoret Biol Med Mod. 2017;1–17.

11 Marini G, Guzzetta G, Rosà R, Merler S. First outbreak of Zika virus in the continental United States: a modelling analysis. Euro Surveill. 2017;1–8.

12 Massad E, Tan S-H, Khan K, Wilder-Smith A. Estimated Zika virus importations to Europe by travellers from Brazil. Global Health Action. 2016;9:31669.

13 Nah K, Mizumoto K, Miyamatsu Y, Yasuda Y, Kinoshita R, Nishiura H. Estimating risks of importation and local transmission of Zika virus infection. Peer J. 2016;4:e1904.

14 Akhtar M, Kraemer MUG, Gardner LM. A dynamic neural network model for predicting risk of Zika in real time. BMC Med. 2019;17:171.

15 Castro LA, Fox SJ, Chen X, et al. Assessing real-time Zika risk in the United States. BMC Infect Dis. 2017;17:284.

16 Chien L-C, Lin R-T, Liao Y, Sy FS, Pérez A. Surveillance on the endemic of Zika virus infection by meteorological factors in Colombia: a population-based spatial and temporal study. BMC Infect Dis. 2018;18:180.

17 Gardner LM, Bóta A, Gangavarapu K, Kraemer MUG, Grubaugh ND. Inferring the risk factors behind the geographical spread and transmission of Zika in the Americas. PLOS Negl Trop Dis. 2018;12:e0006194.

18 Jiang D, Hao M, Ding F, Fu J, Li M. Mapping the transmission risk of Zika virus using machine learning models. Acta Tropica. 2018;185:391–9.

19 Martínez-Bello D, López-Quílez A, Torres Prieto A. Spatio-temporal modeling of zika and dengue infections within Colombia. Int J Environ Res Pub Health. 2018;15:1376.

20 McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. PLOS Negl Trop Dis. 2017;11:e0005295.

21 Messina JP, Kraemer MU, Brady OJ, et al. Mapping global environmental suitability for Zika virus. Elife. 2016;5:1–19.

22 Rajarethinam J, Ong J, Lim S-H, et al. Using human movement data to identify potential areas of Zika transmission: Case study of the largest Zika cluster in Singapore. Int J Environ Res Public Health. 2019;16:1–13.

23 Rocklöv J, Quam MB, Sudre B, et al. Assessing seasonal risks for the introduction and mosquito-borne spread of zika virus in Europe. EBioMedicine 2016;9:250–6.

24 Zhang Q, Sun K, Chinazzi M, et al. Spread of Zika virus in the Americas. Proc Natl Acad Sci USA. 2017;114:E4334–43.

25 Kucharski AJ, Funk S, Eggo RM, Mallet H-P, Edmunds WJ, Nilles EJ. Transmission dynamics of Zika virus in island populations: A modelling analysis of the 2013–14 French Polynesia outbreak. PLOS Negl Trop Dis. 2016;10:e0004726.

26 Mordecai EA, Cohen JM, Evans MV, et al. Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. PLOS Negl Trop Dis. 2017;11:e0005568.

27 Sun K, Zhang Q, Pastore-Piontti A, et al. Quantifying the risk of local Zika virus transmission in the contiguous US during the 2015–2016 ZIKV epidemic. BMC Med. 2018;16:195.

28 Hsieh Y-H. Temporal patterns and geographic heterogeneity of Zika virus (ZIKV) outbreaks in French Polynesia and Central America. Peer J. 2017;5:e3015.

29 Grubaugh ND, Saraf S, Gangavarapu K, et al. Travel surveillance and genomics uncover a hidden Zika outbreak during the waning epidemic. Cell. 2019;178:1057–71.

30 Thézé J, Li T, du Plessis L, et al. Genomic epidemiology reconstructs the introduction and spread of Zika virus in Central America and Mexico. Cell Host Microbe. 2018;23:855–64.

31 Grubaugh ND, Ladner JT, Kraemer MUG, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;546:401–5.

32 Metsky HC, Matranga CB, Wohl S, et al. Zika virus evolution and spread in the Americas. Nature. 2017;546:411–5.

33 Faria NR, Quick J, Claro IM, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406–10.

34 Rees EE, Petukhova T, Mascarenhas M, Pelcat Y, Ogden NH. Environmental and social determinants of population vulnerability to Zika virus emergence at the local scale. Parasites Vectors. 2018;1–13.

35 Teng Y, Bi D, Xie G, et al. Model-informed risk assessment for Zika virus outbreaks in the Asia-Pacific regions. J Infect. 2017;74:484–91.

36 O'Reilly KM, Lowe R, Edmunds WJ, et al. Projecting the end of the Zika virus epidemic in Latin America: a modelling analysis. BMC Med. 2018;1–13.

37 Aguiar BS, Lorenz C, Virginio F, Suesdek L, Chiaravalloti-Neto F. Potential risks of Zika and chikungunya outbreaks in Brazil: A modeling study. Int J Infect Dis. 2018;70:20–9.

38 Alaniz AJ, Bacigalupo A, Cattan PE. Spatial quantification of the world population potentially exposed to Zika virus. Int J Epidemiol. 2017;46:966–75.

39 Bogoch II, Brady OJ, Kraemer MUG, et al. Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study. Lancet Infect Dis. 2016;16:1237–45.

40 Gardner L, Chen N, Sarkar S. Vector status of Aedes species determines geographical risk of autochthonous Zika virus establishment. PLoS Negl Trop Dis. 2017;11:1–18.

41 Cunze S, Kochmann J, Koch LK, Genthner E, Klimpel S. Vector distribution and transmission risk of the Zika virus in South and Central America. Peer J. 2019;7:1–19.

42 McHale TC, Romero-Vivas CM, Fronterre C, et al. Spatiotemporal heterogeneity in the distribution of Chikungunya and Zika virus case incidences during their 2014 to 2016 epidemics in Barranquilla, Colombia. Int J Environ Res Public Health. 2019;16:1490–1500.

43 Giovanetti M, Faria NR, Lourenço J, et al. Genomic and epidemiological surveillance of Zika virus in the Amazon region. Cell Reports. 2020;30:2275–83.

44 Prem K, Lau MSY, Tam CC, Ho MZJ, Ng L-C, Cook AR. Inferring who-infected-whom-where in the 2016 Zika outbreak in Singapore—a spatio-temporal model. J Royal Soc Interface. 2019;16:20180604.

45 Heffernan JM, Smith RJ, Wahl LM. Perspectives on the basic reproductive ratio. J R Soc Interface. 2005;2:281–93.

46 Tyler AD, Mataseje L, Urfano CJ, et al. Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. Sci Rep. 2018;8:10931.

47 Pybus OG, Tatem AJ, Lemey P. Virus evolution and transmission in an ever more connected world. Proc Royal Soc B Biol Sci. 2015;282:20142878.

48 Tesla B, Demakovsky LR, Mordecai EA, et al. Temperature drives Zika virus transmission: evidence from empirical and mathematical models. Proc Royal Soc B Biol Sci. 2018;285:20180795.

49 Wilke ABB, Caban-Martinez AJ, Ajelli M, Vasquez C, Petrie W, Beier JC. Mosquito adaptation to the extreme habitats of urban construction sites. Trends Parasitol. 2019;35:607–14.

50 Zahouli JBZ, Koudou BG, Müller P, Malone D, Tano Y, Utzinger J. Urbanization is a main driver for the larval ecology of Aedes mosquitoes in arbovirus-endemic settings in south-eastern Côte d'Ivoire. PLoS Negl Trop Dis. 2017;11:e0005751.

51 Wesolowski A, Stresman G, Eagle N, et al. Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. Sci Rep. 2015;4:1–7.

52 Lai S, Farnham A, Ruktanonchai NW, Tatem AJ. Measuring mobility, disease connectivity and individual risk: a review of using mobile phone data and mHealth for travel medicine. J Travel Med. 2019;26:1–7.

53 Feng S, Jin Z. Infectious diseases spreading on a metapopulation network coupled with its second-neighbor network. Appl Math Comp. 2019;361:87–97.

54 de Montjoye Y-A, Gambs S, Blondel V, et al. On the privacy-conscientious use of mobile phone data. Sci Data. 2018;5:180286.

55 Anderson JE. The Gravity Model. Annu Rev Econ. 2011;3:133–60.

56 Simini F, González MC, Maritan A, Barabási A-L. A universal model for mobility and migration patterns. Nature. 2012;484:96–100.

57 Kraemer MUG, Faria NR, Reiner RC, et al. Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. Lancet Infect Dis. 2017;17:330–8.

58 Wen T-H, Hsu C-S, Hu M-C. Evaluating neighborhood structures for modeling intercity diffusion of large-scale dengue epidemics. Int J Health Geographics. 2018;1–15.

59 Massaro E, Kondor D, Ratti C. Assessing the interplay between human mobility and mosquito borne diseases in urban environments. Sci Rep. 2019;9:1–15.

60 Bharti N, Tatem AJ. Fluctuations in anthropogenic nighttime lights from satellite imagery for five cities in Niger and Nigeria. Sci Data. 2018;5:180256.

61 Wen T-H, Lin M-H, Teng H-J, Chang N-T. Incorporating the human-Aedes mosquito interactions into measuring the spatial risk of urban dengue fever. Appl Geogr. 2015;62:256–66.

62 Vazquez-Prokopec GM, Bisanzio D, Stoddard ST, et al. Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. PLoS One. 2013;8:e58802.

63 Paz-Soldan VA, Reiner RC, Morrison AC, et al. Strengths and weaknesses of global positioning system (GPS) data-loggers and semi-structured interviews for capturing fine-scale human mobility: Findings from Iquitos, Peru. PLoS Negl Trop Dis. 2014;8:e2888.

64 Strano E, Viana MP, Sorichetta A, Tatem AJ. Mapping road network communities for guiding disease surveillance and control strategies. Sci Rep. 2018;8:1–19.

2.4 References

1. de Lima RAF, Oliveira AA, Pitta GR, de Gasper AL, Vibrans AC, Chave J, et al. The erosion of biodiversity and biomass in the Atlantic Forest biodiversity hotspot. Nature Communications. 2020 Dec 11;11(1):6347.
2. UN. World Population Prospects 2019. UN; 2019. 44 p.
3. Bucciferro JR, Ferreira de Souza PHG. The Evolution of Regional Income Inequality in Brazil, 1872–2015. In: Tirado-Fabregat DA, Badia-Miró M, Willebald H, editors. Time and Space: Latin American Regional Development in Historical Perspective [Internet]. Cham: Springer International Publishing; 2020 [cited 2021 Mar 30]. p. 131–56. (Palgrave Studies in Economic History). Available from: https://doi.org/10.1007/978-3-030-47553-6_6
4. Malta M, Murray L, da Silva CMFP, Strathdee SA. Coronavirus in Brazil: The heavy weight of inequality and unsound leadership. EClinicalMedicine. 2020 Aug;25:100472.
5. UOL. IBGE: Informalidade cresce e atinge 37,3 milhões de trabalhadores em 2017 [Internet]. 2018 [cited 2021 Mar 29]. Available from: https://economia.uol.com.br/empregos-e-carreiras/noticias/redacao/2018/12/05/ibge-trabalhadores-informalidade-brasil-2017.htm
6. Abrantes Pêgo R, Almeida C. Teoría y práctica de las reformas en los sistemas de salud: los casos de Brasil y México. Cadernos de Saúde Pública. 2002 Aug;18(4):971–89.
7. WHO. Flawed but fair: Brazil's health system reaches out to the poor [Internet]. World Health Organization; 2008 [cited 2021 Mar 26]. Available from: https://www.who.int/bulletin/volumes/86/4/08-030408/en/
8. Barreto ML, Barral-Netto M, Stabeli R, Almeida-Filho N, Vasconcelos PFC, Teixeira M, et al. Zika virus and microcephaly in Brazil: a scientific agenda. The Lancet. 2016 Mar 5;387(10022):919–21.
9. Paim J, Travassos C, Almeida C, Bahia L, Macinko J. The Brazilian health system: history, advances, and challenges. The Lancet. 2011 May 21;377(9779):1778–97.
10. Massuda A, Hone T, Leles FAG, Castro MC de, Atun R. The Brazilian health system at crossroads: progress, crisis and resilience. BMJ Global Health. 2018 Jul 1;3(4):e000829.
11. Andrietta LS, Levi ML, Scheffer MC, Alves MTSS de B e, Carneiro Alves de Oliveira BL, Russo G. The differential impact of economic recessions on health systems in middle-income settings: a comparative case study of unequal states in Brazil. BMJ Glob Health [Internet]. 2020 Feb 28 [cited 2020 Nov 20];5(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050378/
12. Nogueira Avelar e Silva R, Russo G, Matijasevich A, Scheffer M. Covid-19 in Brazil has exposed socio-economic inequalities and underfunding of its public health system [Internet]. The BMJ. 2020 [cited 2020 Nov 20]. Available from: https://blogs.bmj.com/bmj/2020/06/19/covid-19-in-brazil-has-exposed-deeply-rooted-socio-economic-inequalities-and-chronic-underfunding-of-its-public-health-system/
13. Ponce D. The impact of coronavirus in Brazil: politics and the pandemic. Nature Reviews Nephrology. 2020 Sep;16(9):483–483.
14. Faria NR, Claro IM, Candido D, Moyses Franco LA, Andrade PS, Coletti TM, et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. Virological org. 2021;586.

15. Fiocruz. COVID-19 Observatory points to the biggest health system collapse in the history of Brazil [Internet]. Fiocruz; 2021 [cited 2021 Mar 29]. Available from: https://portal.fiocruz.br/en/news/covid-19-observatory-points-biggest-health-system-collapse-history-brazil
16. Mapa da vacinação contra Covid-19 no Brasil Acompanhe a evolução da imunização [Internet]. G1 (Globo); 2021 [cited 2021 Nov 5]. Available from: https://especiais.g1.globo.com/bemestar/vacina/2021/mapa-brasil-vacina-covid/
17. Mota MT de O, Terzian AC, Silva MLCR, Estofolete C, Nogueira ML. Mosquito-transmitted viruses – the great Brazilian challenge. Brazilian Journal of Microbiology. 2016 Dec 1;47:38–50.
18. Lowe R, Lee S, Lana RM, Codeço CT, Castro MC, Pascual M. Emerging arboviruses in the urbanized Amazon rainforest. BMJ. 2020 Nov 13;371:m4385.
19. Paixão ES, Teixeira MG, Rodrigues LC. Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases. BMJ Global Health. 2018 Jan;3(Suppl 1):e000530.
20. Kazazian L, Neto ASL, Sousa GS, Nascimento OJ do, Castro MC. Spatiotemporal transmission dynamics of co-circulating dengue, Zika, and chikungunya viruses in Fortaleza, Brazil: 2011–2017. PLOS Neglected Tropical Diseases. 2020 Oct 26;14(10):e0008760.
21. Ministério da Saúde. Brasil Boletins Epidemiológicos. [Internet]. Ministério da Saúde. 2020 [cited 2021 Mar 30]. Available from: https://www.gov.br/saude/pt-br/assuntos/boletins-epidemiologicos-1/por-assunto
22. Magalhaes T, Chalegre KDM, Braga C, Foy BD. The Endless Challenges of Arboviral Diseases in Brazil. Trop Med Infect Dis [Internet]. 2020 May 9 [cited 2021 Jan 18];5(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7345859/
23. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020 Mar 17;1–3.
24. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020 Feb 20;382(8):727–33.
25. de Souza WM, Buss LF, Candido D da S, Carrera J-P, Li S, Zarebski AE, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nature Human Behaviour. 2020 Aug;4(8):856–65.
26. Castro MC, Kim S, Barberia L, Ribeiro AF, Gurzenda S, Ribeiro KB, et al. Spatiotemporal pattern of COVID-19 spread in Brazil. Science [Internet]. 2021 Apr 14 [cited 2021 Apr 14]; Available from: https://science.sciencemag.org/content/early/2021/04/13/science.abh1558
27. Sabino EC, Buss LF, Carvalho MPS, Prete CA, Crispim MAE, Fraiji NA, et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. The Lancet. 2021 Feb 6;397(10273):452–5.
28. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science [Internet]. 2021 Apr 14 [cited 2021 Apr 14]; Available from: https://science.sciencemag.org/content/early/2021/04/13/science.abh2644
29. Domingues CMAS, Maranhão AGK, Teixeira AM, Fantinato FFS, Domingues RAS. The Brazilian National Immunization Program: 46 years of achievements and challenges. Cad Saúde Pública [Internet]. 2020 Oct 26 [cited 2021 Jun 8];36. Available from: https://www.scielo.br/j/csp/a/XxZCT7tKQjP3V6pCyywtXMx/?lang=en

30. Fonseca EM da, Shadlen KC, Bastos FI. The politics of COVID-19 vaccination in middle-income countries: Lessons from Brazil. Social Science & Medicine. 2021 Jul 1;281:114093.
31. World Health Organization. WHO SAGE values framework for the allocation and prioritization of COVID-19 vaccination, 14 September 2020 [Internet]. World Health Organization; 2020 [cited 2021 Nov 5]. Report No.: WHO/2019-nCoV/SAGE_Framework/Allocation_and_prioritization/2020.1. Available from: https://apps.who.int/iris/handle/10665/334299
32. Lopes, Rachel. Estados alteram ordem do plano nacional de vacinação e priorizam forças de segurança e professores. 2021 Mar 29 [cited 2021 Nov 5]; Available from: https://www1.folha.uol.com.br/equilibrioesaude/2021/03/estados-alteram-ordem-do-plano-nacional-de-vacinacao-e-priorizam-forcas-de-seguranca-e-professores.shtml
33. Brown AL, Sperandio M, Turssi CP, Leite RMA, Berton VF, Succi RM, et al. Vaccine confidence and hesitancy in Brazil. Cad Saúde Pública. 2018 Sep 21;34:e00011618.
34. Sato APS. What is the importance of vaccine hesitancy in the drop of vaccination coverage in Brazil? Rev Saúde Pública [Internet]. 2018 Nov 29 [cited 2021 Nov 5];52. Available from: http://www.scielo.br/j/rsp/a/CS5YRcMc3z4Cq4QtSBDLXXG/?format=html&lang=en
35. Wouters OJ, Shadlen KC, Salcher-Konrad M, Pollard AJ, Larson HJ, Teerawattananon Y, et al. Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment. The Lancet. 2021 Mar 13;397(10278):1023–34.
36. Moore DCBC, Nehab MF, Camacho KG, Reis AT, Junqueira-Marinho M de F, Abramov DM, et al. Low COVID-19 vaccine hesitancy in Brazil. Vaccine. 2021 Oct 8;39(42):6262–8.
37. Gramacho WG, Turgeon M. When politics collides with public health: COVID-19 vaccine country of origin and vaccination acceptance in Brazil. Vaccine. 2021 May 6;39(19):2608–12.
38. May JM. Cultural Aspects of Tropical Medicine 1. The American Journal of Tropical Medicine and Hygiene. 1954 May 1;3(3):422–30.
39. Meade MS. Medical Geography as Human Ecology: The Dimension of Population Movement. Geographical Review. 1977 Oct;67(4):379.
40. Pavlovskiĭ EN (Evgeniĭ N, 1884-1965, Levine ND, Pavlovskiĭ EN (Evgeniĭ N, 1884-1965. Natural nidality of transmissible diseases, with special reference to the landscape epidemiology of zooanthroponoses. 1966 [cited 2021 Jan 21]; Available from: https://agris.fao.org/agris-search/search.do?recordID=US201300601889
41. Audy JR. Medical ecology in relation to geography. Br J Clin Pract. 1958 Feb;12(2):102–10.
42. Meade MS, Emch M. Medical geography. 3rd ed. New York: Guilford Press; 2010. 498 p.
43. Prior L, Manley D, Sabel CE. Biosocial health geography: New 'exposomic' geographies of health and place. Progress in Human Geography. 2019 Jun 1;43(3):531–52.
44. Springer KW, Mager Stellman J, Jordan-Young RM. Beyond a catalogue of differences: a theoretical frame and good practice guidelines for researching sex/gender in human health. Soc Sci Med. 2012 Jun;74(11):1817–24.
45. Galuzo IG. Landscape epidemiology (epizootiology). Adv Vet Sci Comp Med. 1975;19:73–96.

46. Reisen WK. Landscape Epidemiology of Vector-Borne Diseases. Annual Review of Entomology. 2010 Jan;55(1):461–83.
47. Meentemeyer RK, Haas SE, Václavík T. Landscape Epidemiology of Emerging Infectious Diseases in Natural and Human-Altered Ecosystems. Annual Review of Phytopathology. 2012 Sep 8;50(1):379–402.
48. Rosenberg MW, Waldbrook N. Creating new geographies of health and health care through technology. In: Warf B, editor. Handbook on geographies of technology. Cheltenham, UK: Edward Elgar Publishing; 2017. (Research handbooks in geography).
49. Keeler C, Emch M. Infectious-disease geography. In: Crooks VA, Andrews GJ, Pearce J, editors. Routledge handbook of health geography. London New York: Routledge; 2018. p. 45–51. (Routledge handbooks).
50. Pybus OG, Tatem AJ, Lemey P. Virus evolution and transmission in an ever more connected world. Proceedings of the Royal Society B: Biological Sciences. 2015 Dec 22;282(1821):20142878.
51. Dellicour S, Rose R, Pybus OG. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. BMC Bioinformatics [Internet]. 2016 Dec [cited 2018 Sep 21];17(1). Available from: http://www.biomedcentral.com/1471-2105/17/82
52. Kamel Boulos MN, Geraghty EM. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. International Journal of Health Geographics. 2020 Mar 11;19(1):8.
53. Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Scientific Data. 2020 Mar 24;7(1):106.
54. Xu B, Kraemer MUG, Xu B, Gutierrez B, Mekaru S, Sewalk K, et al. Open access epidemiological data from the COVID-19 outbreak. The Lancet Infectious Diseases. 2020 May 1;20(5):534.

3 The physical environment and yellow fever virus emergence

Environmental changes can have substantial impact on the emergence and resurgence of mosquito-borne diseases. Climate change and anthropogenic activities have induced alterations in temperature, precipitation, and land-use, which could expand the spatial extent of virus-infected vectors and have redefined the relationships between vector, host(s), and pathogens in the natural environment. The following paper which has been accepted for publication in PLOS Neglected Tropical Diseases covers research conducted to predict the spatial distribution of sylvatic mosquito vectors responsible for transmitting yellow fever virus in Brazil, and assesses the potential risk of virus spillover into human populations.

Mapping environmental suitability of *Haemagogus* and *Sabethes spp.* mosquitoes to understand sylvatic transmission risk of yellow fever virus in Brazil

Short title: Mapping suitability of *Haemagogus* and *Sabethes spp.* in Brazil

Sabrina L. Li [1#], André L. Acosta [2], Sarah C. Hill [3], Oliver J. Brady [4,5], Marco A. B. de Almeida [6], Jader da C. Cardoso [6], Arran Hamlet [7], Luis F. Mucci [8], Juliana Telles de Deus [8], Felipe C. M. Iani [9], Neil S. Alexander [10], G. R. William Wint [10], Oliver G. Pybus [3,11], Moritz U.G. Kraemer [11], Nuno R. Faria [7,11,12], Jane P. Messina [1,13#]

1. School of Geography and the Environment, University of Oxford, Oxford, United Kingdom.
2. Departamento de Ecologia, Instituto de Biociências, Laboratório de Ecologia de Paisagens e Conservação - LEPAC, Universidade de São Paulo, São Paulo, Brazil.
3. Department of Pathobiology and Population Sciences, Royal Veterinary College London, London, United Kingdom.
4. Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom.
5. Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom.
6. State Centre of Health Surveillance, Rio Grande do Sul State Health Secretariat, Rio Grande do Sul, Brazil.
7. MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom
8. Superintendence for Endemic Diseases Control, São Paulo State Health Secretariat, São Paulo, Brazil.
9. Fundação Ezequiel Dias - Lacen/MG, Minas Gerais, Brazil.
10. Environmental Research Group Oxford, c/o Department of Zoology, University of Oxford, Oxford, United Kingdom.
11. Department of Zoology, University of Oxford, Oxford, United Kingdom.

12. Departamento de Molestias Infecciosas e Parasitarias & Instituto de Medicina Tropical da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.
13. Oxford School of Global and Area Studies, University of Oxford, Oxford, United Kingdom.

#Corresponding authors:

1. Sabrina L. Li, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom. E-mail: lisabrinaly@gmail.com

2. Jane P. Messina. School of Geography and the Environment, University of Oxford, Oxford, United Kingdom. E-mail: jane.messina@gtc.ox.ac.uk

Abstract

Background

Yellow fever (YF) is an arboviral disease which is endemic to Brazil due to a sylvatic transmission cycle maintained by infected mosquito vectors, non-human primate (NHP) hosts, and humans. Despite the existence of an effective vaccine, recent sporadic YF epidemics have underscored concerns about sylvatic vector surveillance, as very little is known about their spatial distribution. Here, we model and map the environmental suitability of YF's main vectors in Brazil, *Haemagogus spp*. and *Sabethes spp.,* and use human population and NHP data to identify locations prone to transmission and spillover risk.

Methodology/Principal Findings

We compiled a comprehensive set of occurrence records on *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp.* from 1991-2019 using primary and secondary data sources. Linking these data with selected environmental and land-cover variables, we adopted a stacked regression ensemble modelling approach (elastic-net regularized GLM, extreme gradient boosted regression trees, and random forest) to predict the environmental suitability of these species across Brazil at a 1x1 km resolution. We show that while suitability for each species varies spatially, high suitability for all species was predicted in the Southeastern region where recent outbreaks have occurred. By integrating data on NHP host reservoirs and human populations, our risk maps further highlight municipalities within the region that are prone to transmission and spillover.

Conclusions/Significance

Our maps of sylvatic vector suitability can help elucidate potential locations of sylvatic reservoirs and be used as a tool to help mitigate risk of future YF outbreaks and assist in vector surveillance. Furthermore, at-risk regions identified from our work could help disease control and elucidate gaps in vaccination coverage and NHP host surveillance.

Author Summary

Yellow fever virus (YFV) is an arbovirus transmitted to humans from mosquitoes and can lead to severe disease and death. Recent sporadic outbreaks coupled with low vaccination coverage have highlighted the importance of mosquito surveillance for preventing future outbreaks and potential virus spillover into dense urban areas. Yet, very little is known about the spatial distribution of mosquitoes known to transmit YFV and the factors that contribute to their environmental suitability in Brazil. We compiled an occurrence database of primary and secondary mosquito vectors belonging to *Haemagogus* and *Sabethes* species' collected between 1991-2019 and integrated this data with environmental and land-use data to predict their spatial suitability at 1x1km resolution. Using this information, we identified suitable regions for their co-existence. We overlaid this information with human population density and locations of non-human primate host reservoirs to identify areas at risk of transmission and spillover. Our study provides high-resolution mapping tools to assist with mosquito and arbovirus surveillance which is especially useful in low-resource settings.

Introduction

In Brazil, several sylvatic mosquito species' belonging to the genera *Haemagogus* and *Sabethes* are the primary and secondary vectors of yellow fever virus (YFV) (1,2). YFV is an arbovirus belonging to the family *Flaviviridae* that causes acute infectious disease in humans. Literature states that among non-vaccinated populations from Africa and South America, about 55% (95% credible interval, CI: 37%-74%) of individuals infected with YFV are asymptomatic, while about 33% (95% CI: 13%-52%) develop mild disease. Among the approximate 12% (95% CI: 5%-26%) that develop severe disease, case fatality can reach 47% or greater (95% CI: 31%-62%) (3). In peri-urban and urban areas, transmission of YFV can also be maintained between *Aedes aegypti* (Linnaeus, 1972) and humans; however, this transmission cycle has not been documented in Brazil since 1942 (4). YFV is endemic in the Brazilian Amazon Basin and mainly remained there until the 1940s, when sporadic outbreaks were observed in the Northeast, Southeast, and Southern states (4–6). Recent re-emergence of YFV in Southeastern Brazil between 2016 and 2018 caused the worst YF outbreaks in the last 70 years (7).

Despite the existence of an effective YF vaccine since 1937, eradication efforts in Brazil have been undermined by an enzootic sylvatic transmission cycle maintained in rural areas by reservoirs of *Haemagogus spp.* and *Sabethes spp.* belonging to the Culicidae family. These species feed on non-human primates (NHP) (e.g. *Alouatta sp*.) that dwell in high canopy forests, and occasionally on humans that live in or travel to these areas. During the 2016-2018 outbreaks, *Hg. janthinomys,* which can be naturally infected with YFV (8), was found to be the primary vector, while infected *Hg. leucocelaenus* were also abundant in outbreak areas (1). *Sa. chloropterus* was found to play a local or secondary role in sylvatic transmission due to limited abundance and distribution (1). These mosquitoes are vertically dispersed in forest areas (9) and naturally prefer to bite and breed at high canopy levels (10,11). Due to environmental disturbances, contact between sylvatic mosquitoes and human populations have increased (12). While the occurrence of YF cases follows a seasonal pattern, during which most cases are reported between December to May, outbreaks are still highly dependent on environmental conditions that facilitate transmission (7). Given the frequency and severity of recent epidemics and the ubiquity of human travel to sylvatic

transmission zones, there is increasing concern for a sylvatic-to-urban cycle spillover into *Ae. aegypti* infested areas with low vaccination coverage, especially those in near proximity to sylvatic environments. Thus, a greater understanding of the distribution of sylvatic vectors of YFV and its implications for a spillover is needed to help prevent future epidemics.

Species distribution modelling (SDM) techniques are useful for mapping the distribution of arbovirus vectors by focusing on their ecology as predicted from a combination of data on climate, environmental, and population factors (13,14). To date, only two studies have attempted to map the distribution of sylvatic vectors of YFV. Childs et al. (15) applied SDM using a MaxEnt model to determine the combined distribution of *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sa. chloropterus* in Brazil, by integrating data on climatic and environmental factors and occurrence data retrieved from the Global Biodiversity Information Facility (GBIF) and published literature. Similarly, Almeida et al. (16) predicted the distribution of *Hg. leucocelaenus* in the state of Rio Grande do Sul using occurrence data from the 2008 - 2009 epizootics.

However, these results alone may not accurately determine the ecological niche of each vector and their locations given use of confined occurrence data and limited presumptions about the environmental conditions that influence their distribution in Brazil. Random sampling assumptions adopted for identifying their potential absence via the MaxEnt model may reflect unbiased surveillance efforts, which is rarely the case in the field. While Childs et al. (15) adopted a pseudo-absence selection method consisting of subsampling from occurrence data on mosquito species other than *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sa. chloropterus*, this method may incorrectly identify potential pseudo-absence locations as it classifies excluded species of *Haemagogus* and *Sabethes* sharing similar ecology with the aforementioned primary vectors as pseudo-absences. Given that little is known about these vectors and their niches, these limitations should be carefully considered. Moreover, neither study assessed the environmental suitability of *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp.* individually across Brazil, which is important given differences in their ecological niche and distributions across space.

Building on these previous efforts, we modelled the distribution of *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp.* in Brazil by compiling the most comprehensive dataset of *Haemagogus* and *Sabethes spp.* to date, which contains not only occurrence data but also records from GBIF database, published literature, and primary entomological surveys collected between 1991 and 2019. Using covariate data representative of environmental conditions (daytime land surface temperature, precipitation, relative humidity, and land-cover classes) and human activity (population density and deforestation) conducive to sylvatic mosquito activity (as confirmed by lab and field studies), we modelled the probability of environmental suitability for each species using a stacked regression ensemble approach consisting of three different models to increase predictive performance. By combining the resulting predicted species suitability from this study with modelled presence probability of non-human primate host reservoirs, within and surrounding the Southeastern Atlantic forest biome (17), we further mapped the risk of YFV transmission and spillover in highly populated areas of the region. These maps could inform policy makers the spatial distribution of current and emerging YFV risk zones, which can be targeted for transmission prevention and vaccination programmes.

Materials and Methods

We adopted a stacked binomial regression ensemble modelling approach to quantify the probability of environmental suitability of *Hg. janthinomys, Hg. leucocelaenus*, and *Sabethes spp.,* respectively. A stacked regression ensemble is a diverse combination of base models (level-0), with potentially different assumptions about the predictions, into a meta-model (level-1), which has equal or better predictive performance than each individual model (18). The meta-model is trained on predictions made by the base models and is adjusted based on an optimal weighted average of these models. Previously, this ensemble approach was adopted to map *Plasmodium falciparum* malaria parasite prevalence in sub-Saharan Africa, and was found to outperform any individual method (19). Here we used three models in an ensemble: a lasso or elastic-net regularized generalized linear model (GLM), an extreme gradient boosted regression tree model, and a random forest model. We chose these models because they were suitable for binary prediction and have relatively good generalizability for disease mapping (20,21).

We used the following set of input variables to run the ensemble: (i) a comprehensive dataset of geocoded occurrence locations for *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp.* recorded from 1991-2019; (ii) selected environmental and land-cover covariate datasets that help to explain the ecological niche of each species; and (iii) a dataset outlining the geocoded locations of pseudo-absence records which refines the geographical range of species occurrence and reduces sampling bias (21). This is described in more detail in the following sections and in the Supplementary Information, alongside full details of the data and each analysis stage. An outline of the methodology from data assembly, modelling approach, to outputs is described in Fig 1.

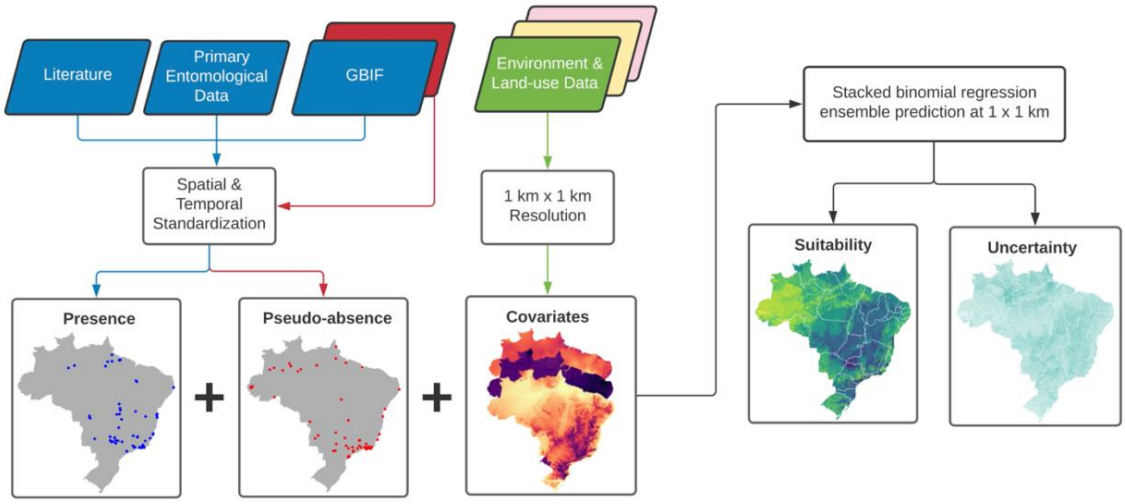


Fig 1. Schematic diagram of the modelling approach adopted to produce maps of environmental suitability for *Haemagogus spp*. and *Sabethes spp.* at 1km x 1km resolution.

Presence records

To assess contemporary distributions, we extracted all records for *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp.* between the years 1991-2019 from our compiled dataset, which contains the geographical coordinates of species occurrence. Due to sparse records, we extracted all species of *Sabethes spp.* that have been confirmed of transmitting, capable of being naturally infected, or are found before and during previous and recent YFV outbreaks in Brazil. This includes *Sa. chloropterus*, *Sa. albiprivus*, *Sa. soperi*, *Sa. glaucodaemon*, *Sa. cyaneus*, *Sa. purpureus*, *Sa. melanonymphe*, *Sa. undosus/fabrici/ignotus*, *Sa. intermedius*, *Sa. belisarioi*, *Sa.*

220 *aurescens*, *Sa. identicus*, *Sa. xyphydes*, and *Sabethes sp.* (1,2,8,11). To our
221 knowledge, our occurrence dataset is the most comprehensive for these species in
222 Brazil to date, which contains the geocoded information on the occurrence of pupae,
223 larvae or eggs, and adults of each species recorded between 1922 – 2019 in Brazil.
224
225 The sources of these data were from (i) The Global Biodiversity Information Facility
226 (GBIF; https://www.gbif.org/) database, (ii) published literature (see list of papers in
227 dataset), and (iii) primary and unpublished occurrence data from entomological
228 surveys collected by the Health Secretariat of the State of Rio Grande do Sul and
229 Projeto Cantareira as part of São Paulo State Superintendence for Endemic Diseases
230 Control (SUCEN). To retrieve occurrences from the literature, we conducted searches
231 on Scopus and PubMed using the search strings “haemagogus AND Brazil” and
232 “sabethes AND Brazil” (Figure 2). We used the geographical coordinates (latitude and
233 longitude) of each observed occurrence when available. In cases where only an area
234 location was mentioned (e.g. name of town or city), we used Google Maps to extract
235 the coordinates of the centroid of the cited location. The number of records retrieved
236 by year is presented in the Supplementary Information – Section 1. To permit
237 reproducibility, all occurrence data containing the aforementioned information are
238 made freely accessible online on https://github.com/sabrinalyli/YFVMappingBrazil.git.
239

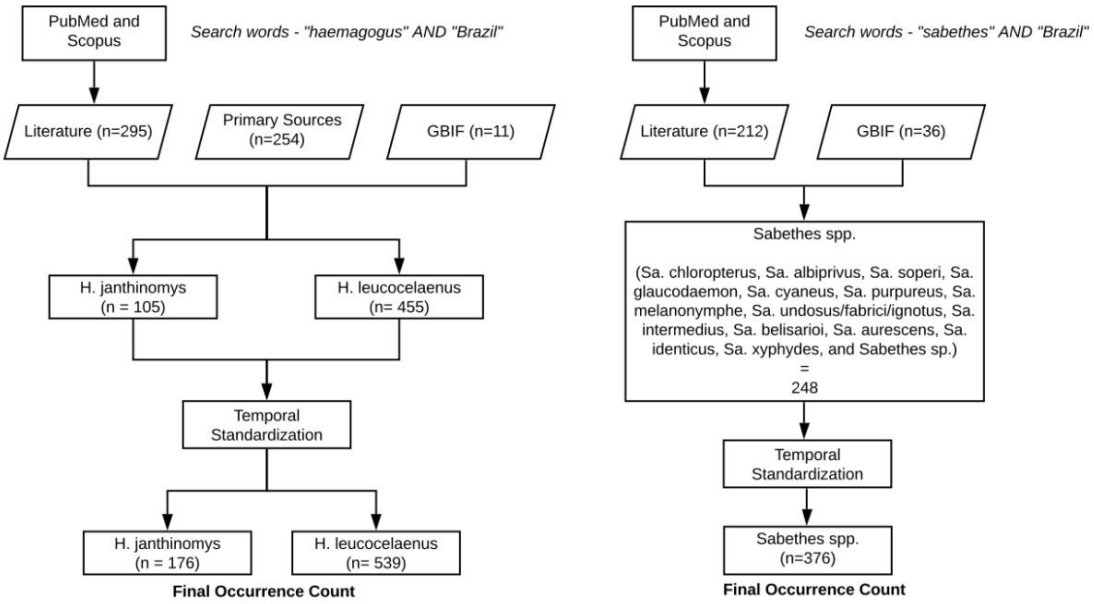


Fig 2. Number of occurrence records for *Hg. janthinomys, Hg. leucocelaenus, and Sabethes spp.* (1991-2019) prior and post standardization.

We performed spatial and temporal standardization using methods described elsewhere (14,22). This creates a list of occurrences in which each occurrence is individually linked with a unique location and year. This ensures that occurrences from regions where multiple surveys are conducted are not overly-represented compared to occurrences from less well sampled regions. To validate the geo-positioning of occurrence locations, we overlaid the occurrence points with a map of Brazil to ensure that points were geocoded to land only. Point locations that were incorrectly geocoded to a location over water were adjusted using an algorithm described elsewhere (23) or manually checked for validity.

Environment and land-use variables

Data on the spatial and temporal distribution of environmental conditions are extracted from high-resolution satellite raster imagery. We standardized all covariates to a 1 km x 1 km spatial resolution by resampling each raster using the bilinear method (24) and have outlined the rationale for inclusion in our models below. A summary of metadata for our covariate datasets along with details of covariate sensitivity analysis can be found in the Supplementary Information – Section 2. All data processing was done in R (Vers. 4.0.2) using the rgdal, raster, and sf packages.

Temperature, Precipitation, and Relative Humidity

Climatic factors such as temperature, relative humidity, and precipitation influence the abundance and activity of *Haemagogus spp.* and *Sabethes spp.* (25–27). These mosquitoes exhibit daytime biting behaviour (28,29); in particular, *Hg. leucocelaenus* only travel closer to the ground to feed during the day (30). To account for temperature in our model, we adopted mean annual daytime land surface temperature layers extracted from MODIS v6 time series processed using the Temporal Fourier Analysis (TFA) described elsewhere (31). For precipitation, we calculated yearly mean total precipitation, yearly minimum precipitation, and yearly maximum precipitation, averaged between years 2001-2019 using data retrieved from ERA5 dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Minimum annual relative humidity from 2015 was determined as a percentage of water vapour needed to saturate the air given surrounding temperature (32), and was extracted for each 1 km x 1 km pixel.

Enhanced Vegetation Index (EVI) and Vegetation Type

Haemagogous and *Sabethes* species have an affinity for forest canopies and inhabit tree holes, bamboo nodes, bromelias, and coconut husks in vegetated areas (33,34). Many complete their larval development in temporarily flooded forest areas (35). *Haemagogus spp.* and *Sabethes spp.* are primarily found in rain forests and deciduous forests (25,36). To characterize vegetation cover, we used EVI retrieved from MODIS, a measure of vegetation canopy greenness ranging from 0-1, where values near 1 indicate lush and dense vegetation. We also included consensus land-cover layers processed by Tuanmu et al. (37) detailing the prevalence of evergreen and deciduous trees.

Forest Cover and Deforestation

Sylvatic mosquito abundance could be influenced by increasing deforestation activities as a result of regular landscape transformation to accommodate agricultural activities and cattle grazing in recent years (4,38). To characterize forest cover and deforestation, we adopted gross forest cover loss for the years 2000-2019, using data processed by Hansen et al. (39). We also included a consensus land cover on

cultivated and managed vegetation to account for land that has been anthropogenically modified for agricultural purposes.

Elevation

While *Haemagogus* and *Sabethes spp.* are known to prefer canopy habitats within the forest strata, they can also be found near the ground depending on the time of the year and climate conditions (11,40,41). Thus, to account for vertical oviposition, we extracted minimum elevation in meters above sea level from a raster processed by Amatulli et al. (42).

Population Density

Higher population density in cities has the potential to increase physical contact between people and YFV mosquito vectors, which could potentially instigate and maintain another urban YFV transmission pathway in Brazil (43). Furthermore, including population density in our model allows us to account for the possible occurrence of *Hg. janthinomys* and *Hg. leucocelaenus* in forest-city ecotones. To account for population density, we included population density estimates adjusted to match country estimates reported by the United Nations from WorldPop (https://www.worldpop.org/) calculated using methods described elsewhere (44).

Non-human primate reservoirs

We retrieved the presence probability of NHP reservoirs predicted at 1km x 1km resolution for six species within the Southeastern Atlantic forest in Brazil, which was determined using occurrence records from 1992 - 2016 retrieved from Culot et al. (17). Methods for producing these rasters are described elsewhere (45). The six species of monkeys were of the following genera: *Alouatta sp.*, *Brachyteles sp.*, *Callicebu*s sp., *Callithrix sp.*, *Leontopithecus sp.*, and *Sapajus sp*. These species have been found to host and transmit YFV in Brazil (46,47). In particular, Howler monkeys (*Alouatta sp.)* are extremely susceptible to YFV and can develop fatal disease (48).

Modelling Approach

Our stacked binomial regression ensemble consists of an elastic-net regularized GLM, extreme gradient boosted regression trees, and random forest. An elastic-net

regularized GLM can be used as a binary classifier that adopts regularization to shrink the coefficients of the linear model to reduce overlearning and to allow for better generalization (49). Both extreme gradient boosted regression trees and random forest use an ensemble of regression trees to increase prediction accuracy (50). The method of gradient boosting builds trees in a stagewise process, where new trees help to correct errors created by the previous tree, while random forest trains each tree independently using a randomly sampled training dataset that makes the model more robust and less prone to overfitting (50). Boosted regression trees, first described for disease mapping in Elith et al. (51), have been adopted to map several arboviral diseases and vectors (52–54). The method has also been previously adopted to estimate environmental suitability of *Aedes* vectors known to facilitate arbovirus transmission (14,55). Within the ensemble, we adopt extreme gradient boosting, which is designed to be computationally efficient with better model performance (56).

From the ensemble, we retrieved a predicted probability of suitability at 1km x 1km spatial resolution. To obtain uncertainty estimates around the prediction, we fitted an ensemble of 100 sub-models by applying a separate bootstrap sampling of the dataset for each model. We then determined the distribution of pixel-based uncertainty by calculating the 95% confidence interval derived from the 100 sub-models. All data analysis were conducted using raster, gbm, dismo, and seegSDM packages in R (Vers. 4.0.2). All models were fitted using the *Superlearner*, *glmnet*, *xgboost*, and *ranger* packages in R.

Pseudo-absence records for removing sample selection bias

A sample of background data defined by absence records is required by BRT models to understand conditions available to the modelled species in an area. Phillips et al. (21) proposes a pseudo-absence selection approach using background records that share the same environmental characteristics as the presence records. This minimizes sample selection bias while highlighting the differentiation between the distribution of presence records and that of the background. We adopted this method by compiling a pseudo-absence dataset containing records of similar acrodendrophilic mosquitoes that belong to the same Culicidae family and are captured using the same methods and equipment in the field. A total number of 1,056 records including *Culex spp., Limatus spp.*, *Wyeomyia spp., Psorophora spp.*, *Orthopodomyia spp.*, *Mansonia spp.*,

and *Coquillettidia spp.* observed during 1992-2020 were extracted from the GBIF database.

Mapping suitable areas, risk of transmission, and risk of spillover

To better visualize areas that are suitable for vector species, we mapped predicted suitability above a certain threshold. This was done by highlighting pixels where predicted suitability was greater than 0.5 for each species and for all species. We also conducted a sensitivity analysis by changing the threshold to 0.25 and 0.75 (Supplementary Information – Section 4). Using these data, we mapped the risk of transmission and risk of spillover from sylvatic environments by determining areas of overlap between two to three variables: high environmental suitability of vector species, high predicted presence probability of NHPs, and presence of human population (yes/no) based on population density. High environmental suitability was defined by suitable areas filtered by the threshold. Likewise, high predicted presence probability of NHPs was defined by a probability greater than 0.5.

We defined the risk of transmission by the presence of two variables, which interact to maintain transmission. Firstly, a transmission cycle can be maintained by the presence of sylvatic mosquito vectors and NHP hosts. Secondly, although YF transmission can be maintained by *Aedes spp.,* it was not observed during the 2016-2018 outbreaks. Nonetheless, increasing overlap between human populations and NHP host reservoirs could instigate an urban cycle. Lastly, the presence of human populations in areas where sylvatic vectors are present can also foster transmission. Similarly, we defined the risk of spillover by the presence of all three variables in an area, where local transmission is maintained by the presence of mosquito vectors and NHP hosts that feed on nearby human populations. Finally, areas where only one variable is present are defined as low risk. Despite the lack of overlap, transmission could still occur if humans, NHP hosts, and mosquito vectors are assumed to be mobile within given suitable environmental conditions and reasonable travel behaviour.

Model Validation

We calculated the overall predictive accuracy of the model ensemble by quantifying the area under the curve (AUC) statistic and its 95% confidence interval. This was done by running our model ensemble on a randomly sampled subsection of the data

referred to as the training set of the data, which contains 80% of the original data. Cross-validation on each model was performed by evaluating the mean area under the curve (AUC) statistic from 10 separate cross-validation folds. This model was then applied for prediction on the holdout set (20% of the original data). Using 2000 bootstraps, we computed the 95% confidence interval. All analyses were performed using the *ROCR* (57) and *pROC* (58) packages in R.

All covariates used in the model were assessed for their variable importance using the permutation method (59,60), which is measured as a change in the model error. Values of each covariate in the data were randomly permuted and assessed for its effect on the model performance using the difference between a baseline performance measure (e.g. RMSE) before and after permutation. This was performed on the stacked binomial regression ensemble using the *vip* package in R (61).

Results

To characterize the presence of *vectors* we retrieved a total of 1,091 occurrences (*Hg. janthinomys* = 176; *Hg. leucocelaenus* = 539; *Sabethes spp*. = 376) over the time period 1991-2019 after standardization. Out of all the occurrences recorded for *Hg. janthinomys,* most were recorded in the states of Rio de Janeiro (RJ: 27.3%), São Paulo (SP: 17.6%), and Bahia (BA:15.9%) (Fig. 3A-B) likely due to higher surveillance in recent years as a result of recent outbreaks. For *Hg. leucocelaenus*, 55% of all occurrences were recorded in Rio Grande do Sul (RS), while 20.5% and 17.6% were recorded in São Paulo and Rio de Janeiro, respectively (Fig. 3C). The distribution of *Sabethes spp.* occurrences adopted in our study were found in São Paulo (27.8%), Rio de Janeiro (19.4%), Rio Grande do Sul (14.4%) and Bahia (8.2%), (Fig. 3D).

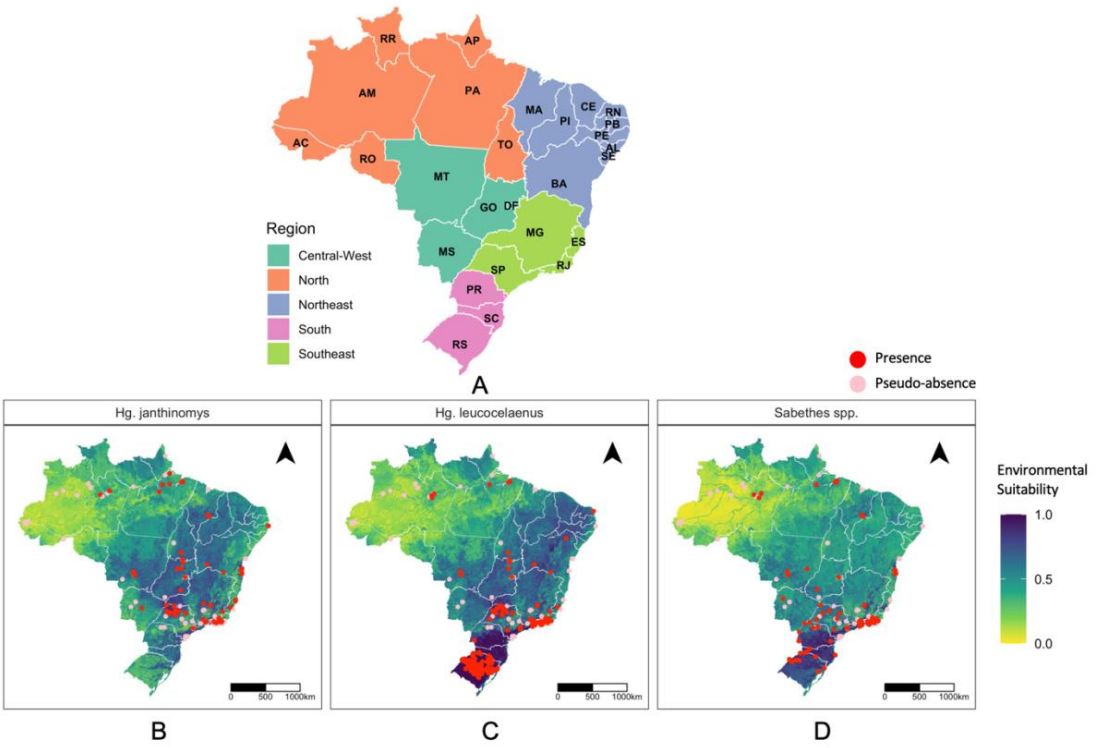


Fig 3. Environmental suitability and occurrence records of *Haemagogus* and *Sabethes spp.* A, Map of Brazil by states (AC – Acre, AL – Alagoas, AP – Amapá, AM – Amazonas, BA – Bahia, CE – Ceará, DF – Distrito Federal, ES – Espírito Santo, GO – Goiás, MA – Maranhão, MT – Mato Grosso, MS – Mato Grosso do Sul, MG – Minas Gerais, PA – Pará, PB – Paraíba, PR – Paraná, PE – Pernambuco, PI – Piauí, RJ – Rio de Janeiro, RN – Rio Grande do Norte, RS – Rio Grande do Sul, RO – Rondônia, RR – Roraima, SC – Santa Catarina, SP – São Paulo, SE – Sergipe, TO – Tocantins) and regions (colour-coded). **B-D,** Maps indicating the point locations (red) of standardized occurrence records of each species collected between 1991-2019 and predicted environmental suitability for *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp*.

Figure 3B-D also shows the predicted suitability of *Hg. janthinomys*, *Hg. leucocelaeunus*, and *Sabethes spp.* To evaluate prediction accuracy of the outputs, we determined model performance using the AUC (area under the curve) on our holdout set and found high accuracy (>0.97). While these species are distributed similarly, there are some differences. The Southeastern region part of the Atlantic forest biome encompassing the states of Minas Gerais (MG), Rio de Janeiro (RJ),

Paraná (PA), and parts of São Paulo (SP) were predicted to be the most suitable locations for *Hg. janthinomys* and *Hg. leucocelaenus*, where the species have been recorded in the last three decades. While few species were recorded in these regions, the Central-West Brazilian states incorporating the *Cerrado* biome of Goiás (GO) and Mato Grosso do Sul (MS), along with some areas within Tocantins (TO), Federal District (DF), and Mato Grosso (MT) were also found to be suitable. Highly suitable areas were also predicted in the Northeast state of Bahia and Piauí (PI), and nearby adjacent areas in the states of Ceará (CE), Paraíba (PB), and Pernambuco (PE), where the species have yet to be reported. The Southern states of Rio Grande do Sul (RS), Paraná (PR), and Santa Catarina (SC) were predicted to be highly suitable for both *Hg. leucocelaenus* and *Sabethes spp*. For all species, suitability was predicted to be low across the Northern region along the Amazon river basin, where contemporary records of species are sparse. Maps summarizing uncertainty associated with predicted estimates are shown in Supplementary Information – Section 3.

These patterns are further reflected in the maps showing suitable areas (suitability probability ≥ 0.5) for each species (Figure 4A-C) and for all species combined (Figure 4D). Our sensitivity analysis shows that 0.5 is an appropriate- threshold to avoid underestimation and overestimation of suitable zones (Supplementary Information – Section 4). States that are suitable for all species are mainly concentrated in highly populated areas of the Southeastern region and along part of Northeast and Central-West.

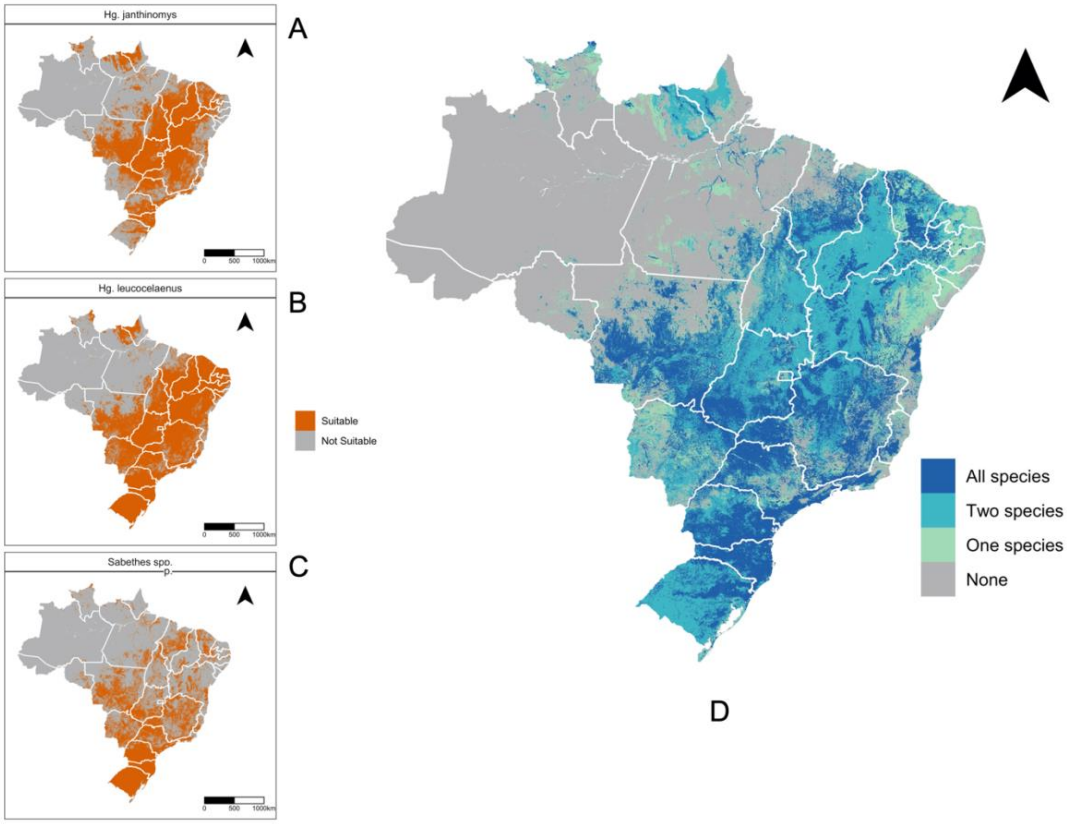


Fig 4. Suitable regions for *Haemagogus* and *Sabethes spp*. A-C, Maps of areas identified as suitable with environmental suitability ≥ 0.5. **D,** Map showing areas suitable for only one species, two species, or all species analyzed in this study.

Likewise, our risk map based on the presence of mosquitoes, NHP host reservoirs, and human populations (Figure 5) shows that areas at risk of YFV spillover are concentrated along highly populated municipalities of Paraná, the cities of São José do Rio Preto and São Paulo, as well as municipalities in Rio de Janeiro and Espírito Santo states. Areas at risk of spillover also include the Federal District and Goiás. These areas are surrounded by comparatively less populated municipalities which are at the risk of transmission. The regions between Goiás and Minas Gerais have the lowest risk given the lack of NHP host reservoirs. We superimposed our spillover risk maps with data on confirmed (PCR, immunohistochemistry and/or IgM ELISAs) YFV cases among humans and non-human primates between 2015 – 2020 (62) retrieved from the Brazilian Ministry of Health National Notifiable Disease System (SINAN; Sistema de Informação de Agravos de Notificação). This shows that previous human

and NHP cases clustered mainly in spillover zones. However, the original site of human infection may be different than the site of reporting attributed to passive surveillance coupled with human movement.

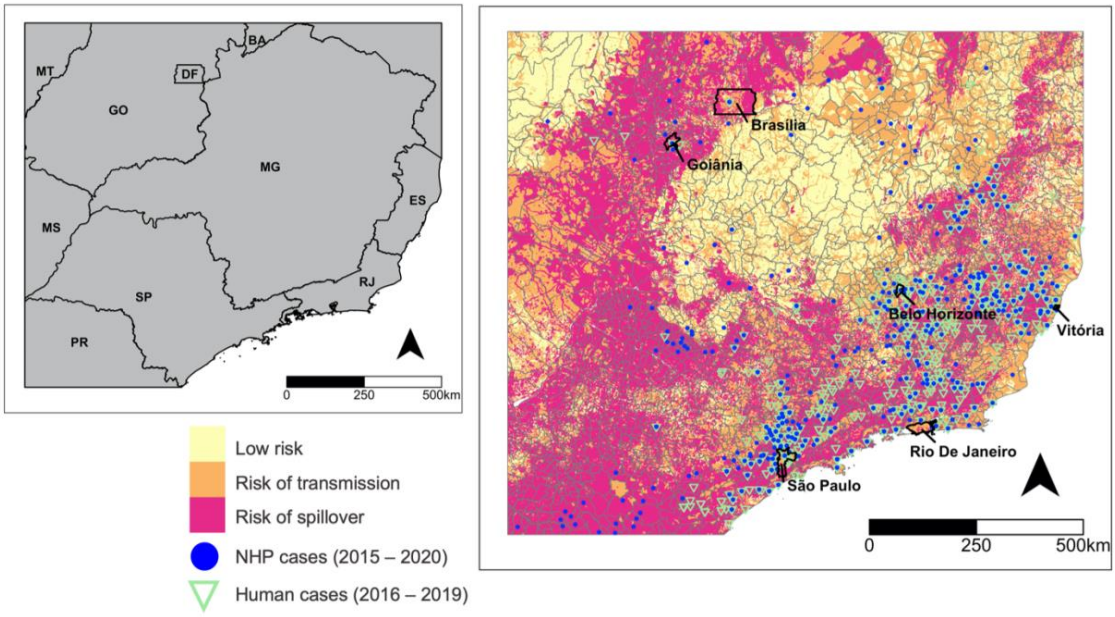


Fig 5. Risk of transmission and spillover along the Southeastern Atlantic Forest Biome. Map showing the distribution in risk status in the Southeastern municipalities of São Paulo, Minas Gerais, Rio de Janeiro, Espírito Santo, and Southern state of Paraná and adjacent municipalities in the Central-West, including the Federal District, Goiás, and Mato Grosso. Risk is defined by the presence of sylvatic mosquito vectors, non-human primate reservoirs, and human population density. Low risk refers to the presence of only one of these factors. Risk of transmission is defined by the overlapping presence of two of these factors, while risk of spillover is defined by the overlapping presence of all three factors. Blue points show the locations of confirmed cases among non-human primates (NHP) recorded between years 2015 – 2020 while green triangles indicates the locations of confirmed human cases recorded between 2016 – 2019.

Predictors with the strongest variable importance on the model ensemble for *Hg. janthinomys* included population density (permuted model error: 0.106±0.001), which shares a negative relationship with species distribution, the prevalence of deciduous tree cover (0.1±0.004), and elevation (0.075±0.001). For modelling *Hg. leucocelaenus*, the most influential predictors were elevation (0.097±0.002) and

minimum daytime land surface temperature (0.069±0.003), followed by population density (0.067±0.002). Similar results were found for modelling *Sabethes spp.*, where the prevalence of deciduous tree cover (0.091±0.002), population density (0.084±0.002), and elevation (0.078±0.002) had the strongest relative influence (Supplementary Information – Section 4). Relative to all covariates, mean forest cover loss had consistently one of the least variable importance for all models. Partial dependence plots for each covariate are presented in the Supplementary Information – Section 4. Overall predictive performance was high for all model ensembles based on computed AUC (Supplementary Information – Section 5).

Discussion

To understand the spatial distribution and patterns of YFV vectors and their risk to humans, we used a stacked regression ensemble modelling approach to map the environmental suitability of main YFV vectors from the *Haemagogus* and *Sabethes* genera at a high spatial resolution of 1km x 1km. This was done using occurrence records reported between 1991-2019 combined with associated environmental and land-cover covariates. Using predicted estimates, we identified suitable regions for each vector and all vectors. We also mapped the distribution of risk of transmission and spillover in regions that overlapped human settlements and the presence probability of NHP host reservoirs.

Previous studies by Childs et al. (15) and Almeida et al. (16) used MaxEnt models with different assumptions about the selection of pseudo-absences. Using a stacked binomial regression model ensemble and pseudo-absence selection based on background records that share the same environmental characteristics as the presence records, we found higher suitability in the Northeast and lower suitability in the North, which is the opposite of the findings presented by Childs et al. (15). Similar to Almeida et al. (16), our model ensemble also predicted high suitability for *Hg. leucocelaenus* in Rio Grande do Sul. Compared to Almeida et al. (16), the integration of additional occurrence points in our model ensemble allowed the prediction of higher suitability for the south of Rio Grande do Sul. Whereas temperature did not play an

important role in model prediction in Childs et al. (15), we found it an important covariate in ranking after elevation, population density, and deciduous forest cover.

Despite the existence of effective vaccines, Brazil experienced the largest sporadic YF outbreaks of recent years during 2016-2018, which resulted in more than 2,153 diagnosed human cases and 744 deaths in highly populated urban areas of Southeast Brazil that were previously considered low risk (63). These outbreaks, which were identified to be of sylvatic origin (1), spilled into the transition zone between the *Cerrado* biome and the Atlantic Forest in Minas Gerais and moved towards the Southeastern states of Minas Gerais, Espírito Santo, Rio de Janeiro, and São Paulo, (61), which corresponds to predicted areas of high environmental suitability for primary and secondary sylvatic vectors in our study. We also identified highly suitable areas in the South and Central-West regions, including Santa Catarina and Rio Grande do Sul, where outbreaks were predicted to occur every six to ten years (4) and have been recently reported for 2019 and 2021 (64,65). These findings are consistent with recent spatial epidemiological patterns of sylvatic YF outbreaks in the Central-West and Southeastern regions (7,66). Low predicted suitability in the Amazon river basin is likely due to decreased entomological reporting in recent years, mainly due to lack of surveillance of sylvatic vectors in the Amazon region. These maps can be used as a public health tool and help guide the surveillance of YFV vectors, especially in areas lacking entomological data but were identified as environmentally suitable.

Our understanding of each species' ecology is confirmed by the relative influence of environmental covariates on the models. The distribution of each species may be dependent on the height of treetop canopies, but further investigation is required. Here we used elevation to characterize vertical oviposition activity. A previous study found that the risk of YFV cases at elevations between 300 – 800 MASL was six times higher compared to other elevations (67), indicating a potential linkage to sylvatic mosquito activity. Population density strongly influenced species' distribution via an inverse relationship, corroborating species' preference for non-urbanized settings. The prevalence of deciduous tree cover had a strong influence on shaping the distribution of both *Hg. janthinomys* and *Hg. leucocelaenus*. Minimum day-time temperature had a strong influence on both *Hg. leucocelaenus* and *Sabethes spp.* species. In particular, higher suitability was detected when temperature was above 27 ℃, which is consistent

572 with recent findings on temperature-population dynamics of *Hg. leucocelaenus*
573 (25,68).
574
575 Forest loss was found to have an overall weak influence on the models for all three
576 species. Evidence on the relationship between deforestation and mosquito abundance
577 is scarce and remains unclear. In Brazil, *Hg. janthinomys* and *Hg. leucocelaenus* are
578 highly dispersed between forest areas, and can be found in small forest patches (e.g.
579 7 hectares) next to urban areas and isolated high-income neighbourhoods located in
580 forest valleys (1), In particular, *Hg. leucocelaenus* are often found in modified forest
581 patches near human activities, including isolated forest fragments in cultivated areas
582 and in pasture grasslands generally associated with small rivers and creeks as well as
583 urban green areas (30,69,70). A systematic review that investigated the effects of
584 deforestation on mosquitoes found that half of mosquito species favoured deforested
585 areas (71). However the loss of mosquito habitats from the progression of
586 deforestation may also reduce transmission risk (72). Moreover, some considerations
587 have also been given to the changing agricultural ecosystem in Brazil, which is a main
588 driver of deforestation. Seasonally varying agricultural activities such as planting and
589 harvesting to accommodate the growth of cash crops have been found to capture
590 seasonal YFV reporting in humans and NHPs (38), but little is known about the impacts
591 of regular land conversion through agriculture on mosquito habitats and transmission
592 activity. Nonetheless, more field observations and investigations are needed to
593 understand the effects of agriculture and deforestation on mosquito abundance.
594
595 By combining predicted estimates of vector suitability, predicted presence of NHP host
596 reservoirs, and population density, we further characterized and mapped the
597 distribution in risk of YFV transmission and spillover within and surrounding the
598 Southeastern Atlantic Forest biome. Our detection of spillover and transmission risk in
599 and around Minas Gerais, Rio de Janeiro, Espírito Santo, and São Paulo, where
600 previous spillover events have been recorded (62), suggests that disease control and
601 prevention should continue to focus in those areas. Furthermore, our map highlighting
602 spillover risk in Paraná also coincides with the occurrence of epizootics in the state
603 between 2019 and early 2021 (73). Given that YFV cases among NHPs are recorded
604 through passive surveillance and are thus often based in areas with human
605 settlements, we cannot rule out the fact that varying levels of risk may subsist in non-

spillover zones that are not identified on our map. Despite this, these results do not reflect the risk of YFV outbreaks as population level vaccination coverage is not considered. Even if vaccination coverage is considered, the heterogeneity of current vaccination data by municipality may not accurately portray the coverage gradient across space at the local level. Furthermore, we also did not consider the dynamics of human mobility, which also contribute to outbreak risk. Nonetheless, our maps can help guide public health officials in detecting at-risk municipalities, where further investigation can be conducted by comparing vaccination coverage and effectiveness.

In September 2018, only a handful of at-risk municipalities within the states in our study region reached 95% vaccination coverage; this included 13% and 19% of municipalities in Paraná (13%) São Paulo, respectively (74). However as of May 2020, coverage in the majority of the municipalities in São Paulo, Paraná, Santa Catarina, Espírito Santo, and Minas Gerais was still less than 60% for individuals aged 15 years and older, and less than 80% for individuals aged 14 years and younger (75). Despite improved coverage, the lack of seroprevalence data for YF presents uncertainty in the disease burden experienced by at-risk populations. A recent seroepidemiological study focusing on the Serro region in Minas Gerais found that about 26% of urban and rural populations did not present neutralizing antibodies against YFV, including some that were vaccinated (76). In particular, children were less likely to present YFV-neutralizing antibodies and were less likely to have received a booster dose. Furthermore, there has been a reduction in childhood immunization in Brazil from 2017 to 2019 (77). This highlights potential underestimation of disease burden and the likelihood of new outbreaks originating among vulnerable populations across all ages living in at-risk areas. Moreover, it is uncertain whether current vaccination efforts have been hindered by the co-existence of the COVID-19 pandemic, which may have redirected outbreak surveillance efforts and the allocation of health resources (78).

Our study has a few limitations. It does not account for seasonal patterns of species distribution, which could influence the distribution in environmental suitability of each species throughout the year. Given that the abundance of *Haemagogus* and *Sabethes spp.* is highest between December and May during the rainy season, the risk of transmission and spillover may also vary throughout the year (79). Despite assembling the most comprehensive database to date, our data comes from a few specific

geographic areas and thus may not cover the full range of ecological conditions that influence each species. This could be addressed by limiting predictions to locations outside of model training regions (80) or using a temperature suitability index (81). Due to limited lab and field evidence on how temperature affects the extrinsic incubation period of these species, their gonotrophic cycle, and oviposition behaviour, future entomological surveys could focus on these observations in various locations to update our knowledge which is mainly based on lab evidence presented from a few decades ago (82–84). Similarly, while our pseudo-absence records on Culicidae species are assembled from GBIF only, future works could explore the addition of records from literature. While doing so may not overcome the issue of differential surveillance efforts conducted by various research teams, it could potentially minimise selection bias and improve modelled outputs.

Our finding of restricted spatial distribution in environmental suitability for *Sabethes spp.* concur with previous field observations. *Sabethes spp.* was observed in low abundance, distribution, and infection rates and played a local or secondary role during the 2016-2018 outbreaks (1). Previous studies have noted the competence of *Sa. albiprivus*, *Sa. petrocchiae*, and *Sa. chloropterus* in transmitting YFV (85), however some species are limited to certain ecological conditions. Among the *Sabethes* genera, *Sa. chloropterus* was found mainly in forest areas and while considered the primary YFV sylvatic vector (2), was found with low infection rates during previous outbreaks (1,86).

While beyond the scope of this paper, the role of *Aedes spp.* mosquitoes and its interactions with the sylvatic cycle in facilitating YFV transmission should be further investigated given conflicting findings. In Brazil, evidence of YFV transmission from *Ae. aegypti* and *Ae. albopictus*, which are sensitive to YFV infection, has not been observed since 1942 (1) but other species such as *Ae. scapularis* and *Ae. taeniorhynchus* were found to be naturally infected during the 2016-2018 outbreaks. While these occasional findings do not define their role as YFV vectors, more investigation is needed on their potential vectorial capacity. In addition, considering human mobility in future studies could also help to identify changes in disease spread facilitated by viremic humans, including those that are asymptomatic, which can be applied to infer the geographical shift in risk over time. Furthermore, current

uncertainties related to our knowledge of NHP host species, their viremia, and the spatial distributions of their reservoirs (87) should be further studied. While most NHP exist in sylvatic environments, some are capable of surviving in degraded environments (88). NHP behaviour of various species across Brazil should be closely monitored to identify reservoir establishment and for epizootics to delimit transmission. In this aspect, the promising application of phylogenetic data and genomic analysis could help to identify the roles of NHP species in facilitating transmission, the variation of YFV infection rates across NHP species, and trends in epizootic cases (89).

In conclusion, our findings expand on our limited understanding on the geographic distribution of potential sylvatic vector reservoirs in Brazil. This work contributes to the growing body of research on vector and vector-borne disease mapping at high spatial resolution in low-and middle-income countries that lack comprehensive surveillance coverage. Our identification of suitable regions for main YFV vectors nationally and areas at risk of YFV spillover in the Southeastern region of Brazil could help guide knowledge users and policy makers working in vector surveillance and vaccine allocation.

References

1. de Abreu FVS, Ribeiro IP, Ferreira-de-Brito A, dos Santos AAC, de Miranda RM, Bonelly I de S, et al. Haemagogus leucocelaenus and Haemagogus janthinomys are the primary vectors in the major yellow fever outbreak in Brazil, 2016–2018. Emerg Microbes Infect. 2019 Feb 1;8(1):218–31.

2. Silva NIO, Sacchetto L, de Rezende IM, Trindade G de S, LaBeaud AD, de Thoisy B, et al. Recent sylvatic yellow fever virus transmission in Brazil: the news from an old disease. Virol J. 2020 Dec;17(1):9.

3. Johansson MA, Vasconcelos PFC, Staples JE. The whole iceberg: estimating the incidence of yellow fever virus infection from the number of severe cases. Trans R Soc Trop Med Hyg. 2014 Aug;108(8):482–7.

4. Monath TP, Vasconcelos PFC. Yellow fever. J Clin Virol. 2015 Mar;64:160–73.

5. Benchimol, JL, editor. Febre amarela: a doença e a vacina, uma história inacabada [online]. Rio de Janeiro: Editora FIOCRUZ, 2001. 470 p. online. Rio de Janeiro: Fiocruz; 2001. 470 p.

6. Taylor, RM. Epidemiology. In: George K. Strode, John C. Bugher, editors. Yellow Fever. New York: McGraw-Hill; 1951. p. 427–59.

7. de Oliveira Figueiredo P, Stoffella-Dutra AG, Barbosa Costa G, Silva de Oliveira J, Dourado Amaral C, Duarte Santos J, et al. Re-Emergence of Yellow Fever in Brazil during 2016–2019: Challenges, Lessons Learned, and Perspectives. Viruses [Internet]. 2020 Oct 30 [cited 2021 Mar 14];12(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7692154/

8. Pinheiro GG, Rocha MN, de Oliveira MA, Moreira LA, Andrade Filho JD. Detection of Yellow Fever Virus in Sylvatic Mosquitoes during Disease Outbreaks of 2017–2018 in Minas Gerais State, Brazil. Insects. 2019 May 10;10(5):136.

9. Lourenço-de-Oliveira R, Failloux A-B. High risk for chikungunya virus to initiate an enzootic sylvatic cycle in the tropical Americas. PLoS Negl Trop Dis. 2017 Jun 29;11(6):e0005698.

10. Alencar J, Morone F, De Mello CF, Dégallier N, Lucio PS, Da Serra-Freire NM, et al. Flight Height Preference for Oviposition of Mosquito (Diptera: Culicidae) Vectors of Sylvatic Yellow Fever Virus Near the Hydroelectric Reservoir of Simplício, Minas Gerais, Brazil. J Med Entomol. 2013 Jul 1;50(4):791–5.

11. Pinto CS, Confalonieri UEC, Mascarenhas BM. Ecology of Haemagogus sp. and Sabethes sp. (Diptera: Culicidae) in relation to the microclimates of the Caxiuanã National Forest, Pará, Brazil. Mem Inst Oswaldo Cruz. 2009 Jul;104(4):592–8.

12. Warren, AJ. Landmarks in the conquest of yellow fever. In: Strode GK, Bugher JC, editors. Yellow fever. New York: McGraw-Hill; 1951. p. 5–37.

13. Elith J, Leathwick JR. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu Rev Ecol Evol Syst. 2009 Dec;40(1):677–97.

14. Kraemer MUG, Sinka ME, Duda KA, Mylne AQN, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. eLife. 2015 Jun 30;4:e08347.

15. Childs ML, Nova N, Colvin J, Mordecai EA. Mosquito and primate ecology predict human risk of yellow fever virus spillover in Brazil. Philos Trans R Soc B Biol Sci. 2019 Sep 30;374(1782):20180335.

16. Almeida MAB, Santos E dos, Cardoso J da C, Silva LG da, Rabelo RM, Bicca-Marques JC. Predicting Yellow Fever Through Species Distribution Modeling of Virus, Vector, and Monkeys. EcoHealth. 2019 Mar 1;16(1):95–108.

17. Culot L, Pereira LA, Agostini I, Almeida MAB de, Alves RSC, Aximoff I, et al. ATLANTIC-PRIMATES: a dataset of communities and occurrences of primates in the Atlantic Forests of South America. Ecology. 2019;100(1):e02525.

18. Laan MJ van der, Polley EC, Hubbard AE. Super Learner. Stat Appl Genet Mol Biol [Internet]. 2007 Sep 16 [cited 2021 Jul 14];6(1). Available from: https://www.degruyter.com/document/doi/10.2202/1544-6115.1309/html

19. Bhatt S, Cameron E, Flaxman SR, Weiss DJ, Smith DL, Gething PW. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. J R Soc Interface. 2017 Sep 30;14(134):20170520.

20. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; 2009. 745 p.

21. Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl. 2009;19(1):181–97.

22. Messina JP, Brady OJ, Pigott DM, Brownstein JS, Hoen AG, Hay SI. A global compendium of human dengue virus occurrence. Sci Data. 2014 May 27;1(1):140004.

23. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013 Apr;496(7446):504–7.

24. Hijmans RJ, van Etten J. raster: Geographic data analysis and modeling. R Package Version. 2016;2(8).

25. Alencar J, Mello CF de, Morone F, Albuquerque HG, Serra-Freire NM, Gleiser RM, et al. Distribution of Haemagogus and Sabethes Species in Relation to Forest Cover and Climatic Factors in the Chapada Dos Guimarães National Park, State of Mato Grosso, Brazil. J Am Mosq Control Assoc. 2018 Jun 1;34(2):85–92.

26. de Araújo WS, Vieira TM, de Souza GA, Bezerra IC, Corgosinho PHC, Borges MAZ. Nocturnal Mosquitoes of Pará State in the Brazilian Amazon: Species Composition, Habitat Segregation, and Seasonal Variation. J Med Entomol [Internet]. 2020 Jun 2 [cited 2020 Oct 2]; Available from: https://academic.oup.com/jme/advance-article/doi/10.1093/jme/tjaa103/5850337

27. La Corte R, Maia PCR, Dolabella SS, Cruz DER, Marteis LS. Mosquitoes of the Caatinga. III. Larval Habitats, Frequency, and Dynamics of Immature and Adult Stages in a Dry Brazilian Forest. J Med Entomol. 2019 08;56(1):120–8.

28. Marteis LS, Natal D, Sallum MAM, Medeiros-Sousa AR, Oliveira TMP, La Corte R. Mosquitoes of the Caatinga: 1. Adults stage survey and the emerge of seven news species endemic of a dry tropical forest in Brazil. Acta Trop. 2017 Feb 1;166:193–201.

29. Santos EB, Favretto MA, Müller GA. When and what time? On the seasonal and daily patterns of mosquitoes (Diptera: Culicidae) in an Atlantic Forest remnant from Southern Brazil. Austral Entomol. 2020;59(2):337–44.

30. Cardoso J da C, de Almeida MAB, dos Santos E, da Fonseca DF, Sallum MAM, Noll CA, et al. Yellow Fever Virus in Haemagogus leucocelaenus and Aedes serratus Mosquitoes, Southern Brazil, 2008. Emerg Infect Dis. 2010 Dec;16(12):1918–24.

31. Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, Wint GRW, et al. Global Data for Ecology and Epidemiology: A Novel Algorithm for Temporal Fourier Processing MODIS Data. PLoS ONE [Internet]. 2008 Jan 9 [cited 2021 Jan 26];3(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2171368/

32. Linacre ET. A simple formula for estimating evaporation rates in various climates, using temperature data alone. Agric Meteorol. 1977 Dec 1;18(6):409–24.

33. Forattini OP. Entomologia médica: 4º volume: psychodidae. phlebotominae. leishmanioses. bartonelose. Entomol Médica 4º Vol Psychodidae Phlebotominae Leishmanioses Bartonelose. 1973;658–658.

34. Silva-Inacio CL, Paiva AAP de, Araújo JMG de, Ximenes M de FF de M, Silva-Inacio CL, Paiva AAP de, et al. Ecological relationships of Haemagogus spegazzinii (Diptera: Culicidae) in a semiarid area of Brazil. Rev Soc Bras Med Trop [Internet]. 2020 [cited 2021 Jan 26];53. Available from: http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0037-86822020000100384&lng=en&nrm=iso&tlng=en

35. Tátila-Ferreira A, Maia D de A, de Abreu FVS, Rodrigues WC, Alencar J. Oviposition behavior of Haemagogus leucocelaenus (Diptera: culicidae), a vector of wild yellow fever in Brazil. Rev Inst Med Trop São Paulo [Internet]. 2017 Aug 7 [cited 2021 Jan 29];59. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553946/

36. Mucci LF, Júnior RPC, de Paula MB, Scandar SAS, Pacchioni ML, Fernandes A, et al. Feeding habits of mosquitoes (Diptera: Culicidae) in an area of sylvatic transmission of yellow fever in the state of São Paulo, Brazil. J Venom Anim Toxins Trop Dis. 2015 Mar 20;21(1):6.

37. Tuanmu M-N, Jetz W. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Glob Ecol Biogeogr. 2014;23(9):1031–45.

38. Hamlet A, Ramos DG, Gaythorpe KAM, Romano APM, Garske T, Ferguson NM. Seasonality of agricultural exposure as an important predictor of seasonal yellow fever spillover in Brazil. Nat Commun. 2021 Jun 15;12(1):3647.

39. Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science. 2013 Nov 15;342(6160):850–3.

40. Alencar J, Mello CF de, Gil-Santana HR, Guimarães AÉ, Almeida SAS de, Gleiser RM. Vertical oviposition activity of mosquitoes in the Atlantic Forest of Brazil with emphasis on the sylvan vector, Haemagogus leucocelaenus (Diptera: Culicidae). J Vector Ecol. 2016;41(1):18–26.

41. Hendy A, Hernandez-Acosta E, Valério D, Mendonça C, Costa ER, Júnior JTA, et al. The vertical stratification of potential bridge vectors of mosquito-borne viruses in a central Amazonian forest bordering Manaus, Brazil. Sci Rep. 2020 Oct 26;10(1):18254.

42. Amatulli G, Domisch S, Tuanmu M-N, Parmentier B, Ranipeta A, Malczyk J, et al. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci Data. 2018 Mar 20;5(1):180040.

43. Possas C, Lourenço-de-Oliveira R, Tauil PL, Pinheiro F de P, Pissinatti A, da Cunha RV, et al. Yellow fever outbreak in Brazil: the puzzle of rapid viral spread and challenges for immunisation. Mem Inst Oswaldo Cruz [Internet]. 2018 Sep 3 [cited 2021 Jan 17];113(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6135548/

44. Lloyd CT, Chamberlain H, Kerr D, Yetman G, Pistolesi L, Stevens FR, et al. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. Big Earth Data. 2019 Apr 3;3(2):108–39.

45. Hamlet, A. Yellow fever in South America: The role of environment and host on transmission dynamics [Unpublished dissertation]. [London]: Imperial Colelge London; 2020.

46. Dietz JM, Hankerson SJ, Alexandre BR, Henry MD, Martins AF, Ferraz LP, et al. Yellow fever in Brazil threatens successful recovery of endangered golden lion tamarins. Sci Rep. 2019 Sep 10;9(1):12926.

47. Sacchetto L, Silva NIO, Rezende IM de, Arruda MS, Costa TA, de Mello ÉM, et al. Neighbor danger: Yellow fever virus epizootics in urban and urban-rural transition areas of Minas Gerais state, during 2017-2018 yellow fever outbreaks in Brazil. PLoS Negl Trop Dis. 2020 Oct;14(10):e0008658.

48. Almeida MAB, Cardoso J da C, dos Santos E, da Fonseca DF, Cruz LL, Faraco FJC, et al. Surveillance for Yellow Fever Virus in Non-Human Primates in Southern Brazil, 2001–2011: A Tool for Prioritizing Human Populations for Vaccination. Gubler DJ, editor. PLoS Negl Trop Dis. 2014 Mar 13;8(3):e2741.

49. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.

50. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees. Boca Raton: Routledge; 1984. 368 p.

51. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008;77(4):802–13.

52. Longbottom J, Browne AJ, Pigott DM, Sinka ME, Golding N, Hay SI, et al. Mapping the spatial distribution of the Japanese encephalitis vector, Culex tritaeniorhynchus Giles, 1901 (Diptera: Culicidae) within areas of Japanese encephalitis risk. Parasit Vectors. 2017 Mar 16;10(1):148.

53. Messina JP, Kraemer MU, Brady OJ, Pigott DM, Shearer FM, Weiss DJ, et al. Mapping global environmental suitability for Zika virus. eLife. 2016 19;5.

54. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. Nat Microbiol. 2019 Sep;4(9):1508–15.

55. Kraemer MUG, Reiner RC, Brady OJ, Messina JP, Gilbert M, Pigott DM, et al. Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. Nat Microbiol. 2019 May;4(5):854–63.

56. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016 Aug 13;785–94.

57. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinforma Oxf Engl. 2005 Oct 15;21(20):3940–1.

58. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12(1):77.

59. Breiman L. Random Forests. Mach Learn. 2001 Oct 1;45(1):5–32.

60. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. ArXiv180101489 Stat [Internet]. 2019 Dec 23 [cited 2021 Jul 16]; Available from: http://arxiv.org/abs/1801.01489

887 61. Greenwell BM, Boehmke BC. Variable Importance Plots—An Introduction to the
888 vip Package. R J. 2020;12(1):343–66.

889 62. Hill SC, Souza R de, Thézé J, Claro I, Aguiar RS, Abade L, et al. Genomic
890 Surveillance of Yellow Fever Virus Epizootic in São Paulo, Brazil, 2016 – 2018.
891 PLOS Pathog. 2020 Aug 7;16(8):e1008699.

892 63. Giovanetti M, Mendonça MCL de, Fonseca V, Mares-Guia MA, Fabri A, Xavier
893 J, et al. Yellow Fever Virus Reemergence and Spread in Southeast Brazil,
894 2016–2019. J Virol [Internet]. 2019 Dec 12 [cited 2021 Mar 13];94(1). Available
895 from: https://jvi.asm.org/content/94/1/e01623-19

896 64. Andrade MS, Campos FS, Campos A a. S, Abreu FVS, Melo FL, Cardoso JC, et
897 al. Real-time genomic surveillance during the 2021 re-emergence of the yellow
898 fever virus in Rio Grande do Sul State, Brazil [Internet]. 2021 Jul [cited 2021
899 Aug 16] p. 2021.07.16.452727. Available from:
900 https://www.biorxiv.org/content/10.1101/2021.07.16.452727v1

901 65. Ministério da Saúde. Saúde Brasil 2019: 14 - Reemergência e manutenção
902 extra-Amazônica da febre amarela no Brasil, 2014 a 2019: principais desafios
903 para a vigilância, a prevenção e o controle. Brasília – DF; 2019 p. 305–27.

904 66. de Rezende IM, Sacchetto L, Munhoz de Mello É, Alves PA, Iani FC de M,
905 Adelino TÉR, et al. Persistence of Yellow fever virus outside the Amazon Basin,
906 causing epidemics in Southeast Brazil, from 2016 to 2018. PLoS Negl Trop Dis
907 [Internet]. 2018 Jun 4 [cited 2021 Mar 14];12(6). Available from:
908 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6002110/

909 67. Hamrick PN, Aldighieri S, Machado G, Leonel DG, Vilca LM, Uriona S, et al.
910 Geographic patterns and environmental factors associated with human yellow
911 fever presence in the Americas. Barker CM, editor. PLoS Negl Trop Dis. 2017
912 Sep 8;11(9):e0005897.

913 68. Couto-Lima D, Andreazzi CS, Leite PJ, Bersot MIL, Alencar J, Lourenço-de-
914 Oliveira R. Seasonal population dynamics of the primary yellow fever vector
915 Haemagogus leucocelaenus (Dyar & Shannon) (Diptera: Culicidae) is mainly
916 influenced by temperature in the Atlantic Forest, southeast Brazil. Mem Inst
917 Oswaldo Cruz [Internet]. 2020 Jul 20 [cited 2020 Sep 30];115. Available from:
918 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7370926/

919 69. Camargo-Neves VLF de, Poletto DW, Rodas LAC, Pachioli ML, Cardoso RP,
920 Scandar SAS, et al. Entomological investigation of a sylvatic yellow fever area
921 in São Paulo State, Brazil. Cad Saúde Pública. 2005 Aug;21:1278–86.

922 70. Montagner FRG, Silva OS, Jahnke SM. Mosquito species occurrence in
923 association with landscape composition in green urban areas. Braz J Biol. 2018
924 May;78(2):233–9.

925 71. Burkett-Cadena ND, Vittor AY. Deforestation and vector-borne disease: Forest
926 conversion favors important mosquito vectors of human pathogens. Basic Appl
927 Ecol. 2018 Feb;26:101–10.

72. Barros FSM, Honório NA. Deforestation and Malaria on the Amazon Frontier: Larval Clustering of Anopheles darlingi (Diptera: Culicidae) Determines Focal Distribution of Malaria. Am J Trop Med Hyg. 2015 Nov 4;93(5):939–53.

73. Coordenação-Geral do Programa Nacional de Imunizações do Departamento de Imunização e Doenças Transmissíveis (CGPNI/DEIDT/SVS), Coordenação-Geral de Laboratórios de Saúde Pública do Departamento de Articulação Estratégica de Vigilância em Saúde (CGLAB/DAEVS/SVS). Boletim Epidemiológico: Informe quinzenal sarampo – Brasil, semanas epidemiológicas 43 de 2020 a 1 de 2021 [Internet]. Brasília - DF: Ministério da Saúde: Secretaria de Vigilância em Saúde; 2021 Feb. Available from: https://www.gov.br/saude/pt-br/media/pdf/2021/fevereiro/11/boletim_epidemiologico_svs_4.pdf

74. WHO. Yellow fever – Brazil [Internet]. WHO. World Health Organization; 2019 [cited 2021 Mar 15]. Available from: http://www.who.int/csr/don/11-february-2019-yellow-fever-brazil/en/

75. Coordenação-Geral de Vigilância de Agravos e Doenças Não Transmissíveis (CGDANT/DASNT/SVS). Boletim Epidemiológico: Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes Aegypti (dengue, chikungunya e zika), Semanas Epidemiológicas 1 a 19, 2020 [Internet]. Brasília/DF: Ministério da Saúde: Secretaria de Vigilância em Saúde; 2020 May. Available from: https://antigo.saude.gov.br/images/pdf/2020/May/20/Boletim-epidemiologico-SVS-20-aa.pdf

76. Stoffella-Dutra AG, Silva de Oliveira J, Barbosa Costa G, Geessien Kroon E, Santos Abrahão J, Desiree LaBeaud A, et al. Absence of YF-neutralizing antibodies in vulnerable populations of Brazil: A warning for epidemiological surveillance and the potential risks for future outbreaks. Vaccine. 2020 Sep 29;38(42):6592–9.

77. Césare N, Mota TF, Lopes FFL, Lima ACM, Luzardo R, Quintanilha LF, et al. Longitudinal profiling of the vaccination coverage in Brazil reveals a recent change in the patterns hallmarked by differential reduction across regions. Int J Infect Dis. 2020 Sep 1;98:275–80.

78. Ribeiro VST, Telles JP, Tuon FF. Arboviral diseases and COVID-19 in Brazil: Concerns regarding climatic, sanitation, and endemic scenario. J Med Virol. 2020;92(11):2390–1.

79. Cavalcante KRLJ, Tauil PL. Epidemiological characteristics of yellow fever in Brazil, 2000-2012. Epidemiol E Serv Saude Rev Sist Unico Saude Bras. 2016 Mar;25(1):11–20.

80. Shearer FM, Longbottom J, Browne AJ, Pigott DM, Brady OJ, Kraemer MUG, et al. Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis. Lancet Glob Health. 2018 Mar 1;6(3):e270–8.

81. Brady OJ, Golding N, Pigott DM, Kraemer MUG, Messina JP, Reiner Jr RC, et al. Global temperature constraints on Aedes aegypti and Ae. albopictus

persistence and competence for dengue virus transmission. Parasit Vectors. 2014 Jul 22;7(1):338.

82. Dégallier N, SÁ Filho GC, Monteiro HAO, Castro FC, Vaz Da Silva O, Brandão RCF, et al. Release–Recapture Experiments with Canopy Mosquitoes in the Genera Haemagogus and Sabeihes (Diptera: Culicidae) in Brazilian Amazonia. J Med Entomol. 1998 Nov 1;35(6):931–6.

83. Galindo P. Bionomics of Sabethes chloropterus Humboldt, a vector of sylvan yellow fever in Middle America. Am J Trop Med Hyg. 1958 Jul;7(4):429–40.

84. Mondet B. [Conditions for the laboratory survival of Haemagogus janthinomys dyar, 1921 (Diptera:Culicidae)]. Rev Soc Bras Med Trop. 1997 Feb;30(1):11–4.

85. Couto-Lima D, Madec Y, Bersot MI, Campos SS, Motta M de A, Santos FB dos, et al. Potential risk of re-emergence of urban transmission of Yellow Fever virus in Brazil facilitated by competent Aedes populations. Sci Rep [Internet]. 2017 Dec [cited 2019 Feb 19];7(1). Available from: http://www.nature.com/articles/s41598-017-05186-3

86. Mascheretti M, Tengan CH, Sato HK, Suzuki A, Souza RP de, Maeda M, et al. Febre amarela silvestre: reemergência de transmissão no estado de São Paulo, Brasil, 2009. Rev Saúde Pública. 2013 Oct;47(5):881–9.

87. Mares-Guia MAM de M, Horta MA, Romano A, Rodrigues CDS, Mendonça MCL, dos Santos CC, et al. Yellow fever epizootics in non-human primates, Southeast and Northeast Brazil (2017 and 2018). Parasit Vectors. 2020 Feb 19;13(1):90.

88. Teixeira B, Hirsch A, Goulart VDLR, Passos L, Teixeira CP, James P, et al. Good neighbours: distribution of black-tufted marmoset (Callithrix penicillata) in an urban environment. Wildl Res. 2015 Jun;42(7):579–89.

89. Jesus JG de, Gräf T, Giovanetti M, Mares-Guia MA, Xavier J, Maia ML, et al. Yellow fever transmission in non-human primates, Bahia, Northeastern Brazil. PLoS Negl Trop Dis. 2020 Aug 11;14(8):e0008405.

4 Differential risk to COVID-19 in the built environment

The health of populations is influenced by human-made environments in which they live, including the places where they were born, live, work, and socialise. The distribution of resources across built environments can modify human behaviour, health, and risk of contracting a disease. Brazil's extreme inequalities have resulted in health outcomes that differ among racial and socioeconomic groups in urban areas. Therefore, to holistically assess population risk to COVID-19 in Brazil, it is essential to understand geographical disparities in access to COVID-19 diagnosis and vulnerability to hospitalisation and death between populations. This chapter discusses differential risk to COVID-19 during the beginning and later stages of the epidemic in Brazil, focusing on the Metropolitan Region of São Paulo (MRSP) and its overarching administrative division, São Paulo state, which is currently the most populous sub-national division in the Americas. The first part of this chapter explores differential access to COVID-19 testing by socioeconomic status (4.1), while the latter part of the chapter explores differential vulnerability to hospitalisation and death by disadvantaged populations (4.2).

4.1 Differential access to diagnosis in the Metropolitan Region of São Paulo

4.1.1 Introduction

Since the detection of the first COVID-19 case in Brazil on 25 February 2020, the Brazilian Ministry of Health has implemented a national COVID-19 reporting system as part of the centralised Sistema de Informação de Vigilância Epidemiológica da Gripe system (SIVEP-Gripe). Cases of laboratory confirmed COVID-19 and severe acute respiratory illnesses (SARI) of unknown aetiology are reported by public and private health services, including primary and emergency care, on this platform. Daily aggregated counts are currently

summarised on Portal do COVID-19 website (https://covid.saude.gov.br/). Until 31 May 2020, a total of 73,648 COVID-19 confirmed cases and 168,001 SARI cases with unknown aetiology were reported in the SIVEP-Gripe system. We hypothesized that the 2.3-fold increase of SARI cases with unknown aetiology was associated with COVID-19 and differential access to healthcare due to socioeconomic factors.

This section explores the geographical risk of COVID-19 infection at the beginning of the epidemic in Brazil, and elucidates important information on spatial inequality in access to COVID-19 testing in the Metropolitan Region of São Paulo (MRSP). The following sections are excerpts of my contributions taken from the paper published in Nature Human Behaviour:

de Souza, W. M., Buss, L. F., da Silva Candido, D., Carrera, J. P., **Li, S.,*** Zarebski, A. E., ... & Faria, N. R. (2020). Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nature Human Behaviour, 4(8), 856-865.

4.1.2 Data and methods

The average household per-capita income for the MRSP was retrieved at the census tract level from the 2010 census (https://censo2010.ibge.gov.br/). We geocoded 24,063 COVID-19 cases and 32,914 SARI cases with unknown aetiology from MRSP, which were reported until 28 May 2020. The geocoding was based on self-reported residential addresses or postal codes using the Galileo algorithm(1) and coordinates were confirmed using Google API.

To elucidate the distribution of COVID-19 cases and SARI cases with unknown aetiology, we mapped the mean relative risk of COVID-19 and SARI with unknown aetiology at the census tract level for MRSP for three epidemiological weeks (12, 16 and 21) (Figure 4.1.1).

As the observation process was a confounding process and without additional assumptions (for example, covariates), we cannot disentangle an increase in prevalence from an increase in case ascertainment. The cumulative number of cases in each tract was modelled as a Poisson random variable with a mean specified by the expected number of cases under a null model adjusted by tract specific risk due to spatial and non-spatial effects: the Besag–York–Mollié model (2). Estimates of the risk of COVID-19 diagnosis or SARI cases with unknown aetiology were obtained using approximate Bayesian methods (integrated nested Laplace approximation). A complete specification of the model and the computational methodology can be found in the Chapter 4.2 Supplementary Information.

The association between final diagnostic category (COVID-19 or SARI with unknown aetiology) and socioeconomic status in the subset of cases in the MRSP with geocoded residential information was evaluated using logistic regression models. We focused on the cases in epidemiological weeks 12, 16 and 22. Within each of those weeks, if a census tract reported any COVID-19 or SARI with unknown aetiology, we calculated the proportion of the number of COVID-19 cases. Since most census tracts reported only one case each week, the proportion of COVID-19 cases for each census tract was mostly either 0 or 1 in a given week. For this reason, we defined two categories: (1) the census tract only reported SARI of unknown aetiology (that is, no COVID-19 cases); or (2) the census tract reported at least one COVID-19 case during the week. We used these two categories as the binary response, and applied logistic regression models to investigate whether income per capita was associated with this response. The analyses were adjusted by the logarithm of the population sizes and the longitude and latitude coordinates of the census tracts. The analyses were performed individually for each of epidemiological weeks 12, 16 and 22. Further analysis details can be found in Chapter 4.1 Supplementary Information.

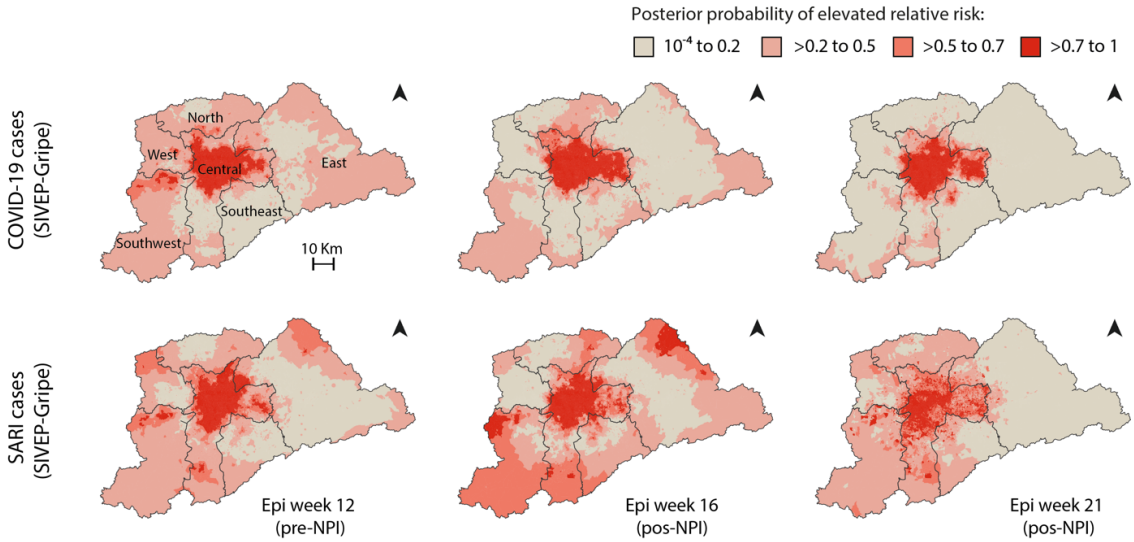


Figure 4.1.1. Posterior probability of elevated relative risk of COVID-19 for confirmed diagnosis (upper panels) and SARI cases with unknown aetiology (lower panels) for epidemiological weeks 12 (pre-implementation of non-pharmaceutical interventions in São Paulo state, and weeks 16 and 21 (post-implementation of non-pharmaceutical interventions in São Paulo state).

4.1.3 Results

We focused on the Metropolitan Region of São Paulo (MRSP), which has a population of 23 million inhabitants across six sub-regions (Central, West, North, East, Southeast and Southwest) and 39 municipalities (Figure 4.1.2A). To test this hypothesis, we obtained per-capita income at the census tract level (typically 150–300 households) in the MRSP, based on the residential address of each case. We then linked this information to each patient's final diagnosis outcome: confirmed case of COVID-19 or SARI with unknown aetiology. While the income distribution of SARI cases with unknown aetiology was similar to that of all residents of the MRSP over the whole period (Figure 4.1.2B), we observed that the income distribution of individuals with COVID-19 confirmed by laboratory and clinical criteria was initially higher than that of all MRSP residents and decreased over time towards similar levels by epidemiological week 21 (Figure 4.1.2B). Importantly, we found that the log odds of one or more confirmed COVID-19 cases per census tract increased with per-capita income in epidemiological weeks 12 and 22 (likelihood ratio test P value < 0.001; Figure 4.1.2B and Chapter 4.1 Supplementary Information Table S1). This provides statistical

evidence of an association between confirmed COVID-19 diagnosis and per-capita income, suggesting a socioeconomic difference in access to COVID-19 diagnosis in the MRSP. For reference, we also provide a map of per-capita income (Figure 4.1.2A) and population density in each census tract (Figure 4.1.3).

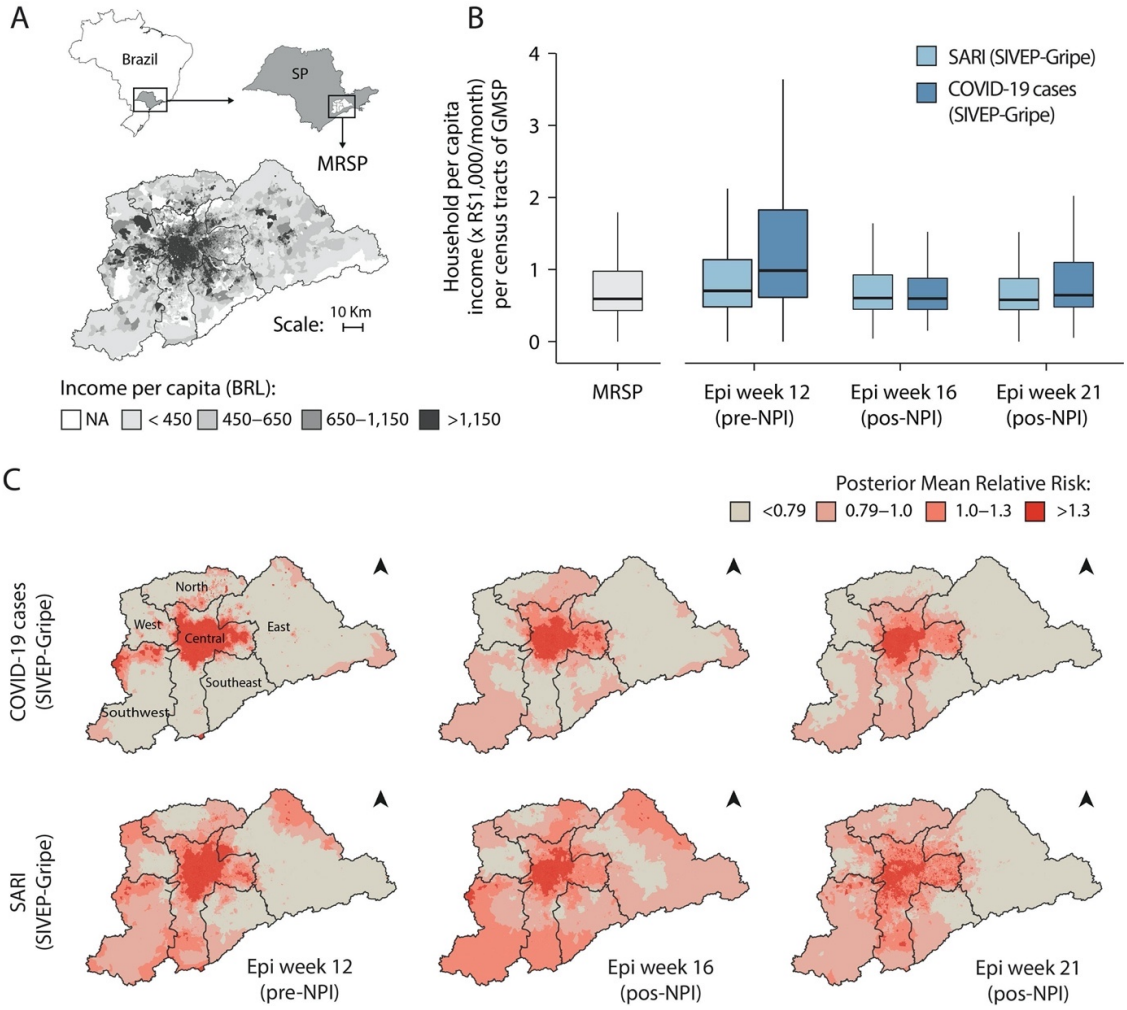


Figure 4.1.2. A, Spatial distribution of income per capita of MRSP based on the census tract of residence. BRL, Brazilian reais. NA, not applicable. **B,** Distribution of household per-capita income based on the census tract of residence for COVID-19 cases and SARI cases with unknown aetiology. The distribution of average per-capita income for MRSP as a whole, weighted by population size, is shown on the left. Per-capita income distributions are presented in box plots, where the horizontal line inside the box represents the median per-capita income level, and the box edges show the per-capita income within the first and third interquartile range. The whiskers represent the per-capita income range. Epi, epidemiological. **C,** Posterior mean relative risk of COVID-19 confirmed diagnosis (top) and SARI cases with unknown aetiology (bottom) for epidemiological week 12 (before implementation of NPI in São Paulo state) and weeks 16 and 21 (after implementation of NPI in São Paulo state) (see Data and Methods for details).

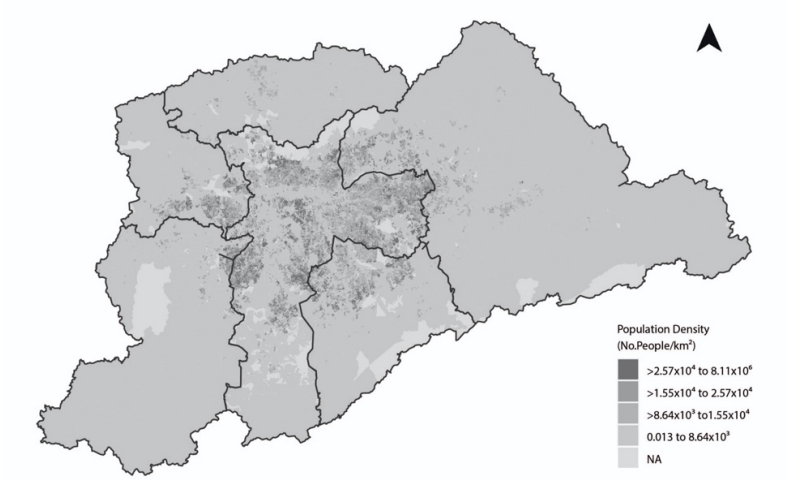


Figure 4.1.3. Map of population density in each census tract in the Metropolitan Region of São Paulo.

We conducted a geospatial analysis to understand the distribution of relative risk of observing a COVID-19 case or SARI case with unknown aetiology in the MRSP, using a Bayesian method and adjusted for spatial and non-spatial effects as defined by the Besag–York–Mollié model (2) (Figure 4.1.2). Our estimates show an increase in the relative risk of COVID-19 diagnosis in higher-income census tracts between epidemiological weeks 12 and 21, especially in the central region of the MRSP (Figure 4.1.2A, C). We observed a similar trend in the relative risk of SARI cases with unknown aetiology among residents of the central region. However, there was also an increased probability of SARI cases with unknown aetiology in the southwest, west, north and south sub-regions, where income per capita is typically lower. Overall, the relative risk of SARI cases with unknown aetiology is more spatially widespread in the MRSP than that of confirmed COVID-19 cases (Figure 4.1.2C).

The relative risk of SARI cases with unknown aetiology compared with confirmed COVID-19 cases in the central region of the MRSP decreased through time, probably as a response to several NPIs implemented throughout the state of São Paulo (see Chapter 4.1 Supplementary Information Table S1). By week 16 (1 month after the start of the NPIs in

São Paulo), we detected an increased risk particularly of SARI cases with unknown aetiology outside the central region of the MRSP, especially in the southwest region. SARI cases with unknown aetiology risk were also high in the east region. By week 21, the risk remained high throughout the central region and the risk of SARI cases with unknown aetiology decreased in the east region, possibly as a result of interventions targeting the reduction of SARS-CoV-2 transmission.

4.1.4 Discussion

Our data uncover a socioeconomic bias in testing and diagnostics in current surveillance guidelines and suggest that the number of reported confirmed case counts may substantially underestimate the number of cases in the general population, particularly in regions of lower socioeconomic status. Socioeconomic differences are associated with access to healthcare (3) and should be taken into account when designing targeted interventions. We found that the proportion of SARI cases with unknown aetiology versus confirmed COVID-19 cases has increased across the entire country (as of 15 June 2020, the number of reported SARI cases with unknown aetiology was nearly twofold greater than the number of confirmed COVID-19 cases). Based on clinical and epidemiological grounds, it is likely that many SARI cases with unknown aetiology are caused by SARS-CoV-2. In order to rigorously establish the contribution of non-SARS-CoV-2 infections to the SARI cases, we would need additional denominator data to understand the level of testing for these viruses (that is, the negative test results). Our findings with regards to socioeconomic bias are likely to apply to other states and regions of Brazil and highlight the importance of scaling up surveillance and laboratory capacity within Latin America. Indeed, the largest Brazilian serosurvey conducted to date suggests that undetected cases may be seven times higher than reported cases (4).

We further show that SARI cases with unknown aetiology are associated with lower socioeconomic status in the MRSP. The socioeconomic disparities observed here were particularly evident at the beginning of the outbreak (Figure 4.1.2B). This can be explained in part by: (1) the high proportion of early cases in returning travellers with higher income and better access to private laboratories for diagnostics; and (2) the more limited access to freely available diagnostic screening. For example, between 25 February and 18 March 2020, two-thirds (586 (66.9%) of 876) of diagnostic tests were performed in private medical laboratories where costs varied typically between 300 and 690 Brazilian Reais (for context, the current minimum monthly salary is 1,045 Brazilian Reais). Thus, the true burden of the epidemic in lower-income neighbourhoods is probably underestimated. In New York City, for example, poorer neighbourhoods have been found to have a higher disease burden, which is driven in part by the movement of essential workers using public transport during the pandemic. Data-driven analyses are urgently needed to help tackle health inequities during the ongoing epidemic in Brazil. Strategies to evaluate and control transmission should consider differential assess to COVID-19 diagnosis for lower-income populations, changes in reporting systems and delays in reporting, which are key to accurately determining rates of epidemic growth (5). Innovative infectious disease surveillance approaches such as those obtained from aggregated mobility data, when used properly, could help support public health actions across the COVID-19 epidemic (6–9).

4.1.5 References

1. Medel CH, Catalán CC, Fernández Vidou MA, Sardón Pérez E. The Galileo Ground Segment Integrity Algorithms: Design and Performance. International Journal of Navigation and Observation. 2008 Jul 13;2008:e178927.
2. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math. 1991 Mar 1;43(1):1–20.
3. Pereira RHM, Braga CKV, Servo LM, Serra B, Amaral P, Gouveia N. Mobilidade urbana e o acesso ao Sistema Único de Saúde para casos suspeitos e graves de COVID-19 nas 20 maiores cidades do Brasil [Internet]. Instituto de Pesquisa Econômica Aplicada (Ipea); 2020. Report No.: 14. Available from:

https://www.ipea.gov.br/portal/index.php?option=com_content&view=article&id=35442&Itemid=7
4. Silveira MF, Barros AJD, Horta BL, Pellanda LC, Victora GD, Dellagostin OA, et al. Population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil. Nature Medicine. 2020 Aug;26(8):1196–9.
5. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. Science [Internet]. 2020 Jul 10 [cited 2021 Mar 28];369(6500). Available from: https://science.sciencemag.org/content/369/6500/eabb9789
6. Buckee CO, Balsari S, Chan J, Crosas M, Dominici F, Gasser U, et al. Aggregated mobility data could help fight COVID-19. Science. 2020 Apr 10;368(6487):145–6.
7. Kraemer MUG, Yang C-H, Gutierrez B, Wu C-H, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science. 2020 May 1;368(6490):493–7.
8. Nouvellet P, Bhatia S, Cori A, Ainslie KEC, Baguelin M, Bhatt S, et al. Reduction in mobility and COVID-19 transmission. Nature Communications. 2021 Feb 17;12(1):1090.
9. Oliveira SB de, Pôrto VBG, Ganem F, Mendes FM, Almiron M, Oliveira WK de, et al. Monitoring social distancing and SARS-CoV-2 transmission in Brazil using cell phone mobility data. medRxiv. 2020 May 5;2020.04.30.20082172.

4.2 Differential vulnerability to hospitalisation and death in São Paulo state

The following paper, which has been accepted for publication in BMJ Global Health, introduces and explores the geographical risk in COVID-19 hospitalisation and death by racial and socioeconomic groups in the state of São Paulo. It will discuss uneven exposure to COVID-19 experienced by disadvantaged populations and its implications for shaping health inequalities.

BMJ Global Health

Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil

Sabrina L Li ,[1] Rafael H M Pereira ,[2] Carlos A Prete Jr ,[3] Alexander E Zarebski ,[4] Lucas Emanuel ,[2] Pedro J H Alves ,[2] Pedro S Peixoto ,[5] Carlos K V Braga ,[2] Andreza Aruska de Souza Santos ,[6] William M de Souza ,[4,7] Rogerio J Barbosa ,[8] Lewis F Buss ,[9] Alfredo Mendrone ,[10] Cesar de Almeida-Neto ,[10,11] Suzete C Ferreira ,[10,12] Nanci A Salles ,[10,12] Izabel Marcilio ,[13] Chieh-Hsi Wu ,[14] Nelson Gouveia ,[15] Vitor H Nascimento ,[3] Ester C Sabino ,[9] Nuno R Faria ,[4,9,16] Jane P Messina [1,17]

To cite: Li SL, Pereira RHM, Prete Jr CA, *et al*. Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil. *BMJ Global Health* 2021;**6**:e004959. doi:10.1136/bmjgh-2021-004959

Handling editor Seye Abimbola

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjgh-2021-004959).

SLL, RHMP and CAPJ contributed equally.

Received 6 January 2021
Revised 8 March 2021
Accepted 6 April 2021

Check for updates

© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Sabrina L Li; lisabrinaly@gmail.com and Dr Rafael H M Pereira; rafael.pereira@ipea.gov.br

ABSTRACT

Introduction Little evidence exists on the differential health effects of COVID-19 on disadvantaged population groups. Here we characterise the differential risk of hospitalisation and death in São Paulo state, Brazil, and show how vulnerability to COVID-19 is shaped by socioeconomic inequalities.

Methods We conducted a cross-sectional study using hospitalised severe acute respiratory infections notified from March to August 2020 in the *Sistema de Monitoramento Inteligente de São Paulo* database. We examined the risk of hospitalisation and death by race and socioeconomic status using multiple data sets for individual-level and spatiotemporal analyses. We explained these inequalities according to differences in daily mobility from mobile phone data, teleworking behaviour and comorbidities.

Results Throughout the study period, patients living in the 40% poorest areas were more likely to die when compared with patients living in the 5% wealthiest areas (OR: 1.60, 95% CI 1.48 to 1.74) and were more likely to be hospitalised between April and July 2020 (OR: 1.08, 95% CI 1.04 to 1.12). Black and *Pardo* individuals were more likely to be hospitalised when compared with White individuals (OR: 1.41, 95% CI 1.37 to 1.46; OR: 1.26, 95% CI 1.23 to 1.28, respectively), and were more likely to die (OR: 1.13, 95% CI 1.07 to 1.19; 1.07, 95% CI 1.04 to 1.10, respectively) between April and July 2020. Once hospitalised, patients treated in public hospitals were more likely to die than patients in private hospitals (OR: 1.40%, 95% CI 1.34% to 1.46%). Black individuals and those with low education attainment were more likely to have one or more comorbidities, respectively (OR: 1.29, 95% CI 1.19 to 1.39; 1.36, 95% CI 1.27 to 1.45).

Conclusions Low-income and Black and *Pardo* communities are more likely to die with COVID-19. This is associated with differential access to quality healthcare, ability to self-isolate and the higher prevalence of comorbidities.

Key questions

What is already known?

- Black and *Pardo* (mixed ethnicity) Brazilians face higher risk of COVID-19 hospitalised death.
- Access to COVID-19 testing has been limited for low-income populations in São Paulo city.

What are the new findings?

- Individual and population-level risk of COVID-19 hospitalisation, death and adherence to non-pharmaceutical interventions vary by race and socioeconomic status.
- Low socioeconomic and/or Black and *Pardo* (Brazilians of mixed ethnic ancestries) communities have lower levels of social isolation and face higher risks of hospitalisation and death.

What do the new findings imply?

- The stark difference in COVID-19 mortality between public and private healthcare settings underscores the need for further investigation on the drivers of mortality in different hospital settings.
- While non-pharmaceutical interventions have been implemented in São Paulo and other states to slow down transmission, the effectiveness of these interventions among population groups varies with socioeconomic status.
- Healthcare workers, disadvantaged groups working in face-to-face occupations in crowded and segregated areas should be prioritised for vaccination.

INTRODUCTION

The COVID-19 pandemic has amplified the effects of social inequalities on exposure and death in low socioeconomic groups,[1] particularly in Brazil, where it has caused significant mortality.[2] The prevalence of COVID-19 mortality is partially driven by pre-existing non-communicable diseases, which are socially

clustered due to entrenched inequalities.[3] These inequalities are shaped by the social determinants of health,[4] which define a population's health based on the environments where they 'grow, live, work, and age', from birth.[5] Even when underlying health conditions are not present, the interactions of these social determinants disproportionately expose disadvantaged groups to COVID-19[4] and other conditions that could induce adverse chronic health conditions.[6] Furthermore, a review has found that disadvantaged groups are the most vulnerable to the psychosocial impacts of COVID-19,[7] which can aggravate the severity of COVID-19.[4]

Several studies that have been conducted in the context of high-income countries have mostly focused on the USA,[8–10] UK[11 12] and European countries,[13 14] which have consistently found that populations identified as non-White, of low socioeconomic status and those living in high poverty were associated with higher SARS-CoV-2 transmission and COVID-19 death. Few studies have addressed the uneven impact of COVID-19 by socioeconomic status and race in low and middle-income countries,[15] in part because national surveillance systems seldom collect or report this information.[16] In Brazil, higher risk of COVID-19 death has been found for Black and *Pardo* (mixed ethnicity) Brazilians, especially those who are identified as male with low socioeconomic status.[17 18] Nonetheless, there is still little information on how the differential health outcomes of COVID-19 are shaped by broader social inequalities that determine the capacity to self-isolate and non-pharmaceutical interventions (NPIs).

It is paramount to understand the potential social drivers of COVID-19 morbidity and mortality, particularly in countries with high inequality such as Brazil.[19] The first COVID-19 cases in Brazil were detected in São Paulo,[20] the most populous state and home to diverse racial groups. In the Brazilian context of politically polarised public health responses,[21] São Paulo has been severely affected by COVID-19[22] and access to testing has been limited for low-income populations.[23]

We conducted a multiscale analysis to investigate the risk of hospitalisation and death from severe acute respiratory infections (SARI), predominantly caused by COVID-19,[23] notified from March to August 2020, in the *Sistema de Monitoramento Inteligente de São Paulo* (SIMI-SP) database, for São Paulo state. We considered all SARI cases instead of only including patients who tested positive for COVID-19 to avoid the bias in access to SARS-CoV-2 testing towards higher socioeconomic classes in Brazil,[23] which allows us to better capture the disproportionate impact of social inequities on racial and socioeconomic groups. We examined differential risk by race and socioeconomic status, by combining multiple high-resolution data from mobile phones, government census and population surveys conducted during the epidemic. We assessed potential drivers of these inequalities by evaluating local levels of self-isolation, access to teleworking and prevalence of comorbidities.

METHODS

Data sources

SARI and patient information

Patient-level information on demographic characteristics, home address, hospitalisation and health outcomes was collected from the São Paulo State Health Secretariat SARI hospitalisations database (SIMI-SP).[24] SARI can be caused by SARS-CoV-2 and is defined by the Brazilian Ministry of Health as influenza-like syndrome plus one of the following: dyspnoea, persistent chest pain or hypoxia. We excluded all SARI cases that were confirmed to be caused by other respiratory viruses. All SARI cases and deaths are notified in the SIMI-SP database, regardless of hospitalisation.

We included all SARI related hospitalisations and deaths notified in São Paulo state between March 15 and August 29, 2020. Given that recent data is incomplete due to reporting delays[25] and to avoid biases, we limited our analysis to patients with symptoms onset between these dates (epidemiological weeks 10 – 35). We also included SARI cases with unknown etiology, as those are likely related to COVID-19 but not lab-confirmed due to low rates of COVID-19 testing in Brazil[26] and socioeconomic bias in testing.[23]

Zip code information was only available for cases reported in São Paulo state. Data were geocoded using the patient's self-reported home address or postal code with Galileo (www.img.com.br) and Google API. For our analysis, we aggregated these data to the census tract level (n=68 296), the smallest administrative unit reported by the Brazilian census for the purpose of spatial statistical analysis. In the state of São Paulo, 95% of the census tracts have a population size between 136 and 1347 individuals (median=724). Information on the health facility where each case was notified was linked to the National Registry of Health Facilities (Cadastro Nacional de Estabelecimentos de Saúde), which includes information on the mode of healthcare provision (public and private).

The race of patients was partially self-declared and partially identified by a health professional. Race was categorised as either 'White', 'Black', 'Asian' (East or Southeast Asian), '*Pardo*' (mixed ethnic ancestries with diverse skin colours)[27] or Indigenous. Race information was missing for 53 480 (23.9%) of retrieved SARI cases, and was imputed using the racial distribution of the census tract of residence (see online supplemental materials for details). About 0.1% of the population in São Paulo self-identified as Indigenous[28] and since only 166 Indigenous patients (0.07%) were recorded, they were not considered in our analysis.

Socioeconomic data

We obtained data on municipality-level socioeconomic factors from the latest population census (2010) compiled by the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística; IBGE). We selected indicators based on their relevance to the social determinants of health as defined

by the WHO's Commission on Social Determinants of Health,[29] which includes income and income distribution, education, employment and job security, and access to healthcare. We included household income per capita, population density and income inequality (Gini Index). We also determined the proportion of residents with a primary education or lower, employment to population ratio and the proportion of the working population without a formal labour market contract or social security. The road network distance from the centroid of each census tract to the nearest healthcare facility was computed considering all 830 facilities that hospitalised patients with SARI via the public healthcare system (Sistema Único de Saúde; SUS). Information on employment status and comorbidities during the epidemic was retrieved from the National Household Sample Survey (Pesquisa Nacional por Amostra de Domicílios (PNAD) COVID-19), a national telephone survey conducted by IBGE with over 1 888 560 interviews between May and September 2020. Details are described in online supplemental materials.

Seroprevalence data

To assess the broader risk of SARS-CoV-2 infection beyond hospitalisation, we adopted seroprevalence data collected as part of the national Covid-IgG study from blood donors aged 16 – 69 living in São Paulo city.[30] Given that samples were taken across the city, a population-weighted cluster sample of approximately 1000 blood donations were tested each month between February and August 2020 using a chemiluminescence assay that detects IgG against the SARS-CoV-2 nucleocapsid (N) protein (Abbott, Chicago, USA). Self-reported race and education level were recorded at the time of blood donation. To correct for differences in the age-sex distribution of blood donors compared with the population of São Paulo, we applied an age-sex normalisation to the measured prevalence. Details about the data collection methods can be found in Buss *et al.* 's [30] study. [31]

Daily mobility and NPIs

To assess the ability of populations to self-isolate at the local level, we used daily mobile phone data provided by In-Loco (https://www.inloco.com.br/covid-19)[32] for the greater metropolitan area of São Paulo (Região Metropolitana de São Paulo; RMSP). These data were aggregated using a hexagonal grid based on the global H3 index at resolution 8. Each cell has an edge of approximately 460 m and an area of 0.74 km^2 (https://h3geo.org/docs/core-library/restable). For each H3 cell, the social isolation index was measured as the number of people who did not leave their cell of residence during the day, divided by the number of residents in that cell. Each mobile phone was assigned to an H3 cell based on the owner's location of residence during the evening and their travel history. The racial composition and income level of each cell were determined using dasymetric interpolation (online supplemental materials).

We also used municipality and state-level data on NPIs from a continuous survey conducted between 13 May and 31 July 2020.[33] The survey had 13 questions related to the implementation and easing of social distancing measures, and responses were obtained from 612 mayors in São Paulo (94.8% of the total).

Data analysis

Probability of hospitalisation and death

We conducted an individual-level analysis to estimate the probability of reporting a SARI hospitalisation given a patient's race and average income level in their census tract of residence. Census tracts were grouped by quantiles of income per capita into six categories as presented in the results. Similarly, we determined the probability of death from SARI given a patient's race, income and administrative type of the health facility where the patient was hospitalised (public or private). Both probabilities were standardised by age and sex to account for demographic differences between groups. They were calculated for every month between March and August 2020. Probabilities for each age-sex group were estimated empirically using relative frequencies. ORs using White patients and the highest income level as reference groups were computed. CIs were calculated using bootstrapping. Details are in online supplemental materials.

Seroprevalence by socioeconomic status

We calculated the proportion of individuals by education and race category with detectable anti-SARS-CoV-2 antibodies between February and October 2020. 95% CIs were calculated by the exact binomial method and corrected for the specificity and sensitivity of the test.[31]

Socioeconomic drivers and hospitalisation risk

An ecological spatiotemporal regression analysis was conducted at the municipality level for São Paulo state (n=645 municipalities) to assess the monthly risk of hospitalisation and its association with socioeconomic factors between 1 March and 29 August 2020. To further understand the association between socioeconomic conditions and COVID-19 risk, we conducted the same analysis at the census tract level (n=30 815) for the RMSP, where the majority of cases were concentrated. The relative risk of hospitalisation was estimated using a hierarchical Bayesian model composed of a generalised log-linear model with spatially structured and unstructured random effects to account for spatial autocorrelation and time-varying random effect. The spatial structure is characterised by population movement between municipalities from 1 March to 15 August 2020 defined by In-Loco mobile geolocation data summarised elsewhere.[32] A detailed description of the model and interpretation, covariates and diagnostics can be found in online supplemental materials.

Population response to NPIs

We used an event study design[34] to examine how different socioeconomic groups changed their daily mobility levels in response to the introduction and relaxation of NPIs in the RMSP. We compared changes in mobility patterns of the population living in H3 cells with predominantly White versus predominantly Black or *Pardo* residents, as well as of the population living in areas of the wealthiest and poorest income quintiles. The daily isolation index from hexagons was regressed on a set of relative time dummies that indicated the number of days before and after the first NPI introduction in São Paulo state. Hexagon fixed effects controlled for unobserved time-invariant determinants of self-isolation while day fixed effects controlled for temporal shocks common to all hexagons. We further included an additional time-varying control variable, representing the number of days relative to the first confirmed SARI case in each hexagon, and a dummy variable indicating the period of NPI relaxation in each municipality. Sensitivity analyses were performed and discussed in online supplemental materials.

We employed a multinomial logistic regression to estimate the probability that employed individuals would be working face to face, teleworking, or taking paid or unpaid leave. Differences in the work status of individuals by race, education and occupation were calculated while controlling for age and sex. Thirty-five groups of employment occupations listed in the PNAD COVID-19 survey were aggregated to 10 SCO-08 one-digit occupational groups defined by the International Labour Organization. We further disaggregated health professionals and health technicians (online supplemental materials).

Comorbidities

Information on patient comorbidities was missing for approximately 61.6% of the cases in the SIMI-SP database. We estimated the incidence of comorbidities for the population of São Paulo state using the PNAD COVID-19 data. A binomial logistic regression was used to estimate the OR of having at least one comorbidity, by race and education attainment (preprimary, primary, secondary and tertiary), while controlling for age and sex for São Paulo state. The comorbidities considered were chronic obstructive pulmonary disease, diabetes, hypertension or cardiovascular disease such as myocardial infarction, angina or heart failure. CIs for the ORs were calculated taking into account PNAD's complex sample design.

RESULTS

SARIs capture COVID-19-related hospitalisations

Between 1 March and 29 August 2020, São Paulo state had the highest number of SARI hospitalisations per 100 000 habitants compared with all states in Brazil (figure 1A). A time series illustrating the number of SARI hospitalisations is presented in figure 1B. During this time, 232 540 patients were notified in the SIMI-SP database (figure 1C), from which 127 434 (54.8%) had a confirmed COVID-19 diagnosis and 103 360 (44.4%) were diagnosed with SARI of unknown or missing aetiology. From these, 223 455 were hospitalised (98.4%) or died without hospitalisation (1.6%). From the non-hospitalised cases, we only selected deaths; 54 108 patients died, of which 52.5% were White, 20.2% were *Pardo*, 6.13% were Black, 1.96% were Asian, 0.052% were Indigenous and 19.6% did not

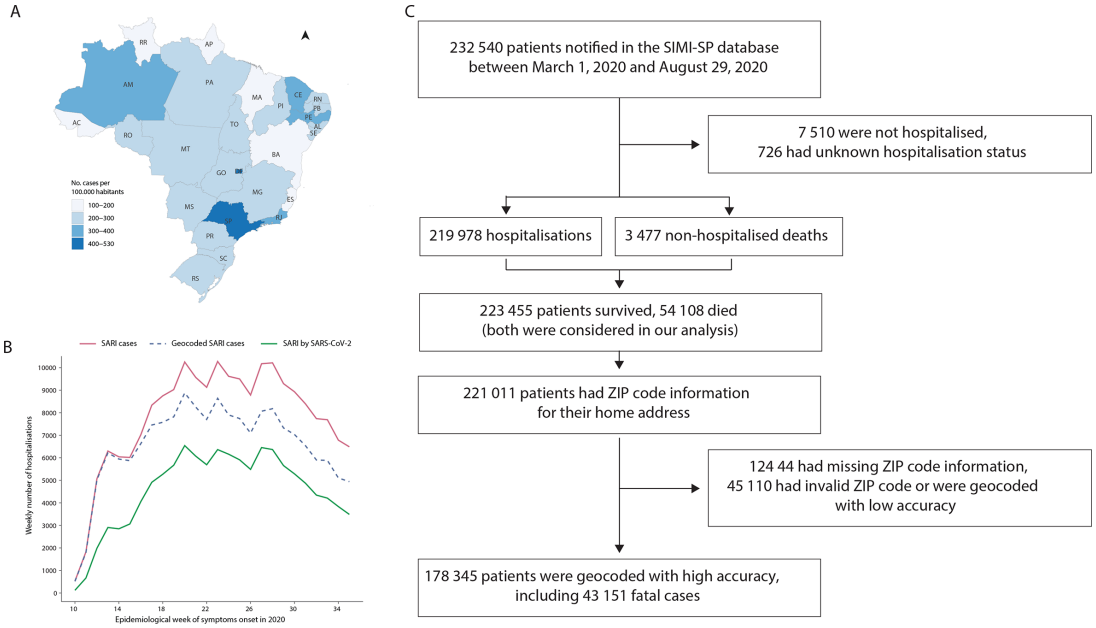


Figure 1 Severe acute respiratory infection (SARI) hospitalisations in São Paulo state. (A) Number of hospitalisations per 100 000 habitants by state in Brazil between 1 March and 29 August 2020. (B) Number of SARI hospitalisations for the state of São Paulo by week of symptom onset. (C) Flow chart of Sistema de Monitoramento Inteligente de São Paulo (SIMI-SP) data processing (Source: https://covid.saude.gov.br).

have race information. We geocoded 178 345 (79.8%) of all SARI cases considered in our analyses with high accuracy at either street address, route or neighbourhood level without compromising personal privacy.

Individual risk to hospitalisation and death varies by race, socioeconomic status and hospital type

During the first month of the COVID-19 epidemic (March) in Brazil, hospitalised patients were more likely to be White or Asian and come from census tracts with higher income per capita (figure 2A,B). During this period, people living in low-income areas were less likely (OR: 0.44, 95% CI 0.42 to 0.46) to be hospitalised with SARI compared with high-income areas. We found that point estimates of inequality levels do not change substantially when observations with missing race information are dropped, though CIs become smaller (online supplemental figure S1). This coincides with the early introduction of COVID-19 in Brazil, when the first infections occurred among higher income travellers returning from overseas.[20 23]

As the epidemic progressed from April onwards, patients were on average more likely to be from low-income census tracts (April to July, OR: 1.08, 95% CI 1.04 to 1.12), except for August, when patients were less likely

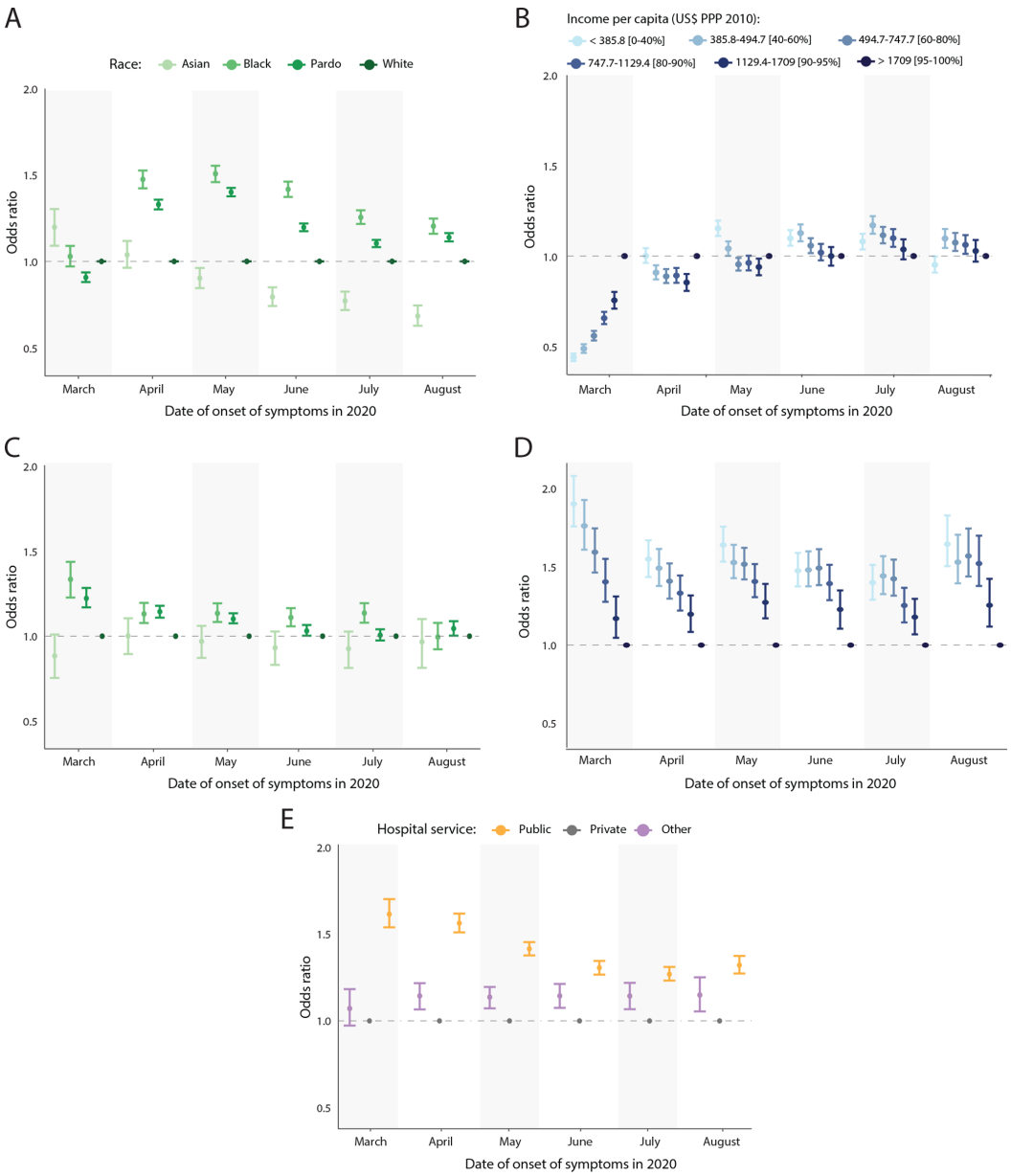


Figure 2 Individual-level hospitalisation and death risk by age-standardised OR. (A) OR for severe acute respiratory infection (SARI) hospitalisation by race. (B) OR for SARI hospitalisation by income. (C) OR for death among patients with SARI by race. (D) OR for death among patients with SARI by income. (E) OR for death among patients with SARI by hospital type. PPP, purchasing power parity.

to be from low-income census tracts (OR: 0.95, 95% CI 0.91 to 1.00). Similarly during this time period, Black Brazilians and *Pardos* became more likely to be hospitalised with SARI than Whites (OR: 1.41, 95% CI 1.37 to 1.46; OR: 1.26, 95% CI 1.23 to 1.28, respectively), while Asians became the least likely to be hospitalised (OR: 0.88, 95% CI 0.82 to 0.94). These results were further confirmed by our seroprevalence findings, where both crude and adjusted prevalence (for age, sex, sensitivity and specificity of the sex) showed that anti-SARS-CoV-2 antibodies were highest in Black blood donors and those with low educational attainment across all age groups (online supplemental figure S2).

Once hospitalised, Black and *Pardo* patients were more likely to die from SARI than White patients between March and August (OR: 1.14, 95% CI 1.07 to 1.21; 1.09, 95% CI 1.05 to 1.13, respectively) (figure 2C). This difference was more pronounced in March and decreased over time. Because our analysis does not control for comorbidities, these results indirectly reflect the differences in the incidence of comorbidities across racial groups. We found that patients living in the poorest census tracts were more likely to die from SARI compared with patients from the wealthiest tracts (OR: 1.60, 95% CI 1.48 to 1.74) (figure 2D). Likewise, patients treated in public hospitals were more likely to die than patients treated in private hospitals throughout the epidemic (OR: 1.40%, 95% CI 1.34% to 1.46%) (figure 2E). Racial differences in the probability of death decreased when considering only patients hospitalised at public health facilities but persisted among patients in private facilities (online supplemental figure S3).

Geographic variation in hospitalisation risk is driven by mobility and socioeconomic status

To understand the geographical variation in SARI hospitalisation, we estimated and mapped the relative risk of SARI hospitalisation at the municipality level (n=645 municipalities) for São Paulo state by month using a model with a spatial structure defined by human movement fluxes derived from anonymised mobile phone data (figure 3A) and covariates related to socioeconomic status (figure 3B). Overall, municipalities with higher levels of movement exchange with the RMSP had higher monthly risk of SARI hospitalisation (figure 3C). We found a lower risk of SARI by SARS-CoV-2 hospitalisation in municipalities with high income per capita (fixed effect=−0.87, 95% CI −1.12 to −0.62) and high proportion of adult residents with a primary education or lower (−0.90, 95% CI −1.12 to −0.68). Municipalities with fewer nearby public health facilities were also found to have lower risk of hospitalisation (−0.31, 95% CI −0.55 to −0.08). We also found a higher risk of SARI hospitalisation in municipalities with higher population density (0.31, 95% credible interval: 0.07–0.54).

We found that over time, the risk of SARI hospitalisation increased particularly in municipalities near and within the RMSP (greater metropolitan area of São Paulo), where 70% of the SARI cases reported for the state are concentrated (figure 4A). By mapping risk at the census tract level (n=30 815) for the RMSP, we detected increasing risk starting from São Paulo city (central region). By June, almost all of the census tracts in and near the city centre were found to have high relative risk, but this risk decreased by August.

Lower ability to self-isolate by disadvantaged groups

Differential risk to SARI in the RMSP was also associated with daily mobility levels. Before the implementation of NPIs on 13 March, mobility levels were similar across all socioeconomic groups (figure 4B,C). However, 14 days after the introduction of NPIs, isolation levels were 8.2% (95% CI 7.2% to 9.2%) higher in predominantly White areas compared with predominantly Black areas. Similarly, 27 days after the introduction of state-level NPIs, isolation levels were 11.2% (95% CI 10.6% to 11.9%) higher in the wealthiest than in the poorest areas. Overall, we detected a decreasing trend in isolation levels over time, and the magnitude of the differences in social isolation levels between areas with predominantly White

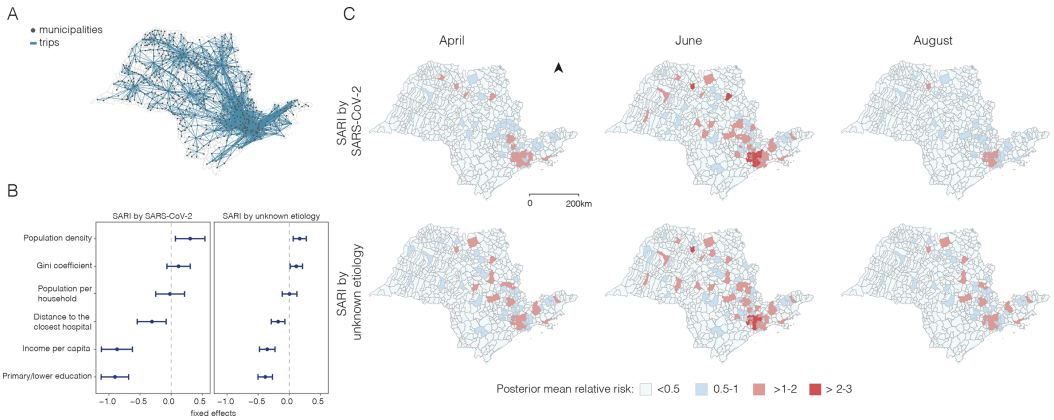


Figure 3 Hospitalisation risk by municipality in São Paulo state. (A) Human movement between municipalities based on InLoco mobile phone data retrieved from March to August 2020. (B) Fixed effects and 95% credible intervals for socioeconomic covariates. (C) Relative risk of severe acute respiratory infection (SARI) hospitalisation at the municipality level.

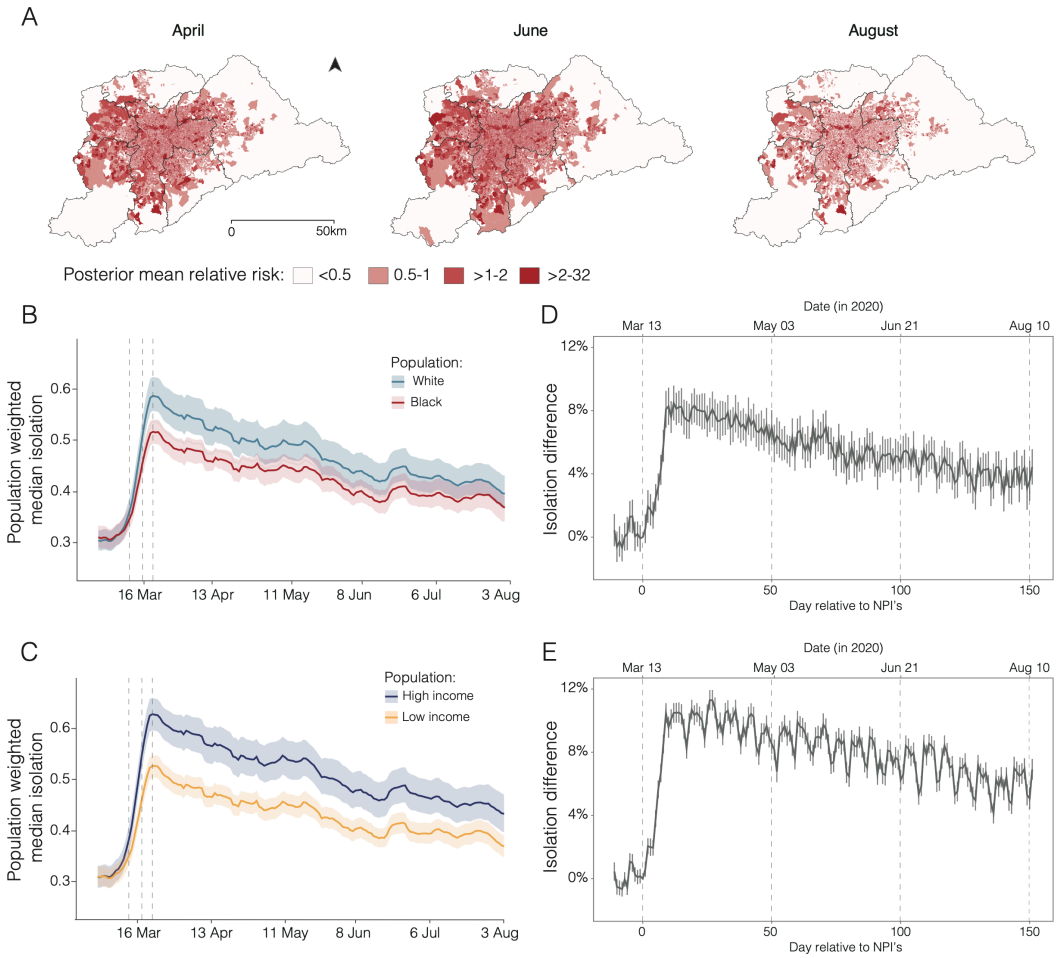


Figure 4 Differential risk based on varying ability to self-isolate in the Região Metropolitana de São Paulo (RMSP). (A) Relative risk of severe acute respiratory infection (SARI) hospitalisation for the RMSP. (B) Seven-day moving average of daily isolation levels by race. (C) Seven-day moving average of daily isolation levels by income. (D) Difference in daily social isolation by race after the introduction of non-pharmaceutical intervention (NPI). (E) Difference in daily social isolation by income after the introduction of NPIs. In panels (B) and (C), solid lines show population-weighted median isolation levels and shaded areas show population-weighted IQR (25%–75%). Dashed vertical lines indicate the dates of NPIs that enabled school closure (13 March was the state NPI) and non-essential activities (18 and 22 March, municipal and state NPIs, respectively).

and Black populations decreased to only 4.4% (95% CI 3.3% to 5.5%) 151 days after the introduction of the NPIs (figure 4D).

Finally, we investigated the differential risk to SARI based on workers' position in the labour market using data from the PNAD COVID-19 survey. After the introduction of NPIs, workers employed in low-skilled jobs or essential services were more likely to keep working face to face than workers in professional or managerial positions (online supplemental figure S4). Workers with pre-primary education were more likely to work in occupations that require in-person contact than workers with tertiary education (probability (PR): 0.89, 95% CI 0.87 to 0.90 compared with PR: 0.58, 95% CI 0.57 to 0.60) and less likely to work in occupations that allow teleworking (PR: 0.005, 95% CI 0.004 to 0.007 vs PR: 0.36, 95% CI 0.35 to 0.37, respectively) (figure 5A). When controlling for education and formal or informal employment, we found no substantial difference between racial groups in the probability of working face to face or teleworking (figure 5B). Nonetheless, because Black and Pardo populations are disproportionally employed in informal and low-skilled jobs, these racial groups were, in general, more likely to be working face to face during our study period.

Disadvantaged groups have more comorbidities

We found that population groups at risk of death from SARI were also more likely to have comorbidities known to aggravate COVID-19 severity. Compared with the population with tertiary education in São Paulo state, individuals with primary education or lower are more likely to have one or more comorbidities (OR: 1.36, 95% CI 1.27 to 1.45) (figure 5C). Similarly, Black individuals were also more likely to have one or more comorbidities than White individuals (OR: 1.29, 95% CI 1.19 to 1.39) (figure 5D). OR estimates for each health condition are summarised in online supplemental figure S5.

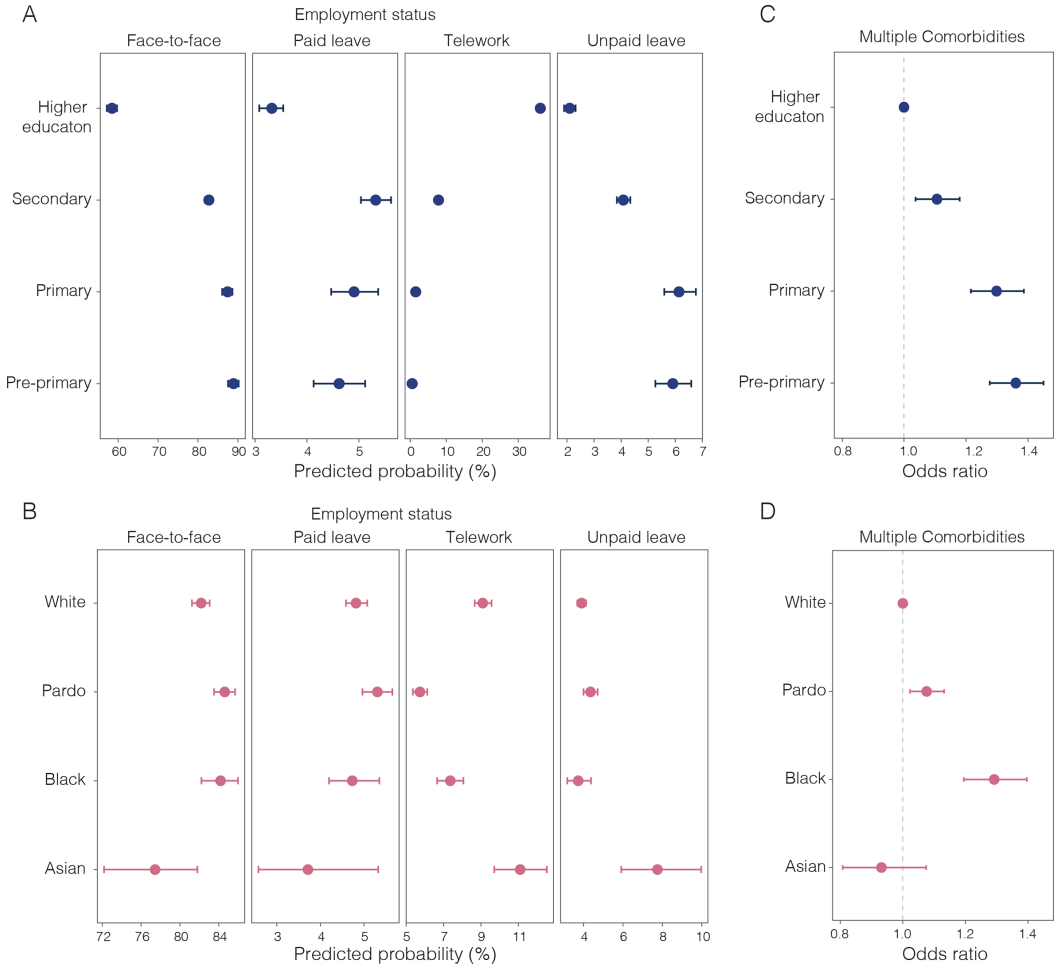


Figure 5 Inequalities in working conditions and comorbidities. (A) Probability of different working conditions by education attainment. (B) Probability of different working conditions by race. (C) OR (OR=1) of having one or more comorbidities by education attainment. (D) OR (OR=1) of having one or more comorbidities by race. Comorbidities considered include chronic obstructive pulmonary disease (COPD), diabetes, hypertension or cardiovascular disease such as infarction, angina and heart failure. Horizontal lines show 95% CIs (Source: Pesquisa Nacional por Amostra de Domicílios (PNAD) COVID-19/Instituto Brasileiro de Geografia e Estatística (IBGE),[17] July to September 2020).

DISCUSSION

Our study shows that socially disadvantaged groups are disproportionately more likely to be hospitalised and die from SARI. We find that the differential health outcomes can be explained by structural inequities linked to the incidence of comorbidities and to socioeconomic conditions, which limit the ability of low-income and non-White populations to socially isolate and reduce their access to quality health services.

Social and racial inequalities shape the risk of SARI hospitalisation and death. After the initial phase of international imports in Brazil,[23] Black or *Pardo* Brazilians and individuals residing in low-income areas were more likely to be hospitalised and die with SARI compared with White individuals and those from wealthier areas, which aligns with recent findings.[17 18] While these results report of severe cases of infection, we find similar results when looking broadly at the population level of COVID-19 infection. Our assessment of anti-SARS-CoV-2 antibodies in blood donors categorised by demographic background further confirms that Black Brazilians and those with lower socioeconomic status are disproportionately exposed to COVID-19.

Patients hospitalised in public health facilities were more likely to die than those in private health facilities. The uneven access to health services explains some but not all the inequality in the risk of death from COVID-19, since racial inequalities in death probability persist among patients within private hospitals. Potential factors influencing this inequality include higher comorbidities among poor Black patients and the lower access to private care among low-income individuals who are disproportionately Black. Other important factors include the disadvantage of having multiple comorbidities, which are more prevalent among Blacks and *Pardos* and those with lower education.

We found that hospitalisation risk is higher for populations living in municipalities that travel to and from

the RMSP, with low income per capita and high population density compared with the rest of São Paulo state. These populations mainly reside in the RMSP, which contains nearly half of the population in São Paulo state and where bias in testing is evident in regions of lower socioeconomic status[23] (online supplemental figure S6). The risk of SARI hospitalisation is particularly elevated in São Paulo city, where seroprevalence estimates from blood donors show that anti-SARS-CoV-2 antibodies were highest in older Black Brazilians and those with lower educational attainment.

We show that inequalities in the risk of SARI hospitalisation were partially explained by differential mobility responses to social distancing guidelines, similar to the UA.[35] In wealthier and predominantly White neighbourhoods, people were able to isolate more, faster, and sustain isolation for long periods of time. We also found that occupational factors played a key role in influencing individuals' ability to physically isolate. Among the working population, low-income and Black workers were less likely to receive a furlough from work or telework. Due to systemic inequalities in education and the labour market, these groups are disproportionately employed in precarious job positions with no social security and dependent on day-to-day income,[36] limiting their ability to socially distance through telework and reductions in daily mobility. The lack of capacity to self-isolate and curtail mobility in these occupations may increase exposure and facilitate SARS-CoV-2 transmission.

Our study has limitations that may have underestimated the level of inequality. First, geocoding cases may have discarded patients from poor census tracts where accuracy is limited.[37] Second, using data aggregated for various administrative levels has inherent limitations due to ecological fallacy and the modifiable areal unit problem.[38] Finally, the 2010 Brazilian population census and PNAD COVID-19 survey may have limited the capture of socioeconomic changes in the last decade and inclusion of extremely wealthy individuals.[39] Additionally, disadvantaged groups can be under-represented in health administrative records because of their lower access to healthcare. Given that São Paulo is the wealthiest state and has the most robust healthcare system in Brazil,[40] it is possible that the impact of inequalities is more severe in other states.

Our findings on the difference of SARI death risk reveal stark inequalities in access to healthcare. Health is a constitutional right and a state responsibility in Brazil.[41] The country's public healthcare system (SUS) is designed to provide universal health coverage without out-of-pocket costs, and regionally planned to improve spatial coverage of services. Yet, previous studies have found that access to COVID-19 health services tends to be lower in less developed regions in the country,[41 42] particularly among low-income and black communities.[43] Only 25% of Brazilians have access to private healthcare via health insurance, reflecting how inequality in access to quality healthcare is largely driven by income.[44] This leaves 75% of the population solely reliant on a chronically underfunded public healthcare system, which highlights a double disadvantage for low-income and non-White populations, who are more likely to be infected and deprived of care. Strengthening healthcare access and its capacity will be critical for reducing health inequities during this and forthcoming public health emergencies.[42]

Our findings on socioeconomic risk factors could help guide vaccine allocation in diverse settings to achieve equitable access. Ensuring that disadvantaged groups, especially those that have in-person occupations and live in crowded and deprived areas, receive vaccination will help prevent and slow down community transmission. While race is not a risk factor in itself, it is critical to consider systemic inequalities that lead Black and *Pardo* communities to be over-represented among low socioeconomic groups, to have higher rates of severe COVID-19 infection, and comorbidities that exacerbate their risk of death. Therefore, including disadvantaged populations among priority groups for vaccination could help reduce health inequities instead of exacerbating them.[45]

As shown in our study, the combination of social, racial and health inequalities in fostering high mortality risk among systematically marginalised communities exemplifies the syndemic nature of the COVID-19 pandemic.[3 46] The negative impact of the COVID-19 pandemic on population health is driven by the accumulation and interaction of two or more adverse conditions, often influenced by the social determinants of health.[47] Therefore, in order to reduce inequalities in COVID-19-related health risk, action is needed to address all factors that contribute to health inequalities. Prior to the start of the pandemic, these determinants have already shaped pre-existing living conditions of disadvantaged groups—such as poor education, precarious work, residential segregation and inadequate housing—which disproportionately impact their access to quality healthcare and expose them to the onset of comorbidities.[47 48] In order to reduce inequalities in COVID-19-related health risk, it is essential to understand what made them vulnerable in the first place by addressing the adverse social determinants that shape an individual's life course. While a key role is played by healthcare systems, action is also required from other stakeholders. COVID-19 control measures affect people differently based on varying levels of financial resources and support available to them, thus governments and industries must work together to address these inequalities by mobilising resources and tools[49] and by developing targeted interventions at both national[50] and local levels.[51]

Our study highlights the need for additional research to comprehend the effects of social and health inequalities during pandemics. First, an assessment of the inequality in access to quality care within public and private health facilities and its risk factors is needed to better understand mortality in different hospital settings. Second, our study shows that population response to NPIs can vary significantly based on social circumstances,

suggesting that future studies should also consider socioeconomic aspects when evaluating the effectiveness of NPIs. Third, more data are needed on whether social safety net programmes that are guaranteeing income for disadvantaged groups during the pandemic (eg, Brazil's emergency cash payment) may have enabled people to reduce their mobility. Nevertheless, our study has shown the impact of social inequities on COVID-19 hospitalisation and death, thus informing future research and policies related to the health impacts of COVID-19 in Latin America.

Author affiliations
[1]School of Geography and the Environment, University of Oxford, Oxford, UK
[2]Institute of Applied Economic Research, Brasília, Brazil
[3]Department of Electronic Systems Engineering, University of São Paulo, São Paulo, Brazil
[4]Department of Zoology, University of Oxford, Oxford, UK
[5]Department of Applied Mathematics, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
[6]Oxford School of Global and Area Studies, Latin American Centre, University of Oxford, Oxford, UK
[7]Virology Research Center, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil
[8]Institute of Social and Political Studies (IESP), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[9]Departamento de Molestias Infecciosas e Parasitarias andInstituto de Medicina Tropical, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil
[10]Fundação Pró-Sangue Hemocentro de São Paulo, São Paulo, Brazil
[11]Disciplina de Ciências Médicas, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil
[12]Laboratory of Medical Investigation in Pathogenesis and Directed Therapy in Onco – Immuno – Hematology (LIM-31) HCFMUSP, University of São Paulo Medical School, São Paulo, Brazil
[13]Hospital das Clinicas da Faculdade de Medicina da Universidade de São Paulo, University of São Paulo, São Paulo, Brazil
[14]Mathematical Sciences, University of Southampton, Southampton, UK
[15]Department of Preventive Medicine, University of São Paulo Medical School, São Paulo, Brazil
[16]MRC Centre for Global Infectious Disease Analysis; and the Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), School of Public Health, Imperial College London, London, UK
[17]Oxford School of Global and Area Studies, University of Oxford, Oxford, UK

Twitter Sabrina L Li @sabrinalyli, Rafael H M Pereira @UrbanDemog and Carlos A Prete Jr @CarlosPrete1

Contributors SLL and RHMP conceived the research questions and designed the study. RHMP, CAPJ, LE, PJHA, CKVB and WMS collected the epidemiological and socioeconomic data. PSP managed mobility data while AASS curated the data on NPIs. LFB, AM, CAN, SCF, NAS and ECS collected the seroprevalence survey samples. SLL, RHMP, CAPJ, AEZ, LE, PJHA, CKVB and RJB conducted exploratory and statistical analyses and interpreted the results. AASS, WMS, RJB, IM, NG and VHN provided guidance on interpreting findings in the political, social and healthcare context of Brazil, while CHW and VHN provided guidance on statistical analysis. SLL, RHMP and CAPJ wrote the manuscript. ECS, NRF and JPM supervised the study. All authors read and revised the final manuscript.

Funding SLL is supported by the Oxford Martin Programme on Pandemic Genomics and the Canadian Social Sciences and Humanities Research Council (SSHRC) Doctoral Fellowship. CAPJ is supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Financial Code 001, Fundação Faculdade de Medicina (FFM), and the São Paulo Research Foundation (FAPESP 2019/21858-0). AEZ is supported by the Oxford Martin Programme on Pandemic Genomics. WMS is supported by the São Paulo Research Foundation (FAPESP 2017/13981-0 and 2019/24251-9). VHN is supported by the Brazilian National Council for Scientific and Technological Development (CNPq: 304714/2018-6). NRF is supported by a Wellcome Trust and Royal Society Sir Henry Dale Fellowship (204311/Z/16/Z). This project was supported by a Medical Research Council-São Paulo Research Foundation (FAPESP) CADDE partnership award (MR/S0195/1 and FAPESP 18/14389-0) (http://caddecentre.org/).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This project was approved by the Brazilian National Research Ethics Committee (CONEP CAAE-30178220.3.1001.0068).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon request. The data sets used and/or analysed during the current study are available from the corresponding authors on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

ORCID iDs
Sabrina L Li http://orcid.org/0000-0002-1183-126X
Rafael H M Pereira http://orcid.org/0000-0003-2125-7465
Carlos A Prete Jr http://orcid.org/0000-0002-3907-423X
Alexander E Zarebski http://orcid.org/0000-0003-1824-7653
Lucas Emanuel http://orcid.org/0000-0002-3754-5755
Pedro J H Alves http://orcid.org/0000-0001-9340-030X
Pedro S Peixoto http://orcid.org/0000-0003-2358-3221
Carlos K V Braga http://orcid.org/0000-0002-6104-7297
Andreza Aruska de Souza Santos http://orcid.org/0000-0003-3585-8683
William M de Souza http://orcid.org/0000-0002-0025-8293
Rogerio J Barbosa http://orcid.org/0000-0002-6796-4547
Lewis F Buss http://orcid.org/0000-0002-9009-9301
Alfredo Mendrone http://orcid.org/0000-0002-3090-4575
Cesar de Almeida-Neto http://orcid.org/0000-0002-8490-4634
Suzete C Ferreira http://orcid.org/0000-0002-4190-0238
Nanci A Salles http://orcid.org/0000-0001-7580-4597
Izabel Marcilio http://orcid.org/0000-0002-2914-6535
Chieh-Hsi Wu http://orcid.org/0000-0001-9386-725X
Nelson Gouveia http://orcid.org/0000-0003-0625-0265
Vitor H Nascimento http://orcid.org/0000-0002-0543-4735
Ester C Sabino http://orcid.org/0000-0003-2623-5126
Nuno R Faria http://orcid.org/0000-0002-9747-8822
Jane P Messina http://orcid.org/0000-0001-7829-1272

REFERENCES

1 Ahmed F, Ahmed Na'eem, Pissarides C, *et al*. Why inequality could spread COVID-19. *Lancet Public Health* 2020;5:e240.
2 Walker PGT, Whittaker C, Watson OJ, *et al*. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science* 2020;369:413–22.
3 Horton R. Offline: COVID-19 is not a pandemic. *Lancet* 2020;396:874.
4 Bambra C, Riordan R, Ford J, *et al*. The COVID-19 pandemic and health inequalities. *J Epidemiol Community Health* 2020;74:964–8.
5 Marmot M, Wilkinson R. *Social determinants of health*. OUP Oxford, 2005.
6 Bartley M. *Health inequality: an introduction to concepts, theories and methods*. John Wiley & Sons, 2016.
7 Dubey S, Biswas P, Ghosh R, *et al*. Psychosocial impact of COVID-19. *Diabetes Metab Syndr* 2020;14:779–88.
8 Emeruwa UN, Ona S, Shaman JL, *et al*. Associations between built environment, neighborhood socioeconomic status, and SARS-

CoV-2 infection among pregnant women in New York City. *JAMA* 2020;324:390–2.
9 Chen JT, Krieger N. Revealing the unequal burden of COVID-19 by income, Race/Ethnicity, and household crowding: US County versus ZIP code analyses. *J Public Health Manag Pract* 2021;27 Suppl 1, COVID-19 and Public Health: Looking Back, Moving Forward:S43–56.
10 Abedi V, Olulana O, Avula V. Racial, economic, and health inequality and COVID-19 infection in the United States. *J Racial Ethn Health Disparities* 2020:1–11.
11 UK Office for National Statistics. Coronavirus (COVID-19) related deaths by ethnic group, England and Wales - Office for National Statistics, 2020. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronavirusrelateddeathsbyethnicgroupenglandandwales/2march2020to10april2020 [Accessed 20 Nov 2020].
12 Niedzwiedz CL, O'Donnell CA, Jani BD, *et al*. Ethnic and socioeconomic differences in SARS-CoV-2 infection: prospective cohort study using UK Biobank. *BMC Med* 2020;18:160.
13 Drefahl S, Wallace M, Mussino E, *et al*. A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden. *Nat Commun* 2020;11:5097.
14 Dragano N, Rupprecht CJ, Dortmann O. Higher risk of COVID-19 hospitalization for unemployed: an analysis of 1,298,416 health insured individuals in Germany. *medRxiv* 2020:2020.06.17.20133918.
15 Argoty-Pantoja AD, Robles-Rivera K, Rivera-Paredez B, *et al*. COVID-19 fatality in Mexico's Indigenous populations. *Public Health* 2021;193:69–75.
16 Pan D, Sze S, Minhas JS, *et al*. The impact of ethnicity on clinical outcomes in COVID-19: a systematic review. *EClinicalMedicine* 2020;23:100404.
17 Baqui P, Bica I, Marra V, *et al*. Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study. *Lancet Glob Health* 2020;8:e1018–26.
18 Ribeiro KB, Ribeiro AF, de Sousa Mascena Veras MA, *et al*. Social inequalities and COVID-19 mortality in the city of São Paulo, Brazil. *Int J Epidemiol* 2021:dyab022.
19 Facundo A, Chancel L, Thomas P. *World inequality report 2018*, 2017.
20 Jesus JGde, Sacchi C, Candido DdaS, *et al*. Importation and early local transmission of COVID-19 in Brazil, 2020. *Rev Inst Med Trop Sao Paulo* 2020;62:e30.
21 Henriques CMP, Vasconcelos W, *et al*. Crises dentro dA crise: respostas, incertezas E desencontros no combate à pandemia dA Covid-19 no Brasil. *Estud Av* 2020;34:25–44.
22 Rezende LFM, Thome B, Schveitzer MC, *et al*. Adults at high-risk of severe coronavirus disease-2019 (Covid-19) in Brazil. *Rev Saude Publica* 2020;54:50.
23 de Souza WM, Buss LF, Candido DdaS, *et al*. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nat Hum Behav* 2020;4:856–65.
24 Secretaria da saúde do estado de São Paulo. SIMI-SP: Pacientes internados POR Síndrome Respiratória Aguda Grave (SRAG). Available: https://www.saopaulo.sp.gov.br/planosp/simi/ [Accessed 25 Oct 2020].
25 Niquini RP, Lana RM, Pacheco AG, *et al*. Srag POR COVID-19 no Brasil: descrição E comparação de características demográficas E comorbidades CoM SRAG POR influenza E CoM a população geral. *Cad. Saúde Pública* 2020;36:e00149420.
26 Hasell J, Mathieu E, Beltekian D, *et al*. A cross-country database of COVID-19 testing. *Sci Data* 2020;7:345.
27 Telles E, Paschel T, Black WI. Who is black, white, or mixed race? how skin color, status, and nation shape racial classification in Latin America. *Am J Sociol* 2014;120:864–907.
28 IBGE - Instituto Brasileiro de Geografia e Estatística. *Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD) COVID-19. Microdados*, 2020.
29 Marmot M, Friel S, Bell R, *et al*. Closing the gap in a generation: health equity through action on the social determinants of health. *The Lancet* 2008;372:1661–9.
30 Buss LF, Prete CA, Abrahim CMM, *et al*. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science* 2021;371:288–92.
31 Diggle PJ. Estimating prevalence using an imperfect test. *Epidemiol Res Int* 2011;2011:1–5.
32 Peixoto PS, Marcondes D, Peixoto C, *et al*. Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil. *PLoS One* 2020;15:e0235732.
33 de Souza Santos AA, Candido DdaS, de Souza WM, *et al*. Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities. *Sci Data* 2021;8:73.
34 Weill JA, Stigler M, Deschenes O, *et al*. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc Natl Acad Sci U S A* 2020;117:19658–60.
35 Chang S, Pierson E, Koh PW, *et al*. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 2021;589:82–7.
36 Lustig N, Pabon VM, Sanz F. The impact of COVID-19 Lockdowns and expanded social assistance on inequality, poverty and mobility in Argentina, Brazil, Colombia and Mexico. ECINEQ, Society for the study of economic inequality, 2020. Available: https://ideas.repec.org/p/inq/inqwps/ecineq2020-558.html [Accessed 20 Nov 2020].
37 Giest S, Samuels A. 'For good measure': data gaps in a big data world. *Policy Sci* 2020;53:559–69.
38 Duranton G, Overman HG. Testing for localization using Micro-Geographic data. *Rev Econ Stud* 2005;72:1077–106.
39 de SP. A distribuição de renda nas pesquisas domiciliares brasileiras: harmonização E comparação entre Censos, PNADs E POFs. *Rev Bras Estud Popul* 2015;32:165–88.
40 Paim J, Travassos C, Almeida C, *et al*. The Brazilian health system: history, advances, and challenges. *The Lancet* 2011;377:1778–97.
41 Castro MC, Massuda A, Almeida G, *et al*. Brazil's unified health system: the first 30 years and prospects for the future. *The Lancet* 2019;394:345–56.
42 Castro MC, de CLR, Chin T. Demand for hospitalization services for COVID-19 patients in Brazil. *medRxiv* 2020:2020.03.30.20047662.
43 Pereira RHM, Braga CKV, Servo LM, *et al*. Geographic access to COVID-19 healthcare in Brazil using a balanced float catchment area approach. *Soc Sci Med* 2021;273:113773.
44 Matijasevich A, Russo G. Covid-19 in Brazil has exposed socio-economic inequalities and underfunding of its public health system. *BMJ* 2020 https://blogs.bmj.com/bmj/2020/06/19/covid-19-in-brazil-has-exposed-deeply-rooted-socio-economic-inequalities-and-chronic-underfunding-of-its-public-health-system/
45 Emanuel EJ, Persad G, Kern A, *et al*. An ethical framework for global vaccine allocation. *Science* 2020;369:1309–12.
46 Singer M, Bulled N, Ostrach B, *et al*. Syndemics and the biosocial conception of health. *Lancet* 2017;389:941–50.
47 Abrams EM, Szefler SJ. COVID-19 and the impact of social determinants of health. *Lancet Respir Med* 2020;8:659–61.
48 Barber S, Diez Roux AV, Cardoso L, *et al*. At the intersection of place, race, and health in Brazil: residential segregation and cardio-metabolic risk factors in the Brazilian longitudinal study of adult health (ELSA-Brasil). *Soc Sci Med* 2018;199:67–76.
49 Ismail SJ, Tunis MC, Zhao L, *et al*. Navigating inequities: a roadmap out of the pandemic. *BMJ Glob Health* 2021;6:e004087.
50 Shadmi E, Chen Y, Dourado I, *et al*. Health equity and COVID-19: global perspectives. *Int J Equity Health* 2020;19:104.
51 Berkowitz RL, Gao X, Michaels EK, *et al*. Structurally vulnerable neighbourhood environments and racial/ethnic COVID-19 inequities. *Cities Health* 2020;31:1–4.

5 Population response to COVID-19 via mass immunisation

Mass vaccination is one of the most effective public health interventions for reducing the number of infections and deaths during infectious disease epidemics. However, eradication of a disease is influenced by human behaviours which drive disparities in vaccine access and coverage. The following paper which has been submitted to a scientific journal for publication discusses the research conducted to understand inequalities in COVID-19 vaccination among socio-demographic groups in Brazil and the impact of vaccination on reducing deaths.

Socio-demographic inequalities in COVID-19 vaccination in Brazil

Sabrina L Li, MSc [1*#], Carlos A Prete Jr, MSc [2*], Alexander E Zarebski, PhD [3*], Andreza A de Souza Santos, PhD [4], Ester C Sabino, MD [5], Vitor H Nascimento, PhD [3], Chieh-Hsi Wu, PhD [6], Jane P Messina, PhD [1,7]

1. School of Geography and the Environment, University of Oxford, Oxford, United Kingdom.
2. Department of Electronic Systems Engineering, University of São Paulo, São Paulo, Brazil.
3. Department of Zoology, University of Oxford, Oxford, United Kingdom.
4. Oxford School of Global and Area Studies, Latin American Centre, University of Oxford, Oxford, United Kingdom.
5. Departamento de Molestias Infecciosas e Parasitarias & Instituto de Medicina Tropical da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.
6. Mathematical Sciences, University of Southampton, Southampton, United Kingdom.
7. Oxford School of Global and Area Studies, University of Oxford, Oxford, United Kingdom.

*These authors contributed equally.

#Corresponding author:

1. Sabrina L. Li, School of Geography and the Environment, University of Oxford, South Parks Rd, Oxford, OX1 3QY, United Kingdom. E-mail: lisabrinaly@gmail.com

Abstract

Background: More than half a million deaths have resulted from the COVID-19 pandemic in Brazil, which was exacerbated by the emergence of the gamma variant in January 2021 and collapsing public healthcare provision. The launch of a vaccination campaign has increased immunisation coverage across Brazil; however, inequalities in coverage vary significantly by state and demographic group. We assessed inequalities in vaccination coverage by age, sex, race, and socioeconomic status to understand their impact on the risk of hospitalisation and death during the vaccination campaign.

Methods: Vaccination and COVID-19 case data were retrieved from the Brazilian Ministry of Health databases *Sistema de Informação do Programa Nacional de Imunizações* (SI-PNI) and *Vigilância de Síndrome Respiratória Aguda Grave* (SRAG), respectively. We analysed vaccination coverage by age, sex, race, and income for both doses. Generalised linear regression was used to investigate the risk of death and hospitalisation by age group, socioeconomic status, and vaccination coverage.

Results: Vaccination coverage varies by region, being higher in the wealthier South and Southeast. Men, people of colour, and low-income groups were more likely to be partially vaccinated as they were less likely to receive a second dose. Increasing vaccination coverage was associated with lower risk of death, especially for older age groups (OR: 10.5-34.8, 95% CI: 10.2, 35.9). Other risk factors such as income per capita (OR: 0.919, 95% CI: 0.917, 0.920), low educational attainment (OR: 0.934, 95% CI: 0.932, 0.936), and unemployment (OR: 1.009, 95% CI:1.007, 1.01) contributed to risk of death. The risk of hospitalisation was higher in areas with higher vaccination, suggesting vaccination is occurring in areas with higher transmission and accessibility to healthcare.

Interpretation: Vaccination reduces the risk of death, especially among older age groups. Despite increasing vaccine uptake, inequalities in coverage still exist amongst states and socio-demographic groups.

Introduction

Since the importation of SARS-CoV-2 in February 2020 (1), Brazil has experienced the third highest number of COVID-19 deaths globally as of October 2021. The gamma (P.1) variant emerged in Manaus in late 2020 and has rapidly spread across the country (2). Despite having a universal healthcare system that is experienced with managing public health emergencies (3), Brazil was unable to contain the spread and mitigate the magnitude of disease severity, especially among socially disadvantaged populations (4). By April 2021, this resulted in intensive care unit (ICU) beds being nearly at capacity in many Brazilian states (5).

A vaccination campaign to promote immunisation against COVID-19 was launched on January 17, 2021 and initially prioritised healthcare workers, elderly individuals aged 90 years or older (gradually decreasing the age limit over time), indigenous communities, and institutionalised individuals. Four vaccines have been offered, including two-dose Coronavac (Sinovac, China) and Covishield AZD1222 (Oxford-AstraZeneca, UK) in the early phase of the campaign, and later Pfizer-BioNTech vaccine (Germany/US) and the single-dose Johnson & Johnson COVID-19 vaccine (Janssen, Netherlands).

The implementation of mass vaccination has enabled a steady increase in vaccination coverage in all states (6). However, the number of doses distributed to each state appears to be driven by the total population of each state instead of considering the different age structures of states. Given that the risk of COVID-19 infection and death increases rapidly with age (7), identifying vaccine uptake by age is critical for reducing inequalities and optimising vaccine distribution and management when faced with vaccine shortages. Continuous vaccine shortages and delays have caused unanticipated halts in the administration of second doses (8). By September 6, 2021, 63.7% of the Brazilian population had received their first dose and 31.8% of the population had been fully vaccinated (6). While overall vaccination coverage is on the rise, little is known about how vaccination impacts age-specific hospitalisations and how it cushions COVID-19-related death rates (9). More investigation is needed to understand whether increasing

vaccination coverage has led to a reduction in the risk of COVID-19 death for all age groups, especially those most vulnerable.

In this study, we investigated inequalities in vaccination coverage across Brazil and assessed how potential driving factors, such as age, sex, race, and socioeconomic status, influenced these inequalities. We hypothesized that vaccination has a protective effect on the population and prevents deaths. We found several states delayed full vaccination among older populations to prioritise allocation for younger individuals. Among the population that was vaccinated, we estimated the proportion of individuals that did not take the second dose and its variation with income, race, and sex, across states. We estimated the effect of vaccination coverage on the risk of hospitalisation and death from COVID-19 and its associations with socioeconomic conditions. We further assessed the effects of vaccination coverage on death across age groups.

Methods

Data sources

Vaccination information

Patient level vaccination data notified between January 17 and September 6, 2021 were retrieved from the Brazilian Ministry of Health's national immunization system, *Sistema de Informação do Programa Nacional de Imunizações* (SI-PNI). For each vaccinated individual we retrieved information on their age, sex, the municipality and state of their home address, location of the vaccination facility, the dose number (i.e., first or second), and the type of vaccine. For a given individual record in the database, we identified the corresponding first and second dose by selecting the vaccination records based on the minimum and maximum vaccination dates for that individual. We did not consider the booster dose campaign in our study, which started on September 15, 2021 with individuals of age 70 years and older.

Cases of severe acute respiratory illness

We retrieved data on severe acute respiratory illness (SARI) hospitalisations and deaths between March 1, 2020 and September 6, 2021, notified in the SRAG (*Vigilância de Síndrome Respiratória Aguda Grave*) database curated by the Ministry of Health. All SARI cases and deaths are notified in the SRAG database regardless of hospitalisation. SARI can be caused by SARS-CoV-2 and is defined by the Ministry of Health as influenza-like illness plus one of the following: dyspnoea, persistent chest pain or hypoxia. We included both confirmed COVID-19 cases and SARI cases with unknown etiology, as those were likely related to COVID-19 but not lab-confirmed due to low rates of COVID-19 testing in Brazil and socioeconomic bias in testing (1,4). All SARI cases confirmed to be caused by other respiratory viruses were excluded from the analysis. From these cases, we filtered for (i) all hospitalisations and, (ii) all deaths (e.g. hospitalised deaths and other). Among these cases, we included patient information on age, sex, and municipality of residence.

Socioeconomic conditions and population size

Socioeconomic data at the municipality level were retrieved from the latest population census (2010) compiled by the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística; IBGE). We selected indicators based on their relevance to the social determinants of health, such as household income per capita, proportion of residents with only a primary education or lower, and unemployment rates. Projected distributions of age and sex in the population for 2020 were also obtained for each municipality and state (10).

Data analysis

Age-group specific vaccination coverage

We estimated the proportion of the population vaccinated with at least one dose using the SI-PNI data and the projected population in each municipality as a whole, and for each age group within that municipality: under 20, 20--29, … 70--79, and 80 years and above

(10). We used the population estimates for 2020 instead of 2021 to partially account for the excess deaths caused by COVID-19 in 2020.

Proportion of the population that missed their second dose

We estimated the proportion of the population that did not receive their second dose within the suggested timeframe (21-28 day interval for Coronavac, and 90 days for all other approved 2-dose COVID-19 vaccines), and defined it as the proportion of individuals that delayed the second dose. We also estimated the proportion of the population that missed the second dose. For this, we determined (i) the proportion of all first dose vaccinated individuals that did not have a second dose recorded, and (ii), from this, selected only those for which the time interval between the first dose and the date of the last entry in the dataset (September 6, 2021) is higher than the suggested time interval between doses for the corresponding vaccine plus a tolerance of 30 days. In other words, this is the proportion of individuals that did not take the second dose and whose second dose is delayed by at least 30 days. Both proportions were estimated for age, sex, race, and municipality. We excluded the single-dose vaccine of Janssen-Cilag from this analysis.

Risk of COVID-19 death by age-group in SARI patients

To estimate the probability that a SARI patient would die from COVID-19, we used logistic regression. Due to missing vaccination status for more than 50% of all SARI patients, we used age group and sex specific vaccination coverage of the individual's municipality of residence. For each SARI patient we used their age group, the socioeconomic conditions, and vaccination coverage of their municipality of residence. We also included a binary indicator for whether the SARI case was hospitalised before the start of the vaccination campaign for each age group. Formally, the equation describing the probability of death among SARI patients is expressed as

$$\begin{aligned} logit = {} & \beta_0 + \beta_1(Age\ group_i) + \beta_2(Age - group\ specific\ vaccination\ coverage_i) \\ & + \beta_3(Age - group\ specific\ vaccination\ campaign_i) \\ & + \beta_4(Unemployment_i) \\ & + \beta_5(Low\ education\ attainment_i) + \beta_6(Income\ per\ capita_i) + \varepsilon_i \end{aligned} \quad (1)$$

where $logit(\Pr(Y_i = 1)) =$ is the log-odds of death for the ith SARI patient (death = 1, alive =0) given intercept β_0 their covariates. $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5,$ and $\beta_6,$ are coefficients for each covariate in the model. For reference, the logit function is defined as $\log\left(\frac{p}{1-p}\right)$.

Risk of COVID-19 hospitalisation by age group

We used quasi-Poisson regression (a GLM with a log-link function) to model the number of SARI hospitalisations in each state adjusting for age group, whether the hospitalisations occurred before or after the start of the vaccination campaign (for each age group and state), and each state's average age group specific vaccination coverage and socioeconomic factors such as the level of unemployment, education, and per capita income. The equation describing the expected number of hospitalisations in this model is

$$\log(\mathrm{E}(Y_i)) = \beta_0 + \beta_1(Age\ group_i) + \beta_2(Age - group\ specific\ vaccination\ coverage_i) + \beta_3(Age - group\ specific\ vaccination\ campaign) + \beta_4(Unemployment) + \beta_5(Low\ education\ attainment) + \beta_6(Income\ per\ capita) + \varepsilon_i \quad (2)$$

with the variance $Var(Y_i) = \varphi\ \mathrm{E}(Y_i))$ where φ is the overdispersion parameter.

Age distribution of death by vaccination coverage

To assess the potential protective effect of vaccination coverage at the population level, we looked at the distribution of ages among COVID-19 deaths in the SARI database. We modelled this distribution as a Dirichlet distribution with parameters alpha = ($1 + D_1$, $1 + D_2$, …, $1 + D_k$), where D_k is the number of deaths observed in age group k. Sampling from the posterior distribution we estimated the marginal credible intervals for the proportion in each age group at four points in the epidemic: Before January 17, January 17 to March 16, March 17 to May 16, and May 17 to September 16, 2021.

Results

Inequalities in vaccination coverage

Between January 17 and September 6, 2021, 133.4 million people were vaccinated with the first dose of a COVID-19 vaccine and notified in Brazil, corresponding to 84.4% of the

213 population above 18. We determined the proportion of the population vaccinated with at
214 least one dose of the COVID-19 vaccine in each municipality and assessed the
215 distribution of coverage across states (Figure 1A). Similar trends can be observed for the
216 proportion of the population that is fully vaccinated (Supplementary Information). Across
217 states, there is a clear regional divide in vaccination coverage, with the South and
218 Southeast states having the highest coverage, and the North and Northeast states having
219 the lowest coverage. Although coverage is relatively high in the South and Southeastern
220 states of São Paulo and Rio Grande do Sul, there is high disparity in coverage between
221 municipalities (Figure 1B). This is also evident for states in the North and Northeast such
222 as the Amazonas and Ceará, where a few municipalities have much higher coverage than
223 the state-level median. Distribution of cumulative vaccination coverage at the national
224 level by age group and sex shows that vaccination coverage is higher with older age
225 groups, and decreases with younger age groups (Figure 1C). A higher proportion of
226 women were vaccinated than men; this difference is more evident among relatively
227 younger age groups below the age of 50 years.
228

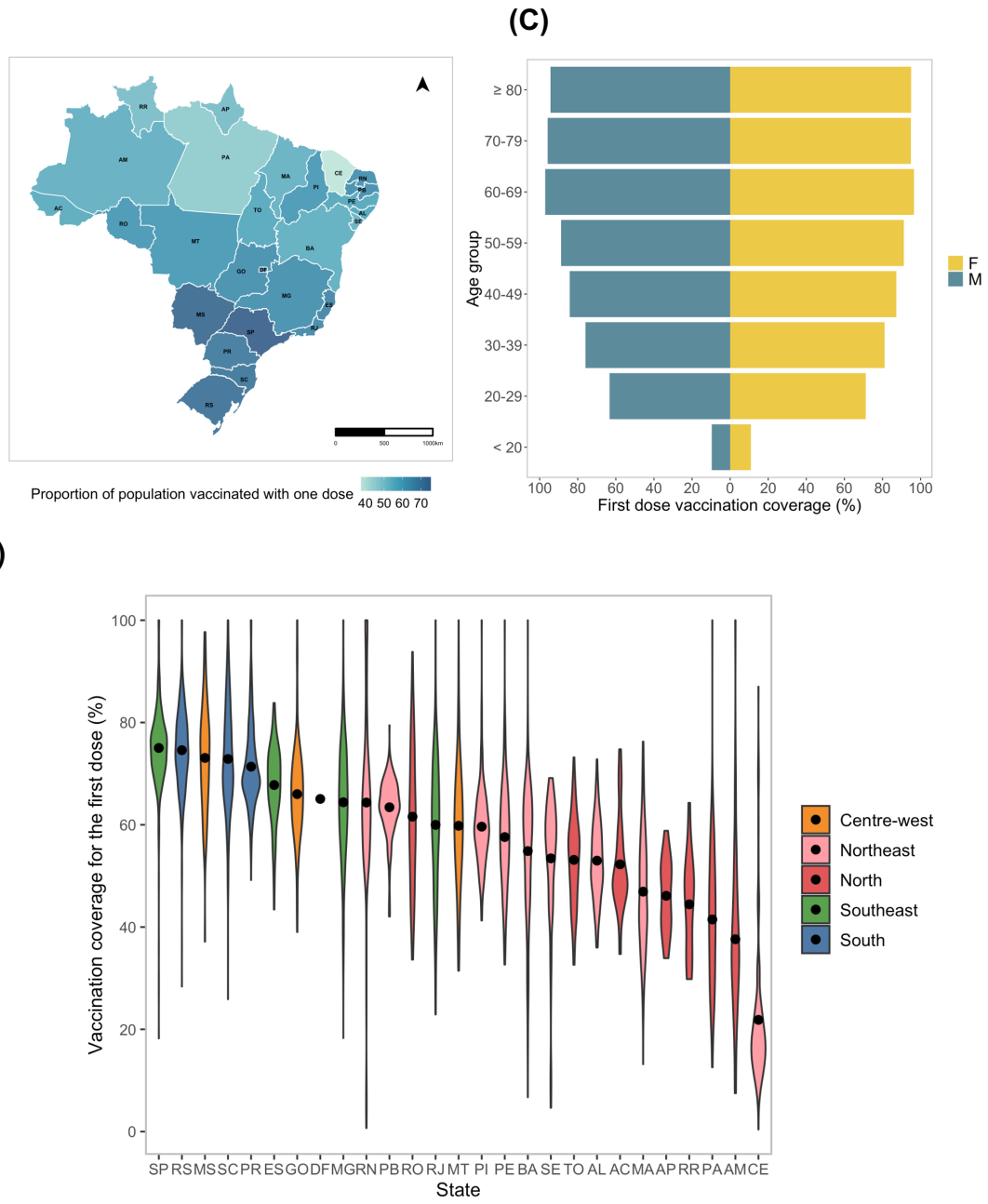


Figure 1. Inequalities in COVID-19 vaccination coverage across Brazil from January 17 - September 6, 2021. (A) Percentage of population vaccinated with at least one dose across states as of September 6, 2021, and (B) across municipalities (black dot represents the mean coverage). (C) Distribution of vaccination uptake for the first dose between men and women by age group.

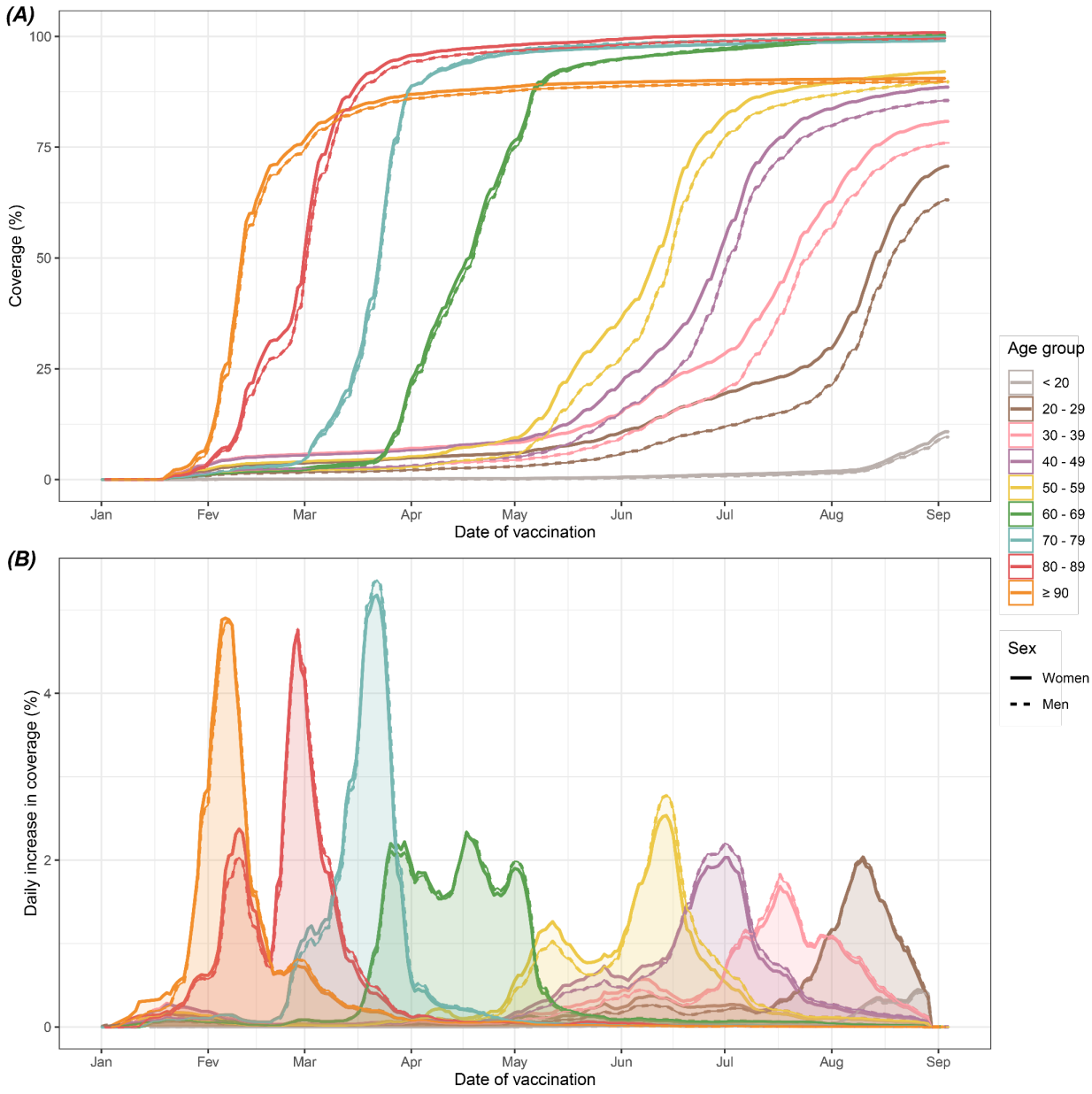


Figure 2. COVID-19 vaccination uptake over time among age groups. (A) Vaccination coverage for the first dose over time. (B) Daily change in vaccination coverage by age group and sex.

Premature advancement of the vaccination campaign towards younger groups

The proportion of the population vaccinated with the first dose in Brazil, by age group and sex, is shown over time in Figure 2A. Because of the uncertainties in the population estimate, there is a small exceedance in coverage over time. For the same reason,

coverages smaller than 100% do not imply that the remaining population is not immunised. At the start of the campaign, both healthcare workers and older populations were prioritised for vaccination, hence the slow increase in the vaccination coverage for younger groups. Looking at vaccination coverage by sex over time, the coverage for men remained consistently lower than women in all age groups. However, this difference between men and women increased abruptly in populations younger than 60 years, creating a substantial difference in coverage between younger and older age groups.

Probability of missing the second dose varies by sex, race, and income

We estimated the probability of an individual missing the second dose, or delaying it by at least 30 days, and determined the variation of this probability by age, sex, race, and income per capita for each municipality in Brazil. Until September 6, 2021, 11.9% and 10.2% of the individuals vaccinated with the first dose had their second dose respectively missed or delayed, respectively. As such, only 14.3% of the individuals that did not take the second dose after 30 days past the recommended date ended up taking a delayed second dose. Among the individuals that were vaccinated with the first dose of Coronavac, 8.78% (95% CI: 8.77%, 8.80%) missed the second dose. This is a smaller proportion compared to the 12.23 (95% CI: 12.25% - 12.29%) of individuals that missed the second dose (for Covishield and Pfizer vaccines), where the optimal interval between the first and second dose is 90 days.

The probability of missing the second dose by race and sex is shown in Figure 3A. We found that men were slightly more likely to miss (RR 1.065, 95% CI: 1.063 - 1.067) the second dose when compared to women. Indigenous Brazilians were also 2.64 (95% CI: 2.62 - 2.66) times more likely to miss the second dose than White Brazilians, followed by East Asian Brazilians (RR 1.68, 95% CI: 1.68 - 1.69), *Pardos* (RR 1.65, 95% CI: 1.65 - 1.65) and Black Brazilians (RR 1.51, 95% CI: 1.50 - 1.51).

Figure 3B shows how the probability of missing the second dose varies with income per capita. This figure demonstrates that municipalities with higher probability of missing the second dose are more likely to have a smaller income per capita. While the median

income per capita of municipalities with probability of missing the second dose smaller than 2% is R$546.80 (IQR: R$421.60 - R$650.50), this value decreased to R$270.40 (IQR: R$212.60 - R$420.80) for municipalities where the probability is higher than 10%.

Figure 3C shows the dependency between the state and the probability of missing the second dose, where the 95% confidence intervals, interquartile ranges, and median for a given state were calculated using the probability of missing the second dose for each municipality.

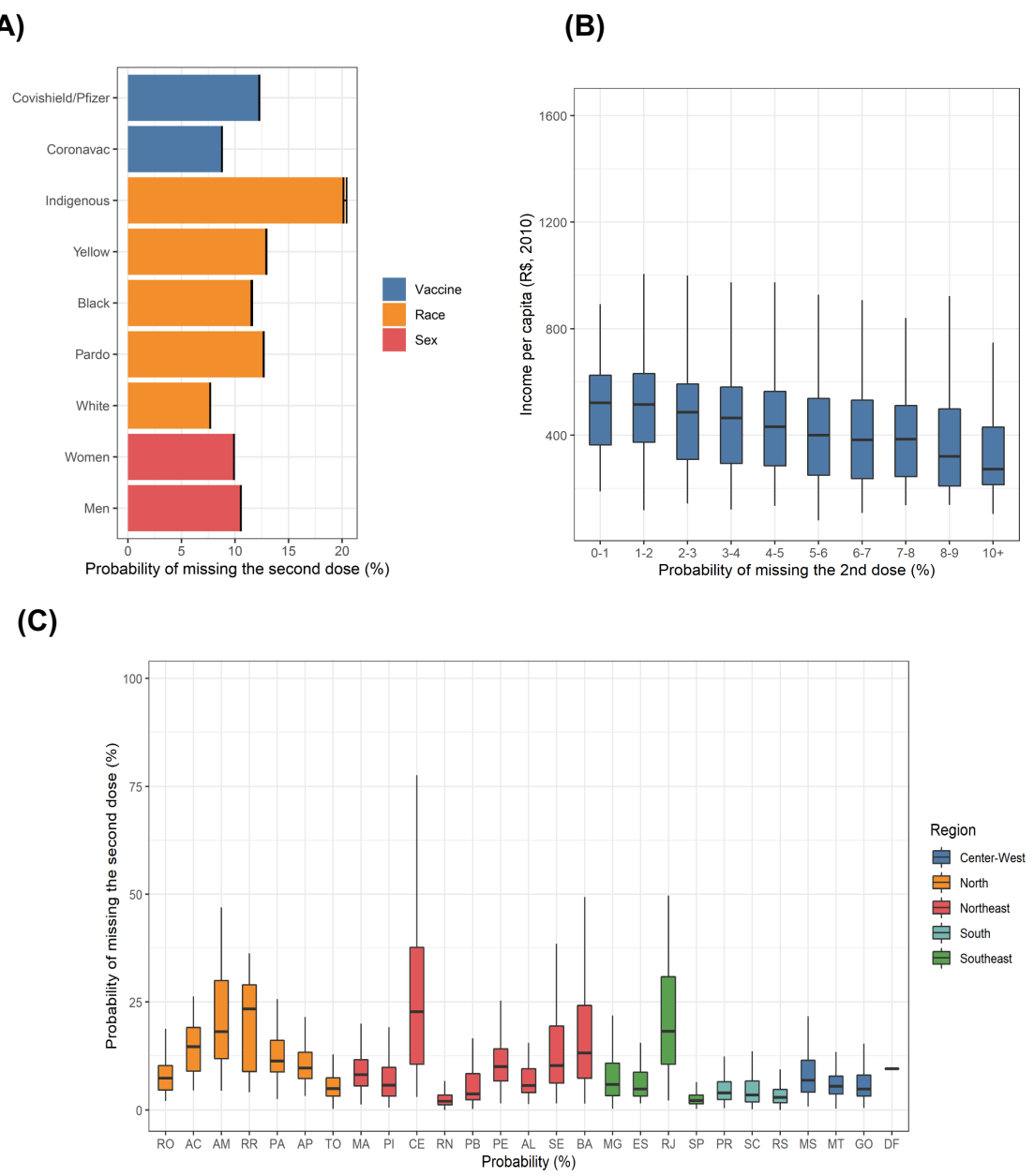


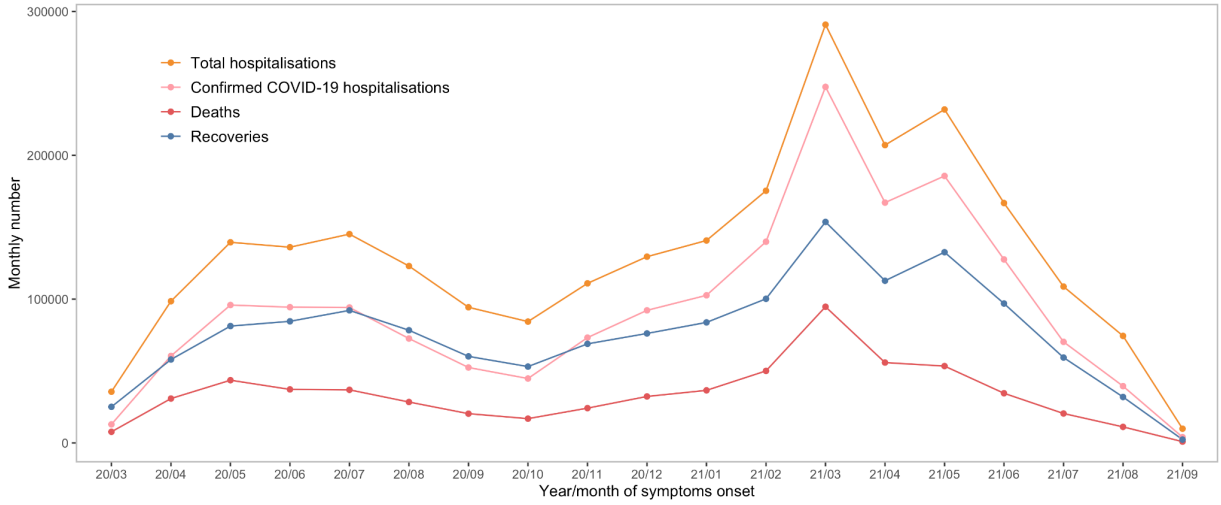
Figure 3. Probability of missing the second dose of the COVID-19 vaccine. (A) Probability of missing the second dose based on three factors: vaccine type, race, and sex. (B) Income per capita calculated at municipality level in terms of the probability of having a second dose missed in the corresponding municipality. (C) Probability of missing the second dose calculated for each municipality and grouped by state.

Vaccination decreases risk of death

We found that vaccination decreases the risk of COVID-19 death while adjusting for age group and socioeconomic conditions. Using the under 20 years age group as a baseline, we found that the risk of COVID-19 death increased as age increased. For instance, the odds of death are at least ten times greater for someone above the age of 50 years (OR: 10.5-34.8, 95% CI: 10.2, 35.9) than an individual under 20 years of age. Patients living in municipalities with higher income per capita and lower education attainment, respectively, had reduced risk of death (OR: 0.919, 95% CI: 0.917, 0.920) and higher education attainment (OR: 0.934, 95% CI: 0.932, 0.936). On the contrary, patients from municipalities with higher unemployment rates were slightly more likely to be at risk of death (OR: 1.009, 95% CI:1.007, 1.01). When the vaccination campaign was accounted for, patients were at a lower risk of death after the start of the vaccination campaign (OR: 0.834, 95% CI: 0.832, 0.837); patients living in municipalities with higher vaccination coverage had a lower risk of death (OR: 0.996, 0.994, 0.997). Despite the overall protective effects of vaccination on death, higher vaccination coverage and lower unemployment were both associated with higher risk of hospitalisation. Detailed information on model outputs can be found in Supplementary Information.

To further verify the relationship between vaccination coverage and risk of death between age groups, we plotted the probability of a patient dying from COVID-19 being in a particular age group prior to the beginning (before January 17, 2021) and after the start of the vaccination campaign (including and after January 17, 2021) during our study period (Figure 4A). Overall, there is a higher probability that those dying before the start of the vaccination campaign were from an older age group. After vaccination began, those dying were more likely to be from relatively younger age groups with ages below 60 years old. This can be attributed to vaccination, where the proportion of vaccination is much higher among people in older age groups due to earlier uptake (Figure 4B). Notably however, those that are under 20 years of age showed a reverse trend where the risk of death was lower prior to the vaccination campaign.

(A)



(B) **(C)**

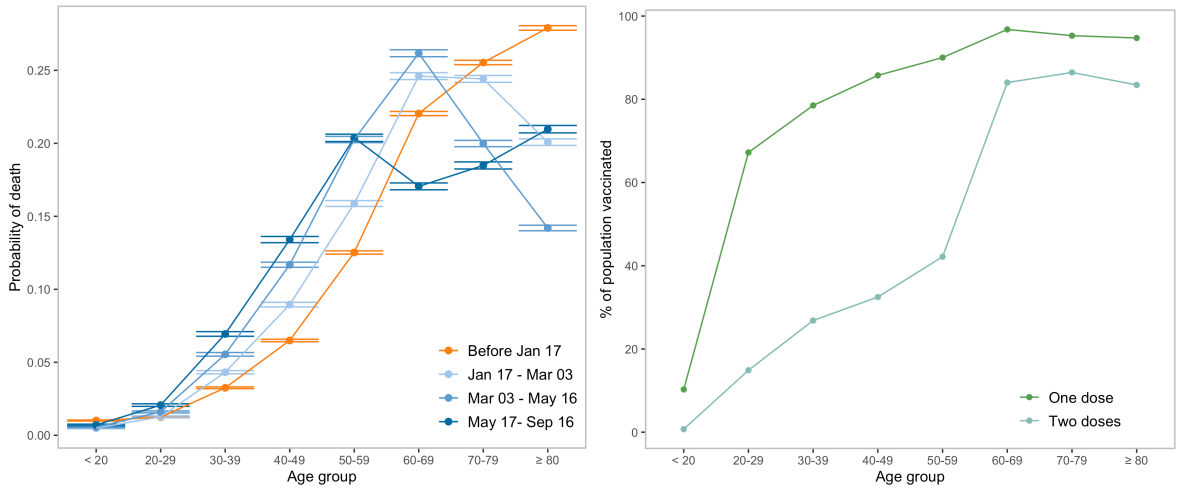


Figure 4. Risk of SARI death and vaccination coverage by age distribution. (A) Distribution of SARI hospitalisations over time (March 2020 - September 2021). (B) Probability of SARI death for each age group across time. (C) Proportion of population vaccinated with at least one dose by age group during study period.

Discussion

Inequalities exist in COVID-19 vaccination coverage, especially across states and among different age groups, sex, and socioeconomic status. Among SARI patients diagnosed

with COVID-19, the risk of hospitalisation and death are significantly greater among older individuals and those with lower socioeconomic status. Increasing vaccination reduces the risk of death, especially among older age groups. Vaccine uptake is also stronger in areas where hospitalisation is high, suggesting better access to healthcare and improved reporting of infections.

At the beginning of the vaccination campaign, variation in coverage between population groups was mainly driven by the unequal distribution of vaccines across states. During the initial stages of the campaign, it was observed that doses were distributed based on the total size of the population in each state. This strategy did not account for the proportion of the population classified as vulnerable, including the elderly, and differences in the age distribution of the population across states. This was compounded by increased vaccination coverage for the first dose among younger age groups when the demand for the COVID-19 vaccine among older populations was still significant.

States from the North and Northeast regions, such as Amazonas, Amapá, and Ceará, were prioritised for vaccination in January 2021 due to high transmission rates of the P.1 (gamma) variant, which was first detected in the capital city of Manaus in the Amazonas in January 2021 (2). Despite this, during our study period, lower overall coverage for the first dose was observed in the North and Northeast regions of Brazil. Relative to southern states, these regions are economically less developed and have found hospital resources scarce during the COVID-19 pandemic (11). Notification delays for the vaccinated have an important impact on differences in coverage between states. For example, even though we determined a 37.5% first-dose coverage rate for the northeastern state of Ceará, official statistics obtained directly from the Brazilian state secretariats on September 6, 2021 found this was 59.2% (12).

Inequalities in coverage are also evident between sex. Women were more likely to have received both doses than men. This may be owing to the early stages of the vaccination campaign, when healthcare workers were prioritised. Women, in particular, make up a significant proportion of healthcare workers in Brazil as they account for 62.5% of all

vaccinated healthcare workers, which corresponds to 10.6% of all vaccinated individuals in our dataset. Moreover, women seek out healthcare more frequently than men. Based on an online survey that received responses from more than 170,000 respondents, a study found that men showed higher hesitancy towards COVID-19 vaccines, indicating the main reason being fear, especially of severe clinical reactions (13). We also found that there was a longer delay for taking the second dose in municipalities with lower income per capita, which is aligned with previous findings that low incomes were associated with higher rates of vaccine hesitancy during the COVID-19 pandemic (13). This has implications for death risk, as our findings show that SARI patients from poor municipalities are already more likely to die from COVID-19 than those from wealthier municipalities.

The association between vaccination and hospitalisation suggests that wealthier areas are likely to have higher vaccination rates, alongside improved reporting of viral transmission and hospitalisation. Nonetheless, vaccination reduces the probability of death. To further investigate whether vaccination can prevent COVID-19 related deaths among different age groups, we considered the age distribution of those that died across multiple periods of time. We observed a reduction in the proportion of old-age mortality, since it was the elderly that were initially prioritised for vaccination. This trend was the least evident with young people under the age of 20. This was likely owing to the delayed vaccination of young people until July 2021. Due to a lack of complete population-level data, it may be premature to surmise the true protective effects of vaccination coverage on younger age groups. While the overall death risk is low for this age group, some young people, with or without pre-existing health conditions, could still die from COVID-19 (14,15).

Our estimates of vaccination coverage by state and age-group are limited by the accuracy of the projected population estimates. Due to missing data, we were unable to retrieve the vaccination status of each SARI patient and thus had to rely on overall vaccination coverage for each age group in each municipality. This may have introduced discrepancies in the modelled relationship between vaccination coverage and risk of

death. Furthermore, it is inevitable that there are inherent limitations to the accuracy of vaccination record-keeping due to human error (e.g. date of vaccination could be replaced by date when the information was entered into the system). Moreover, the lack of healthcare resources in some areas (e.g. Ceará) may have led to underreporting of existing cases and notification delays. Bias may also be prevalent in the self-identification of race, where wealthier Black Brazilians are more inclined to self-identify and contribute race information than poorer Black Brazilians (16).

Despite some vaccine hesitancy among certain demographic groups, the overall intention to receive the COVID-19 vaccine is high (13,17). While a previous study using survey data from late 2020 found that when certain countries of origin for the vaccine were mentioned (e.g. China and Russia), the willingness to vaccinate among Brazilians decreases. However, our study shows that individuals that were vaccinated with the first dose of Coronavac were less likely to miss the second dose when compared to individuals that were vaccinated with other vaccines, such as Covishield or Pfizer. This may be partially attributed to the fact that the optimal interval between the first and second doses for most vaccines is 90 days, while the interval for Coronavac is only 21 to 28 days (18).

Around 90% of all cities in Brazil have already started their third dose or “booster” dose campaign with the elderly population (8) in response to concerns about the Delta variant. The decision to offer a booster dose has been controversial as parts of the population are still awaiting their second dose. Vaccination has a protective effect on populations (19) but the details of how this is communicated to the general public are complex (20). Our findings on spatial inequalities in coverage and on demographic groups that are more willing to miss the second dose should be considered when designing media campaigns and interventions to increase vaccine access and uptake.

References

1. de Souza WM, Buss LF, Candido D da S, Carrera J-P, Li S, Zarebski AE, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nature Human Behaviour. 2020 Aug;4(8):856–65.

2. Faria NR, Kraemer MUG, Hill SC, Jesus JG de, Aguiar RS, Iani FCM, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science. 2018 Aug 31;361(6405):894–9.

3. Domingues CMAS, Maranhão AGK, Teixeira AM, Fantinato FFS, Domingues RAS. The Brazilian National Immunization Program: 46 years of achievements and challenges. Cad Saúde Pública [Internet]. 2020 Oct 26 [cited 2021 Jun 8];36. Available from: https://www.scielo.br/j/csp/a/XxZCT7tKQjP3V6pCyywtXMx/?lang=en

4. Li SL, Pereira RHM, Jr CAP, Zarebski AE, Emanuel L, Alves PJH, et al. Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil. BMJ Global Health. 2021 Apr 1;6(4):e004959.

5. Hallal PC, Victora CG. Overcoming Brazil's monumental COVID-19 failure: an urgent call to action. Nat Med. 2021 Jun;27(6):933–933.

6. Mapa da vacinação contra Covid-19 no Brasil Acompanhe a evolução da imunização [Internet]. G1 (Globo); 2021 [cited 2021 Nov 5]. Available from: https://especiais.g1.globo.com/bemestar/vacina/2021/mapa-brasil-vacina-covid/

7. O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature. 2021 Feb;590(7844):140–5.

8. Confederação Nacional de Municípios (CNM),. Pesquisa CNM – Covid-19 – Edição 28 – de 04 a 07/10 [Internet]. Brasília - DF; 2021 Oct [cited 2021 Nov 5]. Available from: https://www.cnm.org.br/cms/biblioteca/30042021_Relato%CC%81rio_pesquisa_Relampago_Ed_06.pdf

9. Hitchings MDT, Ranzani OT, Torres MSS, Oliveira SB de, Almiron M, Said R, et al. Effectiveness of CoronaVac among healthcare workers in the setting of high SARS-CoV-2 Gamma variant transmission in Manaus, Brazil: A test-negative case-control study [Internet]. 2021 Jun [cited 2021 Nov 5] p. 2021.04.07.21255081. Available from: https://www.medrxiv.org/content/10.1101/2021.04.07.21255081v4

10. Freire FHM de A, Gonzaga MR, Gomes MMF. Projeções populacionais por sexo e idade para pequenas áreas no Brasil. Revista Latinoamericana de Población. 2020;14(26):124–49.

11. Rocha R, Atun R, Massuda A, Rache B, Spinola P, Nunes L, et al. Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. The Lancet Global Health. 2021 Jun 1;9(6):e782–92.

12. Mapa da vacinação contra Covid-19 no Brasil Acompanhe a evolução da imunização [Internet]. G1 (Globo); 2021 [cited 2021 Nov 5]. Available from: https://especiais.g1.globo.com/bemestar/vacina/2021/mapa-brasil-vacina-covid/

13. Moore DCBC, Nehab MF, Camacho KG, Reis AT, Junqueira-Marinho M de F, Abramov DM, et al. Low COVID-19 vaccine hesitancy in Brazil. Vaccine. 2021 Oct 8;39(42):6262–8.

14. Taylor L. Covid-19: Brazil's spiralling crisis is increasingly affecting young people. BMJ. 2021 Apr 1;373:n879.

15. Cunningham JW, Vaduganathan M, Claggett BL, Jering KS, Bhatt AS, Rosenthal N, et al. Clinical Outcomes in Young US Adults Hospitalized With COVID-19. JAMA Internal Medicine. 2021 Mar 1;181(3):379–81.

16. Telles E, Paschel T. Who Is Black, White, or Mixed Race? How Skin Color, Status, and Nation Shape Racial Classification in Latin America. American Journal of Sociology. 2014 Nov 1;120(3):864–907.

17. Gramacho WG, Turgeon M. When politics collides with public health: COVID-19 vaccine country of origin and vaccination acceptance in Brazil. Vaccine. 2021 May 6;39(19):2608–12.

18. Ministério da Saúde. Plano Nacional de Operatcionalização da Vacinação Contra a COVID-19 [Internet]. Brasília - DF; 2021 Oct [cited 2021 Nov 5] p. 124. Report No.: 11. Available from: https://www.gov.br/saude/pt-br/coronavirus/vacinas/plano-nacional-de-operacionalizacao-da-vacina-contra-a-covid-19

19. Rossman H, Shilo S, Meir T, Gorfine M, Shalit U, Segal E. COVID-19 dynamics after a national immunization program in Israel. Nat Med. 2021 Jun;27(6):1055–61.

20. Schwarzinger M, Luchini S. Addressing COVID-19 vaccine hesitancy: is official communication the key? The Lancet Public Health. 2021 Jun 1;6(6):e353–4.

6 Discussion

This thesis elucidates geographical inequalities in infectious disease risk derived from human-environment interactions in Brazil. Guided by the theoretical frameworks of human disease ecology and landscape epidemiology, combined with spatial epidemiological and GIS methods, I investigated how the relationships between human populations and their physical, built, and social environments modified their susceptibility to infectious disease risk. These relationships were explored via recent outbreaks and epidemics in Brazil and were defined at different stages of an infectious disease trajectory starting with "emergence", moving to "spread", and ending with "response".

Here I outline key outcomes from each chapter of this thesis, highlight the strengths and limitations of the methods, and outline how the research findings presented in this thesis could pivot future directions of research.

6.1 Chapter Summary

The ever-changing relationship between humans and the environment have continued to redefine the relationships between human hosts and pathogens and their multi-dimensional impact on public health. In Chapter 1, I contextualised how these interactions have contributed to the emergence of arboviral epidemics and the COVID-19 pandemic in Brazil. Here, I outlined the key research questions of this thesis, focusing on emergence, spread, and response. Namely, (1) how does human-environment interactions influence disease emergence in the physical environment? (2) How do they enable disproportionate disease susceptibility among populations in the built environment? And (3), how can these interactions be used to guide public health response to epidemics?

In Chapter 2, I conducted a literature review to better understand how the physical and social climate, along with demographic characteristics and healthcare, contribute to the emergence, spread, and response to arboviral diseases and COVID-19 epidemics in Brazil. This review elucidates that the foci of infectious diseases between the virus, human host, and vector or animal host, are driven by a multitude of environmental, behavioural, and population factors that interact to facilitate disproportionate infectious disease risk on human populations. Thus, the theoretical frameworks of disease ecology and landscape epidemiology were adopted to conceptualise this dynamic relationship.

The nature of these dynamic interactions across space and time can be quantified using computing tools such as GIS and spatial statistical models, which frames the overarching methodological framework of this thesis. In the methodology section of this chapter, I reviewed the different quantitative spatial modelling methods used to predict and infer the geographic spread, focusing on the Zika virus epidemic as a case study, which was the last WHO Public Health Emergency of International Concern (PHEIC) that severely impacted Brazil. Despite the focus on Zika, the main aim of this review is to summarise the various modelling methods used to study the spatial spread of disease, by focusing on model type, aim, applications, scale, and the feasibility of integrating multidimensional data to improve the model and its outputs. Several methods highlighted in this review have been adopted in the proceeding chapters, such as species distribution modelling, a machine learning technique for predicting the geographical distribution of vector and disease suitability, and statistical models such as Bayesian spatio-temporal models, for quantifying the distribution of disease risk. The frameworks discussed in this chapter are then used to guide the exploration of the three research questions, which are discussed in the subsequent chapters.

To address the first research question, I investigated how the physical environment contributed to the disease emergence of yellow fever virus (YFV) in Chapter 3, and its interactions with human populations to foster disease outbreaks in Brazil. YFV has caused the worst sporadic outbreaks in Southeastern Brazil in the last 70 years (1). In my investigation, I first compiled the most comprehensive database of three primary mosquito vectors known to facilitate the sylvatic transmission of yellow fever, namely *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp*, to date. I modelled and mapped the environmental suitability of these three vectors, respectively, to classify the spatial distribution of their ecological niche in fostering transmission. This was done using a stack regression ensemble, which combines various models with potentially different prediction assumptions to improve predictive performance (2). The ensemble comprised of three binary prediction-based machine learning models (elastic-net regularized GLM, extreme gradient boosted regression tree model, and a random forest model), which were chosen for their relatively good generalizability for disease mapping (3,4). Model performance was evaluated via the AUC on the holdout data, where high prediction accuracy was obtained.

Mapping the ensemble outputs illustrated that the most suitable regions for all species were in the Southeast, and parts of the Northeast and Central-west. Environmental covariates found to have the strongest influence on predicting suitability included the prevalence of deciduous tree cover, elevation, and minimum daytime temperature, which are positively correlated with the likelihood of species suitability, and population density, which is negatively correlated. By superimposing this information with NHP reservoirs and human populations within the Southeastern Atlantic Forest Biome, areas at risk of transmission and spillover were identified in the highly populated municipalities of Paraná, the cities of São José do Rio Preto and São Paulo in São Paulo state, as well as municipalities in Rio de Janeiro and Espírito Santo states. Municipalities near these spillover zones were

identified at risk of transmission. Conversely, regions between Goiás and Minas Gerais were observed with the lowest risk given the lack of NHP host reservoirs. By mapping the locations of reported YFV cases among NHPs from 2015 – 2020, it was observed that cases mainly clustered within spillover zones; however, because YFV cases are reported through passive surveillance, it is possible that levels of spillover risk may exist in non-spillover zones.

While these maps could be used as public health tools to monitor future YFV outbreaks, two key limitations should be noted in this context. Firstly, vaccination coverage was not considered for identifying spillover risk; despite this, current vaccination data may also be limited in accurately capturing the spatial variation of coverage at the local level. Secondly, while the dynamics of mobility was not considered, it should be noted that movement among humans and NHPs can influence the distribution of transmission and spillover from the origin of infection. This is often difficult to capture via passive surveillance systems especially in rural areas, where access to healthcare is poor and reporting is sparse.

The social determinants of health, which encompasses access to healthcare, can lead to health inequalities that influence differential population susceptibility to disease in the built environment (5). This was evident during the COVID-19 epidemic in Brazil, when certain populations were disproportionately impacted due to geographical disparities in access to COVID-19 diagnosis and vulnerability to hospitalisation and death. This concern, which outlines the basis for the second research question of this thesis, is explored in Chapter 4, where I investigated the spread of COVID-19 in this context during the beginning and later stages of the epidemic, focusing first on the Metropolitan Region of São Paulo (MRSP), then later, on its overarching administrative division, São Paulo state.

In Chapter 4.1, I explored the beginnings of the COVID-19 epidemic in Brazil, focusing on the geographical disparity in COVID-19 testing and infection within the MRSP. Since the start of the epidemic, Brazil adopted a nationwide COVID-19 reporting system (e.g. SIVEP-Gripe), which collated both laboratory confirmed COVID-19 and severe acute respiratory illnesses (SARI) of unknown aetiology cases from public and private health services. Using a spatial hierarchical Bayesian model, I modelled and mapped the risk of infection for both laboratory confirmed cases and SARI cases throughout the first 22 weeks of the epidemic at the census tract level (typically 150-300 households).

While an increase in relative risk was observed for both lab-confirmed COVID-19 cases and SARI cases with unknown aetiology in central MRSP, the relative risk of SARI was more widespread across other regions where income per capita was much lower. This was explained partially by the high concentration of initial cases among wealthy Brazilian travellers that were able to access private laboratories for COVID-19 diagnostics and limited access to free COVID-19 diagnostics during the earlier days of the epidemic. This was further evidenced by the fact that during the first month of the epidemic in Brazil, about two-thirds of the tests were performed in private laboratories where the cost varied between 300-690 Brazilians Reais (in comparison, the minimum monthly household salary is 1,045 Brazilian Reais), illustrating that that the true burden of the epidemic is likely underestimated in low-income neighbourhoods.

The findings from this analysis motivated additional investigation into the differential risk of COVID-19 hospitalisations and deaths between social and racial groups in São Paulo state, which was covered in Chapter 4.2. I examined both individual- and population-level risk to hospitalisation and death by race and socioeconomic status, using multiple datasets sourced from mobile phones, population census, and national surveys during the pandemic,

and a combination of quantitative methods. During the later phases of the COVID-19 pandemic in 2020, it became clear that disadvantaged communities such as Black and *Pardo* (Brazilians of mixed ethnic ancestries) communities experienced lower levels of social isolation and faced higher risks of hospitalisation and death.

While non-pharmaceutical interventions were implemented in São Paulo to slow down COVID-19 transmission, it was evident that the effectiveness of these interventions varied with socioeconomic conditions. Here, I developed a spatio-temporal Bayesian hierarchical model to understand the association between socioeconomic conditions and hospitalised deaths over time and across municipalities, while factoring in population movement between cities. It was found that the risk of hospitalisation and death was elevated for those living in high density low-income municipalities that experienced high levels of movement exchange with the RMSP. Stark differences in mortality was noted when factoring in access to public and private healthcare, where patients hospitalised in public health facilities were more likely to die than those in private health facilities. However within private health facilities, racial inequalities in death probability still persisted, mainly influenced by higher comorbidities and lower access to private care among Black patients.

While the findings from these two studies clearly illustrate the disproportionate risk of disease experienced by disadvantaged populations, it is worth noting that both studies used ecological regression which is limited in its ability to adjust for individual-level risk factors and associations, and thus is prone to ecological fallacy. Despite this inherent limitation, this analysis is still useful for making area-level conclusions that are useful for policy-making in public health and can assist with optimising resource allocation to achieve equitable access. The findings from these studies are particularly useful for guiding the prioritisation strategies for the COVID-19 vaccination campaign.

The 2021 COVID-19 vaccination campaign is arguably society's most effective intervention for reducing disease burden and death from the COVID-19 pandemic. Chapter 5 focused on understanding the inequalities in COVID-19 vaccination access between demographic groups and its impact on COVID-19 related deaths, which can help to provide insight on the effectiveness of the current vaccination campaign and identify recommendations for creating more inclusive interventions. This chapter addressed the final research question, focusing on understanding population response to epidemics through immunisation campaigns. Data on both COVID-19 hospitalisations and vaccination from the first vaccination campaign were collated alongside relevant sociodemographic information to understand the distribution of vaccination coverage of one or more doses by area, age, sex, race, and income in Brazil. Generalized linear regression analyses were performed to determine the risk of death and hospitalisation by age group, socioeconomic status, and vaccination coverage.

Given the rapid circulation of the P.1. gamma variant in the state of Amazonas in early 2021, the Brazilian government prioritised vaccination of the first dose in these areas. However, it was observed that later on, overall coverage was much lower compared to the wealthier regions of the South and Southeast which encompasses São Paulo state. Even though higher overall coverage was observed in the South, there was high disparity between municipalities in the region. Men, people of colour, and low-income groups were more likely to be partially vaccinated as these populations were more likely to miss the second dose. This could be partially due to higher hesitancy towards the COVID-19 vaccine among men and lower income communities in Brazil (6). Increasing vaccination coverage was shown to reduce the risk of death, especially for older age groups. Areas with higher vaccination were also observed to be associated with higher hospitalisation, suggesting vaccination was evident in areas with higher transmission and better access to healthcare. Only the

vaccination coverage of the municipality was considered in the model due to substantial missing data on individual-level vaccination status, which was a main drawback of this work.

This study has shown that despite the Brazilian government efforts to increase vaccination by defining priority groups and including disadvantaged populations in the first COVID-19 vaccination campaign, disparities exist in vaccination coverage among different age groups, sex, socioeconomic status and across regions. These findings along with observations made on the demographic groups that were more willing to miss the second dose should be considered when designing awareness-raising campaigns to increase vaccination uptake. This information can be useful as the third or “booster” dose campaign takes off in Brazil amid the increasing concern of the Delta variant transmission.

It is evident from the case studies presented in this thesis that human-environment interactions play a key role in shaping geographical inequalities in infectious disease risk in Brazil. Many of these interactions are influenced by population interconnectivity through the spatial distribution of settlements, movement, and behaviour. While the relationships of these interactions can be discerned using the frameworks of human disease ecology and landscape epidemiology, complexity exists in the human population response to infectious diseases. As illustrated in this thesis, some of this complexity could be detected using a combination of geospatial mapping, machine learning and statistical modelling, and open access large datasets (e.g. environmental, social, and health). As open-access and real-time data becomes more ubiquitous in the near future, more accurate patterns on how humans drive the emergence and spread of infectious diseases can be detected in the era of global pandemics.

6.2 Directions for Future Research

Unprecedented open data sharing of case information during the COVID-19 pandemic in real-time has facilitated timely mathematical predictions and modelling of disease outbreak and spread, enabling the design of suitable interventions and government policies. However, many of these datasets are prone to accuracy issues which is a key limitation. For instance, notification delay in case information and vaccination status may hinder the accurate modelling of disease risk given that some regions may experience higher under notification than others. Additionally, fine level socioeconomical data, which are mainly retrieved from the government censuses, could contain indicators that are out of date due to the infrequency of conducting mass population surveys which are expensive. One solution to address this could be the development of additional data sets using data sourced from healthcare registrations, and perhaps even from smartphones and smart household gadgets coupled with social media accounts. While the potential of adopting open large datasets for modelling infectious disease risk is high in the future, these issues should be further explored, especially when considering the possibility of modelling disease spread in real time.

Reflecting on the studies conducted on Zika virus, yellow fever virus, and COVID-19 in this thesis, it can be acknowledged that an understanding of population interconnectedness at high resolution can help to identify population groups and areas at risk, and enable the prediction of future disease emergence. Thus, improving the integration of near-real time human mobility data with high resolution environmental, socio-economical, and health data at both the individual- and population-level in spatial modelling could elucidate the complex patterns of human interactions. This could help to detect the co-emergence of new and existing infectious diseases. For countries like Brazil which suffers from the coexistence of

multiple ongoing epidemics throughout the year, further improving data integration can help elevate epidemic preparedness and the protection of public healthcare.

This thesis illustrates the importance of human-environment interactions which influence disease risk at each stage of the virus's life trajectory – starting from its "emergence", "spread" to instigating a "response" from human populations. Evident geographical inequalities in disease risk can be explained by physical environmental factors, alongside social determinants of health, including population connectivity and behaviour. Findings and conclusions from the research presented here have been used towards developing policy reports to improve public health and human development in Brazil. The strong policy implications of this work can be used to support the goal of achieving equitable access to good health for all.

6.3 References

1. de Oliveira Figueiredo P, Stoffella-Dutra AG, Barbosa Costa G, Silva de Oliveira J, Dourado Amaral C, Duarte Santos J, et al. Re-Emergence of Yellow Fever in Brazil during 2016–2019: Challenges, Lessons Learned, and Perspectives. Viruses [Internet]. 2020 Oct 30 [cited 2021 Mar 14];12(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7692154/
2. Laan MJ van der, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology [Internet]. 2007 Sep 16 [cited 2021 Jul 14];6(1). Available from: https://www.degruyter.com/document/doi/10.2202/1544-6115.1309/html
3. Bhatt S, Cameron E, Flaxman SR, Weiss DJ, Smith DL, Gething PW. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. Journal of The Royal Society Interface. 2017 Sep 30;14(134):20170520.
4. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; 2009. 745 p.
5. Marmot M, Wilkinson R. Social Determinants of Health. OUP Oxford; 2005. 501 p.
6. Moore DCBC, Nehab MF, Camacho KG, Reis AT, Junqueira-Marinho M de F, Abramov DM, et al. Low COVID-19 vaccine hesitancy in Brazil. Vaccine. 2021 Oct 8;39(42):6262–8.

Appendices

Statement of Contributions

This thesis comprises of six chapters. With the exception of the introductory and concluding chapters, each chapter will comprise of one or more research articles that have been published in peer-reviewed journals, in submission, or will be submitted. Some articles are outputs from a collaborative effort with members from the UK-Brazil Centre for Arbovirus Discovery, Diagnosis, Genomics, and Epidemiology (CADDE) and other international collaborators. While full author contributions are provided at the beginning and end each article, I have also outlined my own contributions below.

Chapter 2 – Literature Review

This chapter consists of one literature review paper which I first-authored. I oversaw all aspects of this paper, including the conception and design of the study (with JPM, MUGK, and LG), collected and summarised the literature, wrote and revised the final manuscript. I prepared the paper for submission and was responsible for the revision process.

> **Li, S.L.,** Messina, J.P., Pybus, O.G., Kraemer, M.U.G., Gardner, L. (2021). A review of models applied to the geographic spread of Zika virus. Transactions of the Royal Society of Tropical Medicine and Hygiene, DOI: https://doi.org/10.1093/trstmh/trab009

Chapter 3 – The physical environment and yellow fever emergence

This chapter consists of one research article that I first-authored. I oversaw all aspects of this paper. I conceived and designed the study (with JPM), collected the data, conducted the modelling analysis, interpreted the results (with JPM), wrote, and revised the final manuscript. I prepared the paper for submission and was responsible for the revision process.

Li, S. L. , Acosta A.L., Hill, S.C., Brady, O.J., … Messina, J.P., et al. (2021). Mapping environmental suitability of Haemagogus and Sabethes to understand sylvatic transmission risk of yellow fever virus in Brazil. *PLOS Neglected Tropical Diseases – Accepted.*

Chapter 4 – Differential risk to COVID-19 in the built environment

This chapter consists of two papers, where I served as a joint first author. In the first paper listed below, I was a significant contributor responsible for modelling, interpretation of results (with AEZ), and writing of the manuscript sections on geographical inequalities in COVID-19 infection. I also participated in the revision process. In the second paper listed, I conceived and designed the study (with RHP), conducted the modelling analysis, interpreted the results (with AEZ), wrote the first draft of the manuscript, which was revised and approved by all authors. I prepared the paper for submission and led the revision process.

De Souza, W. M.*, Buss, L. F.*, da Silva Candido, D.*, Carrera, J. P.*, **Li, S.*,** Zarebski, A., ... & Faria, N. R. (2020). Pidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nature Human Behaviour **4,** 856–865 (2020). DOI: https://doi.org/10.1038/s41562-020-0928-4
*joint first authors

Li, S. L.*, Pereira, R. H.*, Prete, C. A.*, Zarebski, A. E., Emanuel, L., Alves, P. J., ... & Messina, J. P. (2021). Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil. BMJ Global Health 2021;0:e004959. DOI: http://dx.doi.org/10.1136/bmjgh-2021-004959
*joint first authors

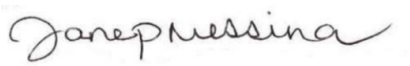
Chapter 5 – Population response to COVID-19 via mass immunisation

This chapter consists of one paper where I served as a joint first author. This paper has been submitted to a scientific journal to be considered for publication.

Li, S. L.*, Prete, C. A.*, Zarebski, A. E.*.. & Messina, J. P. (2021). Socio-demographic inequalities in COVID-19 vaccination in Brazil. *Submitted.*
*joint first authors

I conceived and designed the study (with CAP and AEZ), collected the data (with CAP), conducted the modelling analysis, interpreted the results (with CAP and AEZ), wrote the first draft of the manuscript, which was revised and approved by all authors. I prepared the paper for submission.

As Sabrina's primary supervisor, I hereby certify that the statement of contribution listed above is a fair representation of her work.



Dr. Janey P. Messina, November 5, 2021

Chapter 3 Supplementary Information

Mapping environmental suitability of *Haemagogus* and *Sabethes* to understand sylvatic transmission risk of yellow fever virus in Brazil

Supplementary Information

Sabrina L. Li [1#], André L. Acosta [2], Sarah C. Hill [3], Oliver J. Brady [4,5], Marco A. B. de Almeida [6], Jader da C. Cardoso [6], Arran Hamlet [7], Luis F. Mucci [8], Juliana Telles de Deus [8], Felipe C. M. Iani [9], Neil S. Alexander [10], G. R. William Wint [10], Oliver G. Pybus [3,11], Moritz U.G. Kraemer [11], Nuno R. Faria [7,11,12], Jane P. Messina [1,13#]

1. School of Geography and the Environment, University of Oxford, Oxford, United Kingdom.
2. Departamento de Ecologia, Instituto de Biociências, Laboratório de Ecologia de Paisagens e Conservação - LEPAC, Universidade de São Paulo, São Paulo, Brazil.
3. Department of Pathobiology and Population Sciences, Royal Veterinary College London, London, United Kingdom.
4. Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom.
5. Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom.
6. State Centre of Health Surveillance, Rio Grande do Sul State Health Secretariat, Rio Grande do Sul, Brazil.
7. MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom
8. Superintendence for Endemic Diseases Control, São Paulo State Health Secretariat, São Paulo, Brazil.
9. Fundação Ezequiel Dias - Lacen/MG, Minas Gerais, Brazil.
10. Environmental Research Group Oxford, c/o Department of Zoology, University of Oxford, Oxford, United Kingdom.
11. Department of Zoology, University of Oxford, Oxford, United Kingdom.
12. Departamento de Molestias Infecciosas e Parasitarias & Instituto de Medicina Tropical da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.
13. Oxford School of Global and Area Studies, University of Oxford, Oxford, United Kingdom.

#Corresponding authors:

1. Sabrina L. Li, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom. E-mail: lisabrinaly@gmail.com

2. Jane P. Messina. School of Geography and the Environment, University of Oxford, Oxford, United Kingdom. E-mail: jane.messina@gtc.ox.ac.uk

Table of Contents

1. Presence and pseudo-absence occurrence records 3

2. Covariate data 5

3. Uncertainty maps 7

4. Covariate importance 9

5. Model validation statistics 11

6. References 12

1. Presence and pseudo-absence occurrence records

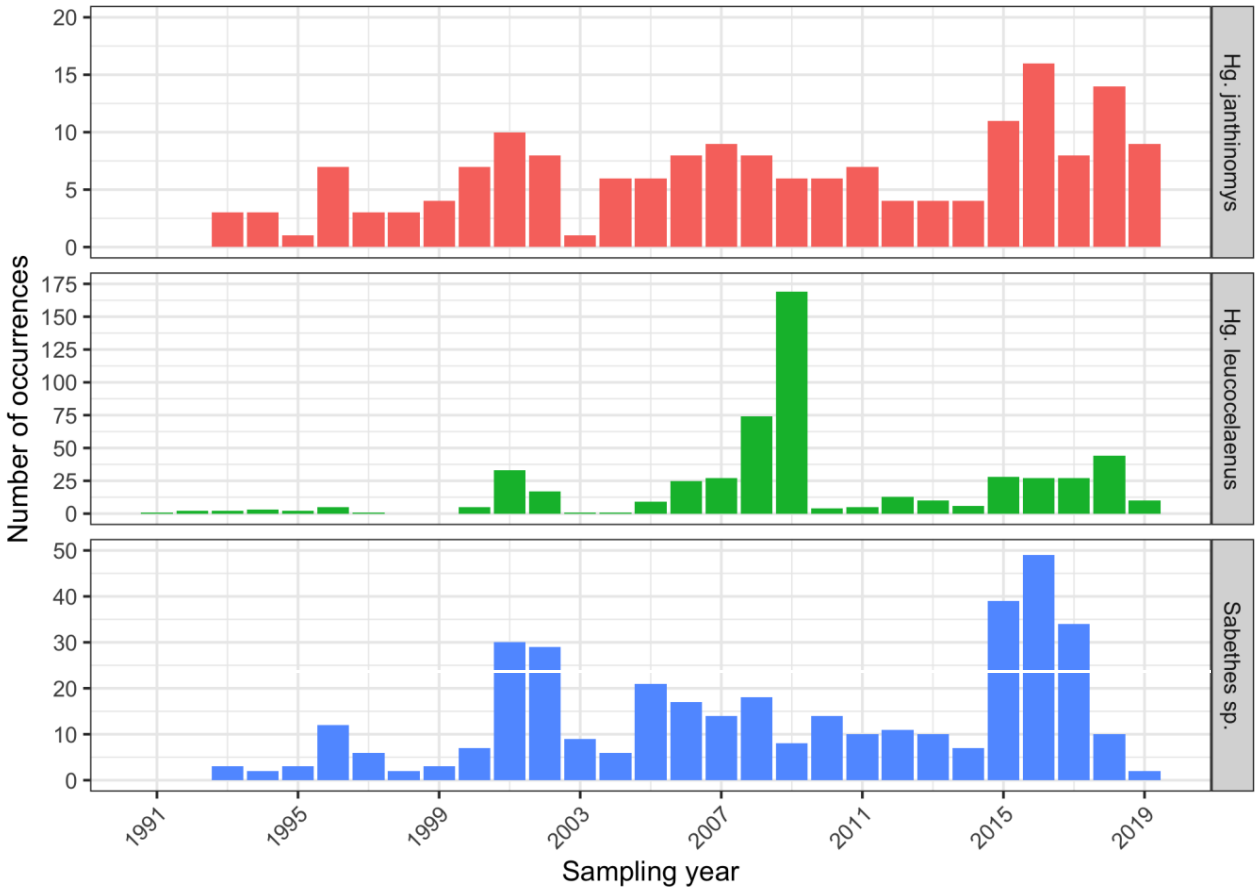


Fig S1. Temporal distribution of occurrence records for *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp.*

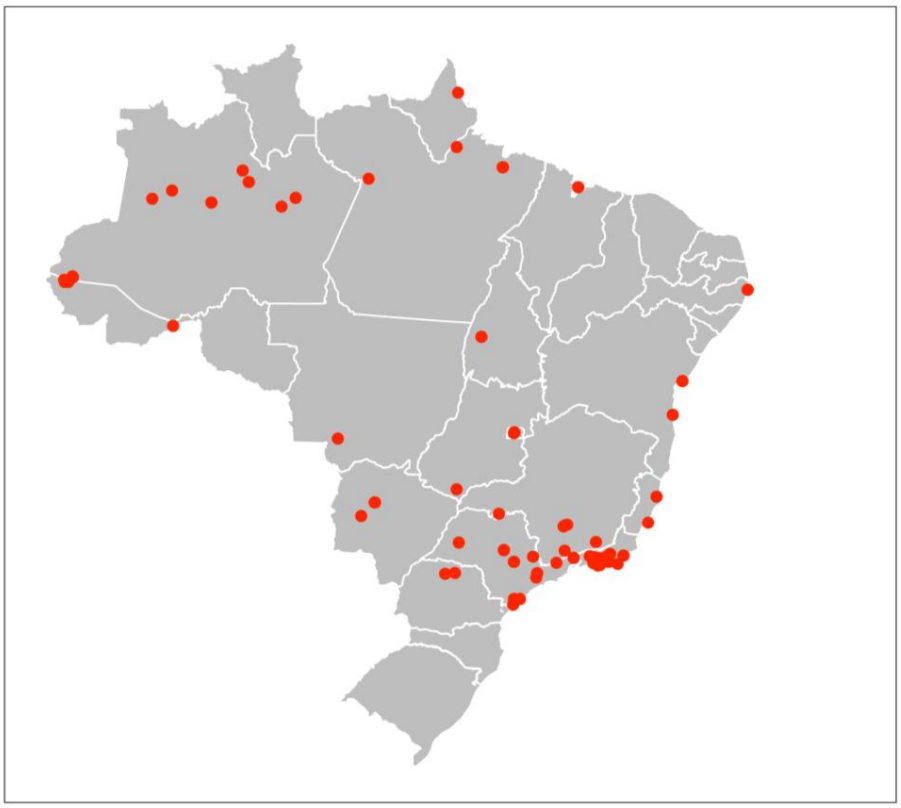


Fig S2. Distribution of pseudo-absence points.

2. Covariate data

Table S1. Metadata of environmental and land-cover covariates used in stacked binomial regression model

Name	Source	Units	Spatial Resolution (Spatial Range)	Temporal Range	Use
Land Surface Temperature	MODIS (1)	Kelvin*50	1 km	2001- 2019	3 variables; Calculate yearly min, mean, and maximum temperature over time range.
Precipitation	ERA5 (European Centre for Medium-Range Weather Forecasts) (2)	Mm/day	1 km	2001-2019	3 variables: Yearly total precipitation, precipitation of the driest month in the year, precipitation of the wettest month in the year averaged between 2001-2019.
Relative Humidity	Messina et al. 2019 (3)	%*10	1 km	2015	Monthly aggregated to yearly.
Elevation	Amatulli et al. 2018 (4)	m	1 km	Static	Minimum elevation in metres above sea level.
Gross Forest Cover Loss	Hansen et al. 2013 (5)	0 (no loss) or 1-17 (loss in year 2001-2019)	30 m	2000-2019	Forest loss during the period 2000–2019, defined as a stand-replacement disturbance, or a change from a forest to non-forest state.
EVI	MODIS (1)	-1 to 1	1 km	16 days (2000-2016)	Mean annual EVI averaged over time period
Population Density	WorldPop (6)	# of people/pixel	1 km	2019	Number of people per pixel, adjusted to match country total from the official UN 2019 population estimates, by the pixel surface area
Consensus Landcover	Tuanmu et al. 2018 (7)	Consensus prevalence in %	1 km	Static	1- Evergreen Broadleaf Trees 2- Deciduous Broadleaf Trees 3- Cultivated and Managed Vegetation

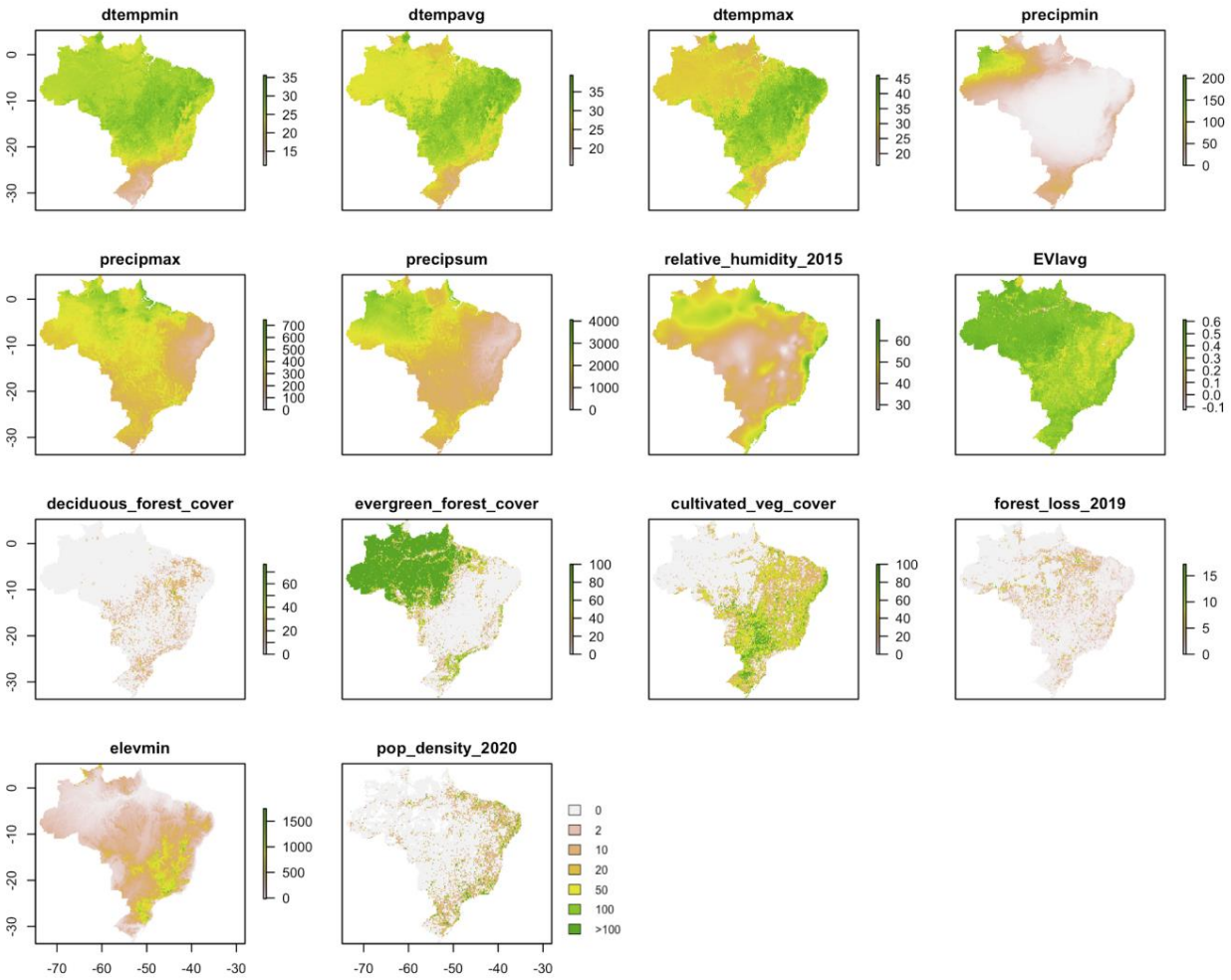


Fig S3. Maps illustrating the distribution of environmental covariate data.

3. Uncertainty maps

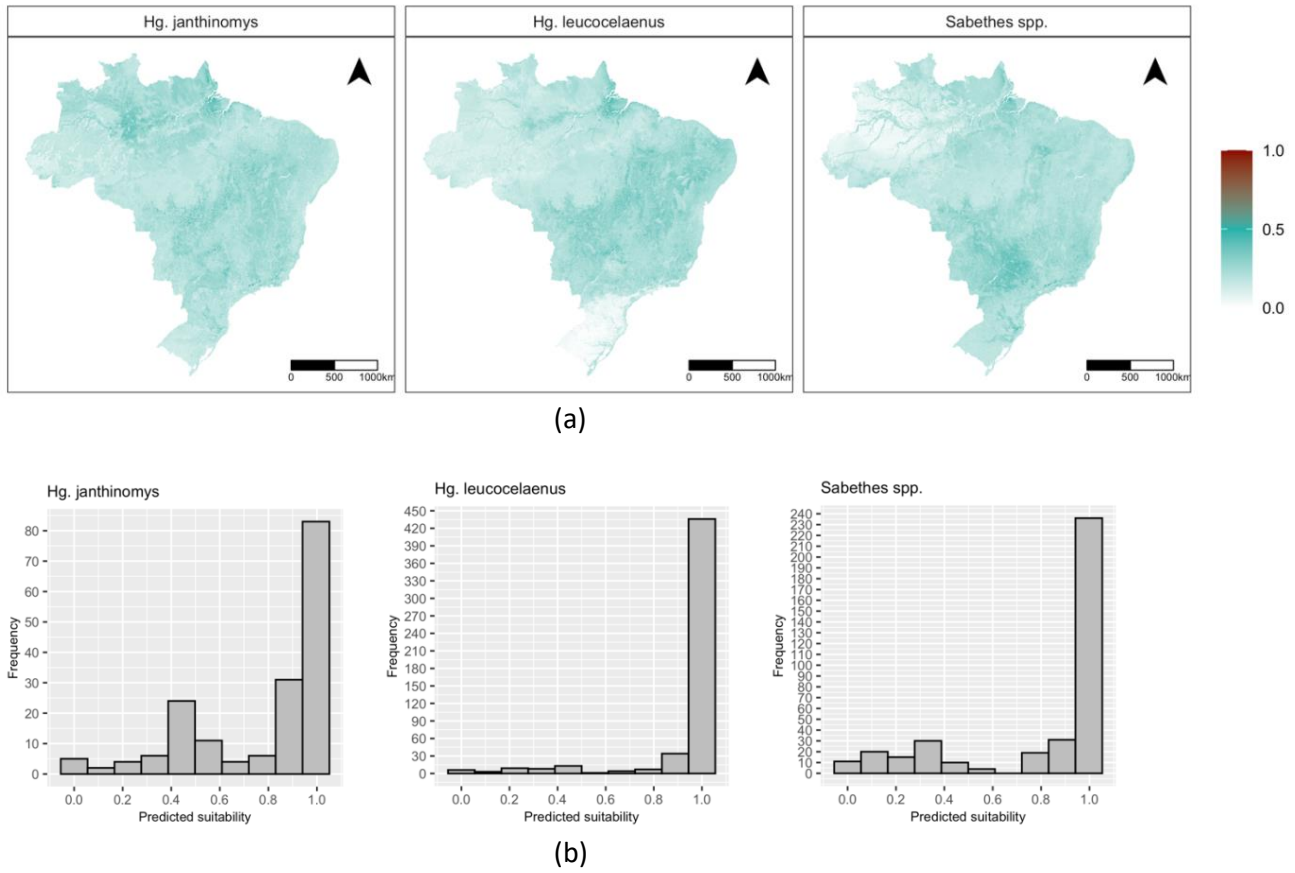


Fig S4. (a) Maps of prediction uncertainty calculated based on the 95% confidence interval and (b) frequency of presence occurrence points falling in each predicted range of suitability

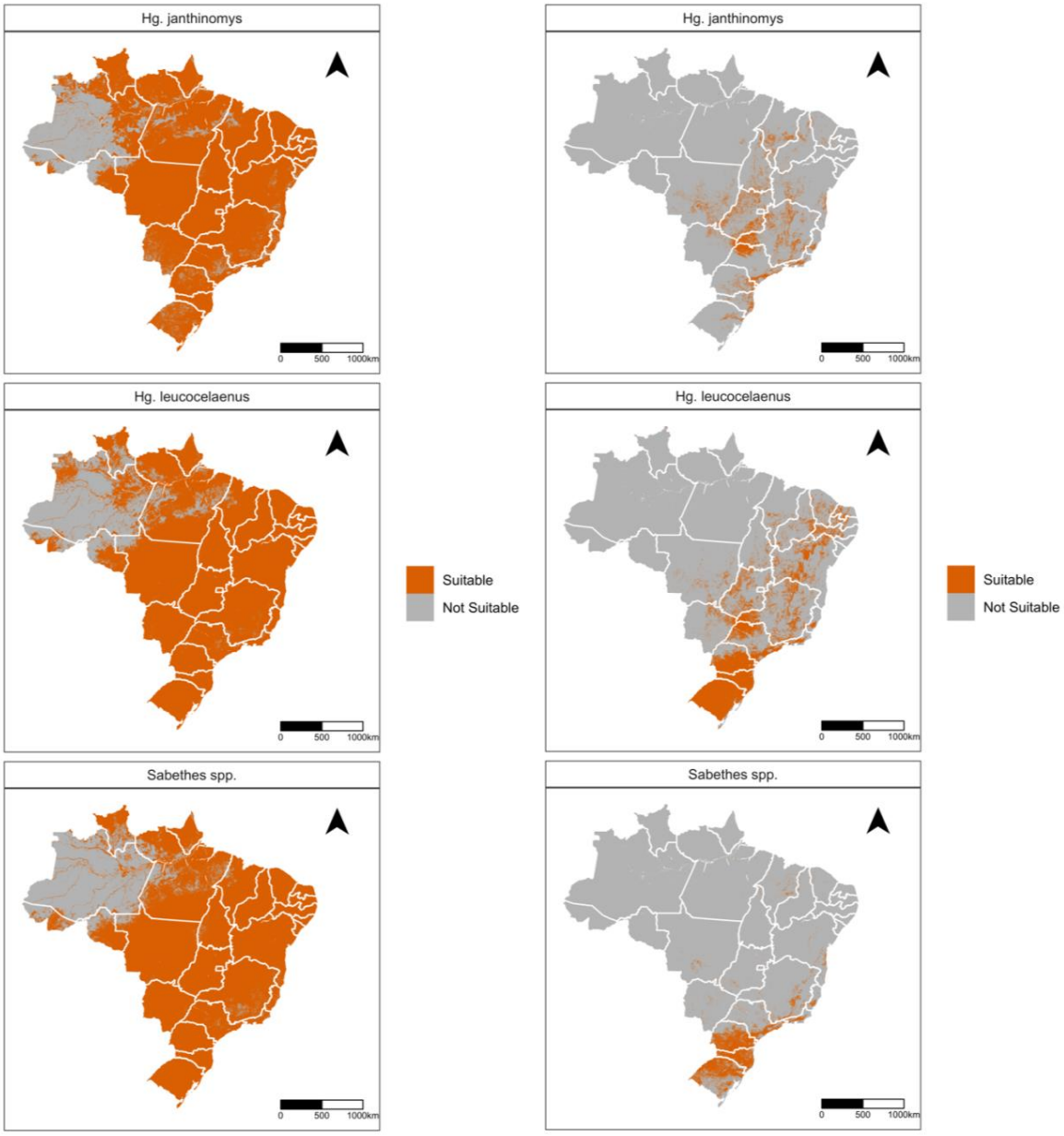


Fig S5. Maps showing suitable regions for *Haemagogus* and *Sabethes spp.* when environmental suitability threshold is set to (a) ≥ 0.25 and (b) ≥ 0.75.

4. Covariate importance

Show 25 entries Search:

	Variable	Importance	StDev	Effect
1	popden_worldpop	0.1055	0.0008	
2	landc03ext	0.1001	0.0039	
3	elev_min	0.0749	0.0012	
4	dtempmax	0.0705	0.0009	
5	precip_max_year	0.0703	0.0021	
6	landc02ext	0.0590	0.0015	
7	landc07ext	0.0580	0.0021	
8	evi_avg_br	0.0491	0.0006	
9	precip_min_year	0.0478	0.0003	
10	dtempavg	0.0438	0.0009	
11	precip_sum_year	0.0425	0.0009	
12	relative_humidity_2015	0.0379	0.0008	
13	forest_loss_2019	0.0278	0.0007	
14	dtempmin	0.0265	0.0004	

Showing 1 to 14 of 14 entries Previous 1 Next

(a)

Show 25 entries Search:

	Variable	Importance	StDev	Effect
1	elev_min	0.0968	0.0020	
2	dtempmin	0.0687	0.0025	
3	popden_worldpop	0.0661	0.0015	
4	evi_avg_br	0.0505	0.0033	
5	precip_sum_year	0.0477	0.0019	
6	precip_min_year	0.0419	0.0011	
7	landc03ext	0.0396	0.0017	
8	relative_humidity_2015	0.0366	0.0011	
9	dtempavg	0.0336	0.0017	
10	landc07ext	0.0335	0.0013	
11	precip_max_year	0.0320	0.0007	
12	forest_loss_2019	0.0238	0.0004	
13	dtempmax	0.0200	0.0008	
14	landc02ext	0.0170	0.0005	

Showing 1 to 14 of 14 entries Previous 1 Next

(b)

Show 25 entries Search:

	Variable	Importance	StDev	Effect
1	landc03ext	0.0911	0.0024	
2	popden_worldpop	0.0836	0.0015	
3	elev_min	0.0780	0.0020	
4	dtempmin	0.0733	0.0027	
5	landc02ext	0.0625	0.0020	
6	precip_min_year	0.0570	0.0007	
7	dtempmax	0.0432	0.0007	
8	precip_sum_year	0.0411	0.0017	
9	relative_humidity_2015	0.0403	0.0007	
10	precip_max_year	0.0400	0.0013	
11	dtempavg	0.0351	0.0020	
12	evi_avg_br	0.0350	0.0025	
13	forest_loss_2019	0.0216	0.0012	
14	landc07ext	0.0121	0.0004	

Showing 1 to 14 of 14 entries Previous 1 Next

(c)

Fig S6. Covariate importance (ranked in descending order) and partial dependence plots for (a) *Hg. janthinomys*, (b) *Hg. leucocelaenus*, and (c) *Sabethes spp.* Land-use variables: landc02ext – evergreen forest cover, landc03ext – deciduous forest cover, landc07ext – cultivated and managed vegetation cover. Interactive version available here*:* https://github.com/sabrinalyli/YFVMappingBrazil

5. Model validation statistics

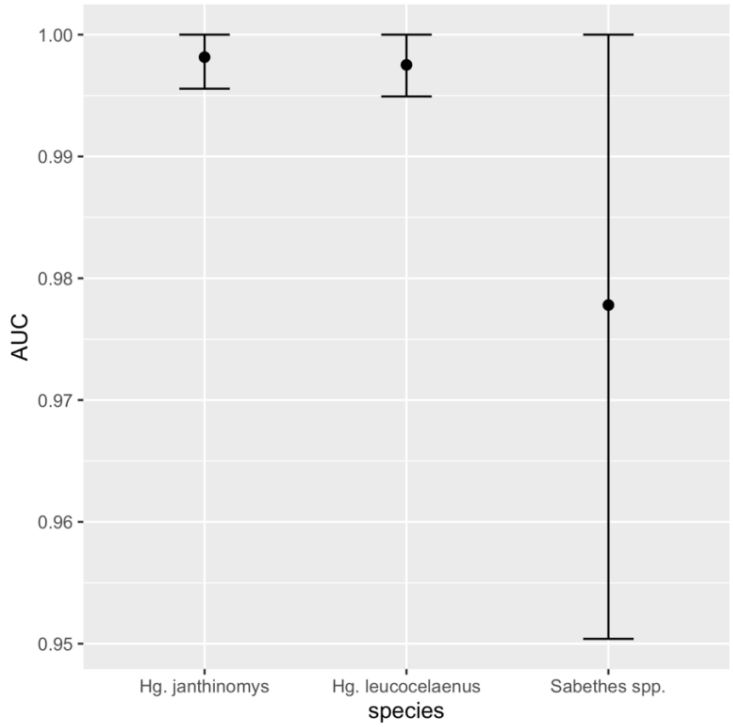


Fig S7. AUC and 95% confidence interval generated for the holdout set for *Hg. janthinomys*, *Hg. leucocelaenus*, and *Sabethes spp.*

6. References

1. Wan, Z., Hook, S., Hulley, G. MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. [Internet]. NASA EOSDIS Land Processes DAAC.; 2015. Available from: https://doi.org/10.5067/MODIS/MOD11A2.006

2. Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. [Internet]. Copernicus Climate Change Service Climate Data Store (CDS); [cited 2020 Jan 27]. Available from: https://cds.climate.copernicus.eu/cdsapp#!/home

3. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. Nat Microbiol. 2019 Sep;4(9):1508–15.

4. Amatulli G, Domisch S, Tuanmu M-N, Parmentier B, Ranipeta A, Malczyk J, et al. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci Data. 2018 Mar 20;5(1):180040.

5. Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science. 2013 Nov 15;342(6160):850–3.

6. Lloyd CT, Chamberlain H, Kerr D, Yetman G, Pistolesi L, Stevens FR, et al. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. Big Earth Data. 2019 Apr 3;3(2):108–39.

7. Tuanmu M-N, Jetz W. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Glob Ecol Biogeogr. 2014;23(9):1031–45.

Chapter 4.1 Supplementary Information

Supplementary information

Geospatial analysis

We adopted a Bayesian hierarchical model to compute relative risk for each census tract, due to the following reasons: (i) there is a large number of census tracts (n=30,815), (ii) there is substantial heterogeneity in the size of census tracts, and (iii) small counts in each tract obscure the spatial distribution of observed cases. The number of observed cases in census tract i is modelled using a Poisson distribution $Y_i = Poisson(\lambda_i)$ with mean $\lambda_i = E_i\,\mu_i$ where E_i is the expected number of cases under a null model in which cases are uniformly distributed among the population. For example, the total number of cases in the MRSP multiplied by the proportion of the population in the census tract $E_{it} = \frac{\sum_i Y_i}{\sum_i pop_i} \times pop_i$. The factor of μ_i describes tract specific risk and models the additional variation in the observation process[1]. A log-linear model is used to estimate the relative risk μ_i. For example, the log relative risk is expressed as a sum of an intercept α, which represents the overall relative risk (in our case, the global relative risk is zero), and random effects (Z_i):

$$\log(\mu_i) = \alpha + Z_i$$

We used a Besag-York-Mollié model (BYM)[2] to separate the random effects into a spatially structured U_i, and independent random effects, V_i, so ($Z_i = U_i + V_i$). In the BYM model, a conditional autoregressive (CAR) process is used to introduce correlation among the U_i for each tract. Given the U_i of neighbouring tracts, the U_i has a normal distribution with mean equal to the average of the neighbours' U_i, and variance $s_i^2 = \frac{1}{\#N(i)\tau_U}$ where $\#N(i)$ is the number of tracts that share boundaries with tract i and τ_U is a precision parameter. The random effect, V_i follows a zero mean normal distribution with unknown precision, $\tau_V = \frac{1}{\sigma_{v^2}}$ (where σ_{V^2} is the variance). Both random effects in the model capture extra-Poisson variability, and were expressed as the following:

$$U_i | U_{j\neq i} \sim Normal\left(m_i, s_i^2\right), \quad V_i \sim N(0, \sigma_v^2)$$

$$m_i = \frac{\sum_{j\in N(i)} U_j}{\#N(i)} \quad , \quad s_i^2 = \frac{\sigma_U^2}{\#N(i)} = \frac{1}{\#N(i)\tau_U}$$

The log of the precision parameters, τ_U and τ_V, follows a gamma distribution with shape 1 and rate 0.0005. These are the default priors used by R-INLA and are minimally informative[3]. The prior default distributions in R-INLA were used for the precision parameters of both U_i and V_i. These are minimally informative and are the recommended settings [4].

To quantify the uncertainty in the point estimates of the mean relative risk estimates, we mapped the posterior probability of elevated relative risk in each census tract (**Figure 4.1.1**). This is the posterior probability, which a tract has an elevated risk of observing cases, formally, this is Prob(μ_i > 1| data). For instance, a probability of 0.6 in a census tract indicates a 60% chance that this census tract is at greater risk of observing cases relative to the rest of the MRSP.

Analysis of the relationship between income per capita and final diagnostic category in the Metropolitan Region of Sao Paulo (MRSP)

We evaluated the relationship between final diagnostic category (COVID-19 or SARI cases with unknown aetiology) and socioeconomic status in the subset of cases in the MRSP with geocoded residential information. We focused on the cases in epidemiological weeks 12, 16 and 22, where the census tracts that reported cases varied across weeks. In each of the three weeks, if a census tract reported any COVID-19 or SARI cases with unknown aetiology with unknown aetiology, we calculated the proportion of the number of COVID-19 cases. Since most census tracts reported only one case each week, the proportion of COVID-19 of each census tract were mostly either 0 or 1 in a given week. Based on this observation and let i index the census tracts, we subsequently defined the binary outcome Y_i of census tract i, where (i) $Y_i = 0$ if census tract i only reported SARI cases with unknown aetiology with unknown aetiology, i.e. no COVID-19 cases, (ii) $Y_i = 1$ if census tract i reported at least one COVID-19 case in the week. Logistic regression models were applied to investigate the association between this binary outcome and the log(X+1) transformed income per capita. The analyses were adjusted by the logarithm of the population sizes. In addition, the census tracts were grouped by their geographic locations using cluster analysis, and the groupings were used as the random effect in the logistic regressions to account for potential spatial autocorrelation. The number of clusters was chosen based on the AIC/BIC values of the logistic regression models. The analysis was performed individually for each of epidemiological weeks 12, 16 and 22.

A likelihood ratio test (LRT) is applied to each analysis to examine whether the log(X+1) transformed income per capita provides information in addition to the information from the log population size and the random effects. The regression coefficients and LRT P-values of income are presented in (**Supplementary Table S1**).

Supplemental Table S1

	Logistic regression	
	Coefficient estimate (95% CI)	**LRT**
Week 12	0.678 (0.534, 0.826)	< 0.001
Week 16	-0.058 (-0.178, 0.062)	0.343
Week 22	0.467 (0.236, 0.693)	< 0.001

Notes:

LRT = likelihood-ratio test.

CI = Confidence Interval

References

1 Lawson, A. B. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiolog*. (2008).

2 Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-20, doi:10.1007/BF00116466 (1991).

3 Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 319-392, doi:10.1111/j.1467-9868.2008.00700.x (2009).

4 Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology* **7**, 39-55, doi:10.1016/j.sste.2013.07.003 (2013).

Chapter 4.2 Supplementary Information

Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil

Supplementary Materials

Sabrina L Li, MSc [1*#], Rafael H M Pereira, DPhil [2*#], Carlos A Prete Jr, MSc [3*], Alexander E Zarebski, PhD [4], Lucas Emanuel, PhD [2], Pedro JH Alves, MSc [2], Pedro S Peixoto [5], PhD, Carlos K Braga, MSc [2], Andreza Aruska de Souza Santos, PhD [6], William M de Souza, PhD [4,7], Rogerio J Barbosa, PhD [8], Lewis F Buss, MD [9], Alfredo Mendrone-Junior, MD [10], Cesar de Almeida-Neto, MD [10,11], Suzete C Ferreira, PhD [10,12], Nanci A Salles, BSc [10,12], Izabel Marcilio, MD [13], Chieh-Hsi Wu, PhD [14], Nelson Gouveia, MD [15], Vitor Heloiz Nascimento, PhD [3], Ester C Sabino, MD [9], Nuno R Faria, PhD [4,9,16], Jane P Messina, PhD [1,17]

1. School of Geography and the Environment, University of Oxford, Oxford, United Kingdom.
2. Institute for Applied Economic Research, Brasília, Brazil.
3. Department of Electronic Systems Engineering, University of São Paulo, São Paulo, Brazil.
4. Department of Zoology, University of Oxford, Oxford, United Kingdom.
5. Department of Applied Mathematics, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil.
6. Oxford School of Global and Area Studies, Latin American Centre, University of Oxford, Oxford, United Kingdom.
7. Virology Research Center, University of São Paulo, Ribeirão Preto, Brazil.
8. Center for Metropolitan Studies, Faculty of Philosophy, Languages and Literature, and Human Sciences, University of São Paulo, São Paulo, Brazil.
9. Departamento de Molestias Infecciosas e Parasitarias & Instituto de Medicina Tropical da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.
10. Fundação Pró-Sangue Hemocentro de São Paulo
11. Disciplina de Ciências Médicas, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.
12. Laboratory of Medical Investigation in Pathogenesis and Directed Therapy in Onco-Immuno – Hematology (LIM-31) HCFMUSP, University of São Paulo Medical School, São Paulo, Brazil.
13. Epidemiologic Surveillance Center, Hospital das Clinicas – University of São Paulo Medical School, São Paulo, Brazil.
14. Mathematical Sciences, University of Southampton, Southampton, United Kingdom.
15. Department of Preventive Medicine, University of São Paulo Medical School

16. MRC Centre for Global Infectious Disease Analysis, J-IDEA, Imperial College London, London, UK.
17. Oxford School of Global and Area Studies, University of Oxford, Oxford, United Kingdom.

*These authors contributed equally.

#Corresponding authors:

1. Sabrina L. Li, School of Geography and the Environment, University of Oxford, South Parks Rd, Oxford, OX1 3QY, United Kingdom. E-mail: sabrina.li@ouce.ox.ac.uk

2. Rafael H.M. Pereira, Institute for Applied Economic Research (Ipea), SBS - Quadra 1 - Bloco J - Ed. BNDES - CEP: 70076-900 - Brasília - DF – Brasil. E-mail: rafael.pereira@ipea.gov.br

Table of Contents

Appendix A: Methodology details....4
1 Probability of hospitalisation and death....4
1.1 Probability of death....4
1.2 Probability of hospitalisation by SARI....5
1.3 Confidence intervals....5
2 Geospatial analysis....6
2.1 Model description....6
2.2 Modelling relative risk....6
2.3 Ecological regression....7
2.4 Computational method....7
2.5 Model evaluation....8
2.6 Covariates....9
2.6.1 2010 census....9
2.6.2 Distance to the nearest health facility....10
2.7 Spatial joins for the GMSP....10
3 Probability of working conditions....11
4 Event-study analysis....12
4.1 Summary statistics....13
4.2 Model sensitivity analysis....14
5 Crosswalk: PNAD COVID-19 Occupations to ISCO-08 (1-digit)....14
6 Classification of non-classified SARI cases as COVID-19 cases....15
7 References....16
Appendix B: Figures....18

Table of Figures

Figure S 1....18
Figure S 2....19
Figure S 3....20
Figure S 4....21
Figure S 5....22
Figure S 6....23
Figure S 7....24

Appendix A: Methodology details

1 Probability of hospitalisation and death

Probability of hospitalisation and death due to COVID-19 were calculated by income quantile, race or administrative type of the health facility (public or private). The same method was used to calculate the estimates for the different grouping criteria (i.e., income quantile, race, administrative type of the health facility). We use the term 'class' to refer to a specific subgroup for each of these criteria, i.e., the higher income class, in the case of analysis by income, or the public health facility class in case of analysis by type of health facility. It is worth noting that only hospitalisations or deaths contained in the SIVEP-Gripe dataset are considered SARI cases. We also discarded cases that were confirmed to be the results of other etiological agents.

The health facility where each patient was hospitalized is defined for all SARI patients in the SIVEP-Gripe dataset. Administrative types of health facilities were extracted from the National Registry of Health Facilities (CNES) database. We classified as "Public" all health facilities whose administrative type is "SUS" (Brazilian public health system), as "Private" all private health facilities, regardless of whether they accept patients with health insurance or not, and as "Other" all non-private health facilities whose administrative type is not marked as "SUS". In most cases, these are public health facilities that accept patients with health insurance. Health facilities with missing administrative type were discarded. 26·7% of all SARI patients were notified by health facilities classified as public, compared to 43·6% for private, 5.4% for "Other" and 21·7% for health facilities with missing administrative type.

Missing races are imputed based on the racial distribution of the corresponding census tracts. It is important to impute missing races instead of excluding them to avoid racial bias because different races may have different probabilities of being missing. Figures S1 shows SARI hospitalisation and death probabilities computed without race imputation. Even though the results with race imputation are more robust, similar probabilities of death and hospitalization are obtained if race imputation is not employed.

1.1 Probability of death

Let w_{ik} be the probability of the patient i belonging to a given class k in the grouping criterion of interest (income, race or administrative type of the health facility). When comparing private and public hospitals, w_{ik} can only be 1 or 0 because the health facility type is known for all patients. Similarly, w_{ik} can only be 1 or 0 for individuals whose races are not missing.

In the case of incomplete information, that is, individuals with unknown race and for the income comparisons, we assign a distribution over the possible values using estimates obtained from census data for each census tract. The probabilities w_{ik} are assigned as the proportion of the individuals in the corresponding census tract that belong to class k (that is, w_{ik} is given by the racial or income distribution of the census tract where the individual lives). In order to compute the probabilities of death and SARI for a given class, the weights w_{ik} are used to divide the contribution of each individual i between all classes k for $k = 1, \dots, K$, where K is the number of classes. Individuals whose census tract is not known were excluded from the income analysis, but only individuals with both unknown race and census tract were excluded from the racial analysis.

Let us denote the probability of an event A as $P(A)$. The probability of death given class, age, and sex $P(\text{death}|\text{class}, \text{age}, \text{sex})$ was estimated by dividing the number of SARI deaths for a given class and age-sex group by the number of SARI cases for the same class, age, and sex group. The probability of death given class (among SARI patients) can be obtained through

$$P(\text{death}|\text{class}) = \sum_{\text{age,sex}} P(\text{death}|\text{class}, \text{age}, \text{sex}) P(\text{age}, \text{sex}|\text{class}),$$

where $P(\text{age}, \text{sex}|\text{class})$ is the conditional age-sex distribution of SARI patients given their class. This distribution may differ significantly for different classes due to differences in the age composition of classes, especially when race is used as class: The proportion of residents of the state of São Paulo that are older than 70 years old is 6·1% for White, 2·9% for *Pardo*, 4·4% for Black and 11·4% for Asians according to the 2010 census. Since the probability of death depends heavily on age, in order to allow a fairer comparison of the burden of the disease according to income and race, we use an age-sex normalised death probability given class, which is the death probability given class that would be obtained if all classes had the same age-sex distribution. If this age-sex normalisation was not applied, the probability of death would be overestimated for white patients and underestimated for black patients, as the white population is older than the black population. This normalised distribution is obtained by substituting $P(\text{age}, \text{sex}|\text{class})$ by $P(\text{age}, \text{sex})$, the proportion of patients from an age-sex group regardless of the class:

$$P_{\text{normalised}}(\text{death}|\text{class}) = \sum_{\text{age,sex}} P(\text{death}|\text{class}, \text{age}, \text{sex}) P(\text{age}, \text{sex})$$

This age-sex standardisation was not employed for the health facility type because patients from public and private hospitals share similar age-sex distributions. The age of the patients was discretized into the following groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+. It is worth noting that weighting estimates that depend on age to produce an age-standardised estimate was also done in other similar studies [1].

1.2 Probability of hospitalisation by SARI

The probability of hospitalisation by severe acute respiratory infection (SARI) given each class for each criterion (only income or race in this analysis) is computed for each age-sex group as the ratio between the number of recorded SARI cases for a given class, age and sex and the number of individuals of that class and age-sex group. The age-sex distributions of for population of São Paulo was extracted from the 2010 census. The same age-sex standardisation employed to compute the probabilities of death was used to estimate the probabilities of SARI, but as in the 2010 census, patients above 70 years old were aggregated into the same age group. Only the race and income quantile classes are used in this analysis.

1.3 Confidence intervals

Confidence intervals were estimated through bootstrapping using 1 000 realizations. For each realization, the same number of individuals in the dataset were randomly selected with replacement from the set of individuals in the dataset. Then, the desired estimates and odds ratios are obtained from each resampled dataset. For the racial analysis each selected individual i with unknown race had its race

imputed for each bootstrap iteration such that the probability of the imputed race being k is w_{ik}. The confidence intervals were obtained from the quantiles of the set of 1 000 bootstrapped parameters.

2 Geospatial analysis

2.1 Model description

We used a Bayesian hierarchical model to compute the relative risk of hospitalisation at the municipality level for São Paulo state (n= 645) and at the census tract level for the greater metropolitan area of São Paulo (GMSP: n=30 815). The number of observed cases Y_i in an area i is modelled using a Poisson distribution $Y_i \sim Poisson(\lambda_i)$ with mean $\lambda_i = E_i\, \mu_i$ where E_i is the expected number of cases in area i under a null model in which cases are uniformly distributed among the population, i.e., the number of cases in a given area is proportional to the population of that area. For each area i, this is given by $E_i = \frac{\sum_i Y_i}{\sum_i pop_i} \times\ pop_i$, where pop_i is the population in area i. The factor of μ_i describes the area-specific risk and models the additional variation in the observation process [2].

To quantify the uncertainty in the point estimates of the mean relative risk estimates, we mapped the posterior probability of elevated relative risk in each area (Appendix B - Figure S7). This is the posterior probability that a tract has an elevated risk of observing cases, formally P(μ_i > 1| data). For instance, a posterior probability of 0.6 in an area indicates a 60% chance that this area is at greater risk of observing cases.

2.2 Modelling relative risk

We fit a log-linear model to estimate the relative risk μ_i, which is modelled as the sum of an intercept and random effects. Random effects are broken into the spatial (A_i) and temporal components (B_i), as shown in Eq. (1·1):

$$\log(\mu_i) = \alpha + A_i + B_i \quad (1{\cdot}1)$$

$$\log(\mu_i) = \alpha + U_i + V_i + \gamma_t + \phi_t \quad (1{\cdot}2)$$

To account for existing spatial autocorrelation, we used a Besag-York-Mollié model (BYM) [3] to separate the spatial component into spatially structured U_i, and non-spatial, unstructured random effects, V_i, so ($A_i = U_i + V_i$), as shown in Eq. 1·2. In the BYM model, a conditional autoregressive (CAR) process is used to introduce correlation among the U_i for each tract. Given the U_i of neighbouring areas, the U_i has a normal distribution with mean equal to the average of the neighbours' U_i, and variance $s_i^2 = \frac{1}{\#N(i)\tau_U}$ where $\#N(i)$ is the number of areas that share boundaries with area i and τ_U is a precision parameter. The random effect, V_i follows a zero mean normal distribution with precision parameter, τ_V=$\frac{1}{\sigma_{v^2}}$ (where σ_{V^2} is the variance). Both random effects in the model capture extra-Poisson variability, and were expressed as the following:

$$U_i|\ U_{j\neq i} \sim \mathcal{N}\left(m_i, s_i^2\right), \quad V_i \sim \mathcal{N}(0, \sigma_v^2)$$

$$m_i = \frac{\sum_{j \in N(i)} U_j}{\#N(i)} \quad , \quad s_i^2 = \frac{\sigma_U^2}{\#N(i)} = \frac{1}{\#N(i)\tau_U}$$

To account for temporal structure in the data, we included the random effect ($B_t = \gamma_t + \phi_t$), which assumes that the number of cases observed in a given area depends on the number of cases observed in the given area in the previous month and a residual [2,4]. The temporal component includes γ_t, a temporally structured effect modelled dynamically using a random walk of order 1, and an unstructured temporal effect ϕ_t to account for independent time effects, which follows a zero mean normal distribution. Both are expressed as the following:

$$\gamma_t | \gamma_{t-1} \sim \mathcal{N}\left(\gamma_{t-1}, \frac{1}{\tau_\gamma}\right)$$
$$\phi_t \sim \mathcal{N}\left(0, \frac{1}{\tau_\phi}\right)$$

We adopted minimally informative prior distributions in R-INLA [3]. The log of the precision parameters adopted for the spatial effects, τ_U and τ_V, follows a gamma distribution with shape 1 and rate 0·0005. The precision parameter for both the structured and unstructured temporal effects τ_γ and τ_ϕ also follows a gamma distribution, with shape 1 and rate 0·001. The prior default distributions in R-INLA, which are the recommended settings [4], were also used for the precision parameters of both U_i, V_i, γ_t, and ϕ_t.

2.3 Ecological regression

To evaluate the effects of socioeconomic covariates on the risk of hospitalisation at the municipality level, we reformulated our model expressed by Eq. 1·3 by adding a fixed effect, which we refer as the ecological regression model. The log of the relative risk is given by the following:

$$\log(\mu_i) = \alpha + U_i + V_i + \gamma_t + \phi_t + X'_{ik}\,\beta \quad (1{\cdot}3)$$

where X'_{ik} is the i th row and kth column of covariates matrix X based on the socioeconomic covariates for each municipality and month, and β is the regression parameter modelled as fixed effects with normal priors ($\beta \sim \mathcal{N}(0, 100)$) .

Here, we define U_i using a graphical structure in R-INLA which describes the connections between municipalities by looking at estimates of the level of human mobility between them in the state of São Paulo. We considered two municipalities to be connected if there were at least 550 journeys between them. This threshold was selected to ensure that the sparsity of the connectivity matrix was similar to the nearest neighbour matrix described in Section 3·2. The number of origin-destination journeys between municipalities are retrieved from processed mobile geo-location data obtained from the In Loco company described elsewhere [5]. The prior default distributions in R-INLA were used for the precision parameters of both U_i, V_i, γ_t, and ϕ_t.

2.4 Computational method

We carried out model fitting with R-INLA which uses an Iterated Nested Laplace Approximation (INLA) based on a combination of analytical approximations and numerical integration to estimate

posterior distributions [6]. INLA was designed as an efficient alternative to Markov Chain Monte Carlo (MCMC), which is both computationally and time-intensive when applied to a large amount of data. It can be suitably applied to latent Gaussian models including generalised linear models to spatial and spatio-temporal models.

2.5 Model evaluation

We evaluated our model by plotting the empirical relative risk in each area against the fitted risk determined by our model (Figure A1). The empirical relative risk was calculated by weighting the total number of observed cases in a given municipality, with municipality level population as a proportion of the entire state of São Paulo (defined as offset). A density plot illustrating the distribution of empirical versus predicted risk was also created to assess model fit (Figure A2).

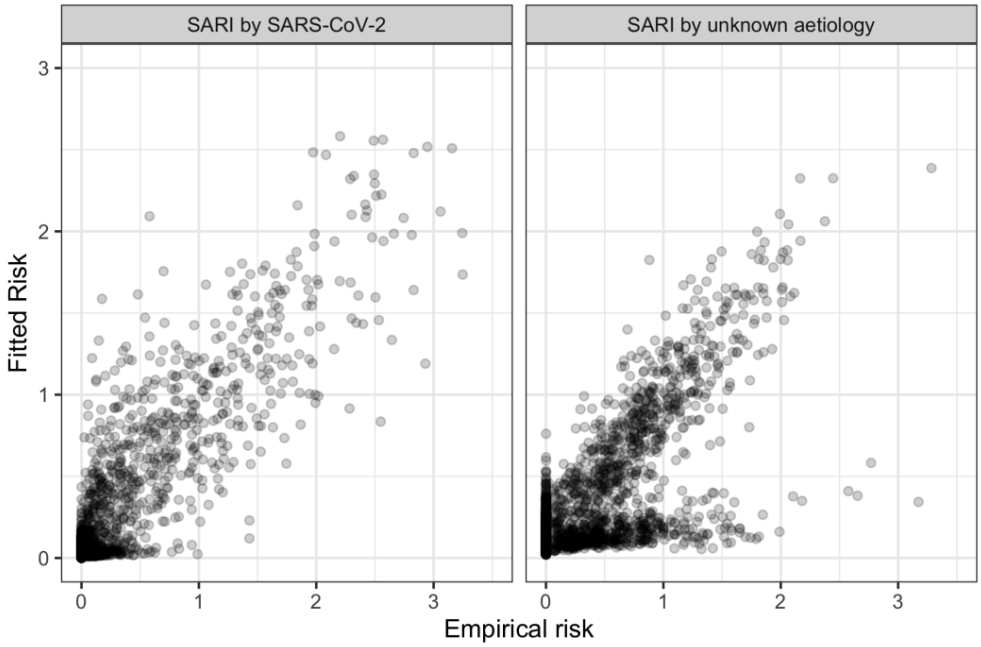


Figure A1. Empirical vs. fitted risk

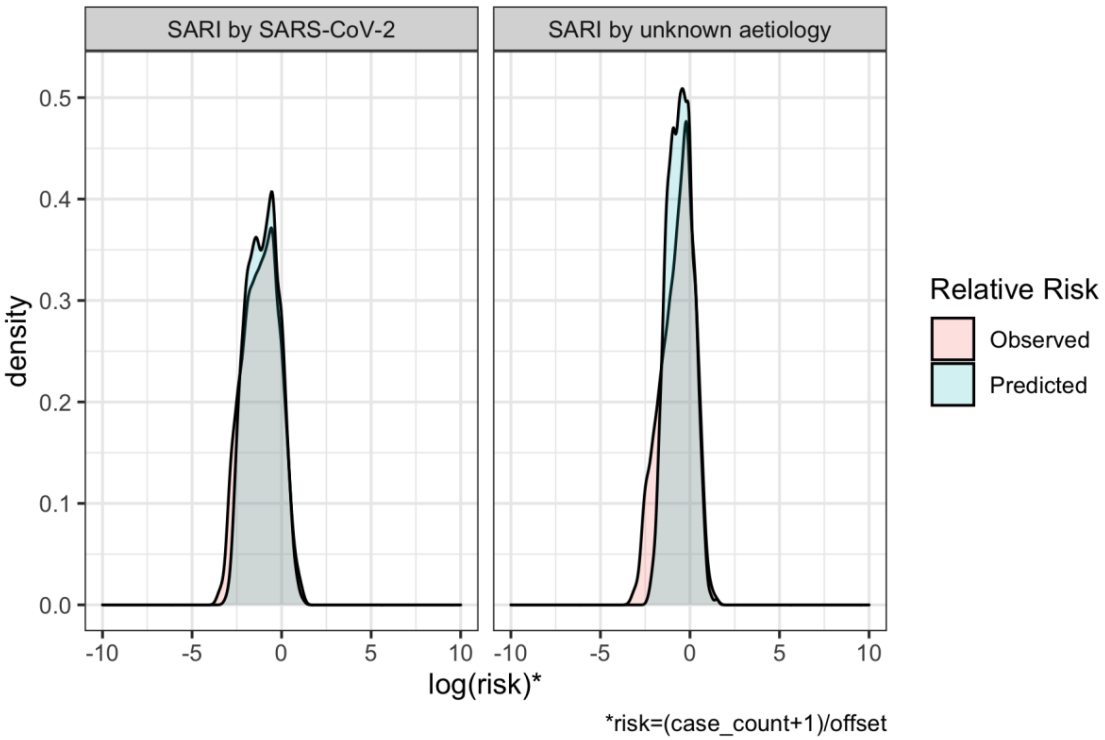


Figure A2. The estimated value of the log-risk for each area closely matches the observed risk in each area.

2.6 Covariates

2.6.1 2010 census

Data on socioeconomic covariates used in the ecological regression at the municipality level were obtained from the 2010 population census compiled by the Brazilian Institute of Geography and Statistics (IBGE) [7].These data include the average household income per capita, housing density (number of private households/km^2), sanitation conditions (proportion of population with access to piped water, access to sewage network, and/or access to septic tank), percentage of urban and rural population, and income inequality measured with the Gini index.

To our knowledge, this is the most complete and recent socioeconomic dataset available for Brazil at our required spatial resolution since no census was carried out in 2020. To assess the variability of some of these covariates over a ten year period, we obtained the same socioeconomic data from the 2000 IBGE census. We compared the same covariates by plotting them at the municipality level for a small sample (in this case, we decided to extract the 39 municipalities in RMSP) (Figure A3). We observed a general trend of increase in income per capita, household density, and number of residents per household, at the municipality level, between 2000 and 2010. Given the inconsistent variation in population access to piped water, septic tank, and sewage network by municipality over the 10-year time frame, we decided to omit these variables from the analysis.

To improve accuracy, we decided to compute population density (people/km^2) at the census tract level, then aggregating it for each municipality, instead of using household density. This was calculated by dividing the total resident population by the area of each census tract using data retrieved from the 2010

census. For large census tracts (with area greater than 0·12 km^2), we considered the occupied area rather than the total area of the census tract. This was done to improve the accuracy of population density estimates in large census tracts, particularly for those located in rural areas, by focusing only on human habitats. Human occupied areas were identified based on population counts from a fine regular grid of 200 meters, which was generated by IBGE for the 2010 census using fieldwork and satellite imagery data.

We checked for multicollinearity by assessing the correlation between variables and computing the variation inflation factor (VIF) of each covariate. Based on these tests, we removed the proportion of informal workers and unemployment from our final model to avoid multicollinearity with the variables on household density and income per capita.

2.6.2 Distance to the nearest health facility

We have also computed mean distance to the nearest health facility of each municipality. To do this, we calculated the road network distance from the centroid of each census tract to the nearest healthcare facility in R with the dodgr package [8]. We considered all the 830 healthcare facilities registered in the SIVE-Gripe database in São Paulo state that hospitalised SARI patients via the Unified Health System (SUS).

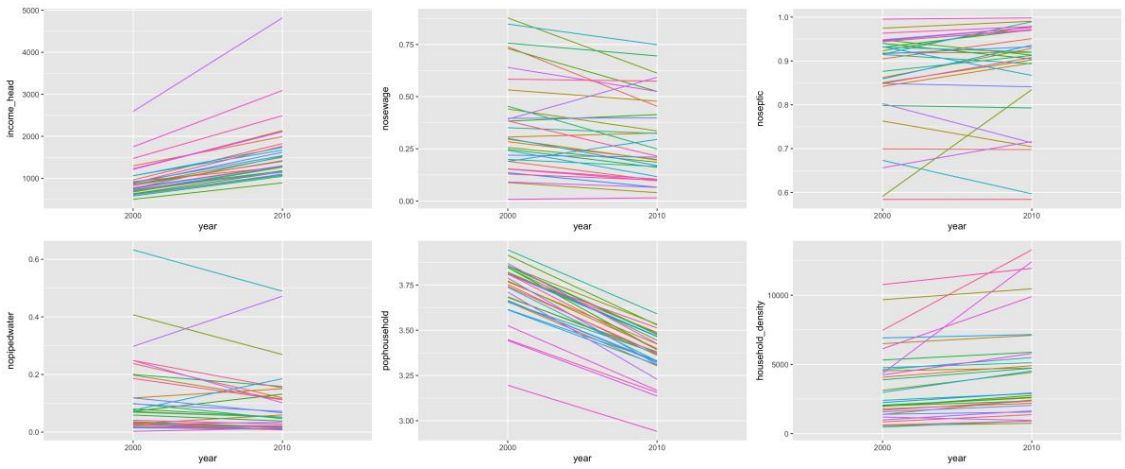


Figure A3. Trends in covariate values between 2000 and 2010. Each coloured line represents a municipality within the GMSP (n= 39).

2.7 Spatial joins for the GMSP

To improve the accuracy of modelling small areas for the RMSP, we performed spatial joins of census tracts to ensure that each census tract had at least one SARI case. Spatial joins were performed by selecting the census tract with the smallest population and joining it to its nearest neighbour (determined by the shortest distance between centroids). This was repeated until all tracts had at least one case of SARI.

3 Probability of working conditions

We used PNAD COVID-19 data from May to September 2020. During this period, 1 888 560 individuals were interviewed, of whom 171 480 were living in São Paulo state. For each individual, the PNAD COVID-19 survey also collects self-reported data on access to COVID-19 testing and comorbidities. From July, August, and September 2020, interviewees were also asked about comorbidities.

In the PNAD COVID-19 survey, work status is a categorical variable that can assume four values: face-to-face, telework, paid leave, and unpaid leave. We model the conditional probabilities using a multinomial logistic regression.

$$\mathrm{P}(y_i = m|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta}_m)}{\sum_{j=1}^{4}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)}$$

$\mathrm{P}(y_i = m|\boldsymbol{x}_i)$ represents the probability that, for the individual i, the variable Work Status (y) will assume a particular value m, given a $k \times 1$ vector $\boldsymbol{x}_i$ of explanatory variables and an intercept. The letter j varies from 1 to 4 and indexes the four categories of Work Status. $\boldsymbol{\beta}_j$ is a vector of coefficients for the category j. The reference category ($j = 1$) is Face-to-Face work and, by construction, $\boldsymbol{\beta}_1 = \mathbf{0}$. So, the estimated vector of parameters $\widehat{\boldsymbol{\beta}}$ is $3k \times 1$ dimensional:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{\boldsymbol{T}} \\ \widehat{\boldsymbol{\beta}}_{\boldsymbol{P}} \\ \widehat{\boldsymbol{\beta}}_{\boldsymbol{U}} \end{bmatrix}$$

Where $\underbrace{\widehat{\boldsymbol{\beta}}_{\mathrm{T}}}_{k\times 1} = \widehat{\boldsymbol{\beta}}_{\text{Telework|Face-to-face}}$, $\underbrace{\widehat{\boldsymbol{\beta}}_{\mathrm{P}}}_{k\times 1} = \widehat{\boldsymbol{\beta}}_{\text{Paid leave|Face-to-face}}$, and $\underbrace{\widehat{\boldsymbol{\beta}}_{\mathrm{U}}}_{k\times 1} = \hat{\beta}_{\text{Unpaid leave|Face-to-face}}$.

We estimated four different nested models, by adding explanatory variables stepwise. Model 0 has race, sex (a dummy variable) and age (a 3rd degree orthogonal polynomial). Model 1 adds education, Model 2 adds occupation (ISCO-08 1-digit groups – see Table **A2**) and a dummy variable for (in)formality, Model 3 adds dummy variables for the months the observations were collected in PNAD COVID-19.

In order to calculate predicted probabilities, we let the variable of interest (e.g., race) to vary and set all other explanatory variables to their grand mean. This way, all the probability differences between categories (e.g. White, Black, *Pardo*, and Asian) are due to changes in the variable of interest.

Models were estimated taking into account the Complex Sample Design of PNAD COVID-19. So, the Variance-Covariance Matrix of the Coefficients already take into account heteroskedasticity and autocorrelation between observations. Confidence intervals for the predicted probabilities were calculated by parametric bootstrapping. By the Central Limit Theorem, the estimated coefficients follow a Multivariate Normal Distribution:

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\beta})$$

Using $\widehat{\boldsymbol{\beta}}$ as the mean and $\underbrace{\widehat{\boldsymbol{\Sigma}}_{\beta}}_{3k\times 3k}$ as the variance-covariance matrix, we simulated 2 000 samples of coefficients. We organized the simulated coefficients for each category of dependent variable in separate matrices: $\underbrace{\widehat{\mathbf{B}}_{\mathrm{T}}^{sim}}_{2000\times k}$, $\underbrace{\widehat{\mathbf{B}}_{\mathrm{P}}^{sim}}_{2000\times k}$, and $\underbrace{\widehat{\mathbf{B}}_{\mathrm{U}}^{sim}}_{2000\times k}$. For a given vector of observations $\boldsymbol{x}$, a distribution of fitted values can be calculated as:

$$\underbrace{\boldsymbol{x}}_{1\times k}\underbrace{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{T}}^{sim}\right)^{T}}_{k\times 2000}=\underbrace{\widehat{\boldsymbol{\mu}}_{\mathrm{T}}^{sim}}_{1\times 2000}$$
$$\underbrace{\boldsymbol{x}}_{1\times k}\underbrace{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{P}}^{sim}\right)^{T}}_{k\times 2000}=\underbrace{\widehat{\boldsymbol{\mu}}_{\mathrm{P}}^{sim}}_{1\times 2000}$$
$$\underbrace{\boldsymbol{x}}_{1\times k}\underbrace{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{U}}^{sim}\right)^{T}}_{k\times 2000}=\underbrace{\widehat{\boldsymbol{\mu}}_{\mathrm{U}}^{sim}}_{1\times 2000}$$

And, by definition $\underbrace{\widehat{\boldsymbol{\mu}}_{\mathrm{F}}^{sim}}_{1\times 2000}=\underbrace{\mathbf{0}}_{1\times 2000}$. A distribution of predicted probabilities given $\boldsymbol{x}$ is given by:

$$\underbrace{\mathbf{P}(y_i=m|\boldsymbol{x}_i)}_{1\times 2000}=\frac{\exp\left(\widehat{\boldsymbol{\mu}}_m^{sim}\right)}{\sum_{j=1}^{4}\exp\left(\widehat{\boldsymbol{\mu}}_j^{sim}\right)}$$

A 95% confidence interval is obtained if we take the quantiles 0,025 and 0,975 of this vector of predicted probabilities.

4 Event-study analysis

We used an event study model to investigate how people from different socioeconomic groups changed their daily isolation levels after the implementation of state non-pharmaceutical intervention (NPI). The model was conducted using an ecological analysis where both socioeconomic characteristics of the population and daily isolation levels are spatially aggregated on the H3 hexagonal grid at resolution 8. Each cell has an edge of approximately 460 meters and an area of 0·74 km^2. Hexagonal H3 cells were then ranked by income based on quintiles of average income per capita. Cells were also categorised as predominantly White when at least 60% of the population self-declared White and predominantly Black when at least 60% of the population self-declared Black or *Pardo*.

The racial composition and income level of each cell were determined using dasymetric interpolation of the 2010 census tract data in two steps. First, data on income and race were passed to a finer regular grid of 200 meters and linked with population count by finding the aerial intersection and population size of each cell. This was reaggregated from the regular grid to the hexagonal grid. Hexagonal H3 cells were then ranked by income based on quintiles of average income per capita. Cells were categorised as predominantly Black when at least 60% of the population self-declared Black or *Pardo*, and likewise for White.

In the event study model, our treated group is composed of hexagons predominantly of White population (race analysis) and the 20% wealthiest hexagons (income analysis). Conversely, the comparison groups were composed of hexagons with predominantly Black population and the 20% poorest hexagons. Our specification includes indicators for pre-and post-treatment effects, as follows:

$$Y_{id} = \left[\sum_{\tau=-12}^{-2} \beta_\tau I(t_{id} - t^* = \tau) + \sum_{\tau=0}^{151} \beta_\tau I(t_{id} - t^* = \tau) \right] + X'_{id}\Theta + \omega_d + \mu_i + \varepsilon_{id}$$

where Y_{id} is the outcome (daily isolation level) observed for hexagon i at day d; the indicator $I(t_{id} - t^* = \tau)$ measures the time (in days) relative to the day of the state NPI implementation on date t^*. We set the coefficient β_{-1} (March 12) equal to zero to use the day immediately prior to the state NPI implementation as the reference. X'_{id} represents the set of hexagons covariates: a dummy variable indicating the beginning of NPI flexibilization period in each municipality, and a time-varying variable with the number of days relative to the first confirmed case of SARI in each hexagon (equates to 0 for days prior to the first case). μ_i is a hexagon fixed effect that controls non-parametrically for time-invariant hexagon factors, such as hexagons' fixed geographical aspects (e.g., urban infrastructure, proximity to healthcare facilities, urban density etc), ω_d is a day fixed effect that controls non-parametrically for aggregate shocks and other policies common to all hexagons at a specific moment in time, and ε_{id} is an idiosyncratic error term. All observations are weighted by the size of resident population in each hexagon [7]. Finally, we clustered standard errors at the hexagon level to make estimations robust to serial correlation and heteroskedasticity [9].

This method allows us to more formally test for pre-trends in outcome variables in the pre-period. The identifying assumption is that the time trend in the mobility level in treated areas would have a similar trend as the one observed in similar nontreated areas in the absence of the policy intervention. Coefficient estimates of β_τ; with $\tau < 0$ (representing the change, in percentage points, in the outcome each day pre-intervention) serves as a direct test of the plausibility of the identifying assumption. If hexagons have similar trends before the date of state declaration and diverge only after policy, it provides strong evidence that such changes were caused by the state NPI adoption rather than an unobservable factor.

4.1 Summary statistics

Table 1 presents summary statistics for the main variables used in our analysis. Columns (1)-(3) shows the number of observations, the mean and the standard deviation for the treated group, while columns (4)-(6) exhibit the same statistics for the comparison group. Panel A shows the summary statistics for the race estimation sample, while Panel B presents the summary statistics for the income estimation sample.

Table A1 - Summary Statistics of hexagon characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A – Race sample estimation					
	Black			White		
	Obs.	Mean	Std. Dev	Obs.	Mean	Std. Dev
Daily isolation level	16 187	0·418	0·084	225 231	0·457	0·100
Days after first SARI case	16 187	45·47	45·23	225 231	59·16	47·14
Days with NPI flexibilization	16 187	0·397	0·489	225 231	0·387	0·487
	Panel B – Income sample estimation					
	Low Income			High Income		
	Obs.	Mean	Std. Dev	Obs.	Mean	Std. Dev

Daily isolation level	97 139	0·419	0·088	90 383	0·486	0·103
Days after first SARI case	97 139	47·55	45·81	90 383	63·40	47·44
Days with NPI flexibilization	97 139	0·396	0·489	90 383	0·381	0·486

Notes. This table polls all days of data per group in each sample estimation (from March 1 to August 11). Data are at the hexagon-by-day level.

4.2 Model sensitivity analysis

We tested how the results of the event study by running the regression with and without covariates and testing for different number of days after the introduction of NPIs. Table A2 shows the results including only one treatment variable so it represents the effect of the average treatment every day. With 99% of statistical significance in all scenarios, the results indicate that the effect tends to increase up until the 120th day. The effects of NPI on isolation levels is consistently positive and significant in all estimates. The inclusion of covariates does not significantly change the coefficients nor the confidence intervals.

Table A2 – Sensitivity analysis of NPI effects on isolation levels

Number of days after NPI	**Analysis by income**					
	Without covariates			With covariates		
	Mean	Min 95	Max 95	Mean	Min 95	Max 95
10	0·055***	0·053	0·058	0·054***	0·052	0·057
30	0·084***	0·082	0·088	0·081***	0·079	0·085
90	0·087***	0·084	0·091	0·084***	0·081	0·088
120	0·083***	0·080	0·087	0·082***	0·079	0·086
150	0·079***	0·076	0·083	0·078***	0·075	0·082
Number of days after NPI	**Analysis by race**					
	Without covariates			With covariates		
	Mean	Min 95	Max 95	Mean	Min 95	Max 95
10	0·039***	0·036	0·044	0·038***	0·035	0·043
30	0·064***	0·060	0·069	0·061***	0·057	0·066
90	0·062***	0·058	0·068	0·059***	0·055	0·064
120	0·059***	0·055	0·064	0·056***	0·053	0·061
150	0·055***	0·051	0·060	0·053***	0·049	0·058

Note. *** p <0·01

5 Crosswalk: PNAD COVID-19 Occupations to ISCO-08 (1-digit)

The crosswalk from PNAD COVID-19 occupational codes to ISCO-08 1-digit codes was done in two steps. First, we applied the conversion rule presented in Table A1 to the occupational codes of occupation categories in PNAD COVID-19 (variable C007C). Then we classified the values 2 and 3 of variable C007 as “Military” (once these occupations are not registered in C007). We further disaggregated Health Professionals and Technicians into separate occupational categories.

Table A2 – Crosswalk: Variable C007C to ISCO-08 1-digit groups

PNAD-COVID-19 Occ. Code	PNAD COVID-19 Occupations (English Label)	ISCO-08 (1-digit)	ISCO-08 Label
1	Domestic worker, daily cleaner, cook (in private households),	9	Elementary Occupations
2	Janitor, cleaning assistant, etc. (in public or private company),	9	Elementary Occupations
3	Office clerk	4	Clerical Support Workers
4	Secretary, receptionist	4	Clerical Support Workers
5	Telemarketing operator	4	Clerical Support Workers
6	Merchant (owner of bars or shops etc.)	5	Services and Sales Workers
7	Store salesperson	5	Services and Sales Workers
8	Home seller, sales representative, catalog seller	5	Services and Sales Workers
9	Street vendors	9	Elementary Occupations
10	Cook and waiter (for restaurants, companies)	5	Services and Sales Workers
11	Baker, butcher and confectioner	5	Services and Sales Workers
12	Farmer, animal breeder, fisherman, forester and gardener	6	Skilled Agricultural, Forestry and Fishery Workers
13	Agricultural labourers	9	Elementary Occupations
14	Drivers (ride hailing apps, taxi, van, mototaxi, bus)	8	Plant and Machine Operators and Assemblers
15	Truck driver	8	Plant and Machine Operators and Assemblers
16	Courier services by motorcycle	9	Elementary Occupations
17	Delivery of goods (restaurant, pharmacy, store, Uber Eats, IFood, Rappy etc.)	9	Elementary Occupations
18	Bricklayer, stonemasons, painter, electrician, carpenter	7	Craft and Related Trades Workers
19	Mechanic of vehicles, industrial machineries etc.	7	Craft and Related Trades Workers
20	Craftsman, dressmaker and shoemaker	7	Craft and Related Trades Workers
21	Hairdresser, manicure and other beauty occupations	5	Services and Sales Workers
22	Machine operator, assembler in the industry	8	Plant and Machine Operators and Assemblers
23	Production assistant, loading and unloading	8	Plant and Machine Operators and Assemblers
24	Teachers and professors (kindergarten, elementary, high school or higher education)	2	Professionals
25	Pedagogue, teacher of languages, music, art and tutoring	2	Professionals
26	Health professionals	2	Professionals
27	Health Technician	3	Technicians and Associate Professionals
28	Babysitters and personal caretakers	5	Services and Sales Workers
29	Security, vigilant, other guard security services	5	Services and Sales Workers
30	Civil police	5	Services and Sales Workers
31	Doorman or Porter	9	Elementary Occupations
32	Artist, religious (priest, pastor, etc.)	2	Professionals
33	Director, manager, political or commissioned position	1	Managers
34	Other higher-level profession (lawyer, engineer, accountant, journalist, etc.)	2	Professionals
35	Other mid-level technician or professional	3	Technicians and Associate Professionals
36	Others	10	OTHER

6 Classification of non-classified SARI cases as COVID-19 cases

We considered all SARI cases with missing or unknown etiology as COVID-19 cases. This approach leads to an estimated number of COVID-19 cases that is closer to the actual number than taking into account only confirmed COVID-19 cases[10]. Figure A4 shows the number of daily SARI cases and deaths that were confirmed as COVID-19, confirmed by cause of a different etiology, or with unknown or missing etiology (non-classified SARI cases) for the state of São Paulo. In March 2020, both confirmed and non-classified SARI cases and deaths increased to a much higher level than what was previously recorded. This suggests that most non-classified SARI cases and deaths are actually caused by COVID-19, and that even severe COVID-19 cases suffer from undernotification in Brazil.



Figure A4. SARI cases and deaths in São Paulo that were confirmed by cause: COVID-19, other etiologies, and unknown etiology.

7 References

1. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. Proc Natl Acad Sci. 2020;117:9696–8.

2. Lawson AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Second Edition. CRC Press; 2013. 398 p.

3. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math. 1991;43:1–20.

4. Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. Stat Med. 2000;19:2555–67.

5. Peixoto PS, Marcondes D, Peixoto C, Oliva SM. Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil. PLOS ONE. 2020;15:e0235732.

6. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B Stat Methodol. 2009;71:319–92.

7. Censo Demográfico | IBGE [Internet]. [cited 2020 Nov 20]. Available from: https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?=&t=o-que-e

8. Padgham M. dodgr: An R Package for Network Flow Aggregation. Findings. 2019;6945.

9. Bertrand M, Duflo E, Mullainathan S. How Much Should We Trust Differences-In-Differences Estimates? Q J Econ. 2004;119:249–75.

10. de Souza WM, Buss LF, Candido D da S, Carrera J-P, Li S, Zarebski AE, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat Hum Behav. 2020;4:856–65.

11. Buss LF, Prete CA, Abrahim CMM, Mendrone A, Salomon T, Almeida-Neto C de, et al. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. Science [Internet]. 2020 [cited 2020 Dec 8]; Available from: https://science.sciencemag.org/content/early/2020/12/07/science.abe9728

Appendix B: Figures

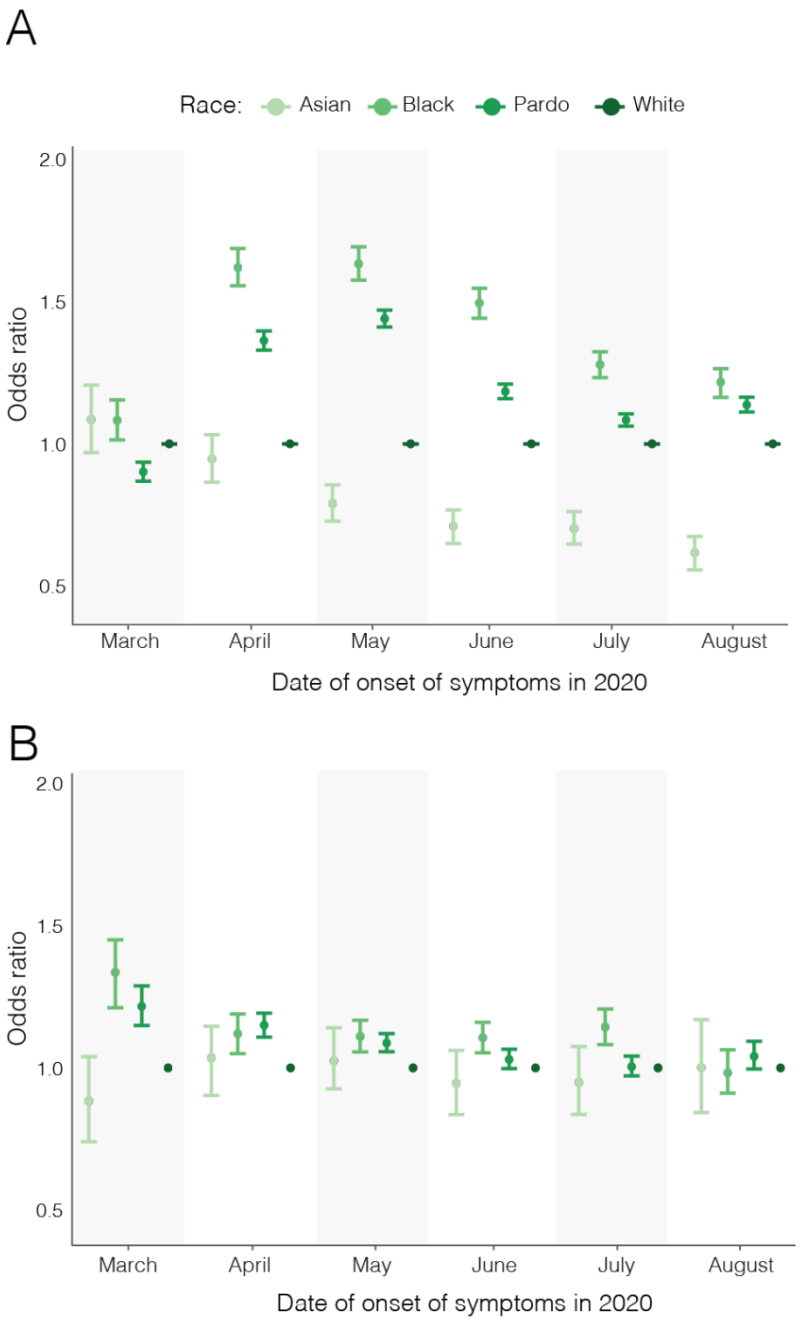


Figure S 1. Individual level hospitalisation and death risk by age-standardised odds ratio (OR) after discarding data with missing race information. **A**, OR for SARI hospitalisation by race and **B**, OR for death among SARI patients by race.

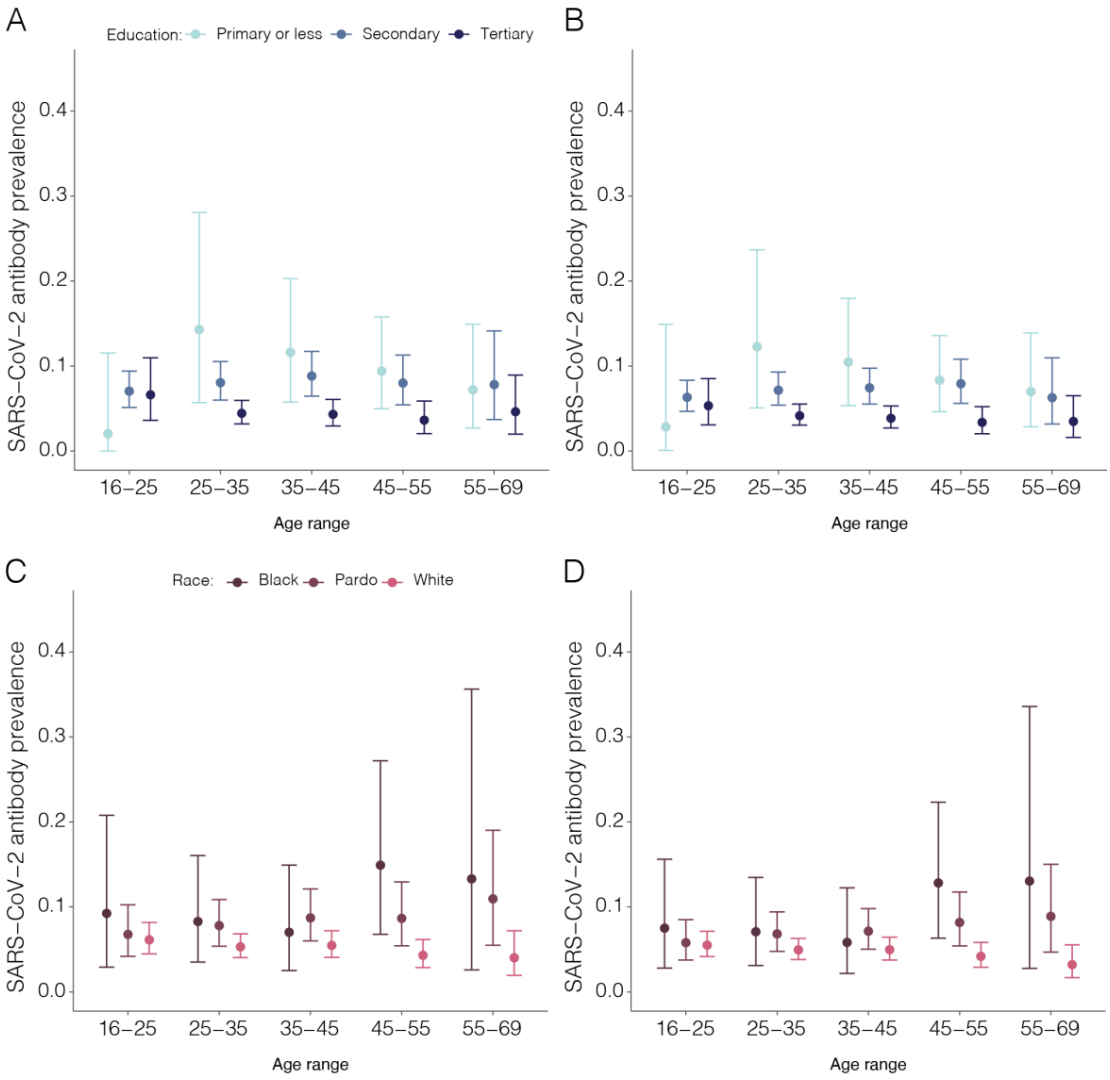


Figure S 2. Prevalence of anti-SARS-CoV-2 IgG antibodies among blood donors in São Paulo, Brazil, according to self-reported race and education attainment. **A and C**, crude prevalence and **B and D**, prevalence corrected by age, sex, specificity and sensitivity are shown. Source: Covid-IgG study [11].

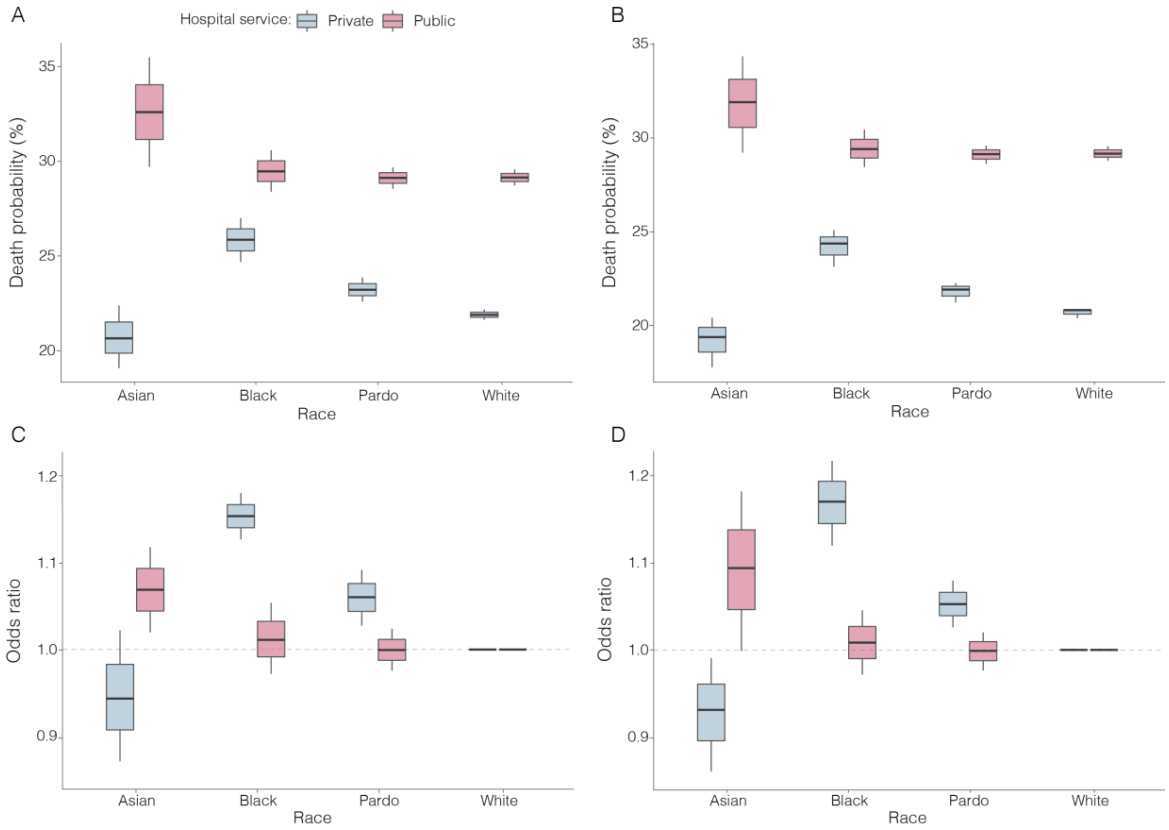


Figure S 3. **A and C,** Death probability and odds ratio of death by race in each hospital type using data without missing race information. **B and D,** with race imputed. Source: SIMI-SP.

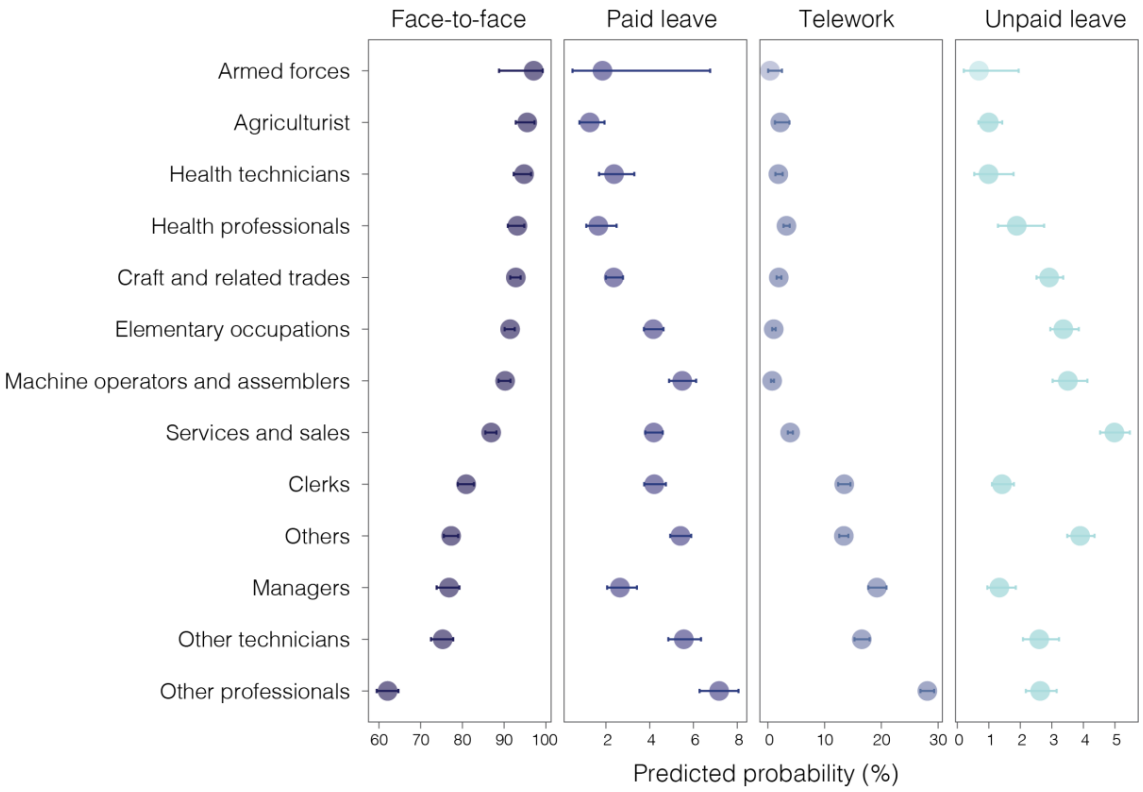


Figure S 4. Probability of working condition by occupation type between May and September 2020 in São Paulo state. Source: PNAD COVID-19 (IBGE).

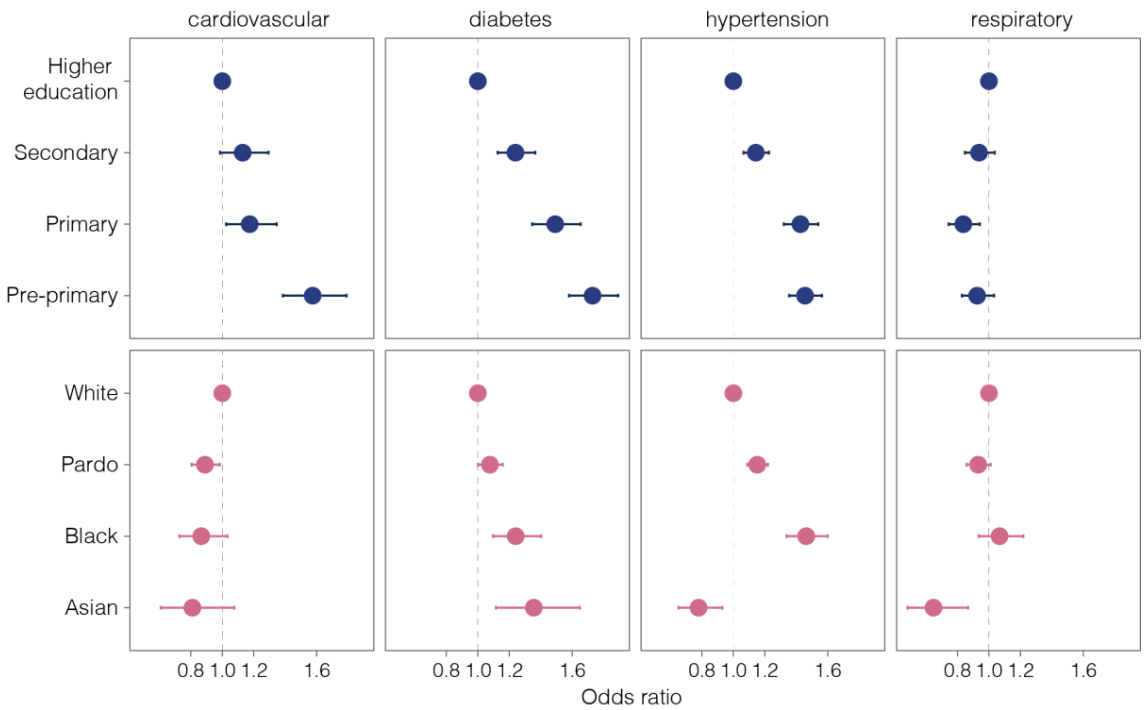


Figure S 5. Odds ratio of having been diagnosed with a comorbidity, by race and education attainment in São Paulo State, 2020. Source: PNAD COVID-19(IBGE).

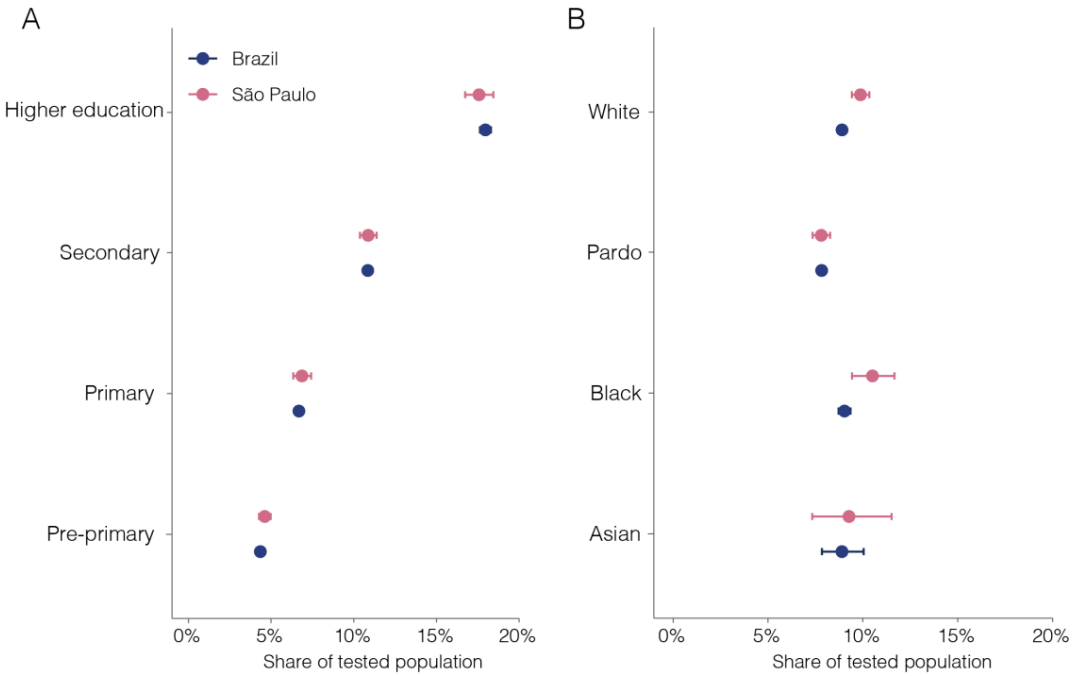


Figure S 6. Proportion of population that reported of being tested for COVID-19 between July and September 2020 by **A,** education attainment and **B,** race in Brazil and São Paulo state. Source: PNAD COVID-19 (IBGE).

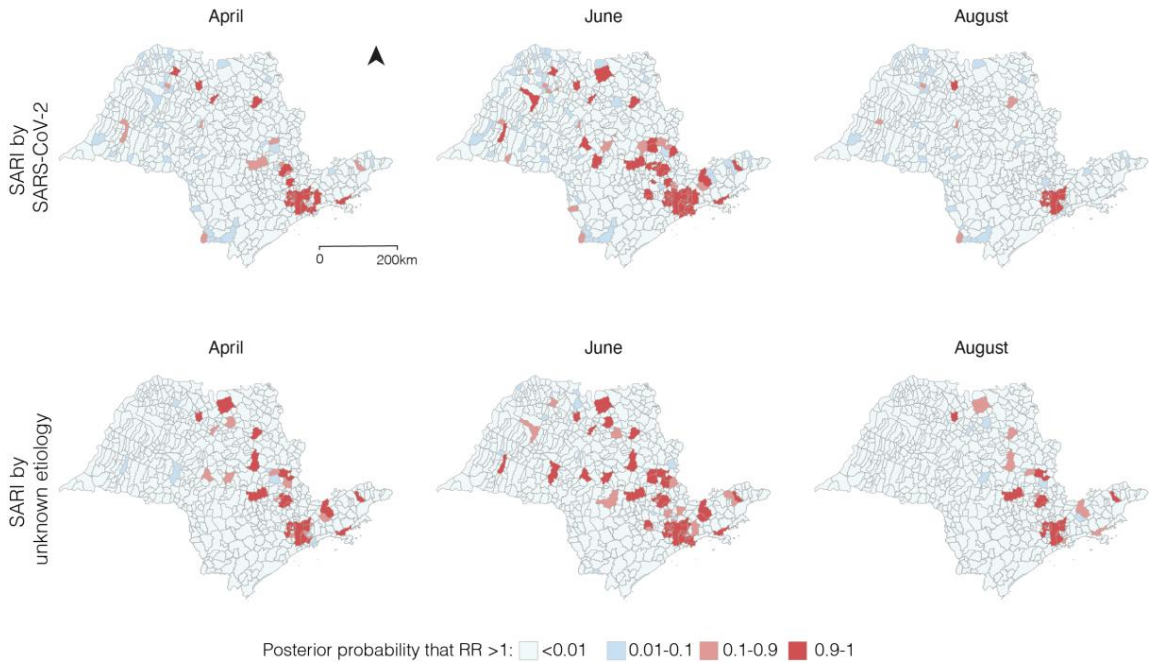


Figure S 7. Posterior probability of elevated relative risk at the municipality level for São Paulo state.

Chapter 5 Supplementary Information

Socio-demographic inequalities in COVID-19 vaccination in Brazil

Supplementary Information

Sabrina L Li, MSc [1*#], Carlos A Prete Jr, MSc [2*], Alexander E Zarebski, PhD [3*], Andreza A de Souza Santos, PhD [4], Ester C Sabino, MD [5], Vitor H Nascimento, PhD [3], Chieh-Hsi Wu, PhD [6], Jane P Messina, PhD [1,7]

1. School of Geography and the Environment, University of Oxford, Oxford, United Kingdom.
2. Department of Electronic Systems Engineering, University of São Paulo, São Paulo, Brazil.
3. Department of Zoology, University of Oxford, Oxford, United Kingdom.
4. Oxford School of Global and Area Studies, Latin American Centre, University of Oxford, Oxford, United Kingdom.
5. Departamento de Molestias Infecciosas e Parasitarias & Instituto de Medicina Tropical da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.
6. Mathematical Sciences, University of Southampton, Southampton, United Kingdom.
7. Oxford School of Global and Area Studies, University of Oxford, Oxford, United Kingdom.

*These authors contributed equally.

#Corresponding authors:

1. Sabrina L. Li, School of Geography and the Environment, University of Oxford, South Parks Rd, Oxford, OX1 3QY, United Kingdom. E-mail: lisabrinaly@gmail.com

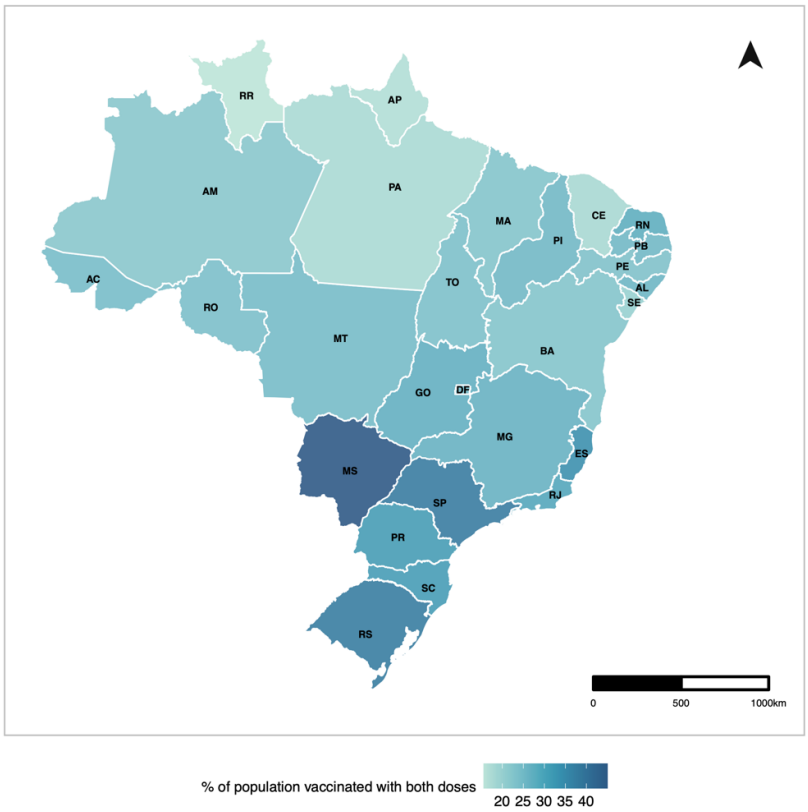


Figure 1. Proportion of population fully vaccinated.

Table 1. Logistic regression model analysis – death risk of SARI patient given age group, socioeconomic status, age group-specific vaccination coverage, and indication of age-specific vaccination campaign

Variable	Coefficient (95% CI)	Significance
Intercept	-3.60 (-3.63, -3.57)	*
Age group[#]: 20-29 years old	1.17 (1.14, 1.21)	*
Age group: 30-39 years old	1.51 (1.14, 1.21)	*
Age group: 40-49 years old	1.91 (1.88, 1.94)	*
Age group: 50-59 years old	2.35 (2.32, 2.37)	*
Age group: 60-69 years old	2.87 (2.84, 2.90)	*
Age group: 70-79 years old	3.21 (3.18, 3.24)	*
Age group 80 years and older	3.55 (3.52, 3.58)	*
Age-group specific coverage	-0.00414 (-0.00568, -0.00261)	*
Age-group specific campaign (yes)	-0.182 (-0.186, -0.178)	*
Unemployed	0.0241 (0.0221, 0.0261)	*
Primary education or lower	-0.0737 (-0.0758, -0.0716)	*
Income per capita	-0.0579 (-0.0602,-0.0556)	*

[#] Baseline age group for comparison is under 20 years old
*statistically significant; p-value < 0.05

Table 2. quasi-Poisson regression model analysis– hospitalisation risk given age group, socioeconomic status, age group-specific vaccination coverage, and indication of age-specific vaccination campaign

Variable	Coefficient (95% CI)	Significance
Intercept	-6.67 (-7.45, -5.89)	*
Age group[#]: 20-29 years old	-0.0427(-0.161, -0.246)	
Age group: 30-39 years old	0.884 (0.721, 1.05)	*
Age group: 40-49 years old	1.39 (1.24, 1.54)	*
Age group: 50-59 years old	1.84 (1.69, 1.98)	*
Age group: 60-69 years old	2.19 (2.04, 2.34)	*
Age group: 70-79 years old	2.62 (2.47, 2.77)	*
Age group 80 years and older	3.16 (3.02, 3.30)	*
Age-group specific coverage	0.00875(0.000612, 0.0169)	*
Age-group specific campaign (yes)	0.0533 (-0.000682, 0.115)	
Unemployed	-3.98 (-5.65, -2.31)	*
Primary education or lower	1.94 (0.502, 3.37)	*
Income per capita	-0.000363 (0.0000902,0.0000982)	*

[#] Baseline age group for comparison is under 20 years old
*statistically significant; p-value < 0.05